Σχολή Οικονομικών Επιστημών και Διοίκησης Επιχειρήσεων - Τμήμα Διοικητικής Επιστήμης και Τεχνολογίας

http://www.c2learn.com/lecture_notes/word_cloud_DM2.gif

Υλοποίηση σε R αλγορίθμων κατηγοριοποίησης και εφαρμογή τους σε δεδομένα του επιχειρηματικού κόσμου

Σοφάκης Ευθύμιος – Δήμου Ελευθερία – Κρινάς Γεώργιος

Χαλκιόπουλος Κωνσταντίνος

2021

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ 3

2 ΚΑΤΗΓΟΡΟΙΟΠΟΙΗΣΗ 5

2..1 Τι είναι η Κατηγοριοποίηση ; Τι είναι η Πρόβλεψη ; 5

2..2 Προετοιμασία των δεδομένων για Κατηγοριοποίηση και Πρόβλεψη 8

2..3 Τεχνικές Κατηγοριοποίησης 13

2..4 Εφαρμογές (γιατί χρειάζεται το classification) 29

3 Εφαρμογή σε R 30

3..1 Περιγραφή της εφαρμογής 30

3..2 ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΘΑ ΑΝΑΛΥΣΟΥΜΕ .. 32

3..3 ΑΠΟΤΕΛΕΣΜΑΤΑ – ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ 33

4 Βιβλιογραφία 42

1 ΕΙΣΑΓΩΓΗ

Στις μέρες μας υπάρχει ένα τεράστιο ποσό δεδομένων τα οποία συλλέγονται και αποθηκεύονται σε βάσεις δεδομένων σε όλα τα μέρη παγκοσμίως. Συνεχίζει να έχει αυξητική τάση κάθε χρόνο. Δεν είναι δύσκολο να βρούμε βάσεις δεδομένων με Terabytes δεδομένων σε επιχειρήσεις και κέντρα ερευνών. Αυτά είναι περισσότερα από 1,099,511,627,776 bytes δεδομένων. Υπάρχει ανεκτίμητη πληροφορία και γνώση ‘’κρυμμένη’’ σε τέτοιες βάσεις δεδομένων˙ και χωρίς αυτοματοποιημένες μεθόδους για την εξόρυξη τέτοιας πληροφορίας είναι πρακτικά αδύνατο να την εξορύξεις. Στην πορεία των χρονών δημιουργήθηκαν πολλοί αλγόριθμοι για την εξόρυξη αυτού που ονομάζεται όγκος γνώσης από μεγάλες συλλογές δεδομένων. Υπάρχουν διάφορες μεθοδολογίες για την προσέγγιση αυτού του προβλήματος: κατηγοριοποίηση(classification), κανόνας σχέσης(association rule), ομαδοποίηση(clustering), κτλ. Η συγκεκριμένη πτυχιακή θα εστιάσει στη κατηγοριοποίηση.

Η κατηγοριοποίηση(classification), βάση μιας συγκεκριμένης εισόδου προβλέπει ένα συγκεκριμένο αποτέλεσμα. Η πρόβλεψη του αποτελέσματος γίνεται με τη χρήση ενός αλγορίθμου πρόβλεψης, ο οποίος επεξεργάζεται ένα σύνολο εκπαίδευσης που περιέχει ένα σύνολο ιδιοτήτων και την αντίστοιχη έκβαση, συνήθως ονομάζεται στόχος ή χαρακτηριστικό πρόβλεψης. Ο αλγόριθμος προσπαθεί να ανακαλύψει σχέσεις μεταξύ των γνωρισμάτων που θα καθιστούσε δυνατόν να προβλεφθεί το αποτέλεσμα. Στη συνέχεια στον αλγόριθμο δίνεται ένα σύνολο δεδομένων που δεν έχουμε ξαναδεί, που ονομάζεται σύνολο πρόβλεψης, το οποίο περιέχει το ίδιο σύνολο χαρακτηριστικών, εκτός από το χαρακτηριστικό πρόβλεψης - δεν είναι ακόμη γνωστό. Ο αλγόριθμος αναλύει την είσοδο και παράγει μια πρόβλεψη. Η ακρίβεια πρόβλεψης καθορίζει πόσο "καλός" ο αλγόριθμος είναι. (Fabricio Voznika, n.d.)

Η κατηγοριοποίηση (classification), η οποία είναι η εργασία ανάθεσης αντικειμένων σε μια από τις διαφορετικές προκαθορισμένες κατηγορίες, είναι ένα ευρέως γνωστό πρόβλημα που περιλαμβάνει πολλές και ποικίλες εφαρμογές. Παραδείγματα εφαρμογών αποτελούν η κατηγοριοποίηση κυττάρων ς κακοήθη ή καλοήθη βασιζόμενη σε αποτελέσματα εξετάσεων MRI, η κατηγοριοποίηση γαλαξιών με βάση το σχήμα τους, η κατηγοριοποίηση αξιόπιστων ή μη δανειοληπτών, η ανίχνευση ανεπιθύμητων ηλεκτρονικών μηνυμάτων βασιζόμενη στην επικεφαλίδα του μηνύματος κ.α. Βασική λειτουργία της κατηγοριοποίησης είναι η δημιουργία και η εκπαίδευση ενός μοντέλου χρησιμοποιώντας ιστορικά δεδομένα, δεδομένα δηλαδή για τα οποία εκ των προτέρων γνωρίζουμε την κατηγορία τους βασιζόμενοι σε επιλεγμένα χαρακτηριστικά. Οι αλγόριθμοι κατηγοριοποίησης χρησιμοποιούνται τόσο στην περιγραφική (descriptive) όσο και στην προβλεπτική (predictive) μοντελοποίηση εργασιών της Εξόρυξης Δεδομένων (Data Mining). Στη βιβλιογραφία έχουν προταθεί αρκετοί αλγόριθμοι κατηγοριοποίησης όπως είναι τα Δένδρα Απόφασης, οι κατηγοριοποιητές κανόνων, οι κατηγοριοποιητές πλησιέστερου γείτονα, οι κατηγοριοποιητές του Bayes κ.α. Η πρόκληση των αλγόριθμων κατηγοριοποίησης είναι η δημιουργία ενός μοντέλου το οποίο να παρουσιάζει τη μεγαλύτερη δυνατή γενίκευση, δηλαδή να κατηγοριοποιεί με τη μέγιστη δυνατή αξιοπιστία άγνωστες περιπτώσεις. Για την αξιολόγηση των κατηγοριοποιητών έχουν προταθεί διάφορα μέτρα απόδοσης, όπως είναι η καμπύλη χαρακτηριστικής λειτουργίας δέκτη, το μέτρο ανάκλησης-ακρίβειας κ.α.

Στην παρούσα πτυχιακή θα υλοποιηθεί μια διαδικτυακή εφαρμογή χρησιμοποιώντας το στατιστικό πακέτο R, στην οποία θα λαμβάνονται επιχειρηματικά δεδομένα, θα εφαρμόζονται αλγόριθμοι κατηγοριοποίησης, θα εκτιμώνται τα αποτελέσματα της ανάλυσης με βάση κριτήρια απόδοσης και θα παρουσιάζονται τα αποτελέσματα.

Στόχοι της πτυχιακής εργασίας είναι η θεωρητική προσέγγιση του προβλήματος της κατηγοριοποίησης, η δημιουργία ενός εύχρηστου περιβάλλοντος όπου θα μπορούν να εφαρμοστούν τεχνικές κατηγοριοποίησης και η αξιολόγησή τους καθώς και η εξοικείωση των σπουδαστών με ένα από τα πλέον δημοφιλή προγραμματιστικά λογισμικά πακέτα, την R. (Maimon & Rokach, 2010) (Everitt & Hothorn, 2010) (Nisbet, Elder, & Miner, 2009) (Han & Kamber, 2006) (Tan, Steinbach, & Kumar, 2006)

Στο δεύτερο κεφάλαιο θα γίνει αναλυτική αναφορά της κατηγοριοποίησης, της πρόβλεψης, τι είναι η κάθε μια, πώς γίνετε η προετοιμασία των δεδομένων για την κατηγοριοποίηση και για την πρόβλεψη, ποιες τεχνικές υπάρχουν και που μας χρησιμεύει μια τέτοια διαδικασία.

Στο τρίτο κεφάλαιο θα γίνει η δημιουργία μιας εφαρμογής σε γλώσσα R. Θα γίνει περιγραφή της εφαρμογής σχετικά με το πως δημιουργήθηκε και πως λειτουργεί, περιγραφή του συνόλου δεδομένων που θα χρησιμοποιηθούν για την ανάλυσή τους χρησιμοποιώντας την εφαρμογή και θα παρουσιάσουμε τα αποτελέσματα και τη συγκριτική ανάλυση.

Τέλος θα βγάλουμε κάποια συμπεράσματα και θα γίνει συζήτηση σχετικά με αυτά.

