ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ
ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ & ΠΑΡΑΓΩΓΗΣ

«Η ΕΠΙΣΤΗΜΗ ΤΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ»

"The science of Machine Learning"

Διπλωματική εργασία της

Λαμπρινής Πεϊβάνη

Τριμελής επιτροπή

1. Παπουτσιδάκης Μιχαήλ
2. Χατζόπουλος Αβραάμ
3. Δρόσος Χρήστος

16 Οκτωβρίου 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ (1η ή 2η σελίδα)

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/Η κάτωθι υπογεγραμμένος/η Πεϊβάνη Λαμπρινή, του Γεωργίου, φοιτητής του Τμήματος Βιομηχανικής Σχεδίασης και Παραγωγής του Πανεπιστημίου Δυτικής Αττικής, πριν αναλάβω την εκπόνηση της Πτυχιακής Εργασίας μου, δηλώνω ότι ενημερώθηκα για τα παρακάτω:

«Η Πτυχιακή Εργασία (ΠΕ) αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο του συγγραφέα, όσο και του Ιδρύματος και θα πρέπει να έχει μοναδικό χαρακτήρα και πρωτότυπο περιεχόμενο.

Απαγορεύεται αυστηρά οποιοδήποτε κομμάτι κειμένου της να εμφανίζεται αυτούσιο ή μεταφρασμένο από κάποια άλλη δημοσιευμένη πηγή. Κάθε τέτοια πράξη αποτελεί προϊόν λογοκλοπής και εγείρει θέμα Ηθικής Τάξης για τα πνευματικά δικαιώματα του άλλου συγγραφέα. Αποκλειστικός υπεύθυνος είναι ο συγγραφέας της ΠΕ, ο οποίος φέρει και την ευθύνη των συνεπειών, ποινικών και άλλων, αυτής της πράξης.

Πέραν των όποιων ποινικών ευθυνών του συγγραφέα, σε περίπτωση που το Ίδρυμα του έχει απονείμει Πτυχίο, αυτό ανακαλείται με απόφαση της Συνέλευσης του Τμήματος. Η Συνέλευση του Τμήματος με νέα απόφασή της, μετά από αίτηση του ενδιαφερόμενου, του αναθέτει εκ νέου την εκπόνηση ΠΕ με άλλο θέμα και διαφορετικό επιβλέποντα καθηγητή. Η εκπόνηση της ενλόγω ΠΕ πρέπει να ολοκληρωθεί εντός του λάχιστον ενός ημερολογιακού μήνου από την ημερομηνία ανάθεσής της.

Ο Δηλών Πεϊβάνη Λαμπρινή

Ημερομηνία 14/10/2020

Περίληψη

Αντικείμενο της παρούσας πτυχιακής εργασίας αποτελεί η μελέτη της επιστήμης της μηχανικής μάθησης. Πιο συγκεκριμένα θα περιγραφεί τι είναι η επιστήμη της μηχανικής μάθησης. Θα αναλυθούν οι διάφορες μεθοδολογίες που χρησιμοποιούνται για την επίλυση των αντικειμένων μελέτης που πραγματεύεται η συγκεκριμένη επιστήμη καθώς επίσης θα παρουσιαστούν και αναλυθούν οι βασικότεροι αλγόριθμοι που εφαρμόζονται. Τέλος, θα χρησιμοποιηθεί το λογισμικό Matlab (MatrixLaboratory) για την υλοποίηση των αλγορίθμων.

Abstract

The subject of this diploma thesis is the study of the science of machine learning. More specifically, it will particularly be described what the science of machine learning is. In addition, the different methodologies that are used for the solution of the objects of study that this science deals with as well as the most important algorithms that are implemented will be further analyzed. Finally, Matlab (Matrix Laboratory) software will be used in order to implement the algorithms.

Πίνακας περιεχομένων

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου κύριο Χρήστο Δρόσο για την στήριξη του και την πολύτιμη βοήθειά του στην προσπάθεια εκπόνησης της παρούσας διπλωματικής εργασίας.

Εισαγωγή

Η εμφάνιση της Μηχανικής Μάθησης

Ο όρος Machine Learning (ML) ο οποίος αποδίδεται στα ελληνικά ως Μηχανική Μάθηση, εισήχθη για πρώτη φορά το 1959 από τον Arthur Samuel[1]. Όπως ισχυριζόταν ο Samuel, «η ML επιτρέπει στα υπολογιστικά συστήματα να μαθαίνουν, χωρίς να έχει προηγηθεί ρητός προγραμματισμός τους». Η δήλωση αυτή συμπληρώθηκε από τον Tom Mitchell[2] ο οποίος περιέγραψε τη διαδικασία ML με πιο αυστηρή μαθηματική τεκμηρίωση ως εξής: «Ένα υπολογιστικό πρόγραμμα **μαθαίνει** από την εμπειρία E σε σχέση με κάποια εργασία T και κάποιο μέτρο απόδοσης P, εάν η απόδοσή του στην T, η οποία μετριέται σε P, **βελτιώνεται** με την εμπειρία E (Mitchell, etal., 1990).

Η ML έχει εξελιχθεί σε αντικείμενο μελέτης καθώς επιτρέπει στο χρήστη είτε αυτός είναι ερευνητής, είτε εμπορική εταιρία, τη βαθύτερη κατανόηση και πρόβλεψη των εξελίξεων των φαινομένων που ερευνώνται αλλά και την αποτελεσματική μόχλευση των υπαρχόντων δεδομένων. Με την κατάλληλη επιλογή μοντέλου ML, ο ερευνητής δύναται να βρίσκεται συνεχώς ένα βήμα μπροστά από τις εξελίξεις. Στο σύγχρονο επιχειρησιακό και ερευνητικό γίγνεσθαι, τα δεδομένα εισρέουν σωρηδόν και γεννάται η ανάγκη συνεχούς επαναπροσδιορισμού και επαναϋπολογισμού των βέλτιστων λύσεων, ανάγκη που καλύπτεται από την ικανότητα των υπολογιστικών προγραμμάτων να βελτιώνουν την απόδοσή τους αποκτώντας εμπειρία, όπως δήλωσε ο Mitchell. Η αξία της παραπάνω ικανότητας είναι αδιαμφισβήτητη, καθώς στο πλαίσιο της ML, εφόσον τα υπολογιστικά προγράμματα τροφοδοτούνται συνεχώς από κατάλληλες και συνεχώς μεταβαλλόμενες πηγές, επιστρέφουν τη δυνατότητα πρόβλεψης του μέλλοντος.

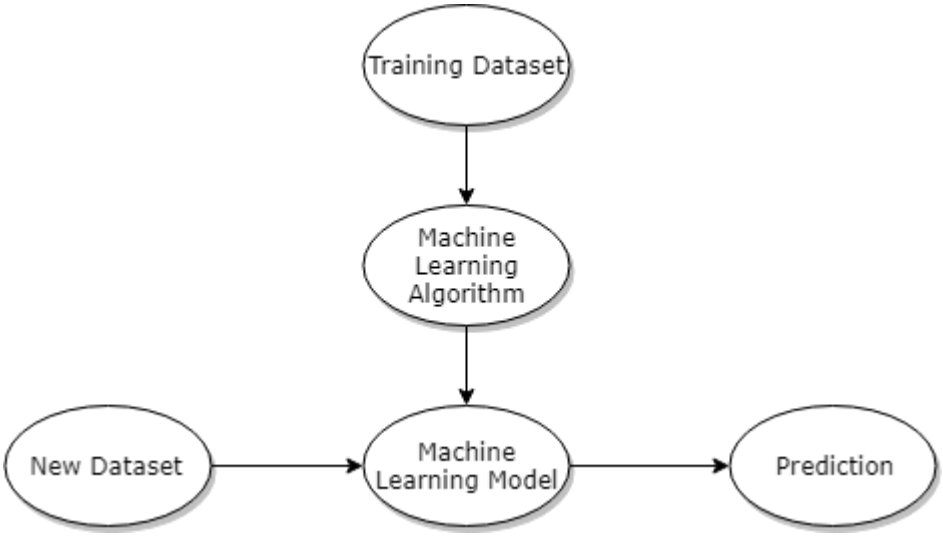
Η ML αποτελεί μία μορφή Artificial Intelligence (AI), (ελληνικά Τεχνητή Νοημοσύνη), η οποία επιτρέπει σε ένα σύστημα να μαθαίνει από την τροφοδότησή του με δεδομένα, χωρίς όμως αυτό να σημαίνει ότι αυτή είναι μια απλή διαδικασία. Χρησιμοποιεί μια πληθώρα αλγορίθμων οι οποίοι επαναληπτικά μαθαίνουν από τα δεδομένα προκειμένου να βελτιώσουν τις προβλέψεις τους. Καθώς οι αλγόριθμοι «καταναλώνουν» δεδομένα, είναι πλέον σε θέση να παράγουν πιο ακριβείς προβλέψεις, βασισμένες στα δεδομένα αυτά. Τα δεδομένα τα οποία τροφοδοτούνται με αντικειμενικό σκοπό την εκπαίδευση των αλγορίθμων καλούνται trainingdata (δεδομένα εκπαίδευσης) (Poole, etal., 1998).

Κατά την εκπαίδευση ενός αλγορίθμου ML με TrainingData, επιστρέφεται ως εξαγόμενο ένα μοντέλο. Το μοντέλο αυτό είναι ένα μοντέλο ML, το οποίο είναι έτοιμο να τροφοδοτηθεί με δεδομένα, για να επιστρέψει τις προβλέψεις του (Bishop, 2006). Για παράδειγμα, ένας ML

[1]Ο Άρθουρ Σάμουελ (5 Δεκεμβρίου, 1901 – 29 Ιουλίου, 1990) ήταν Αμερικανός πρωτοπόρος της Τεχνητής Νοημοσύνης. Ήταν αυτός που επινόησε τον όρο "Μηχανική Μάθηση" το 1959. Το λογισμικό για το παιχνίδι της Ντάμας που ανέπτυξε ήταν ανάμεσα στα πρώτα επιτυχημένα προγράμματα υπολογιστή που μπορούσαν να μαθαίνουν από μόνα τους.

[2]Ο Tom Michael Mitchell (γεννημένος στις 9 Αυγούστου 1951) είναι Αμερικανός επιστήμονας πληροφορικής και Καθηγητής Πανεπιστημίου E. Fredkin στο Πανεπιστήμιο Carnegie Mellon (CMU). Είναι πρώην πρόεδρος του τμήματος Μηχανικής Μάθησης του Carnegie Mellon University. Ο Mitchell είναι γνωστός για τη συμβολή του στην πρόοδο της μηχανικής μάθησης, της τεχνητής νοημοσύνης και της γνωστικής νευροεπιστήμης και είναι ο συγγραφέας του εγχειριδίου Machine Learning. Είναι μέλος της Εθνικής Ακαδημίας Μηχανικών των Ηνωμένων Πολιτειών από το 2010. Είναι επίσης μέλος της Αμερικανικής Ένωσης για την Προώθηση της Επιστήμης και συνεργάτης της Ένωσης για την Προώθηση της Τεχνητής Νοημοσύνης.

προγνωστικός αλγόριθμος, εφόσον τροφοδοτηθεί με το σύνολο TrainingData, θα επιστρέψει το αντίστοιχο προγνωστικό μοντέλο, το οποίο θα είναι έτοιμο να πραγματοποιήσει προβλέψεις για τις μεταβλητές οι οποίες μελετώνται. Επομένως, το αποτέλεσμα, δηλαδή η πρόβλεψη, εξαρτάται άμεσα από την κατάλληλη επιλογή του αλγορίθμου αλλά και από την ορθή επιλογή του TrainingDataset.



Σχήμα 1 - Διάγραμμα ροής πρόγνωσης με χρήση αλγορίθμου μηχανικής μάθησης

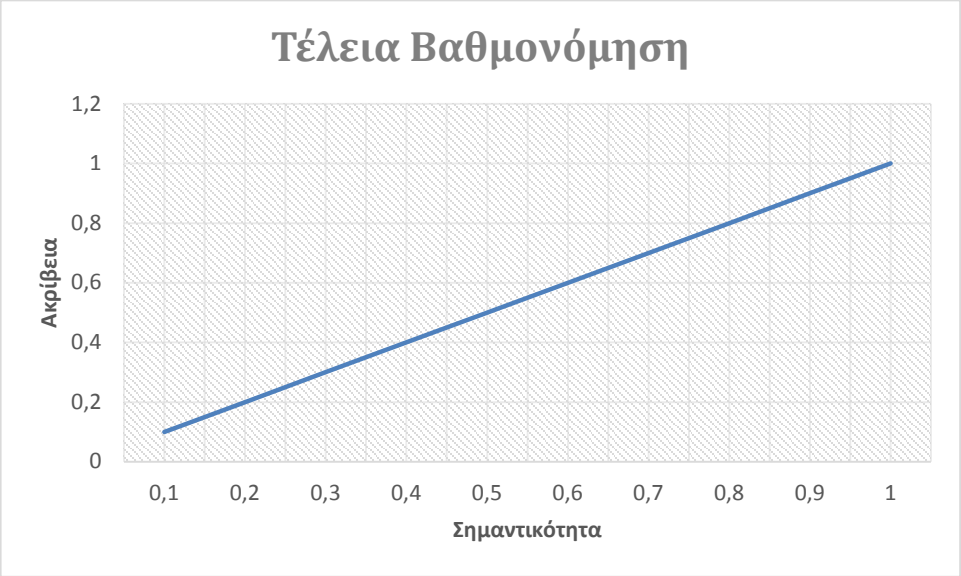
Βαθμονόμηση του μοντέλου

Πριν την τροφοδότηση του ML μοντέλου με τα δεδομένα για την παραγωγή της πρόβλεψης, είναι επιτακτική η ανάγκη της εκτίμησης της ακρίβειας του μοντέλου. Η εκτίμηση αυτή δεν πρέπει να πραγματοποιείται μόνο με απόλυτους όρους, δηλαδή ο υπολογισμός του συνολικού σφάλματος της πρόβλεψης ή του μέσου τετραγωνικού σφάλματος, αλλά είναι χρήσιμη και η εκτίμηση της κατανομής των σφαλμάτων. Υπάρχουν αρκετά παραδείγματα ML μοντέλων τα οποία εκτιμάται ότι έχουν μικρά σφάλματα, αλλά αρκετά κακή κατανομή σφαλμάτων. (Bella, etal., 2010). Για το λόγο αυτό έχουν αναπτυχθεί τεχνικές βαθμονόμησης (calibration) με στόχο τη βελτίωση της ακρίβειας της πρόβλεψης ή ακόμη και την καλύτερη κατανομή των σφαλμάτων. Η βαθμονόμηση του μοντέλου είναι το μέτρο της πιθανότητας η προβλεπόμενη τιμή του μοντέλου να είναι ορθή. Η μαθηματική διατύπωση της παραπάνω πρότασης λαμβάνει την εξής μορφή:

Έστω $X \in \mathcal{X}$και $Y \in \mathcal{Y} = \{1,2,\dots,K\}$ τυχαίες μεταβλητές οι οποίες ακολουθούν την κοινή κατανομή $p(X,Y) = p(Y|X)p(X)$. Έστω h MLμοντέλο για το οποίο $h(X) = (\hat{Y}, \hat{P})$. Το $\hat{Y}$αποτελεί μία κλάση πρόβλεψης και $\hat{P}$ η στάθμη εμπιστοσύνης της. Η τέλεια βαθμονόμηση στην περίπτωση αυτή ορίζεται ως:

$$p(\hat{Y} = T|\hat{P} = p) = p, \forall p \in [0,1]$$

Με άλλα λόγια, ένα μοντέλο είναι τέλεια βαθμονομημένο αν και μόνο αν για κάθε $p \in [0,1]$ η πρόβλεψη μιας κλάσης με σημαντικότητα p είναι ορθή $p \cdot 100\%$ των φορών. Για την οπτικοποίηση του βαθμού της βαθμονόμησης χρησιμοποιείται το διάγραμμα Ακρίβεια-Σημαντικότητα το οποίο όσο πιο κοντά βρίσκεται στην ευθεία $y = x$ τόσο πιο καλή είναι η βαθμονόμηση.



Διάγραμμα 1 - Μέτρηση βαθμονόμησης

Καθώς η $\hat{P}$ είναι συνεχής και είναι αδύνατη η λήψη άπειρων δειγμάτων, η τέλεια βαθμονόμηση είναι αδύνατη στο μέτρο του εφικτού. Για το λόγο αυτό έχουν αναπτυχθεί διαφορετικές στατιστικές τεχνικές, και συγκεκριμένα στατιστικά μέτρα, για τη μέτρηση της βαθμονόμησης ενός μοντέλου.

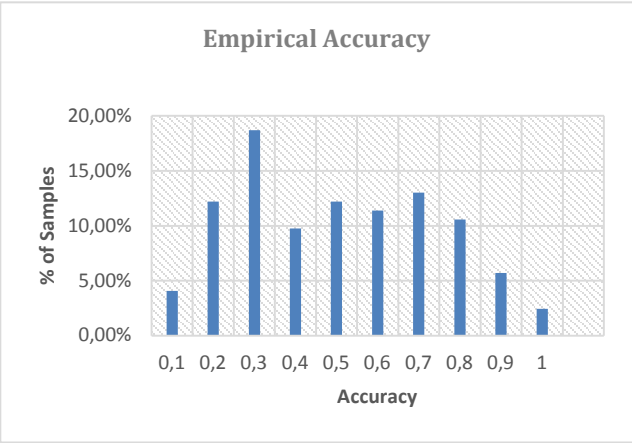
Έστω ένα μοντέλο για το οποίο είναι διαθέσιμος πεπερασμένος αριθμός δειγμάτων. Έστω M προβλέψεις οι οποίες ομαδοποιούνται σε bins(ελληνικά συστοιχίες), η κάθε μία μήκους $1/M$. Έστω B_mτο σετ δεικτών των δειγμάτων των οποίων οι προβλέψεις ανήκουν στο διάστημα $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$για $m \in \{1,2,\dots,M\}$.Ορίζεται ως ακρίβεια του B_m:

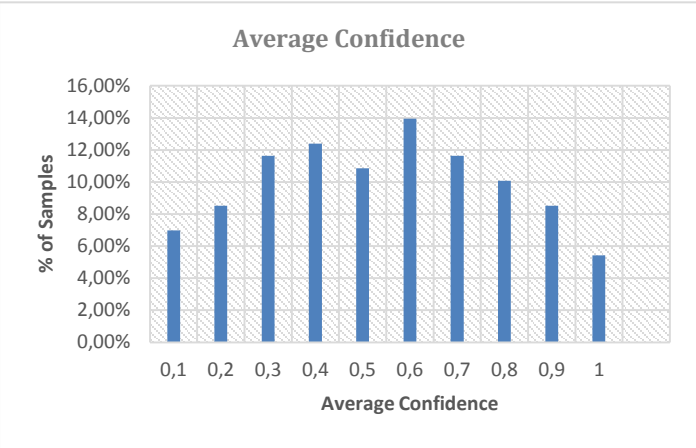
$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\widehat{y_i} = y_i)$$

Ενώ ως μέση σημαντικότητα στο B_m ορίζεται το μέγεθος:

$$conf(B_m) \frac{1}{|B_m|} \sum_{i \in B_m} \widehat{p_i}$$

Όπου $\widehat{p_i}$ είναι η σημαντικότητα για το δείγμα i. Προκειμένου να υπολογιστεί η ακρίβεια του μοντέλου συναρτήσει της σημαντικότητας για δοσμένο πεπερασμένο αριθμό δειγμάτων, αρχικά συσταδοποιούνται τα δείγματα σε διαστήματα ανά εμπιστοσύνη και υπολογίζεται σε κάθε συστάδα η δειγματική ακρίβεια και η εμπειρική εμπιστοσύνη για τα δείγματα.





Διάγραμμα 2–Υπόδειγμα μέσης εμπιστοσύνης και ακρίβειας

Υπολογίζεται ότι το αναμενόμενο σφάλμα βαθμονόμησης (ExpectedCalibrationError or ECE) το οποίο ορίζεται ως:

$$\mathbb{E}_{\hat{p}} = \left[\left|\mathbb{P}\left(\hat{Y} = Y|\hat{P} = p\right) - p\right|\right]$$

Θα είναι ίσο με:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

Αντικειμενικός στόχος είναι η ελαχιστοποίηση του ECE και αυτό επιτυγχάνεται όταν για κάθε συστάδα m ισχύει $acc(B_m) = conf(B_m)$.

Διαχωρισμός σε εποπτευόμενη και μη εποπτευόμενη μηχανική μάθηση

Ο πιο συνήθης διαχωρισμός των ML τεχνικών γίνεται σε Supervised ML (εποπτευόμενη μηχανική μάθηση) και unsupervised ML (μη- εποπτευόμενη), ενώ πολλές βιβλιογραφικές πηγές εισάγουν και τρίτες κατηγορίες όπως πχ semi-supervised ML (υβριδική) (James, etal., 2013).

Η εποπτευόμενη μάθηση χρησιμοποιεί γνωστά δεδομένα για να εκπαιδεύσει ένα νέο μοντέλο, όπως περιγράφηκε στην προηγούμενη παράγραφο για να πραγματοποιήσει προβλέψεις, ενώ η μη εποπτευόμενη μάθηση αναζητά κρυφά μοτίβα ή εσωτερικές δομές σε ένα σετ δεδομένων με στόχο την κατηγοριοποίησή τους. Η μη εποπτευόμενη μάθηση επομένως εξαρτάται μόνο από τα δεδομένα εισόδου και δεν χρησιμοποιεί κανένα trainingdataset.

Η εποπτευόμενη μάθηση, όπως περιγράφηκε προηγουμένως, κατασκευάζει ένα μοντέλο πρόβλεψης χρησιμοποιώντας ένα σύνολο γνωστών δεδομένων. Στόχος είναι η παραγωγή «λογικών» προβλέψεων για τα δεδομένα τα οποία εισάγονται. Για την πραγματοποίηση των προβλέψεων αυτών χρησιμοποιούνται τεχνικές κατηγοριοποίησης (classification) και παλινδρόμησης (regression).

- Classification: Τα μοντέλα κατηγοριοποίησης, ομαδοποιούν τα δεδομένα εισόδου σε κατηγορίες. Παράγονται με τον τρόπο αυτό διακριτές αποκρίσεις ως δεδομένα εξόδου του μοντέλου
- Regression: Τα μοντέλα παλινδρόμησης παράγουν συνεχείς αποκρίσεις και πραγματοποιούν εκτίμηση της απόκρισης της μεταβλητής εισόδου.

Τα μη εποπτευόμενα μοντέλα συνήθως χρησιμοποιούντα για συσταδοποίηση των δεδομένων εισόδου, δηλαδή για την εύρεση ισχυρών δεσμών, δομών και μοτίβων μεταξύ των δεδομένων.

Κεφάλαιο 1 - Γραμμική και λογιστική παλινδρόμηση

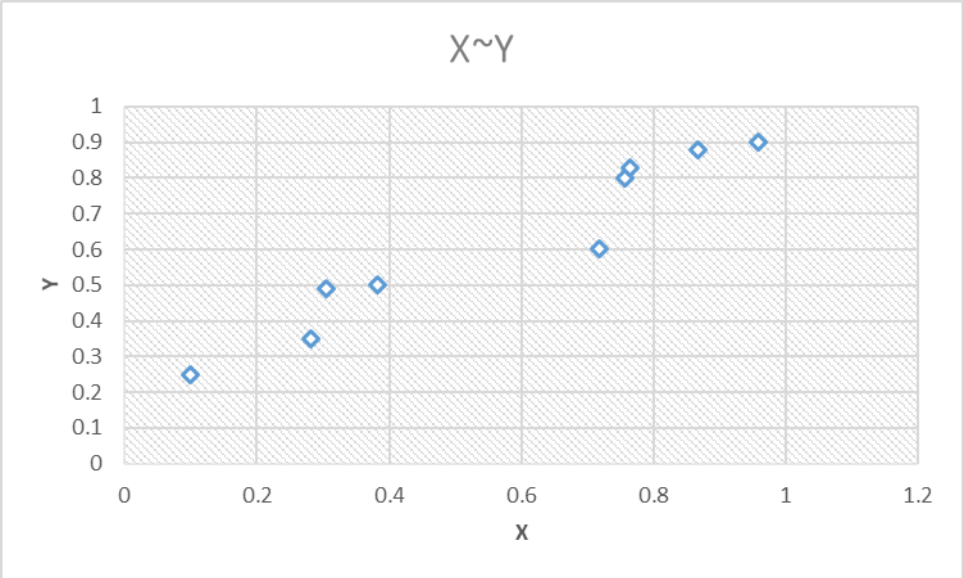
1.1 Γραμμική παλινδρόμηση

1.1.1 Εισαγωγή στη γραμμική παλινδρόμηση

Ο ερευνητής συχνά βρίσκεται στο σημείο όπου έχει συγκεντρώσει επιτυχώς τα δεδομένα του, τα οποία αποτελούνται από διαφορετικές μετρήσεις για δύο ή περισσότερες μεταβλητές, και καλείται να απαντήσει στο ερώτημα πώς σχετίζονται μεταξύ τους οι μεταβλητές αυτές. Η παλινδρόμηση αποτελεί ένα σύνολο τεχνικών οι οποίες εκτιμούν συσχετίσεις μεταξύ μεταβλητών. Η πιο απλή μορφή συσχέτισης είναι η γραμμική η οποία παρά την απλότητά της αποτελεί ένα ισχυρότατο εργαλείο ανάλυσης και εκτίμησης δεδομένων. Η **απλή γραμμική παλινδρόμηση** μοντελοποιεί τη συσχέτιση μεταξύ δύο μεταβλητών υπολογίζοντας μία γραμμική εξίσωση η οποία περιγράφει με τον πιο ακριβή τρόπο τα δεδομένα τα οποία είναι στη διάθεση του ερευνητή. Η μία από τις δύο μεταβλητές καλείται ανεξάρτητη (ή επεξηγηματική μεταβλητή) ενώ η δεύτερη καλείται εξαρτημένη. Σκοπός του ερευνητή είναι η εκτίμηση των τιμών της εξαρτημένης, χρησιμοποιώντας μόνο την τιμή της ανεξάρτητης. Στην περίπτωση που η εξαρτημένη μεταβλητή εμφανίζει σχέσεις εξάρτησης με περισσότερες από μία μεταβλητές, τότε εφαρμόζεται **πολλαπλή γραμμική παλινδρόμηση.**

1.1.2Γραμμική παλινδρόμηση μιας μεταβλητής

Πριν την εφαρμογή της κατάλληλης ευθείας, ο ερευνητής καλείται να συμπεράνει εάν υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών. Για το λόγο αυτό κρίνεται απαραίτητη η εκτέλεση κάποιου ελέγχου συσχέτισης για τις δύο μεταβλητές. Σαν αρχικός και περισσότερο εμπειρικός έλεγχος, πραγματοποιείται η κοινή απεικόνιση των δύο μεταβλητών σε διάγραμμα διασποράς για τον εντοπισμό ή μη υπάρχουσας τάσης (trend).



Διάγραμμα 3 - Υπόδειγμα διαγράμματος διασποράς για συσχετιζόμενες μεταβλητές

Στη συνέχεια επιλέγεται και υπολογίζεται το κατάλληλο μέτρο συσχέτισης (συντελεστές συσχέτισης Pearson, Spearman, Intra-Class, Kendall κ.ο.κ.). Εφόσον εντοπιστεί συσχέτιση (θετική ή αρνητική) υπάρχει πλέον νόημα για την προσαρμογή ευθείας για τη μοντελοποίηση των δεδομένων.

Για την πραγματοποίηση της πρόβλεψης με τη χρήση γραμμικής παλινδρόμησης, το μοντέλο το οποίο τροφοδοτείται είναι μία γραμμική συνάρτηση της μορφής

$$\hat{Y} = \beta_0 + \beta_1 X + \varepsilon$$

Όπου $\hat{Y}$ η εκτίμηση της πραγματικής τιμής της μεταβλητής Y όταν δίνεται η τιμή της μεταβλητής X, και διαφέρει κατά ένα σφάλμα ε τέτοιο ώστε $\hat{Y} - Y = \varepsilon$.

Η ευθεία αυτή δεν είναι μονοσήμαντα ορισμένη, καθώς εξαρτάται από τους συντελεστές της $(\beta_0, \beta_1, \varepsilon)$. Η ευθεία η οποία θα περιγράφει καλύτερα το παραπάνω μοντέλο θα είναι αυτή για την οποία το σφάλμα θα είναι ελάχιστο. Για την ελαχιστοποίηση του σφάλματος έχουν αναπτυχθεί διαφορετικές τεχνικές και παρουσιάζονται συνοπτικά οι πιο διαδεδομένες.

Μέθοδος Ελαχίστων Τετραγώνων

Η πιο διαδεδομένη μέθοδος ελαχιστοποίησης των σφαλμάτων είναι η «Μέθοδος Ελαχίστων Τετραγώνων»(Μπόρα-Σεντα & Μωυσιάδης, 1997) η οποία εμφανίστηκε ολοκληρωμένη για πρώτη φορά από το Γάλλο μαθηματικό Adrien-MarieLegendre[3] και στη συνέχεια εφαρμόστηκε από τον Γερμανό Carl-FriedrichGauss[4] για τον υπολογισμό της τροχιάς του πλανήτη Δήμητρα. Σύμφωνα με τη μέθοδο ελαχίστων τετραγώνων, η ευθεία που προσαρμόζεται με το βέλτιστο τρόπο στα δεδομένα είναι αυτή που ελαχιστοποιεί το άθροισμα των τετραγώνων των σφαλμάτων ε_i. Είναι αυτή για την οποία ελαχιστοποιείται το

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i)\right)^2$$

Συμβολίζουμε με $Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$ το διάνυσμα στήλη το οποίο περιέχει το σύνολο των παρατηρήσεων της εξαρτημένης μεταβλητής. Με $X = \begin{bmatrix} 1 x_{11} \dots x_{1k} \\ \vdots \ \vdots \ \vdots \ \vdots \\ 1 x_{n1} \dots x_{nk} \end{bmatrix}$ τον πίνακα ο οποίος περιέχει το σύνολο των k ανεξάρτητων μεταβλητών X_i, με $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix}$ το διάνυσμα στήλη το οποίο αποτελείται από το σύνολο των συντελεστών του μοντέλου και με $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$ το διάνυσμα στήλη το οποίο αποτελείται από το σύνολο των σφαλμάτων πρόβλεψης. Υποθέτουμε ότι τα σφάλματα είναι ανεξάρτητες τυχαίες μεταβλητές με κοινή κανονική κατανομή $N(0, \sigma^2)$. Επομένως η μέση τιμή του ε θα είναι $E(\varepsilon) = 0$ και η διακύμανση θα

[3]Adrien-MarieLegendre, 18 Σεπτεμβρίου 1752 – 10 Ιανουαρίου 1833) ήταν Γάλλος μαθηματικός ο οποίος πιστώνεται μεγάλο αριθμό συνεισφορών στα μαθηματικά. Οι σημαντικότερες έννοιες οι οποίες εισήγαγε είναι τα πολυώνυμα Legendre και οι μετασχηματισμού Legendre,οι οποίες πήραν το όνομά του.

[4]Ο JohanCarlFriedrichGauss 30 Απριλίου 1777 – 23 Φεβρουαρίου 1855 ήταν Γερμανός μαθηματικός που συνεισέφερε σε πολλά ερευνητικά πεδία της επιστήμης του, όπως η θεωρία αριθμών, η στατιστική, η μαθηματική ανάλυση, η διαφορική γεωμετρία, η γεωδαισία, η αστρονομία και η φυσική. Αποκλήθηκε «ο πρίγκηψ των μαθηματικών» και ο «μεγαλύτερος μαθηματικός μετά τον Αρχιμήδη και τον Ευκλείδη». Σε ηλικία 21 ετών είχε ολοκληρώσει το κύριο έργο του στα καθαρά μαθηματικά, το DisquisitionesArithmeticae,. Αυτό το έργο διαδραμάτισε θεμελιώδη ρόλο στην εδραίωση της θεωρίας αριθμών ως αυτοδύναμου κλάδου των μαθηματικών.

είναι $V(\varepsilon) = \sigma^2 I_n$ όπου I_n ο μοναδιαίος διαγώνιος πίνακας διάστασης n. Σύμφωνα με τους παραπάνω συμβολισμούς το σύστημα προς επίλυση γράφεται

$$Y = X\beta + \varepsilon$$

Συμβολίζοντας το άθροισμα τετραγώνων των σφαλμάτων με

$$SSE = S_\varepsilon(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y_i - X\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$

Για την ελαχιστοποίηση της $S_\varepsilon(\beta)$ πρέπει οι μερικές παράγωγοι ως προς τα β_iνα είναι ίσες με 0, επομένως

$$\frac{\partial S_\varepsilon(\beta)}{\partial \beta} = 2(X'X\beta - X'Y) \Leftrightarrow 0 = 2(X'X\beta - X'Y) \Leftrightarrow X'X\beta = X'Y$$

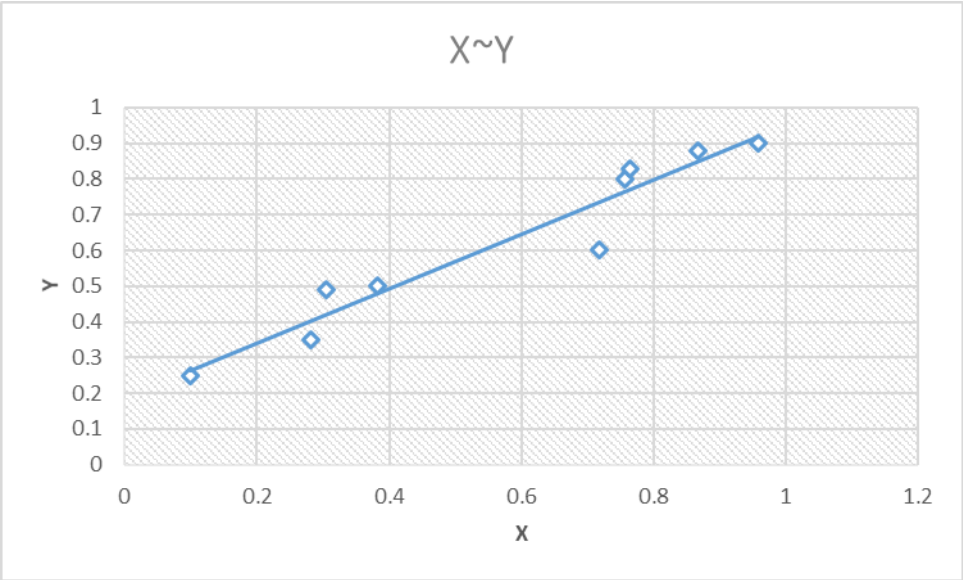
Η τελευταία εξίσωση πινάκων παριστά εξίσωση $k+1$ εξισώσεων $k+1$ αγνώστων, οι οποίοι είναι οι παράμετροι $\beta_0, \beta_1, \dots, \beta_k$. Το σύστημα επομένως είναι γραμμικό και η επίλυσή του δίνεται από τον υπολογισμό του γινομένου $(X'X)^{-1}X'Y$. Η μοναδική του λύση επομένως (εάν $X'X$ αντιστρέψιμος) είναι η

$$\beta = (X'X)^{-1}X'Y$$

Σε περίπτωση που $|X'X| = 0$, τότε ο β μπορεί να εκτιμηθεί μόνο εάν το σύστημα $X'X\beta = X'Y$ είναι συνεπές, δηλαδή εάν $r(X'X) = r(X'X, X'Y)$, όπου $r(X)$ ο βαθμός του πίνακα X. Τότε, αν $(X'X)^-$είναι ένας γενικευμένος αντίστροφος του $X'X$ θα είναι:

$$\beta = (X'X)^-X'Y$$

Η ευθεία η οποία προσαρμόζεται καλύτερα στα δεδομένα του διαγράμματος 4 θα έχει την εξής μορφή



Διάγραμμα 4 - Ευθεία παλινδρόμησης σε διάγραμμα διασποράς δύο συσχετιζόμενων μεταβλητών

Το σύνολο των προβλέψεων θα ανήκει στην ευθεία. Οι κουκίδες είναι οι δειγματικές τιμές, ενώ οι αποστάσεις των κουκίδων από την ευθεία παλινδρόμησης είναι τα σφάλματα.

Αλγόριθμος απότομης καθόδου

Ο αλγόριθμος αυτός έχει ως στόχο τον υπολογισμό της ευθείας η οποία προσαρμόζεται βέλτιστα στις δειγματικές τιμές των μεταβλητών X και Y, μέσω της ελαχιστοποίησης μίας συνάρτησης, η οποία συνήθως καλείται συνάρτηση κόστους (Lossfunction or Costfunction). Έστω h_θ το ζητούμενο γραμμικό μοντέλο, έτσι ώστε $h_\theta(x) = \theta_0 + \theta_1 x$. Στόχος είναι ο προσδιορισμός των (θ_0, θ_1) για τα οποία ελαχιστοποιείται η συνάρτηση κόστους ελαχίστων τετραγώνων η οποία είναι η

$$\min_{\theta_0,\theta_1}\big(J(\theta_0,\theta_1)\big) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

Όπου m το πλήθος των δειγματικών τιμών (x_i, y_i).

Τα βήματα του αλγορίθμου είναι τα εξής με ρυθμό μάθησης L :

1. Απόδοση αρχικών τιμών στα (θ_0, θ_1). Οι συντελεστές (θ_0, θ_1) μεταβάλλονται σε κάθε επανάληψη του αλγορίθμου σύμφωνα με το ρυθμό μάθησης L. Για υψηλή ακρίβεια, επιλέγεται χαμηλός ρυθμός μάθησης.
2. Υπολογισμός της μερικής παραγώγου $\frac{\partial J}{\partial \theta_0}$ και υπολογισμός της τιμής της για τα $(x, y, \theta_0, \theta_1)$.
$$\frac{\partial J(\theta_0,\theta_1)}{\partial \theta_0} = \frac{-2}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)\frac{\partial h_\theta}{\partial \theta_0} = \frac{-2}{m}\sum_{i=1}^{m}(\theta_0 + \theta_1 x_i - y_i)x_i$$
$$= \frac{-2}{m}\sum_{i=1}^{m}(\overline{y_i} - y_i)x_i$$
3. Υπολογισμός της μερικής παραγώγου $\frac{\partial J}{\partial \theta_1}$ και υπολογισμός της τιμής της για τα $(x, y, \theta_0, \theta_1)$.
$$\frac{\partial J(\theta_0,\theta_1)}{\partial \theta_1} = \frac{-2}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)\frac{\partial h_\theta}{\partial \theta_1} = \frac{-2}{m}\sum_{i=1}^{m}(\theta_0 + \theta_1 x_i - y_i) = \frac{-2}{m}\sum_{i=1}^{m}\overline{y_i} - y_i$$
4. Ενημέρωση των τιμών (θ_0, θ_1) ως:
$$\theta_0 = \theta_0 - L \cdot \frac{\partial J(\theta_0,\theta_1)}{\partial \theta_0}$$
$$\theta_1 = \theta_1 - L \cdot \frac{\partial J(\theta_0,\theta_1)}{\partial \theta_1}$$
5. Έλεγχος $J(\theta_0, \theta_1) < \varepsilon$ (ιδανικά $J(\theta_0, \theta_1) = 0$, όπου ε η ακρίβεια. Εάν $J(\theta_0, \theta_1) > \varepsilon$ επιστροφή στο βήμα 2.

Γεωμετρικά ο παραπάνω αλγόριθμος μπορεί να περιγραφεί ως η επιλογή μιας ευθείας (συνήθως της $y = 0$) στην οποία μεταβάλλεται σε κάθε επανάληψη η κλίση της και ο σταθερός της όρος, σύμφωνα με το ρυθμό μάθησης, έως ότου συμπέσει με την ευθεία η οποία εφαρμόζεται στα δεδομένα με βέλτιστο τρόπο.

1.1.3Γραμμική παλινδρόμηση πολλών μεταβλητών

Στην περίπτωση που το σύνολο των ανεξάρτητων μεταβλητών περιέχει πάνω από μία το μοντέλο της γραμμικής παλινδρόμηση θα είναι της μορφής

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Επομένως για τον πλήρη προσδιορισμό του μοντέλου αρκεί ο προσδιορισμός των β_ι συντελεστών. Απαραίτητη προϋπόθεση είναι η ύπαρξη περισσοτέρων των k διαφορετικών τιμών της εξαρτημένης μεταβλητής Y για τιμές των μεταβλητών X_i. Οι συντελεστές β_ι θα προσδιοριστούν με τη βοήθεια των υπαρχόντων δεδομένων, όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης.

Έστω ότι τα δεδομένα έχουν την εξής μορφή:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ & & \dots & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon\_n \end{bmatrix}$$

Το μοντέλο μπορεί να γραφεί ως

$$y = \boldsymbol{X}\beta + \varepsilon$$

Εφαρμόζοντας όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης τη μέθοδο ελαχίστων τετραγώνων, οι ζητούμενοι συντελεστές θα είναι οι

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y$$

Και το μοντέλο θα γραφεί αναλυτικά ως:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \cdots + \hat{\beta}_k X_k$$

1.2 Λογιστική παλινδρόμηση

1.2.1 Εισαγωγή στη λογιστική παλινδρόμηση.

Η ανάλυση λογιστικής παλινδρόμησης (**logisticregressionanalysis**) μελετά τη συσχέτιση μεταξύ μιας εξαρτημένης ποιοτικής μεταβλητής και ενός συνόλου ανεξαρτήτων μεταβλητών. Η διαφορά της λοιπόν με τη γραμμική παλινδρόμηση έγκειται στη φύση της εξαρτημένης μεταβλητής η οποία παίρνει πεπερασμένου πλήθους διακριτές τιμές. Η ονομασία **λογιστική παλινδρόμηση** χρησιμοποιείται όταν η εξαρτημένη μεταβλητή έχει μόνο δύο τιμές απόκρισης, οι οποίες συνήθως αντιστοιχίζονται στο 0 και στο 1, ενώ όταν πρόκειται για μελέτη συσχέτισης της ανεξάρτητης μεταβλητής με περισσότερες από μία τιμές απόκρισης η διαδικασία καλείται **πολλαπλή λογιστική παλινδρόμηση**. Παρά την ουσιαστική διαφορά της εξαρτημένης μεταβλητής στις περιπτώσεις της λογιστικής και της γραμμικής παλινδρόμησης, η χρήση των δύο μεθόδων είναι πρακτικά η ίδια. Σε αντίθεση με άλλες τεχνικές ανάλυσης διακριτών κατανομών, η λογιστική παλινδρόμηση καθώς δεν προϋποθέτει την ύπαρξη κανονικότητας, θεωρείται εύχρηστη (Garson, 2014).

1.2.2 Δημιουργία του μοντέλου της Λογιστικής Παλινδρόμησης

Έστω ότι η εξαρτημένη μεταβλητή είναι δυαδική (binary), επομένως οι τιμές απόκρισης αντιστοιχίζονται στο $\{0,1\}$με το μηδέν να αναπαριστά συνήθως αρνητική απόκριση (ή απουσία της μελετώμενης ιδιότητας) ενώ το ένα αναπαριστά θετική απόκριση (ύπαρξη). Η μέση τιμή της εξαρτημένης μεταβλητής Y θα είναι η αναλογία των θετικών αποκρίσεων. Εάν με p συμβολιστεί ο λόγος αυτός, τότε ο λόγος των αρνητικών απαντήσεων θα είναι ίσος με $1-p$. Επομένως για τη μεταβλητή Y θα ισχύει:

$$Y = \begin{cases} 1, & με\ πιθανότητα\ p \\ 0, με\ πιθανότητα\ 1-p \end{cases}$$

Ο λόγος $\frac{p}{1-p}$ καλείται απόδοση (odds) της μεταβλητής Y και $loggit$ ο λογάριθμος της απόδοσης. Επομένως ο μετασχηματισμός $loggit$ για τη δυαδική μεταβλητή Y με πιθανότητα θετικής απόκρισης πιθανότητας p είναι ο

$$l = loggit(p) = \ln\left(\frac{p}{1-p}\right)$$

Ως λογιστικός μετασχηματισμός (logistictransformation) ορίζεται ο αντίστροφος μετασχηματισμός $loggit$ και είναι ο

$$p = logistic(l) = \frac{e^l}{1+e^l}$$

Στη λογιστική παλινδρόμηση, η δεσμευμένη μεταβλητή Y λαμβάνει M διαφορετικές τιμές (εάν είναι δυαδική τότε $M=2$) και παλινδρομείται πάνω σε ένα σύνολο p ανεξαρτήτων μεταβλητών $X_1, X_2, \dots, X_p$. Έστω $X = (X_1, X_2, \dots, X_p)$ και $B_m = (\beta_{m1}\beta_{m2} \dots \beta_{mp})'$ ένα σύνολο M πινάκων διάστασης $p \times 1$. Το μοντέλο λογιστικής παλινδρόμησης δίνεται από ένα σύνολο M εξισώσεων της μορφής:

$$\ln\left(\frac{p_m}{p_1}\right) = \ln\left(\frac{P_m}{P_1}\right) + \beta_{m1}X_1 + \beta_{m2}X_2 + \cdots + \beta_{mp}X_p = \ln\left(\frac{P_m}{P_1}\right) + XB_m$$

Όπου p_m είναι η πιθανότητα οι τιμές των μεταβλητών $X_1, X_2, \dots, X_p$ να έχουνε αποτέλεσμα m, δηλαδή

$$p_m = p(Y = m|X)$$

Ενώ οι ποσότητες $P_1, P_2, \dots, P_M$ αναπαριστούν τις δειγματικές πιθανότητες για τα αντίστοιχα αποτελέσματα. Ένα αποτέλεσμα το οποίο επιλέγεται αυθαίρετα ονομάζεται τιμή αναφοράς. Για το υπόλοιπο του κεφαλαίου ως τιμή αναφοράς επιλέγεται το $m = 1$. Οι συντελεστές παλινδρόμησης $\beta_{11}, \beta_{12}, \dots, \beta_{1p}$ για την τιμή αναφοράς λαμβάνουν την τιμή 0. Έτσι απομένουν $M - 1$ εξισώσεις για τον προσδιορισμό του μοντέλου. Οι υπόλοιποι συντελεστές θα προσδιοριστούν από τα δεδομένα. Οι εξισώσεις αυτές είναι γραμμικές ως προς $logit(p)$ ενώ ως προς τις πιθανότητες p δεν είναι. Οι μη αντίστοιχες μη γραμμικές εξισώσεις είναι

$$p_m = p(Y = m|X) = \frac{e^{XB_m}}{e^{XB_1} + e^{XB_2} + e^{XB_3} + \cdots + e^{XB_M}} = \frac{e^{XB_m}}{1 + e^{XB_2} + e^{XB_3} + \cdots + e^{XB_M}}$$

Καθώς $e^{XB_1} = 1$ αφού οι συντελεστές παλινδρόμησης είναι ίσοι με μηδέν.

Θέτοντας $\pi_{mj} = p(Y = m|X_j)$ ισχύει

$$\pi_{mj} = \frac{e^{X_j B_m}}{e^{X_j B_1} + e^{X_j B_2} + e^{X_j B_3} + \cdots + e^{X_j B_M}} = \frac{e^{X_j B_m}}{\sum_{q=1}^{M} e^{X_j B_q}}$$

Η πιθανοφάνεια ενός δείγματος Q παρατηρήσεων δίνεται από :

$$l = \prod_{q=1}^{Q} \prod_{m=1}^{M} \pi_{mq}^{y_{mq}}$$

Όπου $y_{mq} = 1$ αν η παρατήρηση qανήκει στα αποτελέσματα και 0 διαφορετικά. Καθώς ισχύει ότι:

$$\sum_{m=1}^{M} y_{mq} = 1$$

Ο μετασχηματισμός $\log likelihood\ L$ θα είναι ίσος με

$$L = \ln(l) = \sum_{q=1}^{Q} \sum_{m=1}^{M} y_{mq} \ln(\pi_{mq}) = \sum_{q=1}^{Q} \sum_{m=1}^{M} y_{mq} \ln\left(\frac{e^{X_q B_m}}{\sum_{r=1}^{M} e^{X_s B_q}}\right)$$
$$= \sum_{q=1}^{Q} \left[\sum_{m=1}^{M} y_{mq} X_q B_m - \ln\left(\sum_{m=1}^{M} e^{X_q B_m}\right)\right]$$

Οι εκτιμήσεις μέγιστης πιθανοφάνειας για τους συντελεστές β είναι αυτές οι οποίες μεγιστοποιούν την εξίσωση $\log likelihood$. Αυτό επιτυγχάνεται εξισώνοντας τις μερικές παραγώγους της L με το μηδέν. Προκύπτει λοιπόν ότι

$$\frac{\partial L}{\partial \beta_{ik}} = \sum_{q=1}^{Q} x_{kq}(y_{im} - \pi_{im})$$

Λόγω της μη γραμμικότητας των παραμέτρων, δεν υπάρχει λύση κλειστού τύπου για τις παραπάνω εξισώσεις. Για το λόγο αυτό, προτείνεται η μέθοδος διαδοχικών προσεγγίσεων για την επίλυση του συστήματος. Η πιο συνηθισμένη μέθοδος επίλυσης τέτοιον συστημάτων είναι η Newton-Raphson[5]μέθοδος (Draper&Smith, 1981), η οποία χρησιμοποιεί τον πίνακα $I(\beta)$ ο οποίος σχηματίζεται από τις δεύτερες μερικές παραγώγους, και τα στοιχεία του είναι τα:

$$\frac{\partial^2 L}{\partial \beta_{ik} \partial \beta_{ik'}} = -\sum_{q=1}^{Q} x_{kq} x_{k'q} \pi_{im}(1-\pi_{im}) = \sum_{q=1}^{Q} x_{kq} x_{k'q} \pi_{im} \pi_{i'm}$$

Ο πίνακας I χρησιμοποιείται διότι ο ασυμπτωτικός πίνακας συσχετίσεων των εκτιμήσεων μέγιστης πιθανοφάνειας είναι ίσος με τον αντίστροφο το I. Επομένως

$$V(\hat{\beta}) = I(\beta)^{-1}$$

Από τον πίνακα V υπολογίζονται τα διαστήματα εμπιστοσύνης για τους συντελεστές παλινδρόμησης, οι λόγοι των αποδόσεων και οι προβλεπόμενες πιθανότητες.

1.2.3Ερμηνεία των συντελεστών

Έστω ότι η μεταβλητή Yείναι δυαδική με pπιθανότητα να έχει τιμή 1 και $1-p$ να έχει τιμή 0. Η εξίσωση της λογιστική παλινδρόμησης είναι ίση με

$$\ln\left(\frac{p}{p-1}\right) = \beta_0 + \beta_1 X$$

Εάν η ανεξάρτητη μεταβλητή X αυξηθεί κατά ένα τότε η εξίσωση γίνεται:

$$\ln\left(\frac{p'}{p'-1}\right) = \beta_0 + \beta_1(X+1) = \beta_0 + \beta_1 X + \beta_1$$

Αφαιρώντας τις δύο εξισώσεις απομονώνεται ο συντελεστής β_1 επομένως:

$$\ln\left(\frac{p'}{p'-1}\right) - ln\left(\frac{p}{p-1}\right) = \beta_0 + \beta_1 X + \beta_1 - (\beta_0 + \beta_1 X) \Leftrightarrow \beta_1 = \ln\left(\frac{odds'}{odds}\right)$$

Επομένως ο συντελεστής β_1 παριστά τον λογαριθμισμένο λόγο μεταβολής των αποδόσεων της εξαρτημένης μεταβλητής, σε περίπτωση που η ανεξάρτητη αυξηθεί κατά 1.

Όταν η εξαρτημένη μεταβλητή έχει παραπάνω από μία μεταβλητές απόκρισης τότε οι εξισώσεις παλινδρόμησης είναι πάνω από μία. Επομένως, οι συντελεστές βij παριστούν το λογαριθμισμένο λόγο μεταβολής των αποδόσεων των i ως προς την τιμή αναφοράς για μεταβολή της μεταβλητής X_j κατά ένα αντίστοιχα.

[5] Στην αριθμητική ανάλυση η μέθοδος Newton-Raphson, είναι μία από τις καλύτερες μεθόδους διαδοχικών προσεγγίσεων για την προσεγγιστική εύρεση των ριζών μιας πραγματικής συνάρτησης. Αυτή η μέθοδος όταν συγκλίνει, συγκλίνει ιδιαίτερα γρήγορα και πιο συγκεκριμένα τετραγωνικά. Η γενική αναδρομική σχέση της μεθόδου είναι η $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$

Κεφάλαιο 2 – Νευρωνικά δίκτυα

2.1 Εισαγωγή στα νευρωνικά δίκτυα

2.1.1 Γενικά περί νευρωνικών δικτύων

Τα τεχνητά νευρωνικά δίκτυα (ArtificialNeuralNetworks) αποτελούν κατηγορία υπολογιστικών συστημάτων εμπνευσμένα από τα βιολογικά νευρωνικά δίκτυα τα οποία αποτελούν τον ανθρώπινο εγκέφαλο των ζώων. Το ίδιο το ANN δεν αποτελεί αλγόριθμο από μόνο του, αλλά ένα πλαίσιο για πολλούς διαφορετικούς αλγορίθμους ML για την επεξεργασία πολύπλοκων δεδομένων. Τα συστήματα αυτά "μαθαίνουν" να εκτελούν εργασίες, εξετάζοντας υποδείγματα, γενικά χωρίς να έχει πραγματοποιηθεί εξειδικευμένος προγραμματισμός. (Gerven&Bohte, 2016)

Ένα ANN βασίζεται σε μια συλλογή από συνδεδεμένες μονάδες ή κόμβους που ονομάζονται τεχνητοί νευρώνες, οι οποίοι χωρίς αυστηρότητα μοντελοποιούν τους νευρώνες σε έναν βιολογικό εγκέφαλο. Κάθε σύνδεση, όπως οι συνάψεις σε έναν βιολογικό εγκέφαλο, μπορεί να μεταδώσει ένα σήμα από έναν τεχνητό νευρώνα στον άλλο. Ένας τεχνητός νευρώνας που λαμβάνει ένα σήμα μπορεί να το επεξεργαστεί και στη συνέχεια να σηματοδοτήσει πρόσθετους τεχνητούς νευρώνες που συνδέονται με αυτό.

Στις κοινές εφαρμογές ANN, το σήμα σε μια σύνδεση μεταξύ των τεχνητών νευρώνων είναι ένας πραγματικός αριθμός και η έξοδος κάθε τεχνητού νευρώνα υπολογίζεται από κάποια μη γραμμική συνάρτηση του αθροίσματος των εισόδων του. Οι συνδέσεις μεταξύ τεχνητών νευρώνων ονομάζονται ακμές ή συνάψεις. Οι τεχνητοί νευρώνες και οι ακμές έχουν συνήθως ένα βάρος το οποίο προσαρμόζεται στη μαθησιακή διαδικασία. Το βάρος αυξάνει ή μειώνει τη δύναμη του σήματος σε μια σύνδεση. Οι τεχνητοί νευρώνες μπορεί να έχουν ένα κατώφλι τέτοιο ώστε το σήμα να αποστέλλεται μόνο αν το συνολικό σήμα ξεπεράσει το όριο αυτό. Οι τεχνητοί νευρώνες συνήθως ταξινομούνται σε επίπεδα (layers). Οι νευρώνες διαφορετικών επιπέδων μπορούν να εκτελούν διαφορετικά είδη μετασχηματισμών στις εισόδους τους. Τα σήματα μετακινούνται από το πρώτο επίπεδο (επίπεδο εισόδου) στο τελευταίο επίπεδο (επίπεδο εξόδου), διασχίζοντας τα διαφορετικά επίπεδα πολλαπλές φορές.