2 ΚΑΤΗΓΟΡΟΙΟΠΟΙΗΣΗ

2..1 Τι είναι η Κατηγοριοποίηση ; Τι είναι η Πρόβλεψη ;

Ένα πρωτοφανές ποσό δεδομένων παράγεται σε συνεχώς αυξανόμενο ρυθμό σε πολλούς κλάδους. Κάθε μέρα επιχειρήσεις λιανικής πώλησης συλλέγουν στοιχεία από τις συναλλαγές πωλήσεων, οργανισμοί καταμετρούν πόσα κλικ του ποντικιού γίνονται στις ιστοσελίδες τους, και βιολόγοι παράγουν εκατομμύρια κομμάτια πληροφοριών σχετικά με τα γονίδια. Είναι σχεδόν αδύνατον να βγει κάποιο νόημα των συνόλων δεδομένων που περιέχουν περισσότερο από μια χούφτα δεδομένων χωρίς τη βοήθεια υπολογιστικών προγραμμάτων. (Glenn J. Myatt, 2014)

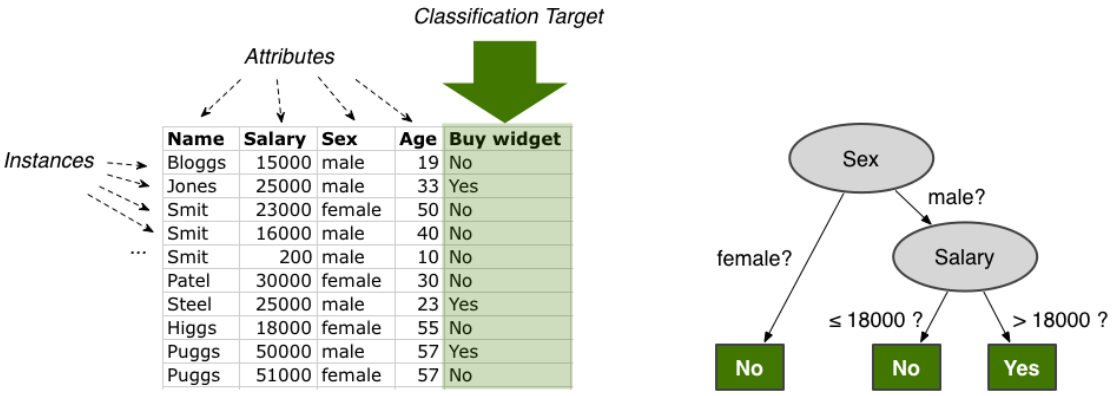
Η εύρεση συγκεκριμένων αρχείων ή πληροφοριών σε ένα τόσο μεγάλο σύνολο δεδομένων είναι ανεπαρκής και συχνά μη ολοκληρωμένη. Οι επιχειρήσεις συχνά αποτυγχάνουν να αναγνωρίσουν το πόσο σημαντικά είναι τα σύνολα δεδομένων τους και την επίπτωση στις καθημερινές λειτουργείες της επιχείρησης. Η διαδικασία της «κατηγοριοποίησης δεδομένων» επιχειρεί να καλύψει αυτό το κενό βοηθώντας τις επιχειρήσεις να κατανοήσουν τί δεδομένα είναι πραγματικά διαθέσιμα, τη θέση τους στην επιχείρηση, πως μπορεί να αποκτηθεί πρόσβαση σε αυτά τα δεδομένα και πώς πρέπει να προστατευτούν για να καλύψουν τις νόμιμες και κανονιστικές απαιτήσεις. (Bigelow, 2005)

Η κατηγοριοποίηση δεδομένων είναι γνωστή ως η διαδικασία οργάνωσης των δεδομένων κατά σχετικών κατηγοριών ώστε να χρησιμοποιηθούν και να προστατευτούν πιο αποτελεσματικά. Η διαδικασία κατηγοριοποίησης όχι μόνο κάνει τα δεδομένα πιο εύκολα ανιχνεύσιμα και ανακτώμενα αλλά είναι και σημαντικής σημασίας όσο αφορά τη διαχείριση κινδύνου, τη συμμόρφωση και την ασφάλεια δεδομένων.

Η κατηγοριοποίηση δεδομένων περιλαμβάνει την απόδοση μιας ετικέτας στα δεδομένα, που τα κάνει εύκολο να ανιχνευτούν και να βρεθούν. Επίσης αφαιρεί πολλαπλές αντιγραφές των δεδομένων, το οποίο μπορεί να μειώσει το αποθηκευτικό και εφεδρικό κόστος, όπως και να επιταχύνει τη διαδικασία αναζήτησης. (Lord, 2016)

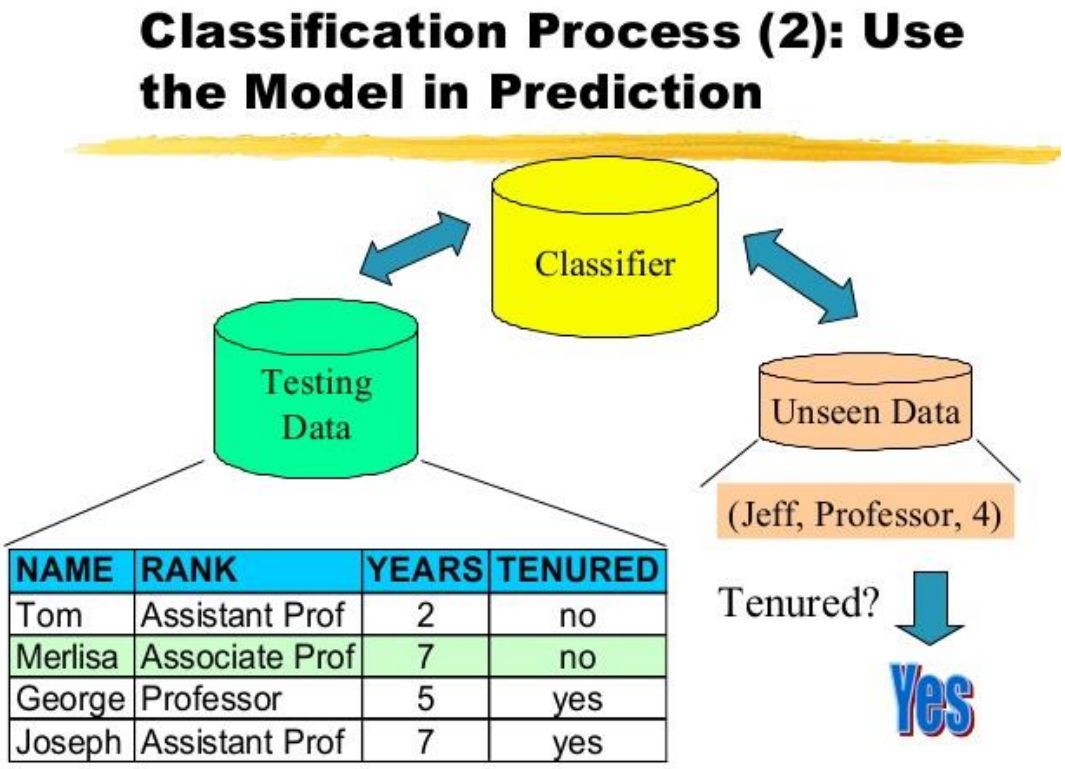
Η ταξινόμηση και η εκτίμηση πιθανότητας κατηγορίας προσπαθούν να προβλέψουν, για κάθε άτομο σε έναν πληθυσμό, σε ποιο (μικρό) σύνολο κατηγοριών αυτό το άτομο ανήκει. Συνήθως οι κατηγορίες είναι αμοιβαία αποκλειστικές. Ένα παράδειγμα ερώτησης κατηγοριοποίησης θα ήταν : «Μεταξύ όλων των πελατών της MegaTelCo, ποιοι είναι οι πιο πιθανοί να ανταποκριθούν σε μια προσφορά;» Σε αυτό το παράδειγμα οι δύο κατηγορίες θα μπορούσαν να ονομαστούν θα αποκριθεί και δεν θα αποκριθεί.

Για έναν στόχο ταξινόμησης, μια διαδικασία εξόρυξης δεδομένων παράγει ένα πρότυπο που, λαμβάνοντας υπόψη ένα νέο άτομο, καθορίζει σε ποια κατηγορία το άτομο αυτό ανήκει. Ένας στενά συνδεδεμένος στόχος είναι η βαθμολόγηση ή η εκτίμηση πιθανότητας κατηγορίας. Ένα πρότυπο βαθμολόγησης που εφαρμόζεται σε ένα άτομο παράγει, αντί μιας πρόβλεψης κατηγορίας, ένα αποτέλεσμα που αντιπροσωπεύει την πιθανότητα (ή κάποιο άλλο προσδιορισμό της ποσοτικοποίησης της πιθανότητας) ότι εκείνο το άτομο ανήκει σε κάθε κατηγορία. Στο σενάριο ανταπόκρισης των πελατών, ένα πρότυπο βαθμολόγησης θα ήταν ικανό να αξιολογήσει κάθε μεμονωμένο πελάτη και να παράγει ένα αποτέλεσμα για το πόσο πιθανό είναι ο καθένας να ανταποκριθεί στη προσφορά. Η ταξινόμηση και η βαθμολόγηση είναι πολύ στενά συνδεδεμένες· όπως θα δούμε, ένα πρότυπο που μπορεί να κάνει ένα μπορεί συνήθως να τροποποιηθεί να κάνει και το άλλο. (Foster Provost, 2013)