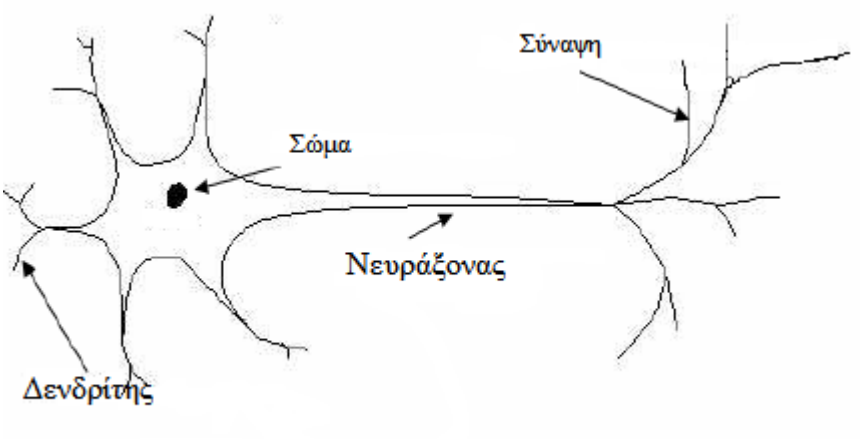
Ο αρχικός στόχος της μοντελοποίησης με τη βοήθεια των ANN ήταν η επίλυση προβλημάτων με παρόμοιο τρόπο με αυτή που θα ακολουθούσε ο ανθρώπινος εγκέφαλος. Ωστόσο, με την πάροδο του χρόνου, τα ANN επικεντρώθηκαν σε συγκεκριμένες δραστηριότητες.

2.1.2Βιολογικός νευρώνας και τεχνητός νευρώνας

Ο εγκέφαλος αποτελείται από μεγάλο αριθμό νευρώνων ο οποίος εκτιμάται ότι ξεπερνάει τα 10 δισεκατομμύρια, τα οποία είναι διασυνδεδεμένα μεταξύ τους σε μεγάλο βαθμό, ξεπερνώντας έτσι το ένα τρισεκατομμύριο συνάψεων. Ο νευρώνας αποτελεί το σημαντικότερο στοιχείο του εγκεφάλου. Συγκεκριμένες ομάδες νευρώνων εκτελούν συγκεκριμένες διεργασίες στον τρόπο με τον οποίο ο εγκέφαλος λειτουργεί. Κάθε νευρώνας αποτελείται από τα εξής τρία βασικά στοιχεία

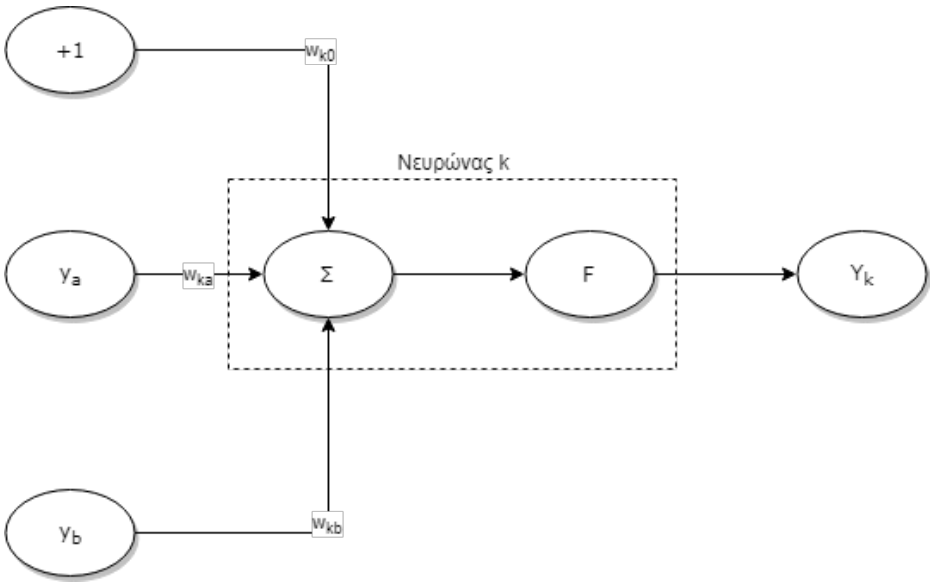
- Δενδρίτες (dendrites) – Λαμβάνουν τα ερεθίσματα ως εισόδους από άλλες πηγές
- Σώμα (Soma) – Επεξεργάζονται τις εισόδους από τους δενδρίτες
- Νευράξονας (Axon) – Μεταφέρουν στις συνάψεις τα αποτελέσματα της επεξεργασίας που πραγματοποιείται στο σώμα.

Στο παρακάτω σχήμα παρουσιάζεται ένας απλοποιημένος βιολογικός νευρώνας.



Σχήμα 2 - Βιολογικός νευρώνας (Εικόνα δημοσιευμένη στο διαδίκτυο, πηγή: https://www.oreilly.com/library/view/deep-learning-with/9781786469786/ea1d0601-2b9b-473b-8860-39a71e5d5429.xhtml)

Στο σχήμα που ακολουθεί παρουσιάζεται η δομή ενός τεχνητού νευρώνα, ο οποίος μοιάζει σε μεγάλο βαθμό με το βιολογικό, αλλά αποτελεί απλούστευσή του.



Σχήμα 3 - Τεχνητός νευρώνας

Οι είσοδοι του νευρώνα αναπαρίστανται με y_aκαι y_b. Κάθε είσοδος πολλαπλασιάζεται με το αντίστοιχο βάρος της σύναψης w_{ki}. Τα γινόμενα αυτά αθροίζονται και αυξάνονται κατά ένα (έτσι ώστε όλοι οι νευρώνες να έχουν θετική έξοδο ακόμη και στην περίπτωση που όλα τα βάρη είναι ίσα με μηδέν). Το άθροισμα αυτό χρησιμοποιείται ως όρισμα στη συνάρτηση ενεργοποίησης F προκειμένου να υπολογιστεί η έξοδος του νευρώνα. Η κατασκευή των περισσότερων τεχνητών νευρώνων είναι παρόμοια, με μικρές δομικές διαφορές.

2.1.3Συναρτήσεις ενεργοποίησης

Η συνάρτηση ενεργοποίησης, η οποία στο σχήμα 3 είναι ο κόμβος F, μετατρέπει το σήμα εισόδου σε σήμα εξόδου. Οι πιο συνηθισμένες συναρτήσεις ενεργοποίησης είναι οι γραμμικές συναρτήσεις ενεργοποίησης, οι συναρτήσεις κατώφλι, οι σιγμοειδείς (λογιστικές και εφαπτομενικές) και η συνάρτηση ενεργοποίησης του Gauss(Urban, 2017).

Γραμμική συνάρτηση ενεργοποίησης

Η γραμμική συνάρτηση ενεργοποίησης για συγκεκριμένη σταθερά cείναι της μορφής

$$y = c \cdot s$$

Μετασχηματίζει γραμμικά την είσοδο του νευρώνα kμε τη βοήθεια σταθεράς c.

Συνάρτηση ενεργοποίησης κατώφλι

Η συνάρτηση κατώφλι επιστρέφει τιμές 0 και 1 ανάλογα εάν η είσοδος είναι μεγαλύτερη από ένα ορισμένο όριοl. Είναι δηλαδή της μορφής

$$y = \begin{cases} +1, s > l \\ 0, s \leq l \end{cases}$$

Λογιστική σιγμοειδής συνάρτηση ενεργοποίησης

Είναι συνάρτηση η οποία επιστρέφει τιμές στο [0,1] και χαρακτηρίζεται από ομαλή μετάβαση. Ο τεχνητός νευρώνας με τον τρόπο αυτό μοιάζει ακόμη περισσότερο στο βιολογικό. Η λογιστική σιγμοειδής είναι της μορφής

$$y = \frac{1}{1 + e^{-s}}$$

Εφαπτομενική σιγμοειδής συνάρτηση ενεργοποίησης

Όμοια με τη λογιστική σιγμοειδή, η εφαπτομενική σιγμοειδής προσδίδει στον τεχνητό νευρώνα ομαλότητα στην ενεργοποίηση με την διαφορά ότι οι τιμές που επιστρέφονται ανήκουν στο διάστημα $[-1,1]$. Είναι της μορφής

$$y = \frac{e^{s} - e^{-s}}{e^{s} + e^{-s}}$$

Συνάρτηση ενεργοποίησης του Gauss

Οι συναρτήσεις ενεργοποίησης του Gausseίναι συναρτήσεις σε σχήμα «καμπάνας», οι οποίες είναι συνεχείς. Επιστρέφουν υψηλές τιμές όταν η είσοδος βρίσκεται κοντά στην ορισμένη μέση τιμή, και χαμηλές όταν βρίσκεται μακριά. Είναι της μορφής

$$y = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

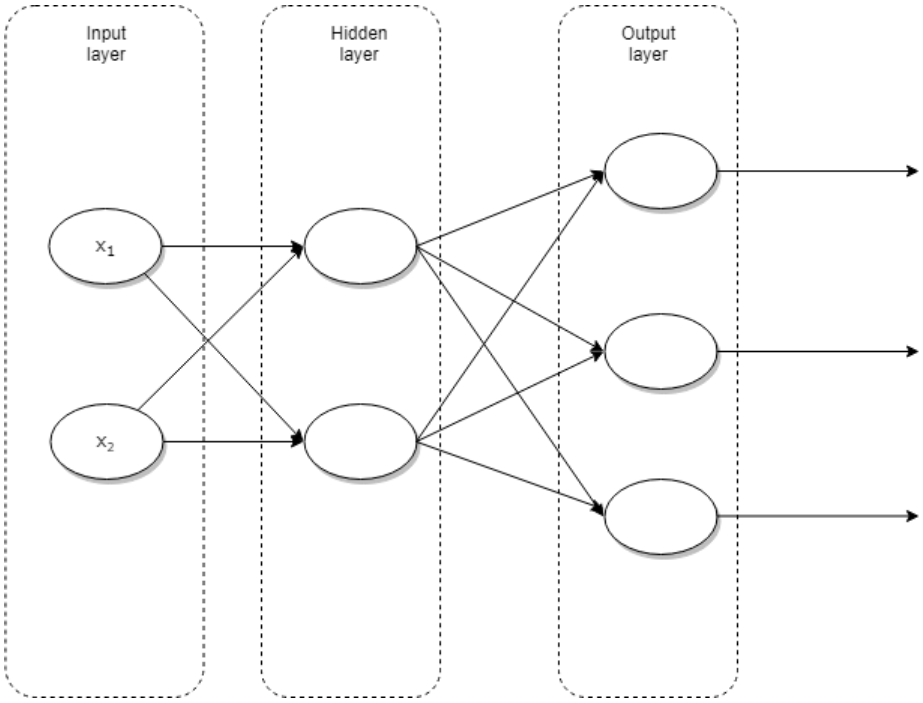
2.2 Αρχιτεκτονική των Νευρωνικών δικτύων

2.2.1Feedforwardνευρωνικά δίκτυα

Τα εμπροσθοτροφοδοτούμενα νευρωνικά δίκτυα (feedforwardNN) αποτελούν την πιο απλή μορφή NN. Είναι νευρωνικά δίκτυα τα οποία εάν παρασταθούν με τη μορφή κατευθυνόμενου γράφου δεν εμφανίζουν κύκλους[6]. Επομένως το επεξεργαζόμενο σήμα ταξιδεύει μόνο προς τη μία κατεύθυνση (προς τα εμπρός) και η είσοδος της χρονικής στιγμής tδεν επηρεάζει ενδεχόμενες εισόδους σε μεταγενέστερο χρόνο. Η βασική αρχή η οποία διέπει τη λειτουργία των feedforwardNN είναι η εξής: Κάθε νευρώνας ενός επιπέδου συνδέεται κατευθυνόμενα μόνο με νευρώνες από επόμενο επίπεδο, καθώς αυτά διατάσσονται από τους νευρώνες εισόδου προς τους νευρώνες εξόδου. Στο σχήμα που ακολουθεί παρουσιάζεται ένα NN το

[6] Κύκλος σε ένα γράφο ονομάζεται η διαδρομή από έναν κόμβο η οποία επιτρέπει την επιστροφή στον κόμβο έναρξης

οποίο αποτελείται από το επίπεδο εισόδου ($1^{ο}$ επίπεδο), τα ενδιάμεσα επίπεδα τα οποία χαρακτηρίζονται ως κρυφά (Hiddenlayer) και το επίπεδο εξόδου (τελευταίο επίπεδο).



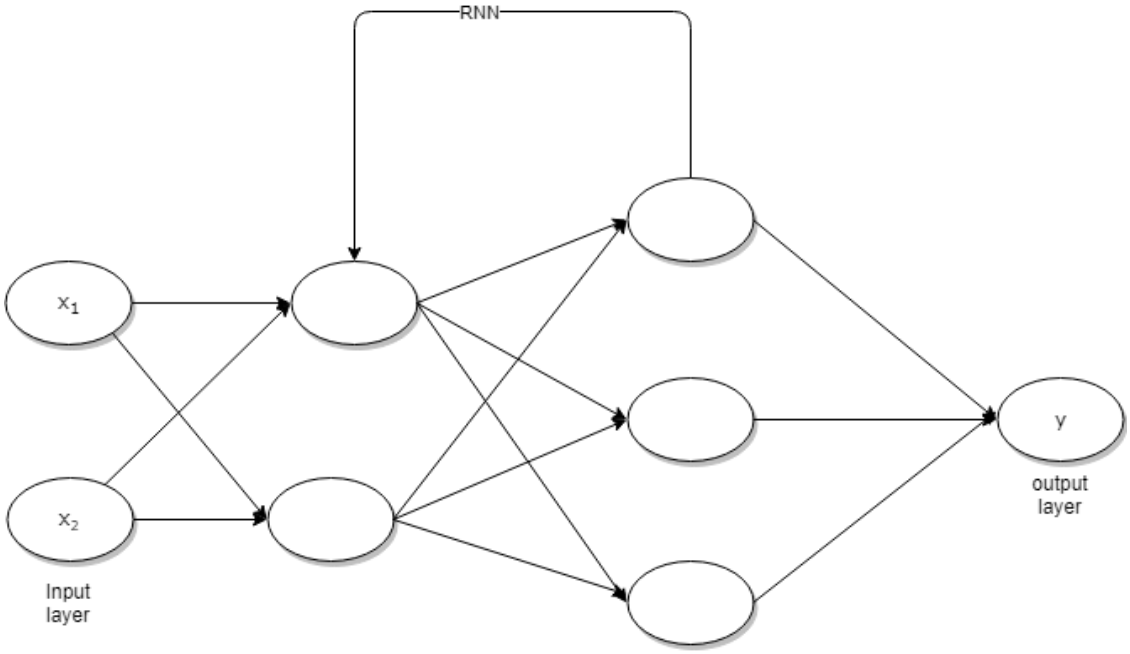
Σχήμα 4 -FeedforwardNN

Τα feedforward NN τα οποία δεν περιλαμβάνουν κρυφό επίπεδο χαρακτηρίζονται ως απλά (singlelayer) ενώ δίκτυα σαν αυτό του σχήματος 4 καλούνται πολύπλοκα (multilayer). Στην περίπτωση του απλού δικτύου, το NN μπορεί να ταυτιστεί με απλά στατιστικά μοντέλα τα οποία μπορούν να παραχθούν από την συνάρτηση ενεργοποίησης που χρησιμοποιεί το NN. Για παράδειγμα, όταν η συνάρτηση ενεργοποίησης ενός απλού feedforwardNN είναι η Λογιστική σιγμοειδής, τότε το NN ταυτίζεται με το μοντέλο λογιστικής παλινδρόμησης.

2.2.2RecurrentΝευρωνικά δίκτυα

ΤαRecurrentΝευρωνικά δίκτυα (RNN) αποτελούν ένα τύπο προηγμένου τεχνητού νευρικού δικτύου που περιλαμβάνει κατευθυνόμενους κύκλους στη μνήμη. Τα RNN χαρακτηρίζονται από την ικανότητα τους να βασίζονται σε παλαιότερους τύπους NN με διανύσματα εισόδου και εξόδου σταθερού μήκους. Μία είσοδος δηλαδή στο χρόνο t θα επηρεάσει την έξοδο του δικτύου για είσοδο η οποία θα εμφανιστεί σε χρόνο μεγαλύτερο από t. Η ιδιότητα αυτή των RNN επιτρέπουν στο δίκτυο να μαθαίνει και να συμπεριφέρεται με παρόμοιο τρόπο όπως ο εγκέφαλος ο οποίος συνηθίζει σε ερεθίσματα και αντιλαμβάνεται τις εισόδους καλύτερα όταν αυτές είναι επαναλαμβανόμενες και δεν εμφανίζονται για πρώτη φορά (Boden, 2001).

Τα RNN μπορούν να παρασταθούν με τη βοήθεια κατευθυνόμενων γράφων στα οποία παρατηρείται η ύπαρξη κύκλων. Ο k νευρώνας του RNN είναι δυνατό να ανατροφοδοτήσει είτε τον εαυτό του (self-feedbackloop) χρησιμοποιώντας την έξοδό του ως είσοδο, ή να τροφοδοτήσει άλλους νευρώνες του ίδιου επιπέδου (layerfeedbackloops). Η ύπαρξη των κύκλων σε συνδυασμό με τη χρήση τελεστών χρονοκαθυστέρησης, επιτρέπει στην έξοδο της εισόδου της χρονικής στιγμής t να προσαρμόσει την είσοδο επόμενης χρονικής στιγμής. Στο σχήμα που ακολουθεί παρουσιάζεται ένα RNN με επαναληπτικό επίπεδο.



Σχήμα 5 - RNN

2.3 Μάθηση Νευρωνικών δικτύων

2.3.1 Διαδικασία μάθησης

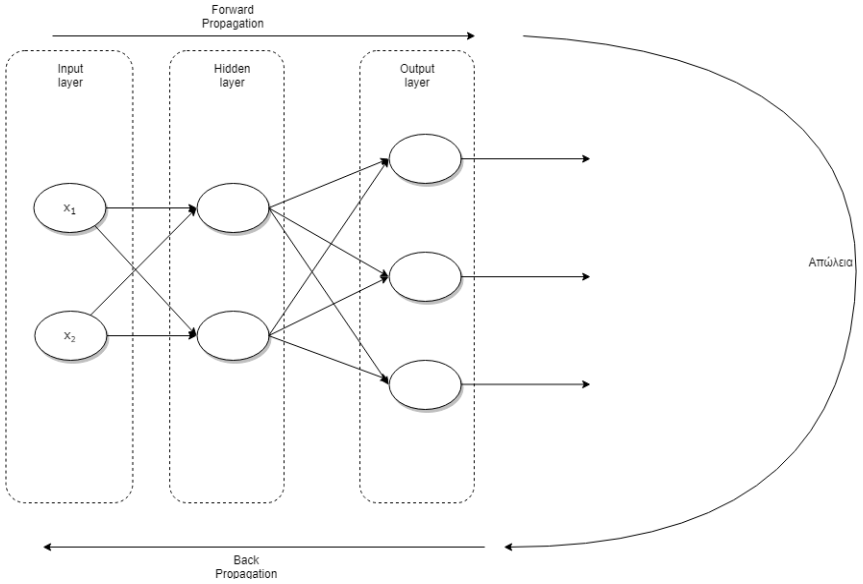
Η εκπαίδευση ενός NN, συνοψίζεται στον υπολογισμό των τιμών των βαρών w_{ij}και των biasesb_j. Η διαδικασία αυτή διαχωρίζει τα NNαπό τους κοινούς αλγορίθμους ελέγχου, καθώς επιτρέπει στο δίκτυο την αυτοπροσαρμογή του ενδεχόμενες αλλαγές σε μελλοντικές εισόδους. Η διαδικασία μάθησης μπορεί να θεωρηθεί ως επαναληπτική διαδικασία «εμπρός μετάδοσης του σήματος» και «πίσω μετάδοσης». Η εμπρός θα καλείται forwardpropagation ενώ η πίσω backwardpropagation(Nazari&Ersoy, 1992).

Η πρώτη φάση της forwardpropagation μάθησης λαμβάνει χώρα όταν στο δίκτυο τροφοδοτούνται τα δεδομένα εκπαίδευσης και αυτά διασχίζουν ολόκληρο το νευρικό δίκτυο προκειμένου να υπολογιστούν οι προβλέψεις τους. Διαβιβάζονται δηλαδή τα δεδομένα εισόδου μέσω του δικτύου με τέτοιο τρόπο, ώστε όλοι οι νευρώνες να εφαρμόζουν τον μετασχηματισμό τους στις πληροφορίες που λαμβάνουν από τους νευρώνες του προηγούμενου επιπέδου και να τις στέλνουν στους νευρώνες του επόμενου επιπέδου. Όταν τα δεδομένα έχουν διασχίσει όλα τα επίπεδα και όλοι οι νευρώνες έχουν πραγματοποιήσει τους υπολογισμούς τους, θα φτάσουν το τελευταίο επίπεδο για να πραγματοποιηθούν οι προβλέψεις τους.

Στη συνέχεια, χρησιμοποιείται μια συνάρτηση απώλειας (lossfunction) για τον υπολογισμό του σφάλματος ή της απώλειας των προβλέψεων που πραγματοποιήθηκαν. Στόχος είναι η ελαχιστοποίηση της συνάρτησης αυτής. Επομένως, καθώς το μοντέλο εκπαιδεύεται, τα βάρη των συνάψεων των νευρώνων θα επαναπροσαρμόζονται έτσι ώστε οι προβλέψεις να συγκλίνουν στις δειγματικές τιμές των δεδομένων εκπαίδευσης.

Όταν ολοκληρωθεί ο υπολογισμός της απώλειας, αυτή μεταδίδεται προς τα πίσω επίπεδα (backpropagation), ξεκινώντας από το επίπεδο εξόδου κατευθυνόμενα προς τα κρυφά επίπεδα εφ’ όσον υπάρχουν. Παρ’ όλα αυτά, οι νευρώνες των κρυφών επιπέδων θα λάβουν μέρος της απώλειας του σήματος ανάλογο με τη συνεισφορά τους στην τελική έξοδο.

Σχηματικά η διαδικασία εκπαίδευσης παρουσιάζεται στο παρακάτω σχήμα



Σχήμα 6 - Διαδικασία μάθησης

2.3.2Ο αλγόριθμος μάθησης

Για την παρουσίαση του αλγορίθμου θα χρησιμοποιηθούν τα παρακάτω μεγέθη:

$w_{ij}^k = το\ βάρος\ της\ σύναψης\ από\ το\ νευρώνα\ i\ στο\ j\ στο\ επίπεδο\ l_k$
$b_i^k\ το\ βάρος\ του\ bias\ \ στο\ νευρώνα\ i\ στο\ επίπεδο\ l_k$
$o_i^k\ η\ έξοδος\ του\ νευρώνα\ i\ στο\ επίπεδο\ l_k$
$a_i^k\ το\ γινόμενο\ των\ αθροισμάτων\ αυξημένο\ κατά\ το\ bias\ για\ τον\ κόμβο\ i\ του\ επιπέδου\ l_k$
$r_k\ ο\ αριθμός\ νευρώνων\ στο\ επίπεδο\ l_k$
$g\ συνάρτηση\ μεταφοράς\ για\ τους\ νευρώνες\ των\ κρυφών\ επιπέδων$
$g_0 συνάρτηση\ μεταφοράς\ για\ τους\ νευρώνες\ του\ επιπέδου\ εξόδου$

Σαν συνάρτηση υπολογισμού κόστους χρησιμοποιείται το μέσο τετραγωνικό σφάλμα

$$E(X) = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y_i} - y_i)^2$$

Όπου (x_i, y_i)το δειγματικό ζεύγος είσοδος-έξοδος.

Βήμα 1- Υπολογισμός προς τα εμπρός

Για κάθε ζεύγος (x_d, y_d) υπολογίζονται τα εξαγόμενα $\widehat{y_d}$, a_j^k, o_j^kγια κάθε νευρώνα j στο επίπεδο k από το επίπεδο 0 στο επίπεδο m το οποίο είναι το επίπεδο εξόδου.

Βήμα 2- Υπολογισμός προς τα πίσω

Για κάθε ζεύγος (x_d, y_d) υπολογίζονται τα εξαγόμενα $\frac{\partial E_d}{\partial w_{ij}^k}$ για κάθε βάρος w_{ij}^k το οποίο συνδέει το νευρώνα iστο επίπεδο $k-1$ με τον jστο επίπεδο k, προχωρόντας από το επίπεδο m προς το επίπεδο 1.

- Υπολογίζεται το σφάλμα $\delta_1^m = g_o'(a_1^m)(\widehat{y_d} - y_d)$
- Υπολογίζονται τα υπόλοιπα σφάλματα $\delta_j^k = g'(a_j^k)\sum_{l=1}^{r^{k+1}} w_{il}^{k+1}\,\delta_l^{k+1}$, από το τελευταίο κρυφό επίπεδο $k = m-1$προς το αρχικό
- Υπολογίζονται οι μερικές παράγωγοι του E_d ως προς w_{ij}^k, $\frac{\partial E_d}{\partial w_{ij}^k} = \delta_j^k o_i^{k-1}$

Βήμα 3 – Συνδυασμός των μερικών κλίσεων για τον υπολογισμό της ολικής κλίσης

$$\frac{\partial E(X)}{\partial w_{ij}^k} = \frac{1}{N}\sum_{d=1}^{N}\frac{\partial E_d}{\partial w_{ij}^k}$$

Βήμα 4 – Ενημέρωση των βαρών

Σύμφωνα με το ρυθμό μάθησης a που έχει επιλεγεί και της ολικής κλίσης υπολογίζονται τα νέα βάρη ως εξής

$$\Delta w_{ij}^k = -\alpha\frac{\partial E(X)}{\partial w_{ij}^k}$$

Επιστροφή στο βήμα 1 και υπολογισμός εκ νέου της συνάρτησης απώλειας. Εφόσον το σφάλμα δεν είναι μηδενικό (ή μικρότερο από το επιλεγμένο κατώφλι) εφαρμόζεται ξανά ο αλγόριθμος.

Κεφάλαιο 3 – Δένδρα αποφάσεων

3.1 Εισαγωγή στα δένδρα απόφασης

3.1.1Τα δένδρα απόφασης

Ένα ευρέως χρησιμοποιούμενο εργαλείο μοντελοποίησης το οποίο αξιοποιεί τις αρχές της ML είναι τα δένδρα αποφάσεων (DecisionTrees – DT). Τα DT όπως και τα περισσότερα μοντέλα τα οποία περιέχονται στην εργασία αυτή αποτελούν μια αλγοριθμική δομή η οποία περιγράφει με υψηλή ακρίβεια το πρόβλημα επίλυσης(Rokach&Maimon, 2015). Τα DTμπορούν να περιγραφούν με τη βοήθεια ενός συνδεδεμένου[7], ακυκλικού και κατευθυνόμενου γράφου, ο οποίος ισοδυναμεί με ένα διάγραμμα ροής (flowchart) το οποίο έχει τα εξής χαρακτηριστικά:

- Υπάρχει ένας και μοναδικός αρχικός κόμβος (Rootnode) του οποίου οι ακμές (στην περίπτωση των δένδρων καλούνται κλαδιά ή branches) είναι ακριβώς δύο και είναι και οι δύο εξερχόμενες. Οι κόμβοι στους οποίους καταλήγουν οι ακμές καλούνται κόμβοι παιδιά (Childrennodes) ενώ οι κόμβοι από τους οποίους ξεκινούν οι ακμές καλούνται γονικοί κόμβοι (Parentnodes).
- Όλοι οι μη αρχικοί κόμβοι έχουν ένα γονέα και δύο παιδιά, εκτός από ένα πεπερασμένο σύνολο τελικών κόμβων οι οποίοι δεν έχουν παιδιά. Οι κόμβοι αυτοί καλούνται φύλλα (leaves)

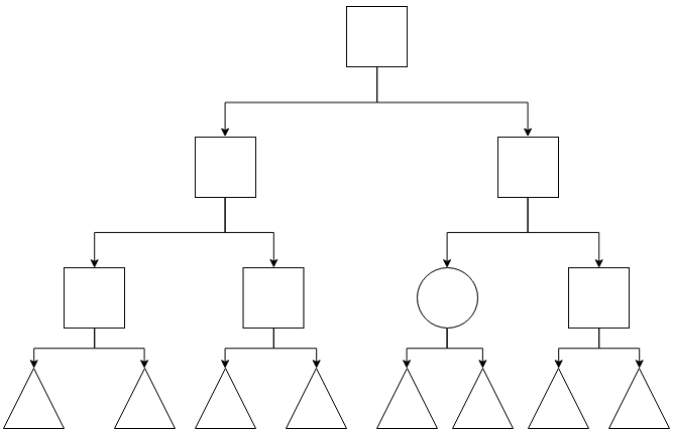
Κάθε κόμβος αποτελεί και μία συνθήκη. Σε κάθε ένα από τα δύο κλαδιά του κόμβου-συνθήκη (ο οποίος καλείται κόμβος απόφασης – DecisionNode) αντιστοιχίζεται και μία τιμή για την επαλήθευση ή μη της συνθήκης. Σε περίπτωση που ο κόμβος παιδί του κόμβου απόφασης είναι φύλλο, τότε η τιμή που έχει λάβει το φύλλο είναι και η τελική απόφαση. Σε αντίθετη περίπτωση, το παιδί είναι και αυτό κόμβος απόφασης ο οποίος θα οδηγήσει σε άλλα δύο παιδιά, μέχρι να οδηγηθεί το δένδρο σε ένα φύλλο. Σε περίπτωση που για έναν κόμβο δεν υπάρχει συνθήκη κλειστού τύπου για την επιλογή της κατάλληλης απόφασης, ανατίθεται σε κάθε ένα κλαδί μία τιμή πιθανότητας επιλογής. Στην περίπτωση αυτή οι κόμβοι καλούνται κόμβοι πιθανοτήτων (ChanceNode).

Έχει επικρατήσει να συμβολίζεται ο κάθε κόμβος ανάλογα με τη λειτουργία του ως εξής:

- Οι κόμβοι απόφασης με τετράγωνο
- Οι κόμβοι πιθανοτήτων με κύκλο
- Τα φύλλα με τρίγωνο

Τα δένδρα όπως περιγράφηκαν μέχρι στιγμής αποτελούν την κατηγορία των ριζωμένων δυαδικών δένδρων (RootedBinaryTree) και αποτελούν την πλέον τυπική μορφή δένδρων. Στο σχήμα που ακολουθεί παρουσιάζεται η τυπική μορφή ενός ριζωμένου δυαδικού δένδρου τεσσάρων επιπέδων. Ως πρώτο επίπεδο ορίζεται ο κόμβος ρίζα. Ως δεύτερο επίπεδο τα παιδιά του κόμβου ρίζα. Ως n-επίπεδο ορίζεται το σύνολο των παιδιών των κόμβων του (ν-1) – επιπέδου.

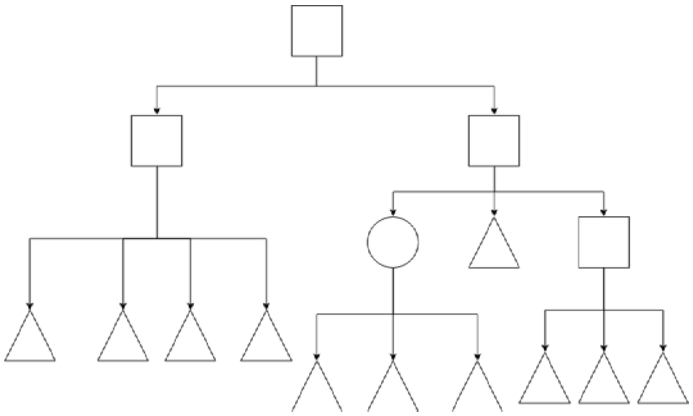
[7] Συνδεδεμένος ονομάζεται ο γράφος για τον οποίο κάθε κόμβος επικοινωνεί με τους υπόλοιπους κόμβους του γράφου



Σχήμα 7 - τυπική μορφή ριζωμένου δυαδικού δένδρου

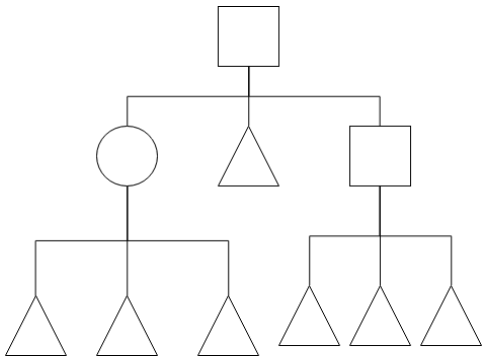
Ανάλογα με τη μορφή του γράφου ο οποίος περιγράφει το δένδρο δημιουργούνται παραλλαγές δένδρων οι οποίες κατηγοριοποιούνται ως εξής:

- Μη δυαδικά δένδρα (Non-binarytree),τα δένδρα για τα οποία δεν ισχύει ο περιορισμός των ακριβώς δύο παιδιών για κάθε γονέα. Παριστάνονται επίσης από μη κυκλικούς, συνδεδεμένους κατευθυνόμενους γράφους.



Σχήμα 8 - Γράφος μη δυαδικού ριζωμένου δένδρου

- Ελεύθερα δένδρα (undirectedtree), τα δένδρα των οποίων τα κλαδιά δεν ορίζουν κατεύθυνση. Ο γράφος ενός ελεύθερου δένδρου είναι μη κατευθυνόμενος, μη κυκλικός και συνδεδεμένος.



Σχήμα 9 - Γράφος ελεύθερου δένδρου

- Δάση (Forest), το σύνολο των δένδρων τα οποία δεν αποτελούν ένα συνδεδεμένο γράφο.

Για το υπόλοιπο της εργασίας ο όρος δένδρο θα χρησιμοποιείται για τα Rootedbinarytrees.Συγκεκριμένα, ένα δένδρο T θα θεωρείται ένα δένδρο πεπερασμένων κόμβων, με ρίζα όπου κάθε κόμβος έχει δύο παιδιά, τα οποία θα χωρίζονται σε αριστερά και δεξιά παιδιά.

Όπως και στις προηγούμενες μεθόδους, το δένδρο θα αποτελέσει τον κανόνα πρόβλεψης των τιμών ενός συνόλου επαλήθευσης , για τις δεδομένες τιμές ενός συνόλου εκπαίδευσης.

3.1.2Κατασκευή των δένδρων απόφασης

Τα δένδρα απόφασης χρησιμοποιούνται για την επίλυση προβλημάτων κατηγοριοποίησης (classification). Η φύση των προβλημάτων τα οποία επιλύονται με τη βοήθεια των DT είναι εντελώς διαφορετική από τα προβλήματα τα οποία επιλύονται με τους αλγορίθμους που έχουν παρουσιαστεί στις προηγούμενες παραγράφους. Η πιο χαρακτηριστική εφαρμογή των DT είναι αυτή της αυτόματης κατηγοριοποίησης του περιεχομένου της εισερχόμενης ηλεκτρονικής αλληλογραφίας σε «Σημαντικό», «Προσφορές», «Εργασία», «Spam» και άλλα. Η περιγραφή της διαδικασίας παρατίθεται στο τέλος του κεφαλαίου. Η δημοτικότητα της κατηγοριοποίησης με τη χρήση DT έχει αυξηθεί καθώς στα DT παρατηρούνται τα εξής πλεονεκτήματα:

- Για την εξαγωγή συμπερασμάτων οι υπολογισμοί που εκτελούνται από τα DT είναι αρκετά απλοί, καθώς εκτελούνται απλές συγκρίσεις σε κάθε κόμβο-απόφασης
- Είναι ικανά να επεξεργαστούν συνεχείς και διακριτές μεταβλητές
- Τα δέντρα απόφασης παρέχουν μια σαφή ένδειξη των τομέων οι οποίοι είναι οι πιο σημαντικοί για την πρόβλεψη ή την ταξινόμηση.

Στην αντίθετη κατεύθυνση, μπορούν να αναγνωριστούν ως μειονεκτήματα των DT τα εξής χαρακτηριστικά:

- Δεν ενδείκνυνται για την πρόβλεψη τιμών συνεχών μεταβλητών
- Τα DT είναι επιρρεπή σε σφάλματα σε προβλήματα ταξινόμησης με πολλές κλάσεις και σχετικά μικρό trainingdataset.
- Τα DT είναι υπολογιστικά δαπανηρά καθώς η διαδικασία ανάπτυξης DTαπαιτεί πολύ μεγάλο αριθμό (απλών) υπολογισμών.

Για την κατασκευή ενός DT έχουν αναπτυχθεί διαφορετικοί αλγόριθμοι οι οποίοι έχουν προσελκύσει πάνω τους το ενδιαφέρον της ερευνητικής κοινότητας, καθώς η υλοποίηση τους είναι απλή αλλά οι εφαρμογές τους εκτείνονται σε ευρύ επιστημονικό φάσμα.

Σύμφωνα με (Wu, etal., 2007), ο αλγόριθμος C4.5 χαρακτηρίστηκε ως ο πιο αποτελεσματικός ανάμεσα στους δέκα πιο δημοφιλείς αλγορίθμους εξόρυξης δεομένων[8] (dataminingalgorithms) με αποτέλεσμα η δημοτικότητά του να εκτοξευθεί. Ο C4.5 αποτελεί την πλέον ενδεδειγμένη μέθοδο κατασκευής ενός DT. Ο C4.5 κατασκευάσει ένα δένδρο απόφασης βασισμένο σε ένα trainingdataset χρησιμοποιώντας την **εντροπία πληροφορίας** (ακολουθεί σύντομη ανάπτυξη της θεωρίας πληροφορίας για την ουσιαστική κατανόηση του αλγορίθμου). Το trainingdataset αποτελείται από ένα πεπερασμένο σύνολο δειγμάτων, όπου το κάθε δείγμα αποτελείται από συγκεκριμένο αριθμό χαρακτηριστικών (attributes). Προκειμένου να κατασκευαστεί το DT, ο αλγόριθμος επιλέγει το χαρακτηριστικό εκείνο το οποίο διαχωρίζει το dataset σε υποσύνολα με αποτελεσματικότερο τρόπο στις κλάσεις κατηγοριοποίησης. Το κριτήριο διαχωρισμού του dataset είναι η διαφορά της εντροπίας πληροφορίας του κάθε χαρακτηριστικού του dataset. Εκείνο το χαρακτηριστικό με τη μέγιστη διαφορά εντροπίας θα αποτελέσει και το χαρακτηριστικό το οποίο θα πραγματοποιήσει το διαχωρισμό.

Θεωρία πληροφορίας

Ως εντροπία πληροφορίας ορίζεται το μέτρο αβεβαιότητας μιας τυχαίας μεταβλητής και αναπτύχθηκε από τον ClaudeShannon[9] το 1948.Η **εντροπία πληροφορίας**(Shannon, 1948), ποσοτικοποιεί το «αναπάντεχο» ενός πειράματος τύχης, και ορίζεται ως

$$H(X) = \sum_i -p_i \log p_i$$

Όπου για την τυχαία μεταβλητή $X = (x_1, x_2, \dots, x_n)$οι αντίστοιχες πιθανότητες εμφάνισης είναι οι $P = (p_1, p_2, \dots, p_n)$ και $\sum_{i=1}^{n} p_i$. Εάν ως βάση του λογαρίθμου επιλέγει το 2, τότε η εντροπία καλείται εντροπία Shannon (η οποία θα χρησιμοποιείται για το υπόλοιπο της εργασίας) και μετράται σε bits. Για κάθε ενδεχόμενο της τυχαίας μεταβλητής Xυπολογίζεται το μέτρο

$$I(x_i) = -p_i \log p_i$$

το οποίο ονομάζεται **πληροφορία** του ενδεχομένου x_i. Σύμφωνα με τα παραπάνω, ως εντροπία της μεταβλητής Xορίζεται η μέση πληροφορία των ενδεχομένων της τυχαίας μεταβλητής.

[8] Τη λίστα με τους 10 πιο δημοφιλείς αλγόριθμος datamining συμπληρώνουν οι: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART

[9] Ο Κλοντ Σάνον (Claude Elwood Shannon, 30 Απριλίου 1916 – 24 Φεβρουαρίου 2001) ήταν Αμερικανός μαθηματικός, ηλεκτρολόγος μηχανικός, και κρυπτογράφος, γνωστός ως «ο πατέρας της θεωρίας πληροφορίας».

Για παράδειγμα, έστω η τυχαία μεταβλητή $X = (x_1, x_2, x_3, x_4)$με $P = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{16}, \frac{3}{16}\right)$, εύκολα υπολογίζονται για το κάθε ενδεχόμενο x_iτα μέτρα $I(x_i)$:

$$I(x_1) = -p_1 log p_1 = -\frac{1}{2}\log\left(\frac{1}{2}\right) = \frac{1}{2}bit$$

$$I(x_2) = -p_2 log p_2 = -\frac{1}{4}\log\left(\frac{1}{4}\right) = \frac{1}{2}bit$$

$$I(x_3) = -p_3 log p_3 = -\frac{1}{16}\log\left(\frac{1}{16}\right) = \frac{1}{4}bit$$

$$I(x_4) = -p_4 log p_4 = -\frac{3}{16}\log\left(\frac{3}{16}\right) = 0.31387bit$$

Επομένως

$$H(X) = \sum_i -p_i \log p_i = \frac{1}{2} + \frac{1}{2} + \frac{1}{4} + 0.317387 = 1{,}70282 \cong 2bit$$

Η τιμή της εντροπίας Shannon συχνά στρογγυλοποιείται στον αμέσως επόμενο ακέραιο καθώς περιγράφει τον αριθμό δυαδικών ερωτήσεων που απαιτούνται (ΝΑΙ-ΟΧΙ) για τον προσδιορισμό του κάθε ενδεχομένου, όταν το κάθε ένα από αυτά μπορεί να θεωρηθεί ως αποτέλεσμα ενός πειράματος τύχης. Η τυχαία μεταβλητή Xγια την οποία υπολογίστηκε η εντροπία, στα προβλήματα κατηγοριοποίησης αντιπροσωπεύει το σύνολο των χαρακτηριστικών του dataset. Επομένως οι πιθανότητες εμφάνισης p_iγια κάθε $x_i \in X$είναι οι σχετικές συχνότητες του κάθε χαρακτηριστικού. Υπολογίζεται λοιπόν η διαφορά της εντροπίας για κάθε ένα χαρακτηριστικό σε σχέση με τα υπόλοιπα και επιλέγεται το χαρακτηριστικό αυτό με τη μέγιστη, ώστε να διαχωριστεί το datasetσε υποσύνολα. Ως κριτήριο διαχωρισμού χρησιμοποιείται ο κανονικοποιημένος ρυθμός απόκτησης πληροφορίας[10]Shannonο οποίος ορίζεται ως

$$IG(X|a) = \frac{H(X) - H(X|a)}{\log(n)}$$

Όπου nείναι το πλήθος των χαρακτηριστικών του datasetκαι a το χαρακτηριστικό το οποίο εξετάζεται, και $H(X|a)$ η δεσμευμένη εντροπία η οποία υπολογίζεται ως

$$H(Y|X) = \sum_{x \in X, y \in Y} -p(x,y) \log \frac{p(x,y)}{p(x)}$$

Ο τρόπος με τον οποίο πραγματοποιείται η διαδικασία διαχωρισμού και ταξινόμησης του συνόλου περιγράφεται στον αλγόριθμο που ακολουθεί(Quinlan, 1993).

Ο αλγόριθμος C4.5

Βήμα 1ο (Έλεγχος των εξής βασικών συνθηκών)

[10] Ο ρυθμός απόκτησης πληροφορίας εισήχθη από του Solomon Kullback και RichardLeibler το 1951 και αποτελεί προσανατολισμένο μέτρο απόκλισης δύο τυχαίων μεταβλητών. Συναντάται στη θεωρία πληροφορίας με την ορολογία απόκλιση Kullback-Leibler.

- Αν όλα τα δείγματα του trainingdataset ανήκουν στην ίδια κλάση, τότε δημιουργείται ένας κόμβος φύλο για την κλάση αυτή.
- Αν κανένα από τα χαρακτηριστικά δεν παρέχει πληροφορία, τότε δημιουργείται ένας κόμβος απόφασης στο αμέσως υψηλότερο επίπεδο χρησιμοποιώντας την αναμενόμενη τιμή της κλάσης.
- Εάν η συναντηθεί παρατήρηση που ανήκει σε προηγούμενη κλάση, τότε δημιουργείται νέος κόμβος απόφασης στο αμέσως υψηλότερο επίπεδο χρησιμοποιώντας την αναμενόμενη τιμή της κλάσης.

Βήμα 2^{ο} (Υπολογισμός πληροφορίας)

Για κάθε χαρακτηριστικό του dataset υπολογίζεται η πληροφορία σε περίπτωση διαχωρισμού του δείγματος βάσει του χαρακτηριστικού αυτού.

Βήμα 3^{ο} (Μέγιστη πληροφορία)

Υπολογισμός του χαρακτηριστικού εκείνου με τη μέγιστη πληροφορία

Βήμα 4^{ο} (Δημιουργία κόμβου απόφασης)

Δημιουργία ενός κόμβου απόφασης για το διαχωρισμό σύμφωνα με το χαρακτηριστικό με τη μέγιστη πληροφορία

Βήμα 5^{ο} (Επαναληπτικό βήμα)

Επιστροφή στο Βήμα 1 και εφαρμογή του αλγορίθμου πάνω στα υποσύνολα που προκύπτουν μετά το διαχωρισμό του datasetσύμφωνα με το χαρακτηριστικό μέγιστης πληροφορίας. Οι κόμβοι που προκύπτουν προστίθενται ως παιδιά του κόμβου που δημιουργήθηκε στο Βήμα 4.

3.1.3Κλάδεμα ενός δένδρου απόφασης

Ένα από τα κύρια μειονεκτήματα των δένδρων απόφασης όπως αναφέρθηκε είναι η πολυπλοκότητα τους. Για το λόγο αυτό αναπτύσσονται τεχνικές οι οποίες έχουν ως στόχο την απλούστευση των DT καταργώντας τμήματα του δένδρου τα οποία δεν αξιοποιούν ουσιαστική δομική πληροφορία του trainingdataset, και πολλές φορές είναι ικανά να οδηγήσουν σε overfitting (υπερπροσαρμογή), δηλαδή το μοντέλο να προσαρμοστεί σε τόσο υψηλό βαθμό στο trainingdataset, που να καταστεί «ανίκανο» να πραγματοποιήσει οποιαδήποτε πρόβλεψη για δεδομένα εκτός trainingdataset. Η διαδικασία αυτή της μείωσης του μεγέθους ενός DT καλείται κλάδεμα (pruning) και έχει ως στόχο τη μείωση της πολυπλοκότητας του DT αλλά και την αύξηση της ακρίβειας της πρόβλεψης(Esposito, etal., 1997).

Το κλάδεμα είναι δυνατό να πραγματοποιηθεί είτε κατά τη δημιουργία του DT, έτσι ώστε το δένδρο που θα κατασκευαστεί να έχει περιορισμένο μέγεθος και να έχει αποφευχθεί το overfitting, είτε να πραγματοποιηθεί μετά την κατασκευή του δένδρου. Η δεύτερη περίπτωση είναι και η πιο αποτελεσματική διότι δεν είναι πάντα εύκολο να προσδιοριστεί ταυτόχρονη ολοκλήρωση της κατασκευής του δένδρου και του κλαδέματος.

Για το επιτυχημένο κλάδεμα ενός δένδρου, το σημαντικότερο βήμα είναι η κατάλληλη επιλογή του κριτηρίου ολοκλήρωσης της διαδικασίας μείωσης του μεγέθους του δένδρου. Για την επίτευξη του, χρησιμοποιούνται οι εξής μέθοδοι:

- Πριν την κατασκευή του δένδρου, αποσπάται από το trainingdataset ένα υποσύνολο το οποίο θα χρησιμοποιηθεί ως σύνολο επαλήθευσης (validationdataset). Το μέγεθός του συνήθως επιλέγεται ως 20% του αρχικού dataset. Το σύνολο αυτό θα χρησιμοποιηθεί για την αξιολόγηση του δένδρου που κατασκευάστηκε.
- Χρησιμοποιείται το trainingdataset (το οποίο πλέον είναι το 80% του αρχικού) για την κατασκευή του δένδρου. Επιλέγεται κατάλληλο στατιστικό μέτρο ή κατάλληλος έλεγχος υποθέσεων (Μέσο τετραγωνικό σφάλμα, έλεγχος - X^2,κλπ) για την εκτίμηση της βελτίωσης της πρόβλεψης το ενδεχόμενο της επέκτασης ενός συγκεκριμένου κόμβου.

Ο αλγόριθμος C4.5 ακολουθεί τεχνική κλαδέματος η οποία αντικαθιστά ένα ολόκληρο τμήμα του δένδρου με φύλλο σε περίπτωση που η αφαίρεσή του μειώνει το ρυθμό σφάλματος ταξινόμησης. Ως ρυθμός σφάλματος ταξινόμησης ορίζεται το εξής:

Έστω ότι ο κόμβος ένας κόμβος ταξινομεί N στοιχεία με Eσφάλματα. Ο λόγος $f = \frac{N}{E}$καλείται ρυθμός σφάλματος ταξινόμησης.Η διαδικασία κλαδέματος που ακολουθεί ο C4.5 είναι η εξής:

Βήμα 1^{ο} (Υπολογισμός του ρυθμού σφάλματος ταξινόμησης)

Υπολογισμός του ρυθμού σφάλματος ταξινόμησης για κάθε κόμβο του τελευταίου επιπέδου.

Βήμα 2^{ο}(Υπολογισμός του ρυθμού σφάλματος ταξινόμησης για το γονικό κόμβο)

Μετατροπή των κόμβων του προ-τελευταίου επιπέδου σε φύλλα και υπολογισμός των ρυθμών σφαλμάτων ταξινόμησης για κάθε ένα φύλλο.

Βήμα 3^{ο}(Σύγκριση των ρυθμών σφαλμάτων ταξινόμησης)

Σύγκριση των ρυθμών σφαλμάτων του κλαδεμένου δένδρου με αυτούς του πλήρους. Σε περίπτωση που το κλάδεμα ενός κόμβου μειώνει το ρυθμό σφάλματος, τότε ο κόμβος αυτός κλαδεύεται.

Βήμα 4^{ο}(Επαναληπτικό βήμα)

Επιστροφή στο Βήμα 1 εφόσον έχει πραγματοποιηθεί κλάδεμα τουλάχιστον ενός κόμβου.

3.2 Εφαρμογή των δένδρων απόφασης στην ταξινόμηση αλληλογραφίας

3.2.1E-mailSPAMfiltering

Τα δένδρα απόφασης λόγω της απλότητας της κατασκευής τους και της λειτουργίας τους εφαρμόζονται σε μια πληθώρα τομέων. Στο διαδίκτυο, υπάρχει μεγάλος όγκος πληροφορίας ο η οποία είναι αποθηκευμένη σε διάφορες μορφές (εικόνες, ή ακολουθίες γραμμάτων – λέξεις) οι οποίες εάν μεταφραστούν στη γλώσσα της στατιστικής θα αποτελούν ποιοτικές μεταβλητές. Συνεπώς η μοντελοποίηση των δομών του διαδικτύου συχνά ανάγεται σε πρόβλημα κατηγοριοποίησης, για τα οποία τα δένδρα απόφασης αποτελούν μία εκ των καταλληλότερων τεχνικών επίλυσης.

Ένα τέτοιο πρόβλημα είναι και το πρόβλημα ταξινόμησης της εισερχόμενης αλληλογραφίας στο διαδίκτυο, σε κλάσεις ανάλογα με το περιεχόμενο της. Η λογική πίσω από τη διαδικασία ταξινόμησης είναι η εξής:

Ο αρχικός στόχος είναι ο διαχωρισμός της χρήσιμης αλληλογραφίας από την άχρηστη. Επομένως αναγνωρίζεται το κείμενο της εισερχόμενης αλληλογραφίας, λέξη προς λέξης, και εάν εντοπισθούν κάποιες λέξεις οι οποίες χρησιμοποιούνται συχνά σε mailπροωθητικού υλικού, τότε το mailχαρακτηρίζεται ως πιθανό spam (ηλεκτρονικό μήνυμα το οποίο έχει αποσταλεί μαζικά με σκοπό την προώθηση προϊόντων ή ιδεών).

Το χαρακτηρισμένο mail περνάει από επόμενες φάσεις ελέγχου, ανάλογα με την ευαισθησία και την ακρίβεια του αλγορίθμου, οι οποίες μπορεί να αφορούν σε έλεγχο για τον αριθμό των εμφανίσεων των λέξεων οι οποίες αφορούν σε spamαλληλογραφία, είτε στον αριθμό των παραληπτών, είτε στο αν ο αποστολέας ανήκει στο βιβλίο διευθύνσεων του παραλήπτη.

Με παρόμοια διαδικασία τα mail τα οποία δεν χαρακτηρίστηκαν ως spamμπορούν να ταξινομηθούν ως επιβλαβή (να περιέχουν κάποιο κακόβουλο λογισμικό) χρησιμοποιώντας ως φίλτρο ελέγχου την ύπαρξη κάποιου εκτελέσιμου αρχείου (.exe) ή σε mailεπαγγελματικά, δηλαδή το περιεχόμενο τους να αφορά. Στην περίπτωση αυτή αναζητούνται λέξεις συναφείς με εργασιακό περιβάλλον (πχ συνεργάτης, συνάντηση, project,κλπ.).

Η παραπάνω διαδικασία βρίσκει εφαρμογή στο διαδίκτυο στην προσωποποιημένη διαφήμιση, στον τρόπο δηλαδή με τον οποίο εμφανίζονται οι διαφημίσεις που συναντά ο χρήστης κατά την περιήγησή του. Για κάθε χρήστη, συλλέγονται πληροφορίες για τις σελίδες που έχει επισκεφτεί στο παρελθόν με σκοπό τη δημιουργία ενός διαδικτυακού προφίλ. Στη συνέχεια, σε περίπτωση που ο χρήστης βρεθεί σε μέρος όπου υπάρχει διαφήμιση, ο ιστοχώρος επιλέγει την πιο κατάλληλη, αναζητώντας λέξεις κλειδιά οι οποίες να ταιριάζουν στο περιεχόμενο του προφίλ του χρήστη.