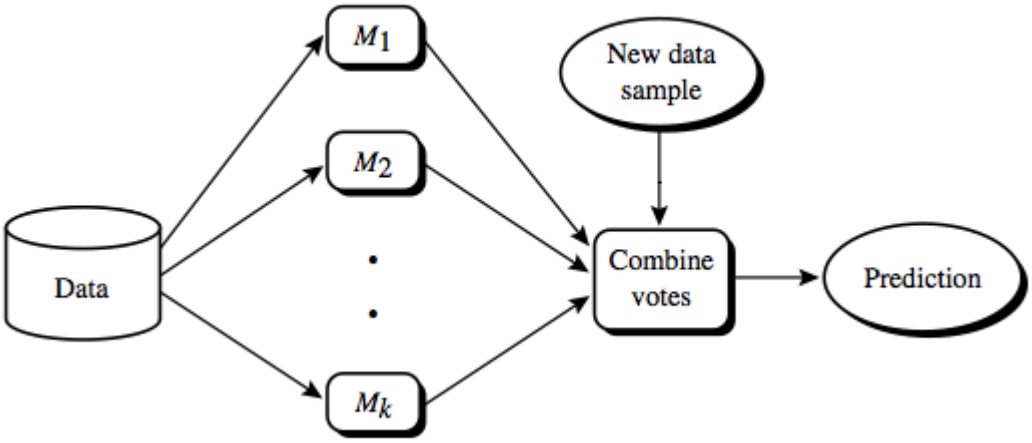
http://fluxicon.com/blog/wp-content/uploads/2012/01/DM-Example.png

Τυπικά ένας ταξινομητής είναι ένα πρότυπο ή μια λειτουργεία M που προβλέπει την ετικέτα κλάσης ˆ y για ένα δεδομένο παράδειγμα εισαγωγής x , δηλαδή, ˆ Y = M (X), όπου ˆ Y ϵ {c1, c2…, ck} και κάθε ci είναι μια ετικέτα κατηγορίας (μια κατηγορική αξία ιδιοτήτων). Για να χτίσουμε το πρότυπο απαιτούμε ένα σύνολο σημείων με τις σωστές ετικέτες κατηγορίας τους, το οποίο καλείτε σύνολο εκπαίδευσης. Αφού μάθουμε το πρότυπο M, μπορούμε αυτόματα να προβλέψουμε την κατηγορία για κάθε καινούργιο σημείο. Πολλοί διαφορετικοί τύποι προτύπων έχουν προταθεί όπως τα δέντρα απόφασης, πιθανολογικοί ταξινομητές, διανυσματικές μηχανές υποστήριξης, και ούτω καθεξής. (Mohammed J. Zaki, 2014) π.χ.



http://image.slidesharecdn.com/6class-150916153909-lva1-app6892/95/data-mining-6-638.jpg?cb=1442418055

Η πρόβλεψη καλύπτει ένα μεγάλο εύρος τεχνικών στατιστικής από το μοντέλο πρόβλεψης, την εκμάθηση μηχανών, και την εξόρυξη δεδομένων που εξετάζουν τα τωρινά και τα ιστορικά συμβάντα για να κάνουν προβλέψεις για μελλοντικά ή ειδάλλως άγνωστα δεδομένα. (Nyce, 2007) (Eckerson, 2007)



https://sites.google.com/a/kingofat.com/wiki/_/rsrc/1242170634319/data-mining/classification/Picture%2022.png

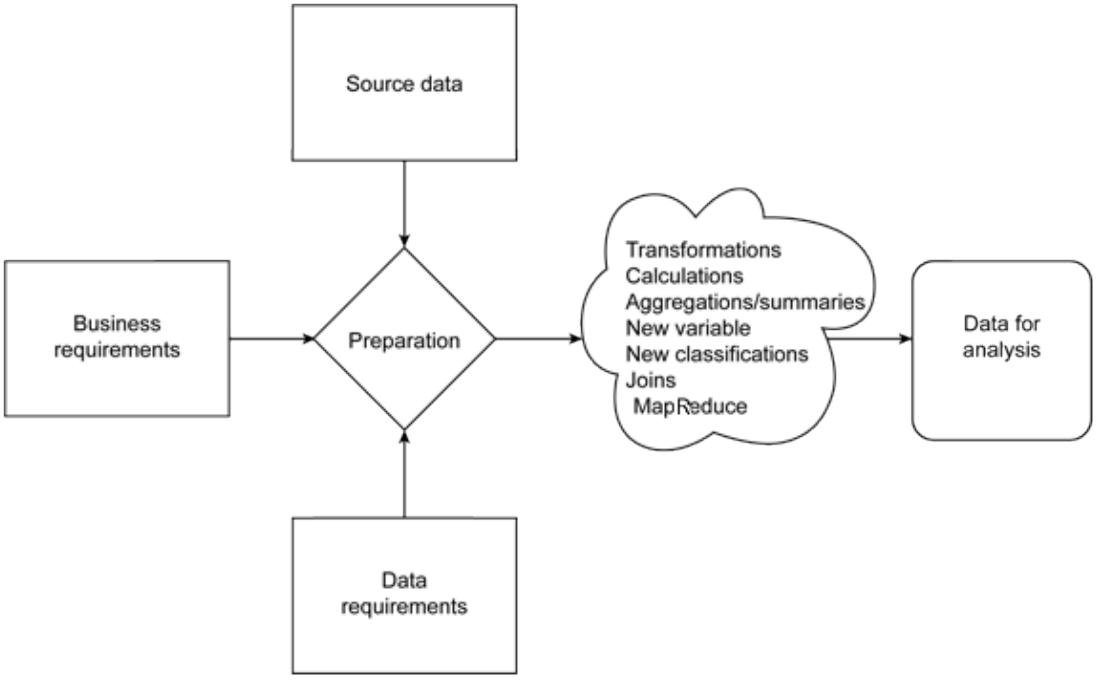
Η πρόβλεψη είναι ένας τομέας της εξόρυξης δεδομένων που εξετάζει την εξαγωγή πληροφοριών από τα δεδομένα και τα χρησιμοποιεί για να προβλέψει τις τάσεις και πρότυπα συμπεριφοράς. Συχνά το άγνωστο γεγονός ενδιαφέροντος είναι στο μέλλον, αλλά η πρόβλεψη μπορεί να εφαρμοστεί σε οποιοδήποτε είδος άγνωστου είτε είναι στο παρελθόν, στο παρών ή στο μέλλον. Παραδείγματος χάριν, ο προσδιορισμός των υπόπτων μετά από ένα έγκλημα έχει δεσμευτεί, ή η απάτη πιστωτικών καρτών όταν συμβαίνει. (Finlay, 2014)

Γενικά η πρόβλεψη εννοείτε ως τα πρότυπα πρόβλεψης, «σημειώνοντας» δεδομένα με τα μοντέλα αυτά, και το προσχεδιασμό. Παρόλα αυτά, οι άνθρωποι χρησιμοποιούν όλο και περισσότερο τον όρο για να αναφερθούν στις σχετικές πειθαρχίες ανάλυσης, όπως η περιγραφική διαμόρφωση και η διαμόρφωση απόφασης ή η βελτιστοποίηση απόφασης. Αυτές οι πειθαρχίες περιλαμβάνουν επίσης την αυστηρή ανάλυση δεδομένων, και χρησιμοποιούνται ευρέως στις επιχειρήσεις για την κατάτμηση και την διαδικασία απόφασης, αλλά έχουν διαφορετικούς σκοπούς και οι βαθύτερες τεχνικές στατιστικής μπορεί να ποικίλουν. (wikipedia, Wikipedia, n.d.)

2..2 Προετοιμασία των δεδομένων για Κατηγοριοποίηση και Πρόβλεψη

Η εξόρυξη δεδομένων στηρίζεται επάνω στην οικοδόμηση ενός κατάλληλου μοντέλου δεδομένων και της δομής του που μπορεί να εφαρμοστεί για την επεξεργασία, τον προσδιορισμό, και το χτίσιμο των πληροφοριών που χρειάζονται. Ανεξάρτητα από τη πηγή και τη δομή των δεδομένων, η οργάνωση και η δόμηση των πληροφοριών σε ένα σχήμα που επιτρέπει την εξόρυξη δεδομένων να πραγματοποιηθεί σε ένα όσο το δυνατόν αποδοτικότερο πρότυπο είναι απαραίτητο. Έχων υπόψη το συνδυασμό των επιχειρησιακών απαιτήσεων για την εξόρυξη δεδομένων, τον προσδιορισμό των υπαρχουσών μεταβλητών (πελάτης, τιμές, χώρα) και την απαίτηση να δημιουργηθούν οι νέες μεταβλητές που θα εφαρμοστούν για την επεξεργασία των στοιχείων στο βήμα προετοιμασιών. Μπορεί να γίνει σύνθεση των αναλυτικών μεταβλητών των δεδομένων από πολλές διαφορετικές πηγές σε μια ενιαία ευπροσδιόριστη μορφή (παραδείγματος χάριν, η δημιουργία μιας κατηγορίας ενός ιδιαίτερου βαθμού και μιας ηλικίας του πελάτη, ή έναν ιδιαίτερο τύπο λάθους).

Ανάλογα με την πηγή δεδομένων, το πως θα χτιστούν και θα μεταφραστούν αυτές οι πληροφορίες είναι ένα σημαντικό βήμα, ανεξάρτητα από την τεχνική που θα χρησιμοποιηθεί για την τελική ανάλυση των δεδομένων. Αυτό το βήμα οδηγεί επίσης σε μια πιο σύνθετη διαδικασία αναγνώρισης, συσσωμάτωσης, απλούστευσης, ή επέκτασης των πληροφοριών για να ταιριάζει τα δεδομένα εισόδου. (Brown, 2012)



https://www.ibm.com/developerworks/library/ba-data-mining-techniques/fig05.gif

Σε πολλά προβλήματα με έναν δυαδικό στόχο, μια αξία στόχων εξουσιάζει στη συχνότητα. Παραδείγματος χάριν, οι θετικές απαντήσεις για μια τηλεφωνική εκστρατεία μάρκετινγκ μπορούν να είναι 2% ή λιγότεροι, και το περιστατικό της απάτης στις συναλλαγές πιστωτικών καρτών μπορεί να είναι λιγότερο από 1%. Ένα πρότυπο ταξινόμησης που στηρίζεται στα ιστορικά στοιχεία αυτού του τύπου μπορεί να μην παρατηρήσει αρκετές θετικές περιπτώσεις για να είναι σε θέση να

διακρίνει τα χαρακτηριστικά των δύο κατηγοριών το αποτέλεσμα θα μπορούσε να είναι ένα πρότυπο που όταν εφαρμόζεται στα νέα στοιχεία προβλέπει την αρνητική κατηγορία για κάθε περίπτωση. Ενώ ένα τέτοιο πρότυπο μπορεί να είναι υπερβολικά ακριβές, μπορεί να μην είναι πολύ χρήσιμο. Αυτό διευκρινίζει ότι δεν είναι μια καλή ιδέα να στηριχθεί απλώς στην ακρίβεια κατά την κρίση ενός προτύπου. Μια λύση σε αυτό το πρόβλημα περιλαμβάνει τη δημιουργία ενός πίνακα πηγής για τη λειτουργία κατασκευής που περιέχει τους περίπου ίσους αριθμούς κάθε αξίας στόχων. Εντούτοις, ο αλγόριθμος θα πάρει την παρατηρηθείσα διανομή ως ρεαλιστική, και θα χτίσει ένα πρότυπο που θα προβλέψει κάθε μια από τις τιμές στόχων στους ίσους αριθμούς εκτός αν καθοδηγείται ειδάλλως. Παρέχοντας την πραγματική διανομή των τιμών στόχων, τα προηγούμενα, στη διαδικασία κατασκευής μπορούν να οδηγήσουν σε ένα αποτελεσματικότερο πρότυπο. Σημειώστε ότι το πρότυπο πρέπει να εξεταστεί ενάντια στο στοιχείο που έχει την πραγματική διανομή των τιμών στόχων. Παραδείγματος χάριν, 98% αρνητικοί και θετικό 2% για τη εκστρατεία μάρκετινγκ. (Margaret Taft, 2005)

Όσο περισσότερη πειθαρχεία υπάρχει στο χειρισμό των δεδομένων, τόσο συνεπέστερα και καλύτερα αποτελέσματα είναι πιθανόν να επιτύχουμε. Η διαδικασία της προετοιμασίας των δεδομένων για έναν αλγόριθμο εκμάθησης μηχανής μπορεί να συνοψιστεί σε τρία βήματα:

- Βήμα 1: Επιλογή των δεδομένων
- Βήμα 2: Προεπεξεργασία των δεδομένων
- Βήμα 3: Μεταμόρφωση των δεδομένων

Μπορείτε να ακολουθήσετε αυτήν τη διαδικασία κατά γραμμικό τρόπο, αλλά είναι πολύ πιθανό να είναι επαναληπτική με πολλούς βρόχους.

Βήμα 1: Επιλογή των δεδομένων

Αυτό το βήμα έχει σχέση με την επιλογή του υποσυνόλου όλων των διαθέσιμων δεδομένων με τα οποία θα εργάζεστε. Υπάρχει πάντα μια ισχυρή επιθυμία για τη συμπερίληψη όλων των δεδομένων που είναι διαθέσιμα, που το αξίωμα «το περισσότερο είναι καλύτερο» θα κρατήσει. Αυτό μπορεί να ισχύει μπορεί και όχι.

Πρέπει να εξεταστεί ποια δεδομένα χρειάζονται πραγματικά για να εξεταστεί η ερώτηση ή το πρόβλημα που εργάζεται κάποιος. Γίνονται μερικές υποθέσεις για τα δεδομένα που απαιτούνται και απαιτείτε προσοχή για να καταγραφούν εκείνες οι υποθέσεις ώστε να είναι δυνατή η δοκιμή τους αργότερα εάν είναι απαραίτητο.

Παρακάτω είναι μερικές ερωτήσεις για να σας βοηθήσουν να σκεφτείτε μέσω αυτής της διαδικασίας.

- Ποια είναι η έκταση των δεδομένων που είναι διαθέσιμα; Παραδείγματος χάριν μέσω του χρόνου, πίνακες βάσεων δεδομένων, συνδεδεμένα συστήματα. Εξασφαλίστε ότι έχετε μια σαφή εικόνα όλων όσων μπορείτε να χρησιμοποιήσετε.
- Ποια δεδομένα δεν είναι διαθέσιμα που επιθυμείτε να είχατε; Παραδείγματος χάριν δεδομένα που δεν καταγράφονται ή δεν μπορούν να καταγραφούν. Μπορεί να είστε σε θέση να παράγετε ή να μιμηθείτε αυτά τα δεδομένα.
- Ποια δεδομένα δεν χρειάζονται για την εξέταση του προβλήματος; Ο αποκλεισμός των δεδομένων είναι πάντα πιο εύκολος από την συμπερίληψη τους. Σημειώστε ποια δεδομένα αποκλείσατε και γιατί.

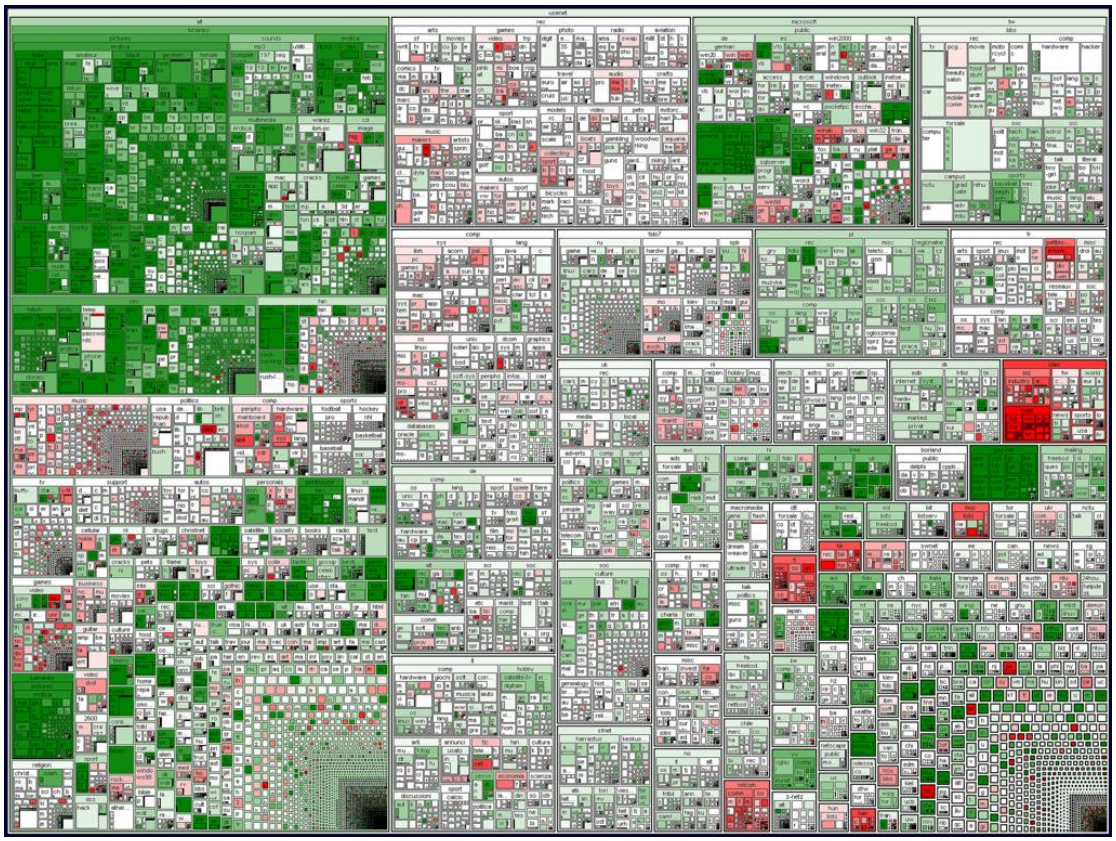
Είναι μόνο σε μικρά προβλήματα, όπως ο ανταγωνισμός ή οι βάσεις δεδομένων παιχνιδιών όπου τα δεδομένα έχουν ήδη επιλεχτεί για σας.