3.2.2 Ο αλγόριθμος ταξινόμησης

Κάθε εισερχόμενο email αντιστοιχίζεται σε μία λίστα με τα εξής 4 στοιχεία (Priyatharsini, etal., 2017)

- Ταυτότητα αποστολέα (ηλεκτρονική διεύθυνση αποστολέα)
- Θέμα (Ακολουθία λέξεων)
- Σώμα (Ακολουθία λέξεων)
- Συνημμένα αρχεία (όνομα αρχείου)

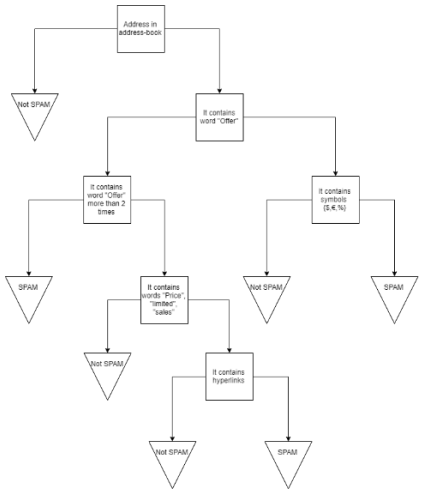
Ένα τυπικό spamemail έχει τα εξής χαρακτηριστικά: Ο αποστολέας δεν βρίσκεται στη λίστα διευθύνσεων του χρήστη, το θέμα του και το κυρίως σώμα του περιλαμβάνει λέξεις οι οποίες προτρέπουν τον παραλήπτη να αγοράσουν κάποιο προϊόν (πχ προσφορά, τιμή, έκπτωση, ευκαιρία) καθώς και νομισματικούς χαρακτήρες (€, $ κλπ.), και περιέχει εξωτερικούς συνδέσμους προς τη σελίδα πώλησης ή προς κάποιο προϊόν. Επίσης, δεν περιέχονται συνημμένα αρχεία.

Για τον αλγόριθμο επιλέγονται οι εξής συνθήκες:

- Βρίσκεται η διεύθυνση του αποστολέα στη λίστα διευθύνσεων του παραλήπτη.
- Υπάρχει υπερσύνδεσμος στο σώμα του κειμένου.
- Υπάρχει η λέξη προσφορά στο σώμα του κειμένου.

- Υπάρχει η λέξη προσφορά τρεις ή περισσότερες φορές στο σώμα του κειμένου.
- Υπάρχει κάποια από τις λέξεις τιμή, έκπτωση ή ευκαιρία στο σώμα του κειμένου.
- Υπάρχει νομισματικός χαρακτήρας στο σώμα του κειμένου.
- Υπάρχουν συνημμένα αρχεία.

Σύμφωνα με τα παραπάνω, ένα δένδρο απόφασης θα έχει την εξής μορφή



Σχήμα 10 - Δένδρο ταξινόμησης εισερχόμενης αλληλογραφίας

Κεφάλαιο 4 SupportVectorMachine

4.1 Εισαγωγή

4.1.1 Εισαγωγή στην SupportVectorMachine

Οι Μηχανές Διανυσμάτων υποστήριξης (SupportVectorMachines – SVM) αποτελούν ένα σύνολο αλγορίθμων ML οι οποίοι χρησιμοποιούνται ευρέως στην επίλυση προβλημάτων ταξινόμησης. Κύριο πλεονέκτημα των SVMείναι η ικανότητά του να ταξινομεί αντικείμενα τα οποία αποτελούνται από μεγάλο αριθμό χαρακτηριστικών(Vapnik, 1995).

Η λειτουργία των SVMμπορεί να περιγραφεί από τον προσδιορισμό του υπερεπιπέδου εκείνου το οποίο διαχωρίζει με τον βέλτιστο δυνατό τρόπο τα trainingdata πάνω στα οποία θα προσαρμοστεί το μοντέλο.

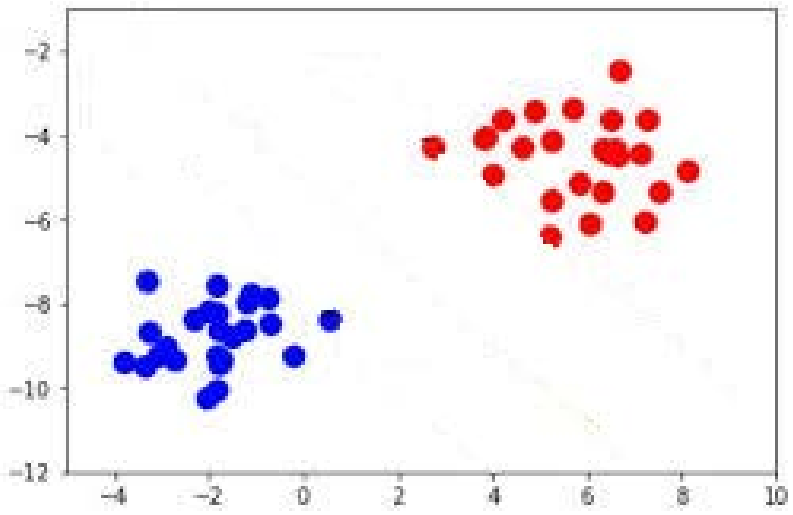
Με τον όρο υπερεπίπεδο ενός διανυσματικού χώρου V διάστασης d ορίζεται κάθε υποχώροςH με διάσταση $d-1$. Σύμφωνα με τον ορισμό που προηγήθηκε:

- Τα υπερεπίπεδα μίας ευθείας (Διανυσματικός χώρος με διάσταση 1) είναι τα σημεία της (Υποχώροι με διάσταση 0 – Τετριμμένοι υποχώροι)
- Τα υπερεπίπεδα του επιπέδου (Διανυσματικός χώρος με διάσταση 2) είναι οι ευθείες οι οποίες «κείνται» στο επίπεδο (υποχώροι με διάσταση 1) και γενικότερα οι καμπύλες του επιπέδου.
- Τα υπερεπίπεδα του χώρου τριών διαστάσεων είναι τα επίπεδα που ανήκουν σε αυτόν
- κ.ο.κ

Έστω ότι το πρόβλημα κατηγοριοποίησης περιλαμβάνει τον χαρακτηρισμό ενός αριθμού αντικειμένων σε «Μπλε» και «Κόκκινα». Κάθε αντικείμενο περιγράφεται από δύο χαρακτηριστικά (X, Y)όπου το Xλαμβάνει τιμές σε ένα διάστημα D_Xκαι το Yσε ένα διάστημα

D_Y. Σε κάθε αντικείμενο αντιστοιχεί μοναδική θέση στο επίπεδο XYανάλογα με τις τιμές των δύο χαρακτηριστικών του.

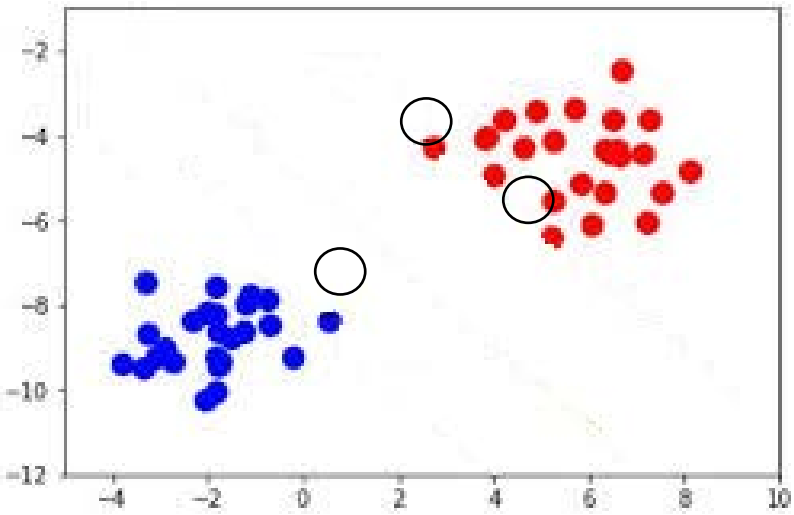
Έστω ότι το trainingdataset όπως φαίνονται στο σχήμα



Σχήμα 11 - Trainingdatasetγια εφαρμογή SVM

Στόχος είναι η ανάπτυξη ενός μοντέλου το οποίο θα αποφαίνεται εάν ένα μη χρωματισμένο αντικείμενο είναι «Μπλε» ή «Κόκκινο».

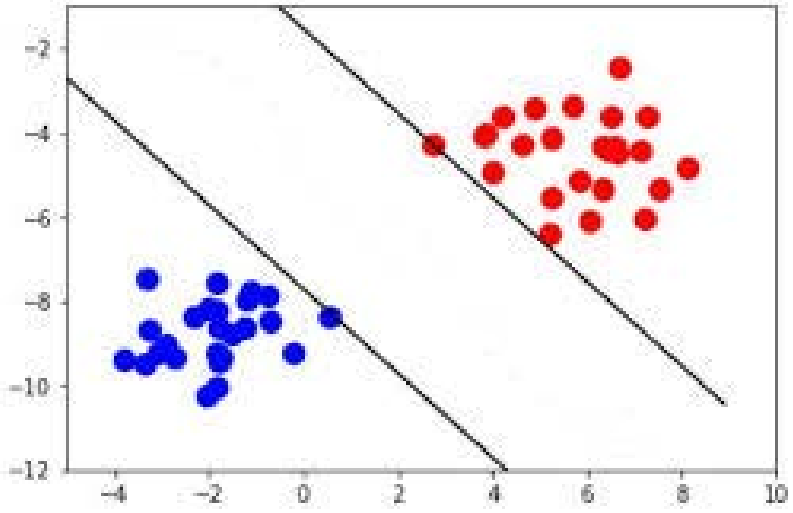
Αρχικά προσδιορίζονται οι εκπρόσωποι των δύο κατηγοριών οι οποίοι βρίσκονται πιο κοντά στην κατηγορία που δεν ανήκουν. Τα αντικείμενα αυτά καλούνται support vectors (Διανύσματα υποστήριξης)



Σχήμα 12–Επιλογή SupportVectors

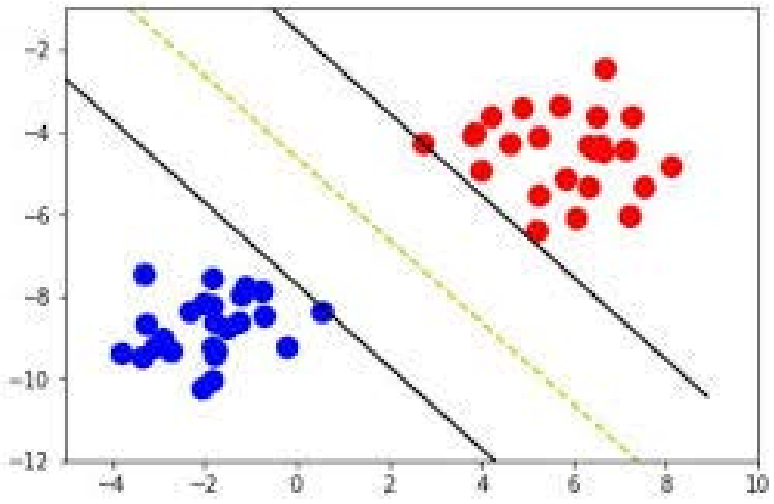
Ως επόμενο βήμα υπολογίζονται τα όρια στα οποία θα βρίσκεται το υεπερεπίπεδο το οποίο θα διαχωρίζει τα «Μπλέ» από τα «Κόκκινα». Προσαρμόζεται στη συνέχεια μία ευθεία η οποία διέρχεται από τα δύο supportvectors της μίας ομάδας και στη συνέχεια προσαρμόζεται η

παράλληλή της η οποία διέρχεται από το supportvector της άλλης ομάδας. Οι δύο παράλληλες ευθείες διαχωρίζουν τις δύο κατηγορίες όπως φαίνεται στο σχήμα 13.



Σχήμα 13 - Προσαρμογή των ορίων του υπερεπιπέδου

Το υπερεπίπεδο το οποίο θα διαχωρίζει τις δύο κατηγορίες αντικειμένων με τον καλύτερο δυνατό τρόπο θα βρίσκεται εντός των δύο αυτών ευθειών. Ως τελικό στάδιο επιλέγεται το υπερεπίπεδο το οποίο βρίσκεται εντός των ορίων σύμφωνα με ένα κριτήριο επιλογής ανάμεσα σε μια πληθώρα διαδικασιών ανάπτυξης αλγορίθμων κατηγοριοποίησης SVM.



Σχήμα 14 - Κατηγοριοποίηση με SVM

Εάν σχεδιαστεί ένα νέο μη κατηγοριοποιημένο αντικείμενο με χαρακτηριστικά (X_1, Y_1), σε περίπτωση που βρίσκεται από την πλευρά του επιπέδου πουορίζει η διακεκομμένη ευθεία και τα κόκκινα αντικείμενα τότε θα είναι «Κόκκινο», ενώ αν βρίσκεται από την διακεκομμένη ευθεία και προς τα μπλε αντικείμενα θα είναι «Μπλε». Στην περίπτωση που ένα αντικείμενο ανήκει στην διακεκομμένη ευθεία ορίζεται εξ' αρχής εάν θα είναι «Κόκκινο» ή «Μπλε». Η απόσταση των δύο μαύρων ευθειών του σχήματος καλείται περιθώριο, ενώ η διακεκομμένη ευθεία όριο απόφασης (Decisionboundary).

Η διαδικασία η οποία περιγράφηκε είναι ίσως η πιο απλή διαδικασία ταξινόμησης με χρήση SVM. Η πολυπλοκότητα του προβλήματος έγκειται στους εξής παράγοντες:

- Στο πλήθος των χαρακτηριστικών του κάθε αντικειμένου τα οποία θα ορίσουν και τη διάσταση του χώρου αναπαράστασης των αντικειμένων
- Στο πλήθος των κλάσεων των αντικειμένων, καθώς η ύπαρξη περισσοτέρων των δύο κλάσεων επιτάσσει τον προσδιορισμό περισσοτέρων των δύο υπερεπιπέδων
- Στη γραμμικότητα ή μη των decisionboundaries

4.1.2 Προσδιορισμός του ορίου απόφασης για γραμμικώς διαχωριζόμενα δεδομένα – Το βέλτιστο υπερεπίπεδο

Το σημαντικότερο ρόλο στον προσδιορισμό του υπερεπιπέδου κατηγοριοποίησης παίζουν τα supportvectors καθώς από προσδιορίζουν το υπερεπίπεδο απόφασης, και οποιαδήποτε μεταβολή τους μεταβάλει και τα όρια απόφασης. Καθώς στα γραμμικός διαχωριζόμενα δεδομένα υπάρχουν άπειρα υπερεπίπεδα τα οποία μπορούν να διαχωρίσουν το χώρο σε δύο υποχώρους τέτοιους ώστε έκαστος να περιέχει μία κατηγορία δεδομένων, η SVMπροσδιορίζει το βέλτιστο υπερεπίπεδο. Για να επιτευχθεί αυτό, μεγιστοποιείται το περιθώριο στο οποίο κινείται το βέλτιστο υπερεπίπεδο και ο προσδιορισμός του πραγματοποιείται με χρήση κλασικών μεθόδων βελτιστοποίησης.

Έστω $T = \{w_1, w_2, \dots, w_k\}$το trainingdataset πάνω στο οποίο θα προσδιοριστεί ο αλγόριθμος SVM, και $w_i = \{w_{i,x_1}, w_{i,x_2}, \dots, w_{i,m}, w_{i,y}\}$όπου$\{x_1, x_2, \dots, x_m\}$ τα χαρακτηριστικά που περιγράφουν κάθε αντικείμενο του trainingdataset και yοι κλάσεις στις οποίες μπορεί να ανήκει. Έστω ότι το πρόβλημα είναι πρόβλημα δυαδικής ταξινόμησης, δηλαδή έστω ότι οι κλάσεις είναι δύο (στην περίπτωση περισσοτέρων των δύο κλάσεων το πρόβλημα ταξινόμησης ανάγεται σε διαδοχικά προβλήματα δυαδικής ταξινόμησης). Επιλέγονται ως ετικέτες των κλάσεων οι τιμές $y \in \{-1,1\}$.Στη συνέχεια υπολογίζονται τα supportvectors $W_1, W_2, \dots, W_{m+1}$ως εξής: Για κάθε στοιχείο του trainingdataset το οποίο ανήκει στην κλάση -1 υπολογίζεται η ελάχιστη απόστασή του από την κλάση 1 και επιλέγεται αυτό με την ελάχιστη

$$W_1 = \min \min_{\mathrm{w_i}} \|w_i - w_j\|$$

Στη συνέχεια επιλέγεται το αντικείμενο της κλάσης 1 το οποίο απέχει ελάχιστη απόσταση από την κλάση -1, και ορίζεται ως W_2. Σαν supportvector W_3επιλέγεται ανάμεσα στα δεύτερα πλησιέστερα διανύσματα στην αντίθετη κλάση της κάθε κλάσης, αυτό με την ελάχιστη απόσταση.Τα αντικείμενα της κλάσης στη οποία ανήκουν τα δύο supportvectors ταξινομούνται κατά αύξουσα απόσταση από την άλλη κλάση, και ως $W_4, \dots, W_{m+1}$επιλέγονται τα πλησιέστερα

Έστω ότι $Y_{W_1} = 1$και $Y_{W_2} = Y_{W_3} = \cdots Y_{W_{m+1}} = -1$. Υπολογίζεται το μοναδικό επίπεδο

$$u^T x + b = -1$$

το οποίο διέρχεται από τα $W_2, W_3, \dots, W_{m+1}$, και το παράλληλό του

$$u^T x + b = +1$$

το οποίο διέρχεται από το W_1. Επομένως, έχουν οριστεί δύο υποχώροι (επίπεδα)

$$H_1: u^T x + b = 1$$

$$H_2: u^T x + b = -1$$

Οι οποίοι είναι τα όρια δύο υπερεπιπέδων για τα οποία ισχύει ότι:

- $u^T x_i + b \geq 1$ όταν $y_i = 1$
- $u^T x_i + b \leq -1$ όταν $y_i = -1$

Καθώς τα H_1και H_2είναι παράλληλα, υπάρχει το διάμεσο επίπεδο H_0 το οποίο είναι παράλληλο στα H_1και H_2 και ισαπέχει από αυτά.Η απόσταση του H_0 από το πλησιέστερο αντικείμενο της κλάσης +1 συμβολίζεται με d_+και απόσταση του H_0 από το πλησιέστερο αντικείμενο της κλάσης -1 συμβολίζεται με d_-. Το περιθώριο του υπερεπιπέδου απόφασης συμβολίζεται με d και είναι ίσο με $d_+ + d_-$.

Συνοψίζοντας, ο αλγόριθμος SVM ισοδυναμεί με τον υπολογισμό του διανύσματος βάρους u^Tκαι της μεροληψίας (bias) bγια τον πλήρη προσδιορισμό του H_0. Ο υπολογισμός αυτός εξαρτάται αποκλειστικά από το σύνολο των supportvectors, επομένως στην περίπτωση όπου τα αντικείμενα μπορούν αν διαχωριστούν γραμμικά, αρκεί ο υπολογισμός των supportvectors για τον προσδιορισμό του ζητούμενου υπερεπιπέδου.

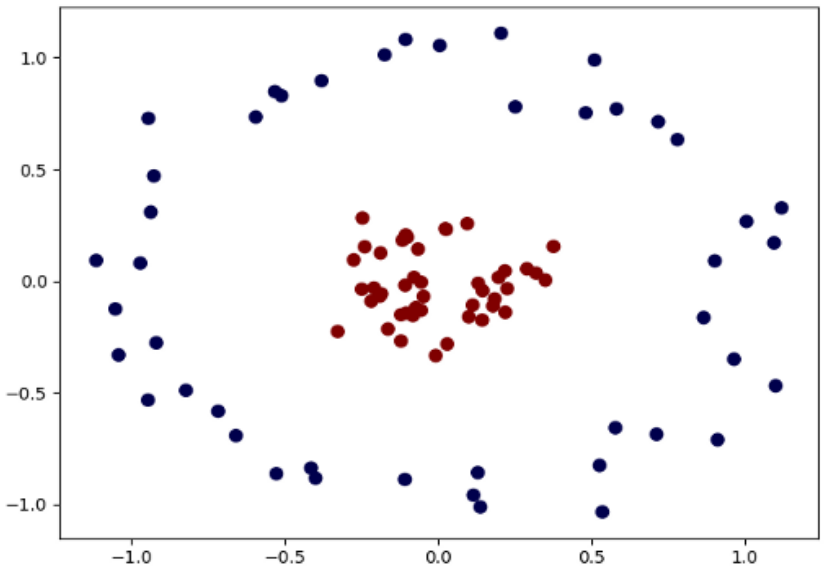
Για να πραγματοποιηθεί διαχωρισμός σε αντικείμενα τα οποία διαχωρίζονται σε παραπάνω από μία κλάσεις $y \in \{-1,1,2,3,..,n\}$, η διαδικασία επέκτασης είναι απλή. Αρχικά οι κλάσεις 1,2,3,... ενωποιούνται σε μία κλάση την κλάση 1. Στη συνέχεια υπολογίζεται το υπερεπίπεδο απόφασης $H_{0,1}$, όπως περιγράφηκε στην παράγραφο αυτή. Έπειτα, για τις κλάσεις $\{1,2,3,\dots,n\}$επαναλαμβάνεται η ίδια διαδικασία ενοποιώντας τις κλάσεις $\{2,3,\dots,n\}$ μέχρι να υπολογίστουν $n-1$ υπερεπίπεδα. Τέλος για να προβλεφθεί η κλάση ενός μη ταξινομημένου αντικειμένου x_t, συγκρίνεται διαδοχικά με τα επίπεδα που υπολογίστηκαν, με τη σειρά που υπολογίστηκαν. Στο πρώτο επίπεδο για το οποίο θα ισχύει$u^T x_t + b \leq -1$θα είναι και αυτό το οποίο θα ορίζει και την κλάση του. Η μη δυαδική κατηγοριοποίηση λοιπόν γραμμικώς διαχωριζόμενων αντικειμένων με τη μέθοδο SVMμπορεί να περιγραφή με τη βοήθεια ενός δένδρου απόφασης, στο οποίο κάθε κόμβος απόφασης είναι και η σύγκριση των χαρακτηριστικών του προς ταξινόμηση αντικειμένου με το υπερεπίπεδο απόφασης.

4.2 Οι Πυρήνες για την ταξινόμηση μη γραμμικώς διαχωριζόμενων δεδομένων

4.2.1 Μη γραμμικώς διαχωριζόμενα δεδομένα

Ο προσδιορισμός του ζητούμενου υπερεπιπέδου στην περίπτωση που οι κλάσεις είναι γραμμικώς διαχωριζόμενες είναι διαδικασία της οποίας η πολυπλοκότητα και η δυσκολία εξαρτάται από τη διάσταση του χώρου, του αριθμού δηλαδή των χαρακτηριστικών του κάθε αντικειμένου.

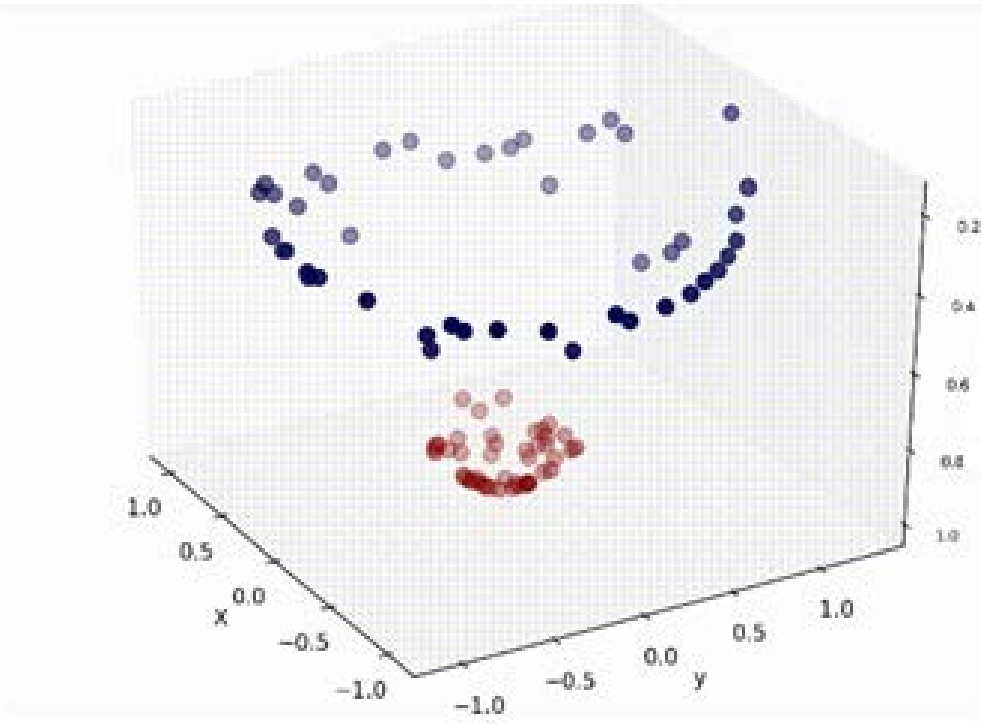
Εάν τα αντικείμενα όταν αναπαρίστανται στο χώρο δεν είναι γραμμικώς διαχωριζόμενα, δηλαδή όταν δεν μπορούν να διαχωριστούν στις ζητούμενες κλάσεις με το σχεδιασμό ενός υπερεπιπέδου, όπως στο Σχήμα 15,η εφαρμογή του αλγορίθμου SVMδεν μπορεί να πραγματοποιηθεί έμμεσα. Για την εφαρμογή του, και προκειμένου να πραγματοποιηθεί ο επιθυμητός διαχωρισμός δημιουργείται η ανάγκη εισαγωγής ενός νέου τρόπου απεικόνισης, τέτοιου ώστε η εισαγωγή ενός μοναδικού υπερεπιπέδου επιτρέπει το διαχωρισμό στις δύο κλάσεις.



Σχήμα 15 - Αντικείμενα τα οποία δεν είναι δυνατό να διαχωριστούν με την εισαγωγή υπερεπιπέδου (ευθείας)[11]

Ο μετασχηματισμός ο οποίος θα χρησιμοποιηθεί έχει ως βασική ιδέα τη θεώρηση της παραπάνω απεικόνισης των δεδομένων ως την προβολή των ίδιων δεδομένων ενός ανώτερου χώρου, στο παραπάνω επίπεδο. Στόχος είναι η ταξινόμησή τους σε έναν χώρο μεγαλύτερης διάστασης έτσι ώστε τα αντικείμενα των δύο διαφορετικών κλάσεων να ανήκουν σε διαφορετικούς υποχώρους. Σχηματικά, το ίδιο σύνολο δεδομένων από το Σχήμα 15, εάν παρασταθεί σε χώρο ανώτερης διάστασης (στην προκειμένη περίπτωση εάν τα σημεία του επιπέδου αναπαρασταθούν στο χώρο τριών διαστάσεων) με τέτοιο τρόπο ώστε τα αντικείμενα της κόκκινης κλάσης καταλάβουν τα κατώτερα επίπεδα του χώρου ενώ τα αντικείμενα της μπλε κλάσης καταλάβουν τα ανώτερα επίπεδα του χώρου, ο διαχωρισμός τους θα μπορέσει να πραγματοποιηθεί εύκολα με την εισαγωγή ενός ενδιάμεσου επιπέδου. Έστω ότι τα δεδομένα αυτά λάβουν τη μορφή όπως φαίνεται στο Σχήμα 16, τότε είναι δυνατή η εφαρμογή του αλγορίθμου SVM, άρα και ο διαχωρισμός τους.

[11] Το σχήμα παρήχθη με τη χρήση της βιβλιοθήκης numpy της Python, και του datasets sklearn



Σχήμα 16 - Δεδομένα διαχωρισμένα

Το ερώτημα το οποίο εγείρεται στην προκειμένη περίπτωση είναι, το πως θα πραγματοποιηθεί αυτή η απεικόνιση, χωρίς να πραγματοποιηθεί ταξινόμηση; Η απάντηση δίνεται μέσα από την οικογένεια των συναρτήσεων των πυρήνων (KernelFunctions[12]).