Βήμα 2: Προεπεξεργασία των δεδομένων

Ύστερα από την επιλογή των δεδομένων, πρέπει να ληφθεί υπόψη πως πρόκειται να χρησιμοποιηθούν τα δεδομένα. Αυτό το βήμα προεπεξεργασίας είναι για να έρθουν τα δεδομένα σε μια μορφή όπου είναι δυνατόν να δουλέψεις με αυτά. Τρία κοινά βήματα προεπεξεργασίας δεδομένων είναι η μορφοποίηση, ο καθαρισμός και η δειγματοληψία.

- **Μορφοποίηση:** Τα δεδομένα που έχουν επιλεχθεί μπορεί να μην είναι σε μια μορφή που τα καθιστά κατάλληλα για εργασία. Τα δεδομένα μπορεί να είναι σε μια σχεσιακή βάση δεδομένων ενώ θα ήταν προτιμότερο να ήταν σε επίπεδο αρχείο, ή τα δεδομένα μπορεί να είναι σε ιδιόκτητη μορφή ενώ θα ήταν προτιμότερο να είναι σε μια σχεσιακή βάση δεδομένων ή σε αρχείο κειμένου.
- **Καθαρισμός:** Ο καθαρισμός είναι η αφαίρεση ή η διόρθωση δεδομένων που λείπουν. Μπορεί να υπάρχουν περιπτώσεις δεδομένων που είναι ελλιπείς και δεν περιέχουν τα δεδομένα που θεωρείτε ότι χρειάζονται για την επίλυση του προβλήματος. Αυτές οι περιπτώσεις θα πρέπει να αφαιρεθούν. Επιπλέον, μπορεί να υπάρχουν ευαίσθητες πληροφορίες σε μερικές από τις ιδιότητες και αυτές οι ιδιότητες μπορεί να χρειαστεί να γίνουν ανώνυμες ή να αφαιρεθούν από τα δεδομένα εξολοκλήρου.
- **Δειγματοληψία:** Μπορεί να υπάρχουν πολύ περισσότερα επιλεγμένα δεδομένα διαθέσιμα από αυτά που χρειάζονται για να εργαστείτε. Περισσότερα δεδομένα μπορεί να έχουν ως αποτέλεσμα πολύ μεγαλύτερους χρόνους εκτέλεσης των αλγορίθμων και μεγαλύτερες υπολογιστικές απαιτήσεις και απαιτήσεις μνήμης. Μπορείτε να πάρετε ένα μικρότερο αντιπροσωπευτικό δείγμα των επιλεγμένων δεδομένων που μπορεί να είναι πολύ γρηγορότερα για την εξερεύνηση και την προτυποποίηση λύσεων πριν παρθεί υπόψη όλο το σύνολο δεδομένων.

Είναι πολύ πιθανό τα εργαλεία εκμάθησης μηχανών που θα χρησιμοποιηθούν στα δεδομένα να επηρεάσουν την προεπεξεργασία που θα απαιτηθεί να εκτελεστεί. Θα ξαναεπισκεφτείτε πιθανώς αυτό το βήμα.



http://3qeqpr26caki16dnhd19sv6by6v.wpengine.netdna-cdn.com/wp-content/uploads/2013/12/So-much-data.jpg

Βήμα 3: Μεταμόρφωση των δεδομένων

Το τελικό βήμα είναι η μεταμόρφωση των δεδομένων διεργασίας. Ο συγκεκριμένος αλγόριθμος που χρησιμοποιείτε και η γνώση του τομέα προβλήματος θα επηρεάσει αυτό το βήμα και είναι πολύ πιθανό να χρειαστεί να ξαναεπισκεφτείτε διαφορετικούς μετασχηματισμούς των προεπεξεργασμένων δεδομένων σας όσο εργάζεστε στο πρόβλημά σας.

Τρεις κοινοί μετασχηματισμοί δεδομένων είναι η ιεράρχηση, η αποσύνθεση ιδιοτήτων και η συνάθροιση ιδιοτήτων. Αυτό το βήμα αναφέρεται επίσης ως εφαρμοσμένη μηχανική χαρακτηριστικών γνωρισμάτων.

- **Ιεράρχηση:** Τα προεπεξεργασμένα στοιχεία μπορούν να περιέχουν τις ιδιότητες με μίγματα κλιμάκων για τις διάφορες ποσότητες όπως ο όγκος δολαρίων, χιλιογράμμων και πωλήσεων. Πολλοί επεξεργάζονται τις μεθόδους εκμάθησης όπως τις ιδιότητες στοιχείων στη μηχανή για να έχουν την ίδια κλίμακα όπως μεταξύ 0 και 1 για τη μικρότερη και μεγαλύτερη αξία για ένα δεδομένο χαρακτηριστικό γνώρισμα. Εξετάστε ότι οποιοδήποτε χαρακτηριστικό γνώρισμα κλιμάκωσης εσείς μπορεί να πρέπει να εκτελέσετε.
- **Αποσύνθεση:** Μπορούν να υπάρξουν χαρακτηριστικά γνωρίσματα που αντιπροσωπεύουν μια σύνθετη έννοια που μπορεί να είναι πιο χρήσιμη σε μια μέθοδο εκμάθησης μηχανών όταν διασπάσετε στα ιδρυτικά μέρη. Ένα παράδειγμα είναι μια ημερομηνία που μπορεί να έχει τα τμήματα ημέρας και χρόνου που θα μπορούσαν με τη σειρά να χωριστούν περαιτέρω. Ίσως μόνο η ώρα της ημέρας είναι σχετική με τη λύση του προβλήματος. Εξετάστε ποιες αποσυνθέσεις χαρακτηριστικών γνωρισμάτων μπορείτε να εκτελέσετε.
- **Συσσωμάτωση:** Μπορούν να υπάρξουν χαρακτηριστικά γνωρίσματα που μπορούν να αθροιστούν σε ένα ενιαίο χαρακτηριστικό γνώρισμα που θα ήταν σημαντικότερο στο πρόβλημα που προσπαθείτε να λύσετε. Παραδείγματος χάριν, μπορούν να υπάρξουν περιπτώσεις δεδομένων για κάθε φορά που εισήλθε(login) ένας πελάτης σε ένα σύστημα που

θα μπορούσε να αθροιστεί σε μια αρίθμηση για τον αριθμό εισόδων(logins) που επιτρέπουν στις πρόσθετες περιπτώσεις να απορριφθούν. Εξετάστε ποιος τύπος συναθροίσεων χαρακτηριστικών γνωρισμάτων μπόρεσε να εκτελεστεί.

Μπορείτε να ξοδέψετε πολύ ώρα σχεδιάζοντας γνωρίσματα από τα δεδομένα σας και μπορεί να είναι πολύ ευεργετικό στην απόδοση ενός αλγορίθμου. Ξεκινήστε σιγά και χτίστε στις δεξιότητες που μαθαίνετε. (Brownlee, 2013)

2..3 Τεχνικές Κατηγοριοποίησης

2..3.1 Bayesian Classification

Ο ταξινομητής Bayes είναι μια από τις απλούστερες προσεγγίσεις στη διαδικασία της ταξινόμησης που είναι ακόμα ικανός να παράγει λογική ακρίβεια. Παρόλο που σε πολλές περιπτώσεις δεν μπορεί να ανταγωνιστεί με πιο καθαρούς αλγόριθμους, όπως τα δέντρα απόφασης, μερικές φορές δε μένει μακριά πίσω, και μπορεί να είναι ακόμη και ανώτερος για ορισμένες συγκεκριμένες περιοχές εφαρμογής, με την ταξινόμηση κειμένων να είναι το πιο προεξέχων παράδειγμα. Η απλότητά του - εννοιολογικά, κατά την εφαρμογή, και υπολογιστικά – το καθιστά εύκολο και ανέξοδο να δοκιμαστεί εκτός από ή πριν από τους περιπλοκότερους ταξινομητές. Το συμπέρασμα Bayes, του οποίου ο ταξινομητής Bayes είναι ένα ιδιαίτερα απλό παράδειγμα, είναι βασισμένο στον κανόνα Bayes που αφορά τις υπό όρους και οριακές πιθανότητες. Ακριβέστερα, επιδεικνύει πως η υπό όρους (μεταγενέστερη) πιθανότητα ενός γεγονότος μπορεί να υπολογιστεί με βάση την οριακή (προγενέστερη) πιθανότητα και η αντίστροφη υπό όρους πιθανότητα. Για δύο γεγονότα A και B, ο κανόνας μπορεί να γραφτεί ως

$$P(A|B) = P(A) * P(B|A) / P(B)$$

Όπου

- P(A) είναι η προγενέστερη πιθανότητα του A
- P(A|B) είναι η υπό όρους πιθανότητα του A δεδομένου του B, αποκαλούμενη επίσης ως μεταγενέστερη πιθανότητα του A,
- P(B|A) είναι η υπό όρους πιθανότητα του B δεδομένου του A, και
- P(B) είναι η πιθανότητα του A

Στην πιο κοινή ρύθμιση, ο κανόνας εφαρμόζεται στο συμπέρασμα σχετικά με ένα σύνολο αμοιβαία αποκλειστικών γεγονότων A1, A2, …, Ak που εξαντλούν το διάστημα πιθανότητας, δηλαδή,

$$P(Ai \cap Aj) = 0 \text{ for } i \neq j$$

$$\sum_{i=1}^{k} P(Ai) = 1$$

Κατόπιν από το νόμο της συνολικής πιθανότητας

$$P(B) = \sum_{j=1}^{k} P(Aj)P(B|Aj)$$

Που επιτρέπει σε έναν να αναθεωρήσει τον κανόνα Bayes όπως

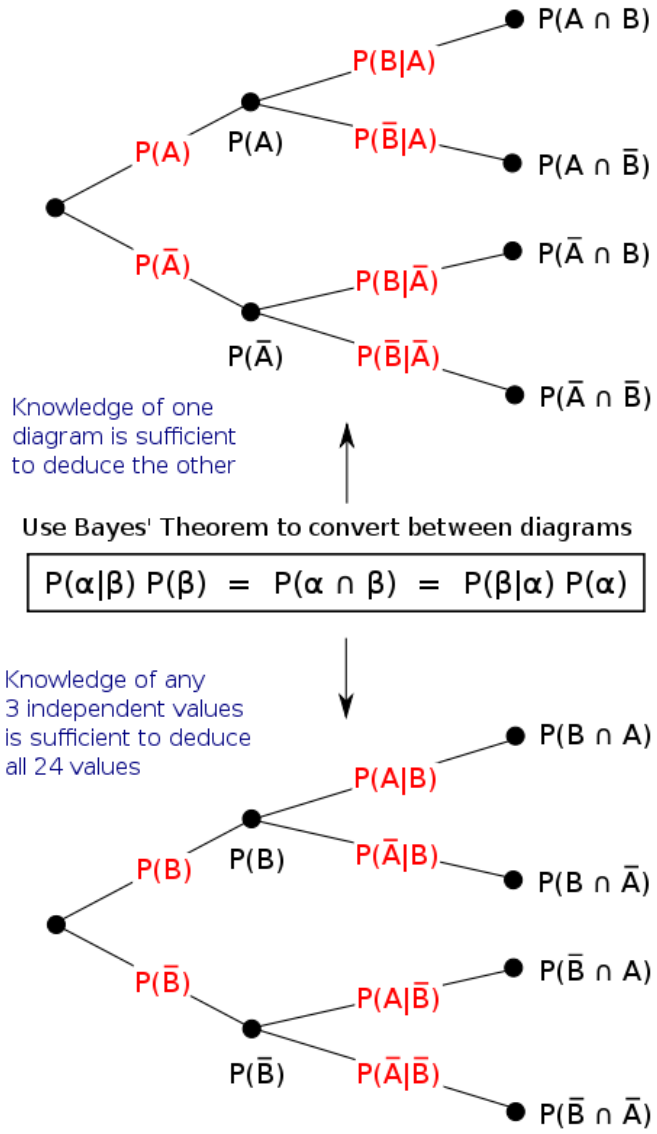
$$P(Ai|B) = P(Ai)P(B|Ai)/ \sum_{j=1}^{k} P(Aj)P(B|Aj)$$

Αυτή η μορφή δείχνει ότι το P(B) πράττει ως σταθερά ομαλοποίησης, εξασφαλίζοντας ότι

$$\sum_{i=1}^{k} P(Ai|B) = 1$$

Σημειώστε ότι, αντίθετα από τις μεταγενέστερες πιθανότητες P(Ai|B), οι αντίστροφες υπό όρους πιθανότητες P(B|Ai) δεν πρέπει και συνήθως δεν συνοψίζουν σε 1.

Χαρακτηριστικά, το A1, A2, …, Ak αντιπροσωπεύει ένα σύνολο εναλλακτικών υποθέσεων, και το B αντιπροσωπεύει κάποια διαθέσιμα στοιχεία που μπορούν να έχουν επιπτώσεις στην πιθανότητα. Χωρίς να παρθούν τα στοιχεία υπόψη, οι υποθέσεις έχουν τις προγενέστερες πιθανότητές τους ορισμένες. Ο κανόνας Bayes επιδεικνύει πως να ενσωματώσει τα στοιχεία και να λάβει τις μεταγενέστερες πιθανότητες υπόθεσης. Κάθε μια από τις αντίστροφες υπό όρους πιθανότητες P(B|Ai) μπορεί να θεωρηθεί μέτρο του βαθμού στον οποίο τα στοιχεία υποστηρίζουν (ή αντικρούουν) την αντίστοιχη υπόθεση Ai.



https://i.stack.imgur.com/J5AuA.png

Υπάρχουν δύο σημαντικές προσεγγίσεις στην εφαρμογή του συμπεράσματος Bayes στο στόχο ταξινόμησης:

Συμπέρασμα πρότυπο-πιθανότητας. Με βάση τον υπολογισμό των μεταγενέστερων πρότυπων πιθανοτήτων δεδομένου ενός συνόλου δεδομένων.

Συμπέρασμα κατηγορία-πιθανότητας. Με βάση τον υπολογισμό των μεταγενέστερων πιθανοτήτων κατηγορίας δεδομένων των τιμών ιδιοτήτων.

Η πρώτη προσέγγιση φαίνεται ελκυστική δεδομένου ότι μπορεί να επιτρέψει τον προσδιορισμό του πιθανότερου προτύπου. Είναι πρακτικό μόνο για ένα περιορισμένο σύνολο υποψηφίων προτύπων, ωστόσο, τα οποία πρέπει να επιλεχτούν εκ των προτέρων είτε χρησιμοποιώντας γνώση υποβάθρου είτε μερικούς άλλους αλγόριθμους. Επιπλέον, η ανάθεση των προγενέστερων πιθανοτήτων σε τέτοια υποψήφια πρότυπα είναι τετριμμένη.

Ο ταξινομητής Bayes ακολουθεί τη δεύτερη προσέγγιση, η οποία δεν υπόσχεται τόσα πολλά, αλλά και δεν υπονοεί τόσες πολλές δυσκολίες. Χωρίς να εξετάσει ρητά οποιαδήποτε υποψήφια πρότυπα και τις

πιθανότητές τους, στην πραγματικότητα δημιουργεί ένα πιθανολογικό πρότυπο που υπολογίζει τις πιθανότητες κατηγορίας για μια περίπτωση βασισμένη στις τιμές ιδιοτήτων του. (Cichosz, 2015)

2..3.2 Decision Trees (ID3, C45) & Regression Trees για πρόβλεψη (CART)

Η εκμάθηση δέντρων αποφάσεων ή η επαγωγή δέντρων αποφάσεων είναι μία από τις προγνωστικές προσεγγίσεις μοντελοποίησης που χρησιμοποιούνται στη στατιστική, την εξόρυξη δεδομένων και τη μηχανική μάθηση. Χρησιμοποιεί ένα δέντρο αποφάσεων (ως προγνωστικό μοντέλο) για να περάσει από παρατηρήσεις σχετικά με ένα στοιχείο (που αντιπροσωπεύεται στους κλάδους) σε συμπεράσματα σχετικά με την τιμή-στόχο του στοιχείου (που αντιπροσωπεύεται στα φύλλα). Τα μοντέλα δέντρων όπου η μεταβλητή-στόχος μπορεί να λάβει ένα διακριτό σύνολο τιμών ονομάζονται δέντρα ταξινόμησης. Σε αυτές τις δομές δέντρων, τα φύλλα αντιπροσωπεύουν ετικέτες κλάσης και οι κλάδοι αντιπροσωπεύουν συνδέσμους χαρακτηριστικών που οδηγούν σε αυτές τις ετικέτες κλάσης. Τα δέντρα απόφασης όπου η μεταβλητή-στόχος μπορεί να λαμβάνει συνεχείς τιμές (συνήθως πραγματικοί αριθμοί) ονομάζονται δέντρα παλινδρόμησης. Τα δέντρα αποφάσεων είναι από τους πιο δημοφιλείς αλγόριθμους μηχανικής μάθησης δεδομένης της κατανόησης και της απλότητάς τους. Στην ανάλυση αποφάσεων, ένα δέντρο αποφάσεων μπορεί να χρησιμοποιηθεί για την οπτική και ρητή αναπαράσταση αποφάσεων και λήψης αποφάσεων. Στην εξόρυξη δεδομένων, ένα δέντρο αποφάσεων περιγράφει δεδομένα (αλλά το δέντρο ταξινόμησης που προκύπτει μπορεί να είναι μια εισροή για τη λήψη αποφάσεων). (wikipedia, https://en.wikipedia.org, 2019)

Ο αλγόριθμος ID3 (Iterative Dichotomiser 3) δημιουργεί ένα δέντρο απόφασης από ένα σταθερό σύνολο παραδειγμάτων. Το προκύπτον δέντρο χρησιμοποιείται για την ταξινόμηση μελλοντικών δειγμάτων. Το παράδειγμα έχει πολλά χαρακτηριστικά και ανήκει σε μια κλάση (όπως ναι ή όχι). Οι κόμβοι των φύλλων του δέντρου αποφάσεων περιέχουν το όνομα της κλάσης, ενώ ένα μη-φύλλο κόμβου είναι κόμβος απόφασης. Ο κόμβος απόφασης είναι ένα τεστ χαρακτηριστικών με κάθε κλάδο (σε άλλο δέντρο απόφασης) να είναι μια πιθανή τιμή του χαρακτηριστικού. Το ID3 χρησιμοποιεί την απόκτηση πληροφοριών για να βοηθηθεί στο να αποφασίσει ποιο χαρακτηριστικό πηγαίνει σε έναν κόμβο απόφασης. Το πλεονέκτημα της εκμάθησης ενός δέντρου αποφάσεων είναι ότι ένα πρόγραμμα, παρά τη σχεδίαση γνώσης, εξάγει γνώση από έναν ειδικό.
(https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm, 2019)