4.2.2 Εισαγωγή στους πυρήνες

Οι πυρήνες είναι θετικά ορισμένες συναρτήσεις οι οποίες επιτρέπουν την απεικόνιση των στοιχείων ενός χώρου $\mathcal{X}$ στο $\mathbb{R}$ν με τον εξής τρόπο:

Ορισμός: Έστω $\mathcal{X}$ ένα μη κενό σύνολο. Μια συμμετρική συνάρτηση $K: \mathcal{X}x\mathcal{X} \rightarrow$ $\mathbb{R}$ονομάζεται θετικά ορισμένος πυρήνας στο $\mathcal{X}$ εάν για κάθε $n \in \mathbb{N}, x_1, x_2, \dots, x_n \in \mathcal{X}, c_1, c_2, \dots, c_n \in \mathbb{R}$ισχύει ότι

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$$

Σημείωση: Η παραπάνω μορφή μπορεί να αναπαρασταθεί και με μορφή πίνακα $\boldsymbol{K_{i\times j}}$ όπου $\boldsymbol{K}_{ij} = K(x_i, x_j)$.

Για τη Μηχανική Μάθηση, ο πυρήνας είναι ένα εργαλείο το οποίο δέχεται ως ορίσματα δύο διανύσματα ενός αρχικού χώρου(Hofmann, 2006), και τα μετασχηματίζει σε διανύσματα του «μελλοντικού» χώρου, αυξάνοντας τη διάσταση κατά 1. Για παράδειγμα, έστω

$$x = (x_1, x_2), y = (y_1, y_2) \in \mathbb{R}^2$$

Και η απεικόνιση $\varphi: x \rightarrow \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right) \in \mathbb{R}^3$.Το εσωτερικό γινόμενο του $\mathbb{R}^3$ θα είναι

[12] Οι συναρτήσεις πυρήνες εμφανίστηκαν πρώτη φορά στις αρχές του 20ου αιώνα από τον JamesMercer (1883–1932). Χρησιμοποιήθηκαν από τον Mercerγια την επίλυση εξισώσεων ολοκληρωτικών τελεστών και σήμερα αποτελούν σημαντικό κομμάτι της θεωρίας τελεστών, της αρμονικής ανάλυσης αλλά και της στατιστικής

$$< \varphi(x), \varphi(y) \geq = < \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right), \left(y_1^2, y_2^2, \sqrt{2}y_1y_2\right) > = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1x_2y_2 = < x, y >^2$$

Επομένως, $k(x, y) = < x, y >^2$, αλλά ο πυρήνας επίσης υπολογίζει το εσωτερικό γινόμενο της απεικόνισης $\varphi(x) = (x_1^2, x_2^2.x_1x_2, x_2x_1) \in \mathbb{R}^4$

Το παραπάνω επιχείρημα αποδεικνύει ότι ο «μελλοντικός χώρος δεν είναι μοναδικός, για δεδομένο πυρήνα.

Παραδείγματα πυρήνων και εφαρμογές τους

1) Πολυωνυμικός πυρήνας (Polynomial Kernel)

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

Όπου, dο βαθμός του πολυωνύμου.Ο πυρήνας αυτός χρησιμοποιείται στην επεξεργασία εικόνας

2) Πυρήνας του Gauss (Gaussian Kernel)

$$k(x, y) = xp\left(\frac{\|x - y\|^2}{2\sigma^2}\right)$$

Ο πυρήνας του Gaussχρησιμοποιείται όταν δεν υπάρχει εκ των προτέρων γνώση για τα δεδομένα

3) Πυρήνας υπερβολικής εφαπτομένης (Hyperbolictangentkernel)

$$k(x_i, x_j) = \tanh(kx_i \cdot x_j + c)$$

Χρησιμοποιείται στα Νευρωνικά δίκτυα

4) ANOVAπυρήνας ακτινικής βάσης (ANOVAradialbasiskernel)

$$k(x, y) = \sum_{k=1}^{n} exp(-\sigma(x^\kappa - y^k)^2)^d$$

Χρησιμοποιείται σε προβλήματα παλινδρόμησης

Η επιλογή του κατάλληλου πυρήνα εξαρτάται κατά κύριο λόγο από το πρόβλημα και τα δεδομένα. Αρκετές φορές, η επιλογή είναι επίπονη διαδικασία. Για παράδειγμα, ο πολυωνυμικός πυρήνας επιτρέπει τη μοντελοποίηση μελλοντικών συζεύξεων έως το βαθμό του πολυωνύμου. Οι πυρήνες ακτινικής βάσης επιτρέπουν την επιλογή κύκλων σε αντίθεση με κάποιο γραμμικό πυρήνα ο οποίος επιλέγει ευθείες. Η επιλογή πραγματοποιείται με κριτήριο το είδος των πληροφοριών των δεδομένων οι οποίες πρόκειται να εξαχθούν.Στην περίπτωση που δεν υπάρχουν αρκετές πληροφορίες για τα δεδομένα και τις συσχετίσεις των χαρακτηριστικών τους με την κλάση στην οποία ανήκουν, δοκιμάζονται διαφορετικοί πυρήνες προκειμένου να αναπτυχθεί το μοντέλο κατηγοριοποίησης, και συγκρίνονται οι αποτελεσματικότητές τους. Εάν, δεν υπάρχει γνώση για την ύπαρξη ή μη της δυνατότητας γραμμικού διαχωρισμού, συνήθως επιλέγεται ο γραμμικός πυρήνας (LinearKernel), και στη συνέχεια συγκρίνεται με κάποιον μη γραμμικό.

Η διαδικασία αυτή μπορεί να πραγματοποιηθεί αυτοματοποιημένα, με τη χρήση του αλγορίθμου Ktree (Howley&Madden, 2006). Ο Ktree αναπτύχθηκε για τους εξής δύο λόγους:

- Για να περιορίσει τον αριθμό των πυρήνων οι οποίοι θα ελεγχθούν ως προς την καταλληλόλητά τους

- Για να δημιουργήσει νέους πυρήνες συγκεκριμένα για τους αλγορίθμους κατηγοριοποίησης SVM

Κατά τον Ktree δημιουργείται μία δομή δένδρου η οποία αναπαριστά τα δεδομένα απεικονιζόμενα μέσω πυρήνων. Στόχος του Ktree είναι η εύρεση εκείνου του δένδρου-πυρήνα το οποίο αναπαριστά με αποτελεσματικότερο τρόπο τα δεδομένα. Ο Ktree μπορεί να περιγραφεί από τα ακόλουθα βήματα:

1. Δημιουργία τυχαίου πληθυσμού από δένδρα
2. Αξιολόγηση κάθε πυρήνα: Έλεγχος του SVM στα trainingdata.
3. Επιλογή των πυρήνων οι οποίοι προσαρμόζονται σωστά για συνδυασμό
4. Τυχαία εναλλαγή των παιδιών
5. Αντικατάσταση του αρχικού πληθυσμού με τα παιδιά
6. Επανάληψη των βημάτων 2 έως 5 μέχρι τη σύγκλιση
7. Κατασκευή του τελικού SVM με το δένδρο-πυρήνα το οποίο προσαρμόζεται καλύτερα στα δεδομένα

Η μέθοδος Ktree βρίσκει ουσιαστική εφαρμογή όταν η γνώση για τις δομές των δεδομένων τα οποία πρόκειται να αναλυθούν είναι μηδενική. Πλην της επιλογής των πυρήνων οι οποίοι θα εξεταστούν, μεγάλη σημασία πρέπει να δίνεται και στην επιλογή των μεθόδων αξιολόγησης του εκάστοτε μοντέλου, καθώς και των μεθόδων κλαδέματος καθώς υπάρχει πάντα ο κίνδυνος του overfitting.

Συμπερασματικά, οι πυρήνες εισήχθησαν στη μηχανική μάθηση για την αντιμετώπιση της αδυναμίας γραμμικού διαχωρισμού των δεδομένων σε έναν χώρο, μέσω της μετάβασης σε χώρο μεγαλύτερης διάστασης. Η διαδικασία αυτή περιπλέκει την ανάπτυξη του αλγορίθμου, καθώς πρέπει να προηγηθεί η ορθή επιλογή του κατάλληλου πυρήνα, προκειμένου να επιτευχθεί ο ζητούμενος διαχωρισμός. Οι αλγόριθμοι ταξινόμησης είναι δυνατό να αξιοποιήσουν δεδομένα από χώρους με μεγάλες διαστάσεις, χωρίς να χρειαστεί να χαρτογραφήσουν τα σημεία στο χώρο των μεγάλων διαστάσεων.

Κεφάλαιο 5 - Μη εποπτευόμενη μάθηση

5.1 Εισαγωγή

Η μάθηση μπορεί να περιγραφεί ως η διαδικασία εύρεσης σχέσεων μεταξύ αντικειμένων και συνόλων τέτοιων αντικειμένων. Η μελέτη αυτών των σχέσεων οδηγεί στην ικανότητα ακριβούς πρόβλεψης των ιδιοτήτων από αντικείμενα τα οποία δεν έχουν παρατηρηθεί προηγουμένως, επιτρέποντας την εξαγωγή συμπερασμάτων βασισμένων σε περιορισμένη πληροφορία. Η μάθηση μπορεί να πραγματοποιηθεί σε διαφορετικά περιβάλλοντα, ανάλογα στις ιδιότητες και τα αντικείμενα τα οποία συμμετέχουν, και στις ιδιότητες που πρόκειται να προβλεφθούν για τα αντικείμενα που θα χρησιμοποιηθούν αργότερα. Για να πραγματοποιηθεί αυτό εύκολα, αρκεί το σύνολο των ιδιοτήτων οι οποίες θα προβλεφθούν, να είναι μικρότερο σε μέγεθος από το σύνολο των ιδιοτήτων οι οποίες έχουν ήδη παρατηρηθεί (trainingset).

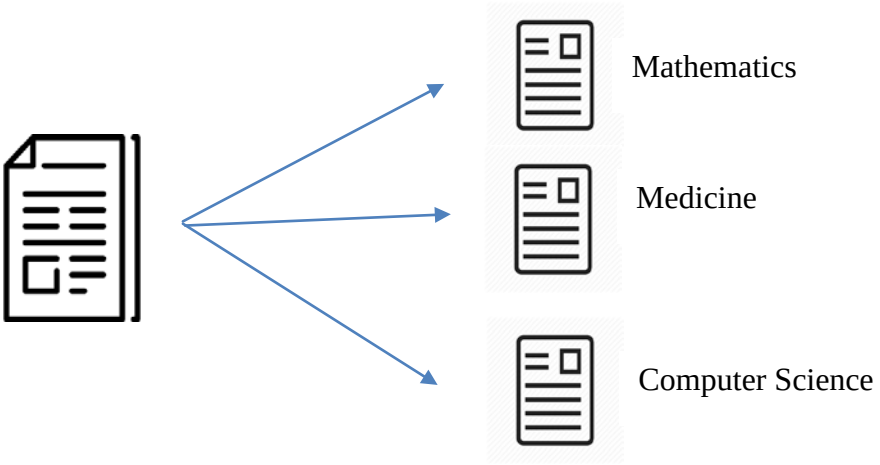
Στην περίπτωση που οι προς πρόβλεψη ιδιότητες (classproperties) είναι υποσύνολο του trainingset, τότε η διαδικασία μάθησης καλείται εποπτευόμενη μάθηση (SupervisedLearning). Χαρακτηριστικά παραδείγματα αλγορίθμων SupervisedLearning αποτελούν οι αλγόριθμοι που αναπτύχθηκαν στα προηγούμενα κεφάλαια. Σε περίπτωση που το σύνολο των classproperties περιέχει ιδιότητες που δεν καλύπτονται από το trainingset, η διαδικασία μάθησης καλείται μη εποπτευόμενη μάθηση, (UnsupervisedLearning).

Το περιβάλλον μάθησης μπορεί να χωριστεί σε δύο επί μέρους κατηγορίες.

- Semi-supervised (ημί-εποπτευόμενο): Στο trainingset υπάρχουν γνωστές τιμές για το σύνολο των classproperties.
- Unsupervised:Στο trainingset δεν υπάρχουν καθόλου γνωστές τιμές για το σύνολο των classproperties

Η μηχανική μη εποπτευόμενη μάθηση χρησιμοποιείται κατά κόρον στην κατηγοριοποίηση αντικειμένων. Κατά την απουσία της εκ των προτέρων γνώσης για τις classproperties, η διαδικασία κατηγοριοποίησης ανάγεται σε συσταδοποίηση των αντικειμένων, σύμφωνα με πιθανά πρότυπα τα οποία αναγνωρίζονται κατά την εφαρμογή των επιλεγμένων αλγορίθμων. Η εφαρμογή τέτοιων προτύπων, ανεπτυγμένων δηλαδή με τη βοήθεια των μεθόδων μηχανικής μάθησης σε πραγματικά δεδομένα για την εξεύρεση πληροφορίας ή την δημιουργία πληροφοριακών δομών κατανοητών από τον άνθρωπο. Η διαδικασία αυτή στη βιβλιογραφία συναντάται με τον όρο εξόρυξη δεδομένων (Datamining). Χαρακτηριστικό παράδειγμα εφαρμογής της εξόρυξης δεδομένων είναι η αναζήτηση προτύπων που αφορούν στις τάσεις των ιστορικών δεδομένων, κατά την προσπάθεια πρόβλεψης αλλαγών στους δείκτες της χρηματιστηριακής αγοράς κατά το προσεχές χρονικό διάστημα. Ευρύ πεδίο εφαρμογών των τεχνικών της μη εποπτευόμενης μηχανικής μάθησης είναι η αναγνώριση της φυσικής γλώσσας. Η κατανόηση της ανθρώπινης γλώσσας είναι ένα περίπλοκο έργο, όπως φαίνεται από τις περιπλοκές της έρευνας για την φυσική γλώσσα. Ωστόσο, είναι δυνατόν να χρησιμοποιηθούν οι παρατηρούμενες ιδιότητες ενός δεδομένου κειμένου για την εκτέλεση ορισμένων προγνωστικών εργασιών χωρίς να χρειάζεται να κατανοηθεί η πλήρης έννοια του εν λόγω κειμένου. Εάν ζητείται ο καθορισμός του εάν ένα επιστημονικό άρθρο δημοσιεύθηκε σε ιατρικό περιοδικό ή σε αστρονομικό περιοδικό, θα ήταν πιθανό να προσδιοριστεί ένα σύνολο λέξεων οι οποίες χρησιμοποιούνται μόνο σε ιατρικό πλαίσιο και ένα σύνολο λέξεων οι οποίες χρησιμοποιούνται μόνο σε αστρονομική θεματολογία, και στη συνέχεια αναζητηθούν αυτές στο συγκεκριμένο άρθρο. Αυτό θα επέτρεπε την ακριβή πρόβλεψη της

κλάσης του αντίστοιχου περιοδικού χωρίς να απαιτείται από τον υπολογιστή να κατανοήσει το περιεχόμενο του εν λόγω άρθρου. Το παραπάνω πρόβλημα λαμβάνει την εξής μορφή:



Εικόνα 1 - Κατηγοριοποίηση επιστημονικού άρθρου

Στο παραπάνω πρόβλημα, οι κλάσεις ταξινόμησης είναι οι επιστημονικοί τομείς, (μαθηματικά, ιατρική, πληροφορική, κ.ά.). Για την ακριβέστερη πρόβλεψη της κλάσης στην οποία ανήκει το έγγραφο, απαιτείται η αναγνώριση των ιδιοτήτων της κάθε κλάσης τις οποίες ικανοποιεί. Καθώς για την κατηγοριοποίηση χρησιμοποιούνται μόνο οι ακολουθίες χαρακτήρων του εγγράφου, αυτές θα αξιοποιηθούν για την αναγνώριση των ιδιοτήτων. Για κάθε κλάση, δημιουργείται ένα λεξικό λέξεων-κλειδιών (Keywords) το οποίο είναι μοναδικό. Στόχος είναι η αναγνώριση των keywords και ο υπολογισμός της συχνότητας εμφάνισής τους. Κατά την ολοκλήρωση της αναγνώρισης, δημιουργείται ένας βαθμός (score) για κάθε κλάση σύμφωνα με τη συχνότητα εμφάνισης των Keywordstου λεξικού της και επιλέγεται αυτή με το μεγαλύτερο σκορ. Κατά την ολοκλήρωση του αλγορίθμου, λέξεις με υψηλή συχνότητα εμφάνισης στο κείμενο εξετάζονται για την πιθανή ένταξή τους στο λεξικό της κλάσης. Στο σημείο αυτό κρίνεται απαραίτητη η δημιουργία ενός ενιαίου λεξικού ευρέως χρησιμοποιούμενων λέξεων, οι οποίες δεν μπορούν να αξιοποιηθούν ως keywordsδιότι στη φυσική γλώσσα εμφανίζονται σε όλες τις εκφάνσεις της. Οι λέξεις αυτές μπορεί να είναι σύνδεσμοι, επιρρήματα, προθέσεις και άλλα μέρη του λόγου χωρίς εννοιολογική ισχύ.

Για την εφαρμογή ενός αλγορίθμου αναγνώρισης φυσικής γλώσσας για την ταξινόμηση επιστημονικών άρθρων, ενδείκνυνται τα εξής στάδια:

Καθαρισμός των δεδομένων

Στο στάδιο αυτό το προς κατηγοριοποίηση κείμενο μετατρέπεται σε αναγνωρίσιμη μορφή από τον αλγόριθμο. Αρχικά αφαιρούνται τα σημεία στίξης. Τα σημεία στίξης προσδίδουν υψηλή γραμματική και εννοιολογική αξία σε ένα κείμενο, αλλά στην περίπτωση της κατηγοριοποίησης κειμένων η οποία πραγματοποιείται σε επίπεδο λέξης και όχι πρότασης, τα σημεία στίξης δεν προσθέτουν αξία. Επομένως στο βήμα αυτό, μία ακολουθία χαρακτήρων της μορφής “Καλησπέρα, πώς είσαι;”, θα λάβει τη μορφή “Καλησπέρα πώς είσαι”.

Στη συνέχεια, όπως και προηγουμένως καθώς ο αλγόριθμος αξιοποιεί μεμονωμένες λέξεις και όχι προτάσεις, η κάθε πρόταση τμηματοποιείται στις επιμέρους λέξεις που την απαρτίζουν. Έτσι, η αρχική φράση "Καλησπέρα, πώς είσαι;" θα μετατραπεί σε "Καλησπέρα, πώς, είσαι".

Τέλος, για την ολοκλήρωση του βήματος αυτού, αφαιρούνται από τις λέξεις που έχουν αναγνωριστεί, εκείνες οι λέξεις οι οποίες ανήκουν στο λεξικό των non-keywords. Στην παραπάνω πρόταση δεν μπορεί να χαρακτηριστεί κάποια λέξη ως keywordεπομένως θα αφαιρεθούν όλες. Σύμφωνα με τους μετασχηματισμούς του σταδίου αυτού, η πρόταση "Τα νευρωνικά δίκτυα χρησιμοποιούνται στην αναγνώριση προτύπων" θα γίνει «νευρωνικά, δίκτυα, αναγνώριση, προτύπων".

Σάρωση του εγγράφου

Στο στάδιο αυτό, καθώς έχουν αναγνωριστεί οι λέξεις οι οποίες απαρτίζουν το κείμενο, για κάθε κλάση πραγματοποιείται σύγκριση των λέξεων του λεξικού της με αυτές του κειμένου για τον υπολογισμό της συχνότητας των εμφανίσεων τους στο κείμενο. Στη συνέχεια υπολογίζεται το σκορ της κλάσης, είτε με απλή πρόσθεση των συχνοτήτων εμφάνισης είτε με πιο περίπλοκο τρόπο (π.χ. με τη χρήση σταθμών για κάθε keyword) και επιλέγεται ως κλάση αντιστοίχισης αυτή με το υψηλότερο σκορ.

Ενημέρωση των keywords της κλάσης

Καθώς έχει ταξινομηθεί το έγγραφο σε μία κλάση, το λεξιλόγιό της ενημερώνεται με τις λέξεις οι οποίες έχουν συχνότητα εμφάνισης μεγαλύτερη από κάποιο επιλεγμένο κατώφλι, και δεν ανήκουν ήδη σε αυτό.

Ο παραπάνω αλγόριθμος εκτελείται σε ημι-εποπτευόμενο περιβάλλον, καθώς χωρίς την ύπαρξη των λεξικών είναι σχεδόν αδύνατη η ταξινόμηση των εγγράφων.

Ο όγκος των διαθέσιμων κειμένων αυξάνεται με πολύ υψηλούς ρυθμούς, επομένως η δυνατότητα χρήσης αυτοματοποιημένων τεχνικών για την εξόρυξη πληροφοριών από τα διαθέσιμα κείμενα επιτρέπει παρέχει μεγάλο όγκο πληροφοριών, χρήσιμων για την ταξινόμηση κειμένων. Σε αυτό το περιβάλλον, η μη εποπτευόμενη μάθηση είναι ζωτικής σημασίας καθώς αφαιρεί κάθε ανάγκη συλλογής δεδομένων τα οποία χρήζουν κατηγοριοποίησης πριν από την εφαρμογή των τεχνικών. Χωρίς τη συμβολή της μη εποπτευόμενης μάθησης, η εξόρυξη πληροφορίας γίνεται ένα δαπανηρό έργο το οποίο βασίζεται στην ανθρώπινη αλληλεπίδραση και κρίσεις. Στη συνέχεια παρουσιάζονται οι πιο δημοφιλείς αλγόριθμοι μη εποπτευόμενης μάθησης.

5.2 K-means clustering

Ο αλγόριθμος k-meansclustering (αλγόριθμος συσταδοποίησης k-μέσων) είναι μία μέθοδος διανυσματικής ποσοτικοποίησης ο οποίος είναι ιδιαίτερα δημοφιλής στην εξόρυξη δεδομένων. Ο στόχος του αλγορίθμου αυτού είναι η διαμέριση ενός συνόλου nπαρατηρήσεων σε kσυστάδες, έτσι ώστε κάθε παρατήρηση να ανήκει στην κλάση με τον πλησιέστερο μέσο όρο. Η διαδικασία αυτή οδηγεί στη δημιουργία μιας κάλυψης από κελιά του χώρου των δεδομένων, με τέτοιον τρόπο ώστε κάθε στοιχείο του χώρου να ανήκει σε μοναδικό κελί.

Το πρόβλημα συσταδοποίησης του χώρου δεδομένων χαρακτηρίζεται από υψηλή πολυπλοκότητα (NP-Hard).Παρόλα αυτά έχουν αναπτυχθεί αρκετοί εμπειρικοί αλγόριθμοι οι οποίοι συγκλίνουν σε βέλτιστες τοπικές λύσεις. Οι περισσότεροι αλγόριθμοι αξιοποιούν το κέντρο της συστάδας για να πραγματοποιήσουν τη ζητούμενη συσταδοποίηση. Το πλεονέκτημα του αλγορίθμου k-means έγκειται στο γεγονός ότι έχει την ιδιότητα να βρίσκει συστάδες όμοιας χωρικής έκτασης ενώ παράλληλα επιτρέπει την διαφοροποίηση στο σχήμα των σχηματιζόμενων συστάδων.

Ο αλγόριθμος k-meansσυχνά συγχέεται με τον αλγόριθμο κατηγοριοποίησης k-nearestneighbor (k-κοντινότερος γείτονας). Ανάμεσα στους δύο αυτούς αλγορίθμους υπάρχει ασθενής σύνδεση καθώς ο συνδυασμός τους συνθέτει νέο αλγόριθμο, τον αλγόριθμο Rocchio.

Ο αλγόριθμος k-means αποτελεί αλγόριθμο μη εποπτευόμενης μάθησης, καθώς χρησιμοποιείται σε περιπτώσει που για τα δεδομένα προς ταξινόμηση δεν υπάρχει γνώση για κατηγορίες ή ομάδες στις οποίες ανήκουν. Η ταξινόμηση των παρατηρήσεων στις *k*συστάδες πραγματοποιείται επαναληπτικά. Η κατανομή των παρατηρήσεων στις συστάδες πραγματοποιείται βάσει των χαρακτηριστικών τους. Τα κέντρα των συστάδων μπορούν πλέον να χρησιμοποιηθούν ως ετικέτες (labels) των νέων δεδομένων.

Το κέντρο της κάθε κλάσης αποτελεί τη συλλογή τιμών για τα χαρακτηριστικά των δεδομένων τα οποία θα ταξινομηθούν. Η διερεύνηση των βαρών των χαρακτηριστικών των κέντρων της κάθε συστάδας μπορεί να χρησιμοποιηθεί προκειμένου να επεξηγηθεί ποιοτικά η φύση της των παρατηρήσεων οι οποίες εκπροσωπούνται από το κέντρο αυτό.

Ο αλγόριθμος k-meanskατά κύριο λόγο χρησιμοποιείται σε περιπτώσεις όπου συναντάται το πρόβλημα συσταδοποίησης δεδομένων τα οποία δεν έχουν ετικέτες. Η περιγραφή τους δηλαδή γίνεται μόνο βάση των χαρακτηριστικών τους. Αυτό βρίσκει εφαρμογή στην επαλήθευση υποθέσεων για το ποιες κατηγορίες δεδομένων υπάρχουν, ή για την αναγνώριση νέων κατηγοριών σε περίπλοκα σύνολα δεδομένων. Μετά την εκτέλεση του αλγορίθμου και τον προσδιορισμό των ομάδων, για πιθανές νέες παρατηρήσεις είναι ιδιαίτερα εύκολο να ταξινομηθούν.

Ο αλγόριθμος αυτός χαρακτηρίζεται από την προσαρμοστικότητά του σε διαφορετικού είδους δεδομένα και ιδιαίτερα σε διαφορετικά χαρακτηριστικά τα οποία συνθέτουν τα κέντρα των συστάδων ταξινόμησης. Μερικά παραδείγματα χρήσης του αλγορίθμου k-means είναι:

- Κατηγοριοποίηση εγγράφων
- Χωρική τοποθέτηση αποθηκών
- Αναγνώριση τοποθεσιών υψηλής εγκληματικότητας
- Κατηγοριοποίηση πελατών
- Εντοπισμός ασφαλιστικής απάτης
- Δημιουργία διαδικτυακού προφίλ εγκληματία
- Ανάλυση ιστορικού κλήσεων Call Record Detail Analysis
- Αυτοματοποιημένες προειδοποιήσεις για πιθανές βλάβες πληροφορικής

Ο αλγόριθμος

Ο αλγόριθμος συσταδοποίησης k-means χρησιμοποιεί τη μέθοδο επαναληπτικής βελτίωση για την επίτευξη του τελικού αποτελέσματος. Ως είσοδο, ο αλγόριθμος δέχεται το πλήθος των

συστάδων στις οποίες θα διαχωριστούν τα δεδομένα, και το σύνολο των δεδομένων το οποίο θα συσταδοποιηθεί. Το σύνολο δεδομένων είναι μια συλλογή παρατηρήσεων, κάθε μία αποτελούμενη από ένα σύνολο χαρακτηριστικών. Ο αλγόριθμος αρχικά πραγματοποιεί εκτιμήσεις για τακ κέντρα των συστάδων, είτε παράγονται με τυχαίο τρόπο, είτε επιλέγονται τυχαία από το σύνολο των δεδομένων. Για την ολοκλήρωσή του εκτελούνται επαναληπτικά τα εξής δύο βήματα:

1. Βήμα ανάθεσης δεδομένων σε συστάδες
   Κάθε κέντρο ορίζει ακριβώς μία συστάδα. Στο βήμα αυτό κάθε παρατήρηση κατανέμεται στη συστάδα η οποία ορίζεται από το πλησιέστερο κέντρο. Η έννοια της εγγύτητας επιτυγχάνεται με την επιλογή της κατάλληλης μετρικής σχέσης. Συνήθως επιλέγεται ως μετρική η Ευκλείδια απόσταση. Συγκεκριμένα, εάνc_iείναι τα κέντρα και Cτο σύνολο τους τότε η παρατήρηση xτου συνόλου δεδομένων αντιστοιχίζεται στην κλάση του κέντρου c_iγια την οποία ισχύει:
   $$\arg_{c_i \in C} \min \|c_i - x\|$$
2. Βήμα επανυπολογισμού των κέντρων
   Ο επανυπολογισμός των κέντρων πραγματοποιείται με τον υπολογισμό του μέσου όλων των παρατηρήσεων της κάθε συστάδας
   $$c_i = \frac{1}{|s_i|} \sum_{x_i \in s_i} x_i$$

Τα βήματα 1-2 επαναλαμβάνονται όταν επαληθεύεται η συνθήκη σύγκλισης. Παραδείγματα τέτοιων συνθηκών είναι:

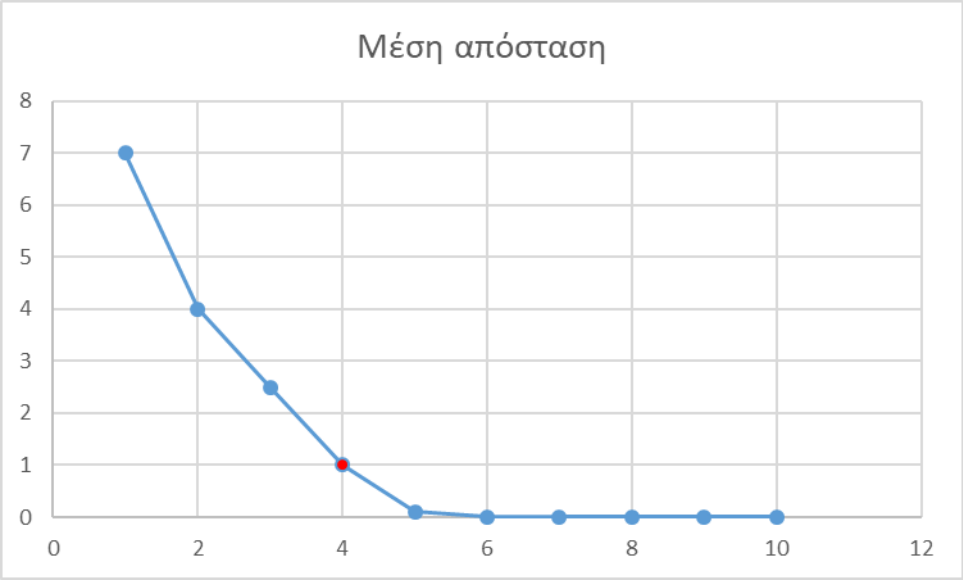
- Καμία παρατήρηση δεν αλλάζει κλάση
- Το άθροισμα των αποστάσεων ελαχιστοποιείται
- Έχει πραγματοποιηθεί συγκεκριμένος αριθμός επαναλήψεων.