Ο ID3 (Iterative Dichotomiser 3) αναπτύχθηκε το 1986 από τον Ross Quinlan.
(https://en.wikipedia.org/wiki/ID3_algorithm, 2019)

Ο πραγματικός αλγόριθμος είναι ο ακόλουθος:

ID3 (Παραδείγματα, Ιδιότητα_Στόχος, Ιδιότητες)

- Δημιούργησε ένα αρχικό κόμβο για το δέντρο
- Εάν όλα τα παραδείγματα είναι θετικά, Επέστρεψε την Ρίζα ως δέντρο με ένα κόμβο, με ετικέτα= +.
- Εάν όλα τα παραδείγματα είναι αρνητικά, Επέστρεψε την Ρίζα ως δέντρο με ένα κόμβο, με ετικέτα= -.
- Εάν ο αριθμός των προβλεπόμενων ιδιοτήτων είναι κενός, τότε Επέστρεψε την Ρίζα ως δέντρο με ένα κόμβο, με ετικέτα = την πιο κοινή τιμή της ιδιότητας-στόχου των παραδειγμάτων.
- Αλλιώς Ξεκίνα
  - Α = Η ιδιότητα που κατηγοριοποιεί καλύτερα τα παραδείγματα.
  - Ιδιότητα Δέντρου απόφασης για τη ρίζα = Α.
  - Για κάθε πιθανή τιμή, N_i , του Α,
  - Πρόσθεσε ένα νέο κλάδο κάτω από τη Ρίζα, που να αντιστοιχεί στη δοκιμή Α = N_i .
  - Θέσε Παραδείγματα(N_i), ως το υποσύνολο των παραδειγμάτων που έχουν την τιμή N_i για το Α

- Αν Παραδείγματα(N_i) είναι κενό
    - Τότε κάτω από αυτό τον νέο κλάδο πρόσθεσε έναν κόμβο-φύλλο με ετικέτα = την πιο κοινή τιμή στόχο στα παραδείγματα
- Αλλιώς κάτω από αυτό τον νέο κλάδο πρόσθεσε το υποδέντρο ID3 (Παραδείγματα(N_i), Ιδιότητα_Στόχος, Ιδιότητες – {A})
- Τέλος
- Επέστρεψε Ρίζα

(https://el.wikipedia.org/wiki/%CE%91%CE%BB%CE%B3%CF%8C%CF%81%CE%B9%CE%B8%CE%BC%CE%BF%CF%82_ID3, 2019)

Ο ID3 είναι ένας μη αυξητικός αλγόριθμος, που σημαίνει ότι παράγει τις κλάσεις του από ένα σταθερό σύνολο εκπαιδευτικών περιπτώσεων. Ένας αυξητικός αλγόριθμος αναθεωρεί τον ορισμό της τρέχουσας έννοιας, αν είναι απαραίτητο, με ένα νέο δείγμα. Οι κλάσεις που δημιουργούνται από το ID3 είναι επαγωγικές, δηλαδή, δεδομένης μιας μικρής σειράς εκπαιδευτικών παρουσιών, οι συγκεκριμένες κλάσεις που δημιουργούνται από το ID3 αναμένεται να λειτουργήσουν για όλες τις μελλοντικές περιπτώσεις. Η κατανομή των άγνωστων πρέπει να είναι η ίδια με τις περιπτώσεις δοκιμών. Οι κλάσεις επαγωγής δεν μπορούν να αποδειχτούν ότι λειτουργούν σε κάθε περίπτωση, αφού μπορούν να ταξινομήσουν έναν άπειρο αριθμό περιπτώσεων. Σημειώστε ότι ο ID3 (ή οποιοσδήποτε επαγωγικός αλγόριθμος) υπάρχει περίπτωση να ταξινομήσει δεδομένα εσφαλμένα.

Τα δεδομένα δείγματος που χρησιμοποιούνται από τον ID3 έχουν ορισμένες απαιτήσεις, οι οποίες είναι:

- Περιγραφή χαρακτηριστικού-τιμής - τα ίδια χαρακτηριστικά πρέπει να περιγράφουν κάθε παράδειγμα και να έχουν έναν σταθερό αριθμό τιμών.
- Προκαθορισμένες τάξεις - τα χαρακτηριστικά ενός παραδείγματος πρέπει να έχουν ήδη καθοριστεί, δηλαδή να μην τα μάθει ο ID3.
- Διακριτές τάξεις – οι τάξεις πρέπει να οριοθετηθούν. Οι συνεχείς τάξεις που χωρίζονται σε ασαφείς κατηγορίες, όπως ένα μέταλλο που είναι "σκληρό, αρκετά σκληρό, εύκαμπτο, μαλακό, αρκετά μαλακό" είναι ύποπτες.
- Επαρκή παραδείγματα - δεδομένου ότι χρησιμοποιείται επαγωγική γενίκευση (δηλ. Μη αποδεδειγμένη) πρέπει να υπάρχουν αρκετές περιπτώσεις δοκιμής για να διακριθούν έγκυρα μοτίβα από τυχαίες περιπτώσεις.

Πώς αποφασίζει ο ID3 ποιο χαρακτηριστικό είναι το καλύτερο; Χρησιμοποιείται μια στατιστική ιδιότητα, που ονομάζεται κέρδος πληροφοριών. Το κέρδος μετράει πόσο καλά μια δεδομένη ιδιότητα χωρίζει τα παραδείγματα εκπαίδευσης σε στοχευμένες κατηγορίες. Εκείνο με τις υψηλότερες πληροφορίες (οι πληροφορίες που είναι οι πιο χρήσιμες για ταξινόμηση) επιλέγονται. Για να ορίσουμε το κέρδος, δανειζόμαστε πρώτα μια ιδέα από την θεωρία των πληροφοριών που ονομάζεται εντροπία. Η εντροπία μετράει το ποσό των πληροφοριών σε ένα χαρακτηριστικό.
(https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm, 2019)

Η **εντροπία** H(S) είναι η μέτρηση του ποσού αβεβαιότητας στο σετ δεδομένων S.

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Όπου:

- S - Το τρέχον σύνολο δεδομένων για το οποίο υπολογίζεται η εντροπία.
    - Αυτό αλλάζει σε κάθε βήμα του αλγορίθμου ID3, είτε σε ένα υποσύνολο του προηγούμενου σετ σε περίπτωση διαίρεσης ενός χαρακτηριστικού είτε σε ένα "αδελφικό" διαμέρισμα του γονέα σε περίπτωση που η επανάληψη τερματίστηκε προηγουμένως.
- X - Το σύνολο των τάξεων μέσα στο S.
- p(x) - Το ποσοστό του αριθμού των στοιχείων στην κλάση x στον αριθμό των στοιχείων στο σύνολο S.

Όταν $H(S) = 0$, το σύνολο S είναι τέλεια ταξινομημένο (δηλαδή όλα τα στοιχεία στο S είναι της ίδιας κλάσης).

Στον ID3, η εντροπία υπολογίζεται για κάθε χαρακτηριστικό που απομένει. Το χαρακτηριστικό με την μικρότερη εντροπία χρησιμοποιείται για να διαιρέσει το σύνολο S σε αυτή την επανάληψη. Η εντροπία στη θεωρία των πληροφοριών μετρά πόση πληροφορία αναμένεται να αποκτηθεί με τη μέτρηση μιας τυχαίας μεταβλητής, έτσι, μπορεί επίσης να χρησιμοποιηθεί για να ποσοτικοποιήσει την ποσότητα στην οποία οι τιμές της ποσότητας δεν είναι γνωστές. Μια σταθερή ποσότητα έχει μηδενική εντροπία, καθώς η κατανομή της είναι απολύτως γνωστή. Αντίθετα, μια ομοιόμορφα κατανεμημένη τυχαία μεταβλητή (διακριτικά ή συνεχώς ομοιόμορφη) μεγιστοποιεί την εντροπία. Επομένως, όσο μεγαλύτερη είναι η εντροπία σε έναν κόμβο, τόσο λιγότερες πληροφορίες είναι γνωστές για την ταξινόμηση των δεδομένων σε αυτό το στάδιο του δέντρου, και ως εκ τούτου, τόσο μεγαλύτερο είναι το δυναμικό βελτίωσης της ταξινόμησης εδώ. Ως εκ τούτου, ο ID3 είναι ένας άπληστος ευρετικός που διεξάγει μια καλύτερη πρώτη αναζήτηση για τοπικές τιμές βέλτιστης εντροπίας. Η ακρίβειά του μπορεί να βελτιωθεί με την προεπεξεργασία των δεδομένων.