Ο αλγόριθμος συγκλίνει πάντα σε ένα αποτέλεσμα. Για να μπορεί να θεωρηθεί σωστό το αποτέλεσμα αυτό πρέπει να γίνει ορθή επιλογή της συνθήκης σύγκλισης. Το αποτέλεσμα αποτελεί τοπική λύση, υπό την έννοια ότι πιθανή διαφορετική επιλογή αρχικών κέντρων μπορεί να οδηγήσει σε διαφορετική λύση. Συνίσταται λοιπόν η εκτέλεση του αλγορίθμου με διαφορετικές αρχικές επιλογές κέντρων. Επί πλέον, για την βελτίωση της αξιοπιστίας του αλγορίθμου προτείνεται η εκτέλεσή του για διαφορετικό πλήθος συστάδων. Καθώς δεν υπάρχει μέθοδος ακριβούς υπολογισμού του πλήθους k, αρκετές μέθοδοι εκτίμησης της τιμής του κέχουν αναπτυχθεί:

- cross-validation (διασταυρωμένη επαλήθευση)
- μέθοδοι θεωρίας πληροφορίας
- silhouette method
- ο αλγόριθμοςG-μέσων τιμών
- μέθοδος μέσων αποστάσεων (Elbowmethod)

Η τελευταία μέθοδος αποτελεί την πιο συνηθισμένη μέθοδο σύγκρισης των αποτελεσμάτων του αλγορίθμου για τά διαφορετικά πλήθη συστάδων k. Είναι μετρική η οποία υπολογίζει τη μέση απόσταση των παρατηρήσεων κάθε συστάδας από το κέντρο τους. Καθώς η αύξηση του πλήθους kοδηγεί στη μείωση των παρατηρήσεων οι οποίες ανήκουν σε κάθε συστάδα, άμεση συνέπεια είναι και η μείωση της τιμής της μέσης απόστασης. Όταν το πλήθος k φτάσει ή

ξεπεράσει το πλήθος των παρατηρήσεων, δημιουργούνται κενές συστάδες και η μέση απόσταση ασυμπτωτικά λαμβάνει την τιμή μηδέν. Επομένως για την επιλογή της βέλτιστης τιμής για tok, δημιουργείται το διάγραμμα μέσων αποστάσεων ~ κκαι επιλέγεται το σημείο στο οποίο παρατηρείται απότομη πτώση της μέσης απόστασης (elbowpoint).



Σχήμα 17 - Elbowpointγια τη μέθοδο μέσης απόστασης

Στο παραπάνω σχήμα το προτεινόμενο πλήθος κσυστάδων είναι ίσο με 4.

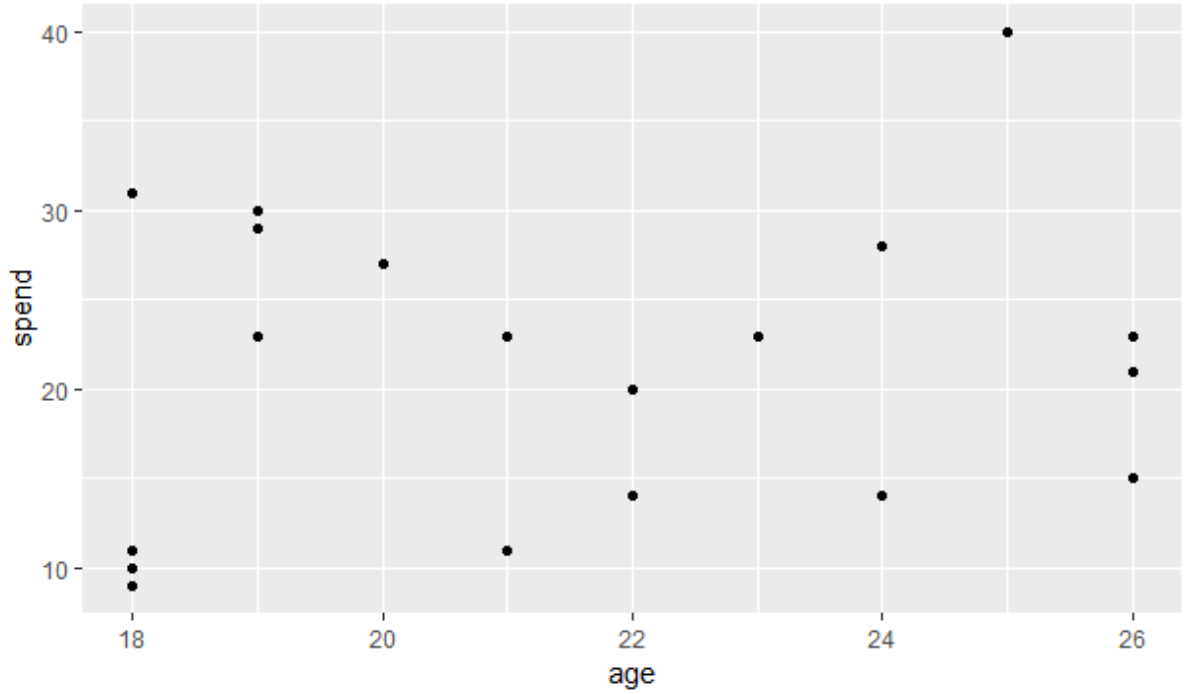
Εφαρμογή: Στον πίνακα που ακολουθεί παρουσιάζονται οι ηλικίες και οι απαντήσεις 20 ατόμων στην ερώτηση «Πόσα χρήματα ξοδέψατε το τελευταίο Σάββατο».

α/α	ηλικία	έξοδα σε €
1	18	10
2	21	11
3	19	30
4	22	14
5	26	15
6	26	23
7	18	11
8	23	23
9	19	29
10	25	40
11	21	23
12	20	27
13	18	9

14	22	20
15	24	28
16	26	21
17	19	30
18	18	31
19	19	23
20	24	14

Πίνακας 1 - Ηλικίες και ποσά σε € που ξόδεψαν 20 διαφορετικά άτομα

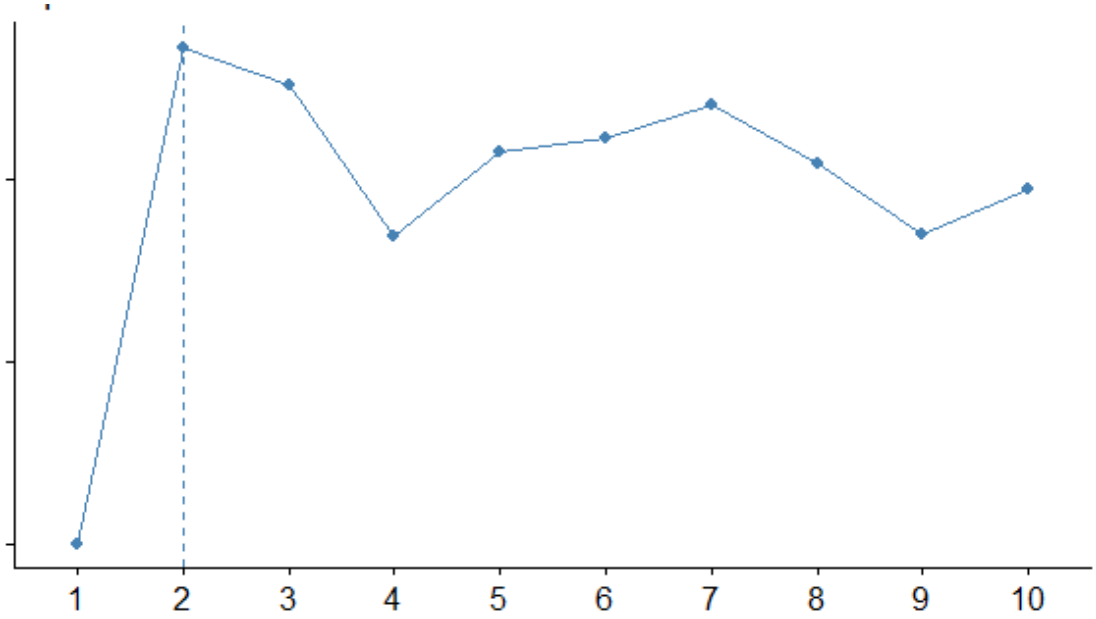
Τα δεδομένα του πίνακα μπορούν να παρασταθούν σε κοινό σύστημα αξόνων όπως φαίνεται στο ακόλουθο διάγραμμα:



Διάγραμμα 5

Από το διάγραμμα δεν είναι δυνατή η εξαγωγή κάποιου μοτίβου διαχωρισμού των δεδομένων. Για παράδειγμα μια υπόθεση της μορφής «Τα νεαρά άτομα ξοδεύουν περισσότερα χρήματα από τους μεγαλύτερους» δεν φαίνεται να είναι αληθής.

Υπολογίζεται αρχικά το elbowpoint για το παραπάνω σύνολο δεδομένων, προκειμένου να προσδιοριστεί ο βέλτιστος αριθμός συστάδων.

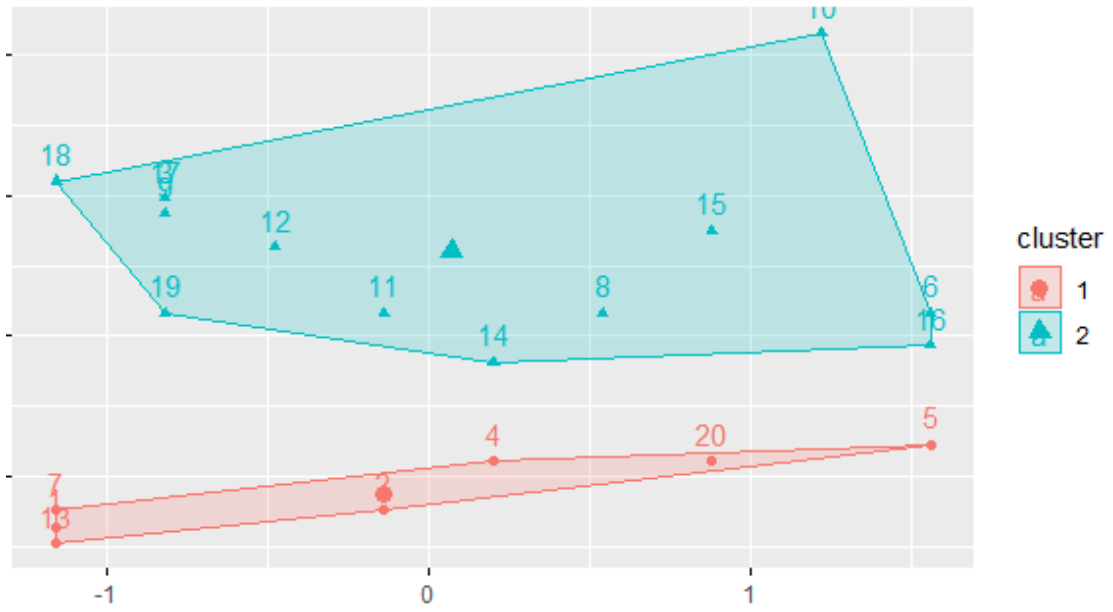


Διάγραμμα 6 - Βέλτιστος αριθμός συστάδων

Όπως φαίνεται και στο σχήμα, ο βέλτιστος αριθμός συστάδων είναι $k = 2$. Επομένως εφαρμόζεται ο 2-meansαλγόριθμος συσταδοποίησης για το συγκεκριμένο dataset. Ο αλγόριθμος εκτελέσθηκε για δέκα επαναλήψεις και η ταξινόμηση η οποία προτάθηκε ως λύση είναι η

Μέλη 1ης συστάδας	Μέλη 2ης συστάδας	
1	3	14
2	6	15
4	8	16
5	9	17
7	10	18
13	11	19
20	12	

Πίνακας 2 - Μέλη κάθε συστάδας



Διάγραμμα 7 - Γραφικός διαχωρρισμός σε δύο συστάδες

Επαληθεύεται λοιπόν η αρχική «εμπειρική» διαπίστωση ότι στο παραπάνω σύνολο δεδομένων δεν υπάρχει συσχέτιση ανάμεσα στο ποσό και την ηλικία. Ο διαχωρισμός που πραγματοποιήθηκε από τον αλγόριθμο μπορεί να μεταφραστεί ως:

- $1^{η}$ κλάση: Άτομα που ξόδεψαν λιγότερα από 15€
- $2^{η}$ κλάση: Άτομα που ξόδεψαν περισσότερα από 15€

Στο σημείο αυτό παρατηρείται ότι εάν εφαρμοστεί διαφορετική ταξινόμηση των δεδομένων, η συσταδοποίηση πιθανόν να είναι διαφορετική.

Στη συνέχεια ακολουθεί ο αλγόριθμος ιεραρχικής συσταδοποίησης (Hierarchicalclustering).

5.3 Αλγόριθμος Ιεραρχικής Συσταδοποίησης

Εισαγωγή

Όπως παρουσιάστηκε και στην προηγούμενη παράγραφο, οι αλγόριθμοι συσταδοποίησης ομαδοποιούν τα δεδομένα σε υποσύνολα ή σε συστάδες. Στόχος των αλγορίθμων αυτών είναι η δημιουργία ομάδων οι οποίες έχουν εμφανείς εσωτερικές σχέσεις ομοιότητας και διαχωρίζονται η μία από την άλλη, υπό την έννοια του ότι στοιχεία που ανήκουν στην ίδια ομάδα έχουν ομοιότητες, ενώ στοιχεία από διαφορετικές ομάδες είναι επουσιωδώς διαφορετικά.

Οι αλγόριθμοι συσταδοποίησης μπορούν να διαχωριστούν σε δύο μεγάλες κατηγορίες σύμφωνα με τον τρόπο λειτουργίας τους:

- Agglomerativeclustering (συσσωρευτική συσταδοποίηση): Κατά την προσέγγιση αυτή κάθε παρατήρηση του dataset ανατίθεται σε ξεχωριστή συστάδα. Στα επαναληπτικά βήματα οι συστάδες συνενώνονται έως μέχρι να ικανοποιηθεί ένα κριτήριο ολοκλήρωσης της διαδικασίας. Η προσέγγιση αυτή είναι “bottom-up”.
- Divisiveclustering (Διαιρετική συσταδοποίηση): Κατά την προσέγγιση αυτή, η οποία είναι μία top-downδιαδικασία, όλες οι παρατηρήσεις του datasetανατίθενται σε μοναδική συστάδα, η οποία κατά τα επαναληπτικά βήματα διαιρείται σε μικρότερες υποσυστάδες της έως ότου ικανοποιηθεί μία συνθήκη ολοκλήρωσης.

Εν γένει, οι διαχωρισμού και οι συνενώσεις πραγματοποιούνται με «άπληστο» (greedy) τρόπο, υπό την έννοια της εξεύρεσης τοπικού βέλτιστου με στόχο τον υπολογισμό του ολικού βέλτιστου. Για το λόγο αυτό η πολυπλοκότητα των αλγορίθμων αυτών είναι ιδιαίτερα υψηλή. Ο τυπικός αλγόριθμος ιεραρχικής συσταδοποίησης έχει χρονική πολυπλοκότητα ίση με$\mathcal{O}(n^3)$η οποία των καθιστά αργό ακόμη και για μικρά σύνολα δεδομένων.

Ομοιότητα

Οι ομοιότητα των στοιχείων εντός των συστάδων συχνά υπολογίζεται με τη χρήση μετρικών ομοιότητας. Οι πιο συνηθισμένες είναι οι:

- Ευκλείδεια απόσταση, $\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
- Απόσταση Manhattan, $\|a-b\|_1 = \sqrt{\sum_i |a_i - b_i|}$
- Μέγιστη απόσταση, $\|a-b\|_\infty = \max_i |a_i - b_i|$
- Απόσταση Mahalanobis $D_M(a,b) = \sqrt{(a-b)^T S^{-1}(a-b)}$, όπου Sείναι ο πίνακας συνδιασποράς των a, b

Η επιλογή της κατάλληλης μετρικής γίνεται με τη βοήθεια της αξιολόγησης των χαρακτηριστικών του dataset.

Προετοιμασία των δεδομένων

Πριν την εφαρμογή του αλγορίθμου πρέπει να πραγματοποιηθούν κάποια προπαρασκευαστικά βήματα. Αρχικά, κανονικοποιούνται οι τιμές για κάθε χαρακτηριστικό προκειμένου να ξεκινήσει η διαδικασία. Ο λόγος για τον οποίο είναι επιτακτική η κανονικοποίηση των δεδομένων έγκειται στο γεγονός ότι τα δεδομένα αναπαρίστανται ως συντεταγμένες ενός n-διάστατου διανυσματικού χώρου, όπου nείναι το πλήθος των χαρακτηριστικών της κάθε παρατήρησης. Η επιλεγμένη μετρική δέχεται ως ορίσματα τις συντεταγμένες από ανά δύο ζευγάρια παρατηρήσεων. Για το λόγο αυτό, η ύπαρξη πολύ

μεγάλων τιμών μπορεί να οδηγήσει σε μη αντιπροσωπευτικά αποτελέσματα, δηλαδή σε συστάδες οι οποίες δεν έχουν σαφείς δομές ομοιότητας για τα στοιχεία τους.

Η πιο συνηθισμένη μέθοδος κανονικοποίησης είναι η μέθοδος min-max κανονικοποίηση, κατά την οποία για κάθε μεταβλητή tεφαρμόζεται ο εξής μετασχηματισμός:

$$s \to t \in X$$

$$s = t - \frac{\min\limits_{x \in X} X}{\max\limits_{x \in X} X - \min\limits_{x \in X} X}$$

Με το μετασχηματισμό αυτό, οι τιμές των χαρακτηριστικών κάθε παρατήρησης απεικονίζονται στο διάστημα $[0,1]$.

Επίσης συχνά χρησιμοποιούμενη μέθοδος κανονικοποίησης είναι και η μέθοδος διασποράς. Κατά τη μέθοδο αυτή σε κάθε μεταβλητή tεφαρμόζεται ο εξής μετασχηματισμός:

$$s \to t \in X$$

$$s = t - \frac{mean(X)}{sd(X)}$$

Όπου $mean(X)$είναι η μέση τιμή του συνόλου Xκαι $sd(X)$είναι η τυπική απόκλιση του συνόλου X.

Συμπλήρωση των NAs

Σε μεγάλα σύνολα δεδομένων, και ιδιαίτερα σε δεδομένα τα οποία συλλέγονται αυτοματοποιημένα πχ από όργανα μέτρησης, είτε συμπληρώνονται από το κοινό μέσω onlineφόρμας, συχνά παρατηρείται ύπαρξη μηδενικών τιμών για ορισμένα χαρακτηριστικά. Για να εκτελεστεί ο αλγόριθμος, οι μηδενικές τιμές θα πρέπει να αντικατασταθούν από τιμές οι οποίες ανήκουν στο εύρος το οποίο ορίζει το dataset$X$. Συνήθως οι μηδενικές τιμές αντικαθίστανται από τη μέση τιμή του Xενώ σε συγκεκριμένες περιπτώσεις επιλέγεται η επικρατούσα τιμή ως τιμή αντικατάστασης των NAs

Αλγόριθμος ιεραρχικής συσταδοποίησης (συσωρευτικός)

Η βασική διεργασία του αλγορίθμου είναι ο επαναληπτικός συνδυασμός ζευγών συστάδων για τη δημιουργία μίας μεγαλύτερης. Οι δύο παράμετροι οι οποίες πρέπει να έχουν προσδιοριστεί πριν την έναρξη του αλγορίθμου, και αφού έχει ολοκληρωθεί η προετοιμασία των δεδομένων είναι οι εξής:

- Ποια μετρική ομοιότητας θα χρησιμοποιηθεί;
- Ποια συνθήκη θα αποτελέσει την συνθήκη ολοκλήρωσης του αλγορίθμου;

Αφού προσδιοριστούν οι παραπάνω δύο παράμετροι, ο αλγόριθμος εκτελείται σύμφωνα με τα παρακάτω βήματα.

Βήμα 0 (Αρχικό βήμα): Δημιουργούνται nκενές συστάδες, όπου nτο πλήθος των παρατηρήσεων του dataset. Κάθε παρατήρηση ανατίθεται σε μοναδική συστάδα.

Βήμα 1 (Επαναληπτικό βήμα):Υπολογίζονται όλες οι αποστάσεις για κάθε πιθανό ζεύγος συστάδων. Δημιουργείται πίνακας αποστάσεων των συστάδων, και για το ζευγάρι αυτό με

την μικρότερη απόσταση, οι συστάδες ενώνονται σε μία μεγαλύτερη συστάδα. Συνέπεια του βήματος αυτού είναι η κατά μία μείωση των συστάδων.

Βήμα 2 (Έλεγχος ολοκλήρωσης):Ελέγχεται εάν πληρείται η συνθήκη ολοκλήρωσης του αλγορίθμου. Εάν όχι, επανεκτελείται το βήμα 1.

Έλεγχος της αποτελεσματικότητας του αλγορίθμου

Κατά την ολοκλήρωση του αλγορίθμου, και για την αξιολόγηση των αποτελεσμάτων και τις αποδοτικότητας του αλγορίθμου, πραγματοποιείται έλεγχος της αποτελεσματικότητάς του αλγορίθμου. Υπάρχουν αρκετές μέθοδοι για τον έλεγχο αυτό. Η πιο συνηθισμένη είναι η χρήση του δείκτη του Dunn. Ο δείκτης του Dunnορίζεται ως ο λόγος της ελάχιστης δια-συσταδικής απόστασης (minimum inter-cluster distance) προς τη μέγιστη ενδοσυσταδική διάμετρο (maximumintra-clusterdiameter). Ως διάμετρος συστάδας ορίζεται η απόσταση των δύο πιο απομακρυσμένων παρατηρήσεων της συστάδας. Προκειμένου να διαχωριστούν τα δεδομένα καλώς και οι συστάδες να είναι συμπαγείς, στόχος είναι ο υψηλός δείκτης Dunn.

Σύγκριση των αλγορίθμων k-means και ιεραρχικής συσταδοποίησης

Οι δύο αλγόριθμοι που αναπτύχθηκαν, παρόλο που είναι εκπρόσωποι διαφορετικών κατηγοριών συσταδοποίησης, καθώς στον αλγόριθμο k-meansδεν υπάρχει η έννοια της υπο-συστάδας, πολλές φορές αποτελούν μέθοδο επίλυσης του ίδιου προβλήματος. Η διαφορές των δύο αλγορίθμων είναι ουσιαστικές καθώς τα συγκριτικά πλεονεκτήματα καθενός αποτελούν σημαντικό παράγοντα για την ορθή επιλογή της μεθόδου συσταδοποίησης.

Ο αλγόριθμος k-meansσυγκεντρώνει τα εξής πλεονεκτήματα:

- Είναι εύληπτος και εύκολος στη χρήση
- Είναι αλγόριθμος χαμηλής πολυπλοκότητας και μπορεί να θεωρηθεί γραμμικός αλγόριθμος

Μειονεκτήματα του αλγορίθμου k-means:

- Σε περιπτώσεις κατηγορικών δεδομένων, οι μέσες τιμές δεν μπορούν να οριστούν. Για το λόγο αυτό η μέση τιμή αντικαθίσταται από τη συχνότητα, και ως κέντρα επιλέγονται οι παρατηρήσεις με τις υψηλότερες συχνότητες
- Το πλήθος των συστάδων δεν επιλέγεται από τον αλγόριθμο
- Ο αλγόριθμος είναι ευαίσθητος στην ύπαρξη απομακρυσμένων τιμών (outliers)

Για τον αλγόριθμο ιεραρχικής συσταδοποίησης παρατηρούνται τα εξής πλεονεκτήματα:

- Είναι εύληπτος και εύκολος στη χρήση
- Οι συνένωση ή ο διαχωρισμός των κλάσεων όταν πραγματοποιείται είναι οριστική διαδικασία. Έτσι, οι εναλλακτικές οι οποίες πρέπει να εξεταστούν σε μετέπειτα βήματα είναι μειώνονται σε αριθμό

Ενώ τα μειονεκτήματα του αλγορίθμου είναι:

- Οι συνένωση ή ο διαχωρισμός των κλάσεων όταν πραγματοποιείται είναι οριστική, επομένως τυχών λανθασμένες επιλογές, δεν είναι δυνατό να διορθωθούν σε μετέπειτα βήματα.
- Η διαδικασίες διαίρεσης ενός συνόλου είναι υπολογιστικά απαιτητικές

- Για μεγάλα σύνολα δεδομένων παρατηρείται δυσκολία κλιμάκωσης των αλγοριθμικών μεθόδων.

Οι διαφορές των δύο αλγορίθμων μπορούν να φανούν στον πίνακα που ακολουθεί

Ιδιότητες	k-means	Ιεραρχική Συσταδοποίηση
Ορισμός	Δημιουργία ξένων μη ιεραρχημένων συστάδων	Δημιουργία ιεραρχημένων συστάδων
Κριτήρια συσταδοποίησης	Απόσταση από το κέντρο	Πίνακας αποστάσεων
Κατηγορικά δεδομένα	Χρειάζεται μετασχηματισμός των κατηγορικών δεδομένων, πριν την εφαρμογή του αλγορίθμου	Χρειάζεται μετασχηματισμός των κατηγορικών δεδομένων, πριν την εφαρμογή του αλγορίθμου
Ευαισθησία σε απομακρυσμένες τιμές	Μεγάλη ευαισθησία του αλγορίθμου στην ύπαρξη απομακρυσμένων τιμών	Δεν παρουσιάζει μεγάλη ευαισθησία στην ύπαρξη απομακρυσμένων τιμών
Πλήθος συστάδων	Αποτελεί παράμετρο την οποία δεν υπολογίζει ο αλγόριθμος	Υπολογίζεται από τον αλγόριθμο και εξαρτάται μόνο από τη συνθήκη ολοκλήρωσης
Υπολογιστικός χρόνος	Ο αλγόριθμος είναι γραμμικός, επομένως ο υπολογιστικός χρόνος είναι χαμηλός	Απαιτεί αρκετό υπολογιστικό χρόνο λόγω υψηλής πολυπλοκότητας
Μέγεθος dataset	Συνίσταται για μεγάλα dataset	Δεν συνίσταται για μεγάλα datasets

Πίνακας 3 - Σύγκριση των δύο αλγορίθμων

Κεφάλαιο 6 – Εφαρμογή

Κατά την ολοκλήρωση της εργασίας πραγματοποιήθηκε εφαρμογή των αλγορίθμων k-means clustering και ιεραρχικής συσταδοποίησης στο σύνολο δεδομένων που προέκυψε έπειτα από δημοσκόπηση η οποία πραγματοποιήθηκε από την GallupWorldPoll για λογαριασμό των Ηνωμένων Εθνών. Το WorldHappinessReport αποτελεί ετήσια έκθεση η οποία συντάσσεται από το Sustainable Development SolutionsNetwork των Ηνωμένων Εθνών και περιλαμβάνει άρθρα και την σχετική κατάταξη 131 εθνών σχετικά με την «Εθνική Χαρά» η οποία συσχετίζεται με τις απαντήσεις των ερωτηθέντων και διάφορες παραμέτρους-δείκτες της διαβίωσης.

Η μελέτη που πραγματοποιήθηκε περιλαμβάνει τη σύγκριση των απαντήσεων των ερωτηθέντων για δύο διαδοχικές χρονιές (2017 και 2018), με δύο διαφορετικές μεθόδους την συσταδοποίησης, τις k-means clustering και ιεραρχικής συσταδοποίησης. Βασικό ερευνητικό ερώτημα το οποίο τέθηκε είναι το κατά πόσο οι δύο αλγόριθμοι συσταδοποίησης ομαδοποιούν με τον ίδιο τρόπο τα έθνη που συμμετείχαν στην έρευνα τις δύο χρονιές.

Οι μεταβλητές οι οποίες λήφθηκαν υπόψιν είναι οι εξής:

1. LogGDPpercapita, μεταβλητή η οποία δείχνει την αγοραστική δύναμη των χωρών που συμμετείχαν. Πηγή: “World Development Indicators”
2. SocialSupport, μέσος όρος των δυαδικών απαντήσεων ΝΑΙ-ΟΧΙ στην ερώτηση «Σε περίπτωση που χρειαστεί να αντιμετωπίσετε κάποιο πρόβλημα, υπάρχει κάποιος συγγενής ή φίλος για να βασιστείτε; »
3. Healthylifeexpectancyatbirth, μεταβλητή η οποία αντιπροσωπεύει το προσδόκιμο ζωής για κάθε έθνος. Τα δεδομένα ανακτήθηκαν από τον Παγκόσμιο Οργανισμό Υγείας “WorldHealthOrganization”. Καθώς διαθέσιμες τιμές υπάρχουν για τα έτη 2000, 2005, 2010, 2015 και 2016 η αναμενόμενη τιμή για τα έτη 2017 και 2018 υπολογίστηκε με τη μέθοδο της γραμμικής παλινδρόμησης.
4. Freedomtomakelifechoices, μέσος όρος των δυαδικών απαντήσεων ΝΑΙ-ΟΧΙ στην ερώτηση «Είστε ικανοποιημένος με την ελευθερία να επιλέγετε εσείς τι θα κάνετε στη ζωή σας; »
5. Generosity, μέσος όρος των δυαδικών απαντήσεων ΝΑΙ-ΟΧΙ στην ερώτηση «Δωρίσατε χρήματα σε φιλανθρωπία τον τελευταίο μήνα; »
6. Perceptionsofcorruption μέσος όρος των δυαδικών απαντήσεων ΝΑΙ-ΟΧΙ στην ερώτηση «Πιστεύετε ότι η διαφθορά έχει επεκταθεί στην καθημερινή ζωή; »
7. Positiveaffect μέσος όρος των δυαδικών απαντήσεων ΝΑΙ-ΟΧΙ στην ερώτηση «Νιώσατε κάποιο από τα παρακάτω συναισθήματα κατά την διάρκεια της προηγούμενης ημέρας; Χαρά, διασκέδαση, γέλιο; »
8. Negativeaffect μέσος όρος των δυαδικών απαντήσεων ΝΑΙ-ΟΧΙ στην ερώτηση «Νιώσατε κάποιο από τα παρακάτω συναισθήματα κατά την διάρκεια της προηγούμενης ημέρας; Λύπη, θλίψη, ανησυχία; »

9. Confidenceinnationalgovernment, μέσος όρος των δυαδικών απαντήσεων ΝΑΙ-ΟΧΙ στην ερώτηση «Πιστεύετε ότι μπορείτε να εμπιστευτείτε την κυβέρνηση για την επίλυση των σημαντικών προβλημάτων της χώρας σας; »

Η ανάλυση των δεδομένων πραγματοποιήθηκε με το λογισμικό ανοιχτού κώδικα OCTAVE και ο κώδικας ο οποίος παράχθηκε είναι συμβατός με την μέχρι σήμερα πιο πρόσφατη εμπορική έκδοση της MatlabR2017a. Οι συναρτήσεις οι οποίες χρησιμοποιήθηκαν ήταν οι:

1 kmeans(X,k), με ορίσματα το παραπάνω datasetX και k=9 το πλήθος των συστάδων
2 linkage(X,method, metrics), με ορίσματα το παραπάνω dataset X, μέθοδο average και μετρική την Chebishev για τη δημιουργία του πίνακα αποστάσεων
3 Cluster(Z,k) με ορίσματα τον παραπάνω πίνακα αποστάσεων Z και k=9 το πλήθος των συστάδων

Κατά την ολοκλήρωση, πραγματοποιήθηκε μετα-ανάλυση των αποτελεσμάτων σε περιβάλλον Microsoft Excel.

Οι 131 χώρες χωρίστηκαν σύμφωνα με τις απαντήσεις των πολιτών τους σε 9 συστάδες οι οποίες συμβολίζονται με $A, B, C, \dots I$. Στη συνέχεια συγκρίθηκαν τα αποτελέσματα της συσταδοποίησης των δύο αλγορίθμων για το έτος 2017 και το έτος 2018 ως εξής:

Καταγράφηκαν το σύνολο των γειτόνων κάθε χώρας (χωρών που ανήκουν δηλαδή στην ίδια κλάση) για τα δύο έτη, και συγκρίθηκαν ένας προς ένας οι γείτονες για τις δύο μεθόδους.

Ως μέτρο ομοιότητας των δύο αλγορίθμων χρησιμοποιήθηκε η εξής κανονικοποιημένη ποσότητα

$$Sim_j = \frac{\sum_{i=1}^{N} X_{ij}}{N}$$

Όπου:

- $j = \eta\ \varepsilon\xi\varepsilon\tau\alpha\zeta\acute{o}\mu\varepsilon\nu\eta\ \chi\acute{\omega}\rho\alpha$
- $N = To\ \pi\lambda\acute{\eta}\theta o\varsigma\ \tau\omega\nu\ \chi\omega\rho\acute{\omega}\nu\ \pi o\upsilon\ \sigma\upsilon\mu\mu\varepsilon\tau\varepsilon\acute{\iota}\chi\alpha\nu\ \sigma\tau\eta\nu\ \acute{\varepsilon}\rho\varepsilon\upsilon\nu\alpha$
- $X_{ij} = \begin{Bmatrix} \delta\upsilon\alpha\delta\iota\kappa\acute{\eta}\ \mu\varepsilon\tau\alpha\beta\lambda\eta\tau\acute{\eta}\ \eta\ o\pi o\acute{\iota}\alpha\ \lambda\alpha\mu\beta\acute{\alpha}\nu\varepsilon\iota\ \tau\eta\nu\ \tau\iota\mu\acute{\eta}\ 1\ \acute{o}\tau\alpha\nu\ o\iota\ \chi\acute{\omega}\rho\varepsilon\varsigma\ i, j \\ \tau\alpha\xi\iota\nu o\mu o\acute{\upsilon}\nu\tau\alpha\iota\ \omega\varsigma\ \gamma\varepsilon\acute{\iota}\tau o\nu\varepsilon\varsigma\ \kappa\alpha\iota\ \alpha\pi\acute{o}\ \tau o\upsilon\varsigma\ \delta\acute{\upsilon}o\ \alpha\lambda\gamma o\rho\acute{\iota}\theta\mu o\upsilon\varsigma\ \acute{\eta}\ \acute{o}\tau\alpha\nu \\ \delta\varepsilon\nu\ \tau\alpha\xi\iota\nu o\mu o\acute{\upsilon}\nu\tau\alpha\iota\ \omega\varsigma\ \gamma\varepsilon\acute{\iota}\tau o\nu\varepsilon\varsigma\ \kappa\alpha\iota\ \alpha\pi\acute{o}\ \tau o\upsilon\varsigma\ \delta\acute{\upsilon}o\ \alpha\lambda\gamma o\rho\acute{\iota}\theta\mu o\upsilon\varsigma \\ \kappa\alpha\iota\ 0\ \delta\iota\alpha\phi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{Bmatrix}$

Για το έτος 2018 οι δύο αλγόριθμοι παρουσίασαν ποσοστό συμφωνίας γειτόνων ίσο με 89,46% ενώ για το έτος 2017 το ποσοστό αυτό ανήλθε στο 93,49%.

Για το έτος 2017 δεκαπέντε χώρες παρουσίασαν ομοιότητα στα αποτελέσματά τους με ποσοστό μεγαλύτερο του 90% ενώ για το 2018, 14 από αυτές εμφάνισαν τέτοια ποσοστά επιτυχίας.

Αντίθετα, για το έτος 2017 σε 6 περιπτώσεις το ποσοστό ομοιότητας γειτόνων είναι μικρότερο του 50% ενώ για το 2018 οι περιπτώσεις με «μικρό» ποσοστό ομοιότητας είναι 16. Αναλυτικά τα ποσοστά ομοιότητας των αποτελεσμάτων των δύο αλγορίθμων είναι για τα δύο έτη, καθώς και οι κλάσεις στις οποίες ανήκει η κάθε χώρα για τις διαφορετικές ταξινομήσεις παρατίθενται στο τέλος της εργασίας.

Βιβλιογραφικές αναφορές

Abdi, H. & Williams , L., 2010. Principal component analysis.. *WIREs Comp Stat,* Τόμος 2, pp. 433-459.

Bella, A., Ferri, C., Hernandesz-Orallo, J. & Ramirez-Quintana, . M. J., 2010. Calibration of Machine Learning Models. Στο: E. S. a. G. J. D. a. M. M. a. M. J. R. a. S. L. A. J. Olivas, επιμ. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques.* s.l.:IGI Global, pp. 128-146.

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning.* 1st επιμ. s.l.:Springer.

Boden, M., 2001. *A guide to recurrent neural networks and backpropagation.* s.l.:School of Information Science, Computer and Electrical Engineering.

Cortes, C. & Vapnik, V., 1995. Support-Vector Networks. *Machine Learning,* Τόμος 20, pp. 273-297.

Dietterich, T. G., 2000. *Ensemble Methods in Machine Learning,* Corvallis: Oregon State University.

Domingos, P., 2012. A Few Useful. *Communications of the ACM,* Volume 55(10).

Draper, N. & Smith, . H., 1981. *Applied Regression Analysis.* 2nd edition επιμ. s.l.:s.n.

Esposito, F., Malerba, D., Semararo, G. & Kay, J., 1997. A Comparative Analysis of Methods for Pruning Decision Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , Τόμος 19, pp. 479-491.

Garson, D., 2014. *LOGISTIC REGRESSION: BINARY AND MULTINOMIAL.* 2014 Edition επιμ. Asheboro, NC 27205 USA: G. David Garson and Statistical Associates Publishing.

Gerven , . M. v. & Bohte, S., 2016. Artificial Neural Networks as Models of Neural Information Processing. *Frontiers in Computational Neuroscience,* 10(94).

Hofmann, M., 2006. *Support Vector Machines — Kernels and the.* Bamberg: University of Bamberg.

Howley, T. & Madden, M. G., 2006. An Evolutionary Approach to Automatic Kernel. *Lecture Notes in Computer Science.*

James, G., Witten, D., Hastie , T. & Tibshirani, R., 2013. *An Introduction to Statistical Learning.* 1st επιμ. s.l.:Springer.

Kittler, J. & Roli, F., 2000. *Multiple Classifier Systems.* Cagliari: Springer.

Kubat, M., 2017. *An introduction to Machine Learning.* 2nd επιμ. Coral Gables: Springer.

Mitchell, T. καισυν., 1990. Machine Learning. *Annual Review of Computer Science,* 4(1), pp. 417-433.

Mitchell, T. M., 1997. *Machine Learning.* New York: McGraw-Hill.

Mohammed, M., Khan, M. B. & Bashier, E. B. M., 2017. *Machine Learning: Algorithms and Applications.* s.l.:CRC Press.

Nazari, J. & Ersoy, O. K., 1992. *Implementation of back-propagation neural networks with MatLab.* s.l.:Electrical and Computer Engineering, Purdue University Libraries.

Nilsson, N., 1998. *Introduction to machine learning*, Stanford: University of Stanford.

Poole, D., Mackworth , A. & Goebel, R., 1998. *Computational Intelligence: A Logical Approach.* 1st επιμ. New York: Oxford University Press..

Priyatharsini, P., Chandrasekar, C. & Phil Scholar, M., 2017. Email Spam Filtering using Classifiers in Data Mining. *International Journal of Engineering Science and Computing,* 7(Issue No.11).

Quinlan, J. R., 1993. *C4.5: Programs for Machine Learning.* s.l.:Copyright © 1993 Elsevier Inc. All rights reserved..

Rokach, L. & Maimon, O., 2015. *Data mining with decision trees. Theory and applications.* 2nd επιμ. Singapore: World Scientific Publishing Co. Pte. Ltd.

Shannon, C. E., 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, Τόμος 27, pp. 349/423, 623-656.

Taiwo, O. A., 2010. *Introduction to Machine Learning, New Advances in Machine Learning.* Rijeka: Yagang Zhang.

Urban, S., 2017. *Neural Network Architectures and Activation Functions: A Gaussian Process Approach.* Munich: Technical University Munich.

Vapnik, V., 1995. *The nature of statistical learning theory.* 1st επιμ. Heidelberg: Springer-Verlag Perlin.

Wu, X. καισυν., 2007. Top 10 algorithms in data mining. *Knowl Inf Syst,* Τόμος 14, p. 1–37.

Μπόρα-Σεντα, Ε. & Μωυσιάδης, Χ., 1997. *Εφαρμοσμένη Στατιστική - Πολλαπλή Παλινδρόμηση, Ανάλυση Διασποράς, Χρονοσειρές.* 2η επιμ. Θεσσαλονίκη: Ζήτη.

Παράρτημα -1

ΚώδικαςOctave/Matlabk-means clustering

```
filename = 'data_2017.csv';%import the data.file

k=9;%select the number of classes

X =readtable(filename,'ReadRowNames',true);%data with names

Y=readtable(filename,'ReadRowNames',false);%data without names

A = table2array(X);%transform to array

Z = linkage(A,'average','chebychev');%selection of similarity metrics

T = cluster(Z,'maxclust',9);%clustering precedure

%scatter3(X(:,1),X(:,2),X(:,3),10,c)

cutoff = median([Z(end-2,3) Z(end-1,3)]);%visualization

dendrogram(Z,'ColorThreshold',cutoff)%

lastTwo = Z(end-1:end,:)%

TT=crosstab(T,Y.CountryName)%

K=table(TT);%

[idx,c] = kmeans(A,k);%

K=table(Y.CountryName,idx);%

K.Properties.VariableNames = {'Country' 'Class'}

writetable(K,'results2017.csv')%save results

%re-calculate for 2018

type 'results2018.csv'

filename = 'data_2017.csv';

X =readtable(filename,'ReadRowNames',true);

Y=readtable(filename,'ReadRowNames',false);

A = table2array(X);

[idx,c] = kmeans(A,10);

K=table(Y.CountryName,idx);

K.Properties.VariableNames = {'Country' 'Class'}

%write(TT,'hresults2018.csv')

%type 'results2017.csv'
```

Παράρτημα -2

ΚώδικαςOctave/Matlabιεραρχική συσταδοποίηση

```
filename = 'data_2017.csv';

k=9;%number of classes

X =readtable(filename,'ReadRowNames',true);

Y=readtable(filename,'ReadRowNames',false);

A = table2array(X);

Z = linkage(A,'average','chebychev');

T = cluster(Z,'maxclust',9);

cutoff = median([Z(end-2,3) Z(end-1,3)]);

dendrogram(Z,'ColorThreshold',cutoff)

lastTwo = Z(end-1:end,:)

TT=crosstab(T,Y.CountryName)

K=table(TT);

write(TT,'hresults2018.csv')

%type 'results2017.csv'
```

Παράρτημα -3

Αποτελέσματα

Country	2017
United Arab Emirates	95%
Guinea	94%
Ukraine	94%
Turkmenistan	93%
Spain	93%
Greece	93%
Malawi	92%
Pakistan	91%
Thailand	91%
Zimbabwe	90%
Russia	90%
Kenya	90%
Germany	90%
Portugal	90%
Nigeria	90%
Tunisia	89%
Ivory Coast	89%
France	88%
Nepal	88%
Israel	88%
Namibia	88%
Rwanda	87%
Romania	86%
Italy	86%
El Salvador	86%
Haiti	86%
Colombia	86%
Guatemala	86%

Country	2017
Gabon	85%
Afghanistan	85%
Montenegro	85%
Bulgaria	85%
Uzbekistan	84%
Botswana	84%
Austria	84%
Albania	83%
Zambia	83%
Sweden	83%
Congo (Brazzaville)	83%
Yemen	83%
Lithuania	83%
Uganda	82%
Tanzania	82%
Algeria	82%
Sri Lanka	82%
Turkey	82%
South Africa	82%
United Kingdom	82%
Ethiopia	82%
Togo	81%
South Korea	81%
Slovakia	81%
Peru	80%
Sierra Leone	80%
Panama	80%
Norway	79%

Country	2017
Slovenia	79%
New Zealand	78%
Nicaragua	78%
Mozambique	78%
Myanmar	78%
Netherlands	78%
Serbia	78%
Madagascar	77%
Malta	77%
Palestinian Territories	76%
Mexico	76%
Moldova	76%
Finland	75%
Mauritius	75%
Iran	75%
Venezuela	75%
Ireland	74%
Niger	74%
Uruguay	74%
Kyrgyzstan	74%
Croatia	74%
Cyprus	74%
Morocco	73%
Vietnam	73%
Dominican Republic	73%
Taiwan Province of China	73%
Jordan	72%

Country	2017
Mongolia	72%
Cameroon	72%
Chile	72%
United States	71%
Senegal	71%
Lebanon	70%
Gambia	70%
Switzerland	70%
Philippines	70%
Kosovo	69%
Canada	69%
Singapore	68%
Macedonia	68%
Georgia	67%
Bolivia	67%
Saudi Arabia	66%
Belgium	66%

Country	2017
Liberia	66%
China	66%
Latvia	66%
Mauritania	65%
Mali	64%
Bosnia and Herzegovina	64%
India	64%
Benin	63%
Luxembourg	62%
Ecuador	62%
Libya	62%
Denmark	62%
Laos	60%
Costa Rica	60%
Kazakhstan	58%
Cambodia	58%
Japan	58%

Country	2017
Bangladesh	57%
Indonesia	56%
Honduras	55%
Armenia	55%
Estonia	54%
Egypt	53%
Czech Republic	51%
Chad	50%
Burkina Faso	48%
Brazil	47%
Belarus	46%
Azerbaijan	45%
Argentina	45%
Australia	45%

Country	2018
Zimbabwe	100%
Turkmenistan	98%
Jordan	98%
Uganda	97%
Spain	95%
Kazakhstan	95%
Sri Lanka	94%
Germany	94%
Greece	93%
France	92%
Gabon	91%
Afghanistan	90%
United States	90%
Albania	90%
Turkey	89%
Mali	89%
United Arab Emirates	88%
Russia	87%
Libya	87%
Togo	86%
Morocco	86%
Kosovo	86%
Switzerland	85%
Saudi Arabia	85%
Zambia	84%
Mauritania	84%
Japan	84%
Sweden	83%
South Korea	83%
Norway	83%

Country	2018
Peru	82%
Slovenia	82%
Luxembourg	82%
Honduras	82%
Guinea	82%
Nepal	82%
South Africa	82%
Palestinian Territories	82%
Singapore	82%
Slovakia	81%
Ireland	81%
Congo (Brazzaville)	80%
Sierra Leone	80%
Myanmar	80%
Niger	80%
Azerbaijan	80%
Yemen	79%
New Zealand	79%
Finland	79%
Vietnam	79%
Uzbekistan	78%
Nicaragua	78%
Estonia	78%
Burkina Faso	78%
Argentina	78%
United Kingdom	78%
Netherlands	78%
Mongolia	78%
Czech	78%

Country	2018
Republic	
Tanzania	77%
Mexico	77%
China	77%
Tunisia	76%
Moldova	76%
Serbia	75%
Mauritius	75%
Laos	75%
Romania	74%
Portugal	74%
Italy	74%
Kyrgyzstan	74%
Kenya	73%
Panama	73%
Pakistan	72%
Madagascar	72%
Georgia	72%
Ivory Coast	71%
Israel	70%
Mozambique	70%
Lebanon	70%
Chad	70%
Montenegro	70%
Guatemala	69%
Bosnia and Herzegovina	69%
Indonesia	69%
Malta	68%
Benin	68%
Gambia	67%

Country	2018
Egypt	67%
Iran	66%
Dominican Republic	66%
Canada	66%
El Salvador	65%
Venezuela	65%
Croatia	65%
Uruguay	64%
Bolivia	64%
Cyprus	63%
Cameroon	63%
Belgium	63%
Ukraine	62%
Chile	62%

Country	2018
Brazil	62%
Thailand	61%
Taiwan PC	60%
Belarus	60%
Senegal	58%
Australia	58%
Rwanda	58%
Philippines	56%
Nigeria	54%
Namibia	53%
Malawi	51%
Macedonia	50%
Lithuania	50%
Liberia	49%
Latvia	48%

Country	2018
India	46%
Haiti	46%
Ethiopia	44%
Ecuador	42%
Denmark	42%
Costa Rica	40%
Colombia	39%
Cambodia	38%
Bulgaria	37%
Botswana	36%
Bangladesh	34%
Austria	34%
Algeria	33%
Armenia	33%

Year – Method

Country	k-means 2017	Hierarchical clustering2017	k-means 2018	Hierarchical clustering 2018
Afghanistan	G	H	G	G
Albania	D	I	D	A
Algeria	A	I	A	A
Argentina	D	I	D	A
Armenia	A	I	A	A
Australia	C	C	C	C
Austria	C	C	C	C
Azerbaijan	H	I	A	A
Bangladesh	H	I	H	A
Belarus	A	F	A	A
Belgium	C	I	C	C
Benin	F	C	F	F

	Year – Method			
Country	**k-means 2017**	**Hierarchical clustering2017**	**k-means 2018**	**Hierarchical clustering 2018**
Bolivia	H	H	H	A
Bosnia and Herzegovina	A	F	A	A
Botswana	F	I	E	F
Brazil	A	E	A	A
Bulgaria	A	I	A	A
Burkina Faso	G	I	F	F
Cambodia	E	F	E	A
Cameroon	G	H	G	G
Canada	C	C	C	C
Chad	G	G	G	G
Chile	D	G	D	C
China	D	I	D	A
Colombia	A	I	A	A
Congo (Brazzaville)	F	E	E	F
Costa Rica	D	H	C	C
Croatia	D	I	D	C
Cyprus	C	I	I	I
Czech Republic	D	C	D	C
Denmark	C	I	C	C
Dominican Republic	A	C	A	A
Ecuador	A	I	D	A
Egypt	E	I	E	A
El Salvador	A	F	A	A
Estonia	D	I	D	A
Ethiopia	F	I	E	F

	Year – Method			
Country	**k-means 2017**	**Hierarchical clustering2017**	**k-means 2018**	**Hierarchical clustering 2018**
Finland	C	E	C	C
France	C	C	C	C
Gabon	E	C	E	F
Gambia	F	E	F	F
Georgia	H	H	H	A
Germany	C	F	C	C
Greece	C	C	C	C
Guatemala	H	E	H	A
Guinea	G	C	F	F
Haiti	F	F	F	F
Honduras	A	H	A	A
India	E	H	E	F
Indonesia	E	I	H	A
Iran	A	C	A	A
Ireland	C	I	C	C
Israel	C	C	C	C
Italy	C	E	C	C
Ivory Coast	G	F	G	G
Japan	C	I	C	C
Jordan	A	E	A	A
Kazakhstan	H	C	H	A
Kenya	E	C	E	F
Kosovo	H	C	I	E
Kyrgyzstan	H	G	H	A
Laos	F	I	E	F
Latvia	A	C	A	A
Lebanon	A	I	A	A

	Year – Method			
Country	**k-means 2017**	**Hierarchical clustering2017**	**k-means 2018**	**Hierarchical clustering 2018**
Liberia	F	F	F	F
Libya	E	E	E	E
Lithuania	A	F	A	A
Luxembourg	C	I	C	C
Macedonia	A	F	A	A
Madagascar	F	E	E	F
Malawi	F	I	F	F
Mali	G	G	G	G
Malta	C	H	I	I
Mauritania	F	F	F	F
Mauritius	A	I	A	A
Mexico	A	C	D	A
Moldova	H	I	H	A
Mongolia	E	E	H	A
Montenegro	D	E	D	A
Morocco	A	H	A	A
Mozambique	G	C	F	F
Myanmar	F	E	E	F
Namibia	F	I	F	F
Nepal	E	I	H	A
Netherlands	C	F	C	C
New Zealand	C	F	C	C
Nicaragua	A	I	A	A
Niger	G	I	G	F
Nigeria	G	H	G	G
Norway	C	E	C	B
Pakistan	F	E	E	C

	Year – Method			
Country	**k-means 2017**	**Hierarchical clustering2017**	**k-means 2018**	**Hierarchical clustering 2018**
Palestinian Territories	B	F	B	F
Panama	D	C	D	B
Peru	A	C	D	C
Philippines	E	I	H	A
Portugal	C	H	C	A
Romania	A	G	A	C
Russia	H	C	H	A
Rwanda	E	E	E	A
Saudi Arabia	A	A	A	A
Senegal	E	I	E	A
Serbia	A	I	D	F
Sierra Leone	G	I	G	A
Singapore	C	F	I	G
Slovakia	D	C	D	A
Slovenia	D	I	C	C
South Africa	F	F	F	F
South Korea	C	E	C	C
Spain	C	I	C	C
Sri Lanka	A	E	A	A
Sweden	C	G	C	C
Switzerland	C	C	C	C
Taiwan Province of China	B	I	B	B
Tanzania	F	I	F	F
Thailand	A	E	A	A
Togo	F	C	F	F

	Year – Method			
Country	**k-means 2017**	**Hierarchical clustering2017**	**k-means 2018**	**Hierarchical clustering 2018**
Tunisia	A	B	A	A
Turkey	A	C	A	A
Turkmenistan	E	I	H	A
Uganda	F	C	F	F
Ukraine	H	C	H	A
United Arab Emirates	A	A	A	A
United Kingdom	C	F	C	C
United States	D	E	D	A
Uruguay	D	I	D	A
Uzbekistan	H	H	H	A
Venezuela	A	F	A	A
Vietnam	A	I	A	A
Yemen	I	I	F	D
Zambia	F	F	F	F
Zimbabwe	F	H	F	F