Το κέρδος πληροφορίας IG (A) είναι το μέτρο της διαφοράς στην εντροπία από πριν έως μετά το σύνολο S χωριστεί σε ένα χαρακτηριστικό A. Με άλλα λόγια, πόση αβεβαιότητα στην S μειώθηκε μετά το διαχωρισμό του σετ S στο χαρακτηριστικό A.

$$IG(S, A) = H(S) - \sum_{t \in T} p(t)H(t) = H(S) - H(S|A).$$

Όπου:

- H(S) – Εντροπία του σετ S.
- T - Τα υποσύνολα που δημιουργήθηκαν διαχωρίζοντας από το σύνολο S το χαρακτηριστικό A έτσι ώστε $\bigcup_{t \in T} t$
- p(t) - Το ποσοστό του αριθμού των στοιχείων του t στον αριθμό των στοιχείων στο σετ S
- H(t) – Εντροπία του υποσυνόλου t

Στον ID3, το κέρδος πληροφορίας μπορεί να υπολογιστεί (αντί για την εντροπία) για κάθε χαρακτηριστικό που απομένει. Το χαρακτηριστικό με το μεγαλύτερο κέρδος πληροφοριών χρησιμοποιείται για να διαιρέσει το σύνολο S σε αυτή την επανάληψη. (https://en.wikipedia.org/wiki/ID3_algorithm, 2019)

Μπορεί ο ID3(Iterative Dichotomiser 3) να είναι ο πιο συνηθισμένος συμβατικός αλγόριθμος δέντρων αποφάσεων αλλά έχει κάποια κωλύματα. Τα χαρακτηριστικά πρέπει να είναι ονομαστικές τιμές, το σύνολο δεδομένων δεν πρέπει να περιλαμβάνει δεδομένα που λείπουν, και τέλος ο αλγόριθμος τείνει να πέσει σε υπερφόρτωση. Εδώ, ο Ross Quinlan, εφευρέτης του ID3, έκανε κάποιες βελτιώσεις για αυτά τα σημεία συμφόρησης και δημιούργησε ένα νέο αλγόριθμο με το όνομα C4.5. Τώρα, ο αλγόριθμος μπορεί να δημιουργήσει ένα πιο γενικευμένο μοντέλο που περιλαμβάνει συνεχή δεδομένα και θα μπορούσε να χειριστεί ελλείποντα δεδομένα. Επιπλέον, ορισμένοι πόροι, όπως ο Weka, ονόμασαν αυτόν τον αλγόριθμο ως J48. Στην πραγματικότητα, αναφέρεται στην εκ νέου εφαρμογή της έκδοσης 8 του C4.5. (Serengil, 2018) Ποιο συγκεκριμένα ο C4.5 αντιμετωπίζει τα ακόλουθα προβλήματα του ID3:

- Αποφυγή υπερφόρτωσης των δεδομένων
    - Προσδιορισμός του πόσο βαθιά θα αναπτυχθεί ένα δέντρο αποφάσεων.
- Μειωμένο κλάδεμα σφαλμάτων.
- Κανόνας μετά το κλάδεμα.
- Διαχείριση συνεχών χαρακτηριστικών.
    - π.χ., θερμοκρασίας.
- Επιλογή κατάλληλου μέτρου επιλογής χαρακτηριστικών.
- Διαχείριση δεδομένων εκπαίδευσης με ελλείπουσες τιμές χαρακτηριστικών.

- Χειρισμός χαρακτηριστικών με διαφορετικό κόστος.
- Βελτίωση της υπολογιστικής αποδοτικότητας.
  (http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html, 2019)

2..3.3 Neural Networks

Ένα νευρωνικό δίκτυο είναι μια σειρά αλγορίθμων που προσπαθούν να αναγνωρίσουν τις υποκείμενες σχέσεις σε ένα σύνολο δεδομένων μέσω μιας διαδικασίας που μιμείται τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Με αυτή την έννοια, τα νευρωνικά δίκτυα αναφέρονται σε συστήματα νευρώνων, είτε οργανικής είτε τεχνητής φύσης. Τα νευρωνικά δίκτυα μπορούν να προσαρμοστούν στις μεταβαλλόμενες εισροές. έτσι ώστε το δίκτυο να παράγει το καλύτερο δυνατό αποτέλεσμα χωρίς να χρειάζεται να επανασχεδιάσει τα κριτήρια εξόδου. Η έννοια των νευρωνικών δικτύων, η οποία έχει τις ρίζες της στην τεχνητή νοημοσύνη, κερδίζει γρήγορα δημοτικότητα στην ανάπτυξη συστημάτων συναλλαγών. (CHEN, 2019)

Το πρώτο νευρωνικό δίκτυο σχεδιάστηκε από τον Warren McCulloch και τον Walter Pitts το 1943. Έγραψαν ένα σημαντικό τεύχος για το πώς οι νευρώνες μπορούν να δουλέψουν και να διαμορφώσουν τις ιδέες τους δημιουργώντας ένα απλό νευρωνικό δίκτυο χρησιμοποιώντας ηλεκτρικά κυκλώματα.

Αυτό το καινοτόμο μοντέλο άνοιξε το δρόμο για την έρευνα νευρωνικών δικτύων σε δύο τομείς:

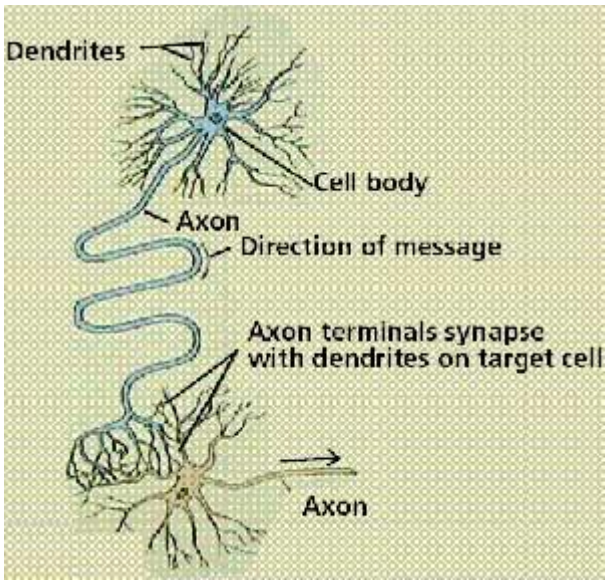
1. Βιολογικές διεργασίες στον εγκέφαλο.
2. Η εφαρμογή των νευρωνικών δικτύων στην τεχνητή νοημοσύνη (AI)

Η έρευνα AI επιταχύνθηκε ταχέα, με τον Kunihiko Fukushima να δημιουργεί το πρώτο πραγματικό, νευρωνικό δίκτυο με πολλά επίπεδα το 1975.

Ξεκινώντας, ο στόχος στη προσέγγιση των νευρωνικών δικτύων ήταν να δημιουργηθεί ένα υπολογιστικό σύστημα που θα μπορούσε να λύσει προβλήματα όπως ο ανθρώπινος εγκέφαλος. Ωστόσο, με την πάροδο του χρόνου, οι ερευνητές στρέφουν την εστίασή τους στη χρήση νευρωνικών δικτύων για να ταιριάζουν με συγκεκριμένα καθήκοντα, οδηγώντας σε αποκλίσεις από μια αυστηρά βιολογική προσέγγιση. Έκτοτε, η υποστήριξη καθηκόντων των νευρωνικών δικτύων έχει εμπλουτιστεί, όπως οραματισμό στον υπολογιστή, αναγνώριση ομιλίας, μηχανική μετάφραση, φιλτράρισμα κοινωνικών δικτύων, παιγνιδιών και βιντεοπαιχνιδιών και ιατρική διάγνωση. (https://www.sas.com, 2019)

Σήμερα, τα νευρικά δίκτυα χρησιμοποιούνται για την επίλυση πολλών επιχειρηματικών προβλημάτων, όπως η πρόβλεψη πωλήσεων, η έρευνα πελατών, η επικύρωση δεδομένων και η διαχείριση κινδύνων. (Shah, 2017)

Βιολογικοί Νευρώνες



(https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/images/neuron.jpg, 2019)

Ο εγκέφαλος αποτελείται κατά κύριο λόγο προσεγγιστικά από 10 δισεκατομμύρια νευρώνες, κάθε ένας από τους οποίους συνδέεται με περίπου 10.000 άλλους νευρώνες. Κάθε μία από τις κίτρινες κηλίδες στην παραπάνω εικόνα είναι τα νευρωνικά κυτταρικά σώματα (soma), και οι γραμμές είναι τα κανάλια εισόδου και εξόδου (δενδρίτες και άξονες) που τα συνδέουν. Κάθε νευρώνας λαμβάνει ηλεκτροχημικές εισροές από άλλους νευρώνες στους δενδρίτες. Εάν το άθροισμα αυτών των ηλεκτρικών εισόδων είναι επαρκώς ισχυρό για να ενεργοποιήσει τον νευρώνα, μεταδίδει ένα ηλεκτροχημικό σήμα κατά μήκος του αξόνου και περνά αυτό το σήμα στους άλλους νευρώνες των οποίων οι δενδρίτες είναι προσαρτημένοι σε οποιοδήποτε από τα άκρα του αξόνου. Αυτοί οι προσκολλημένοι νευρώνες μπορεί τότε να πυροδοτήσουν.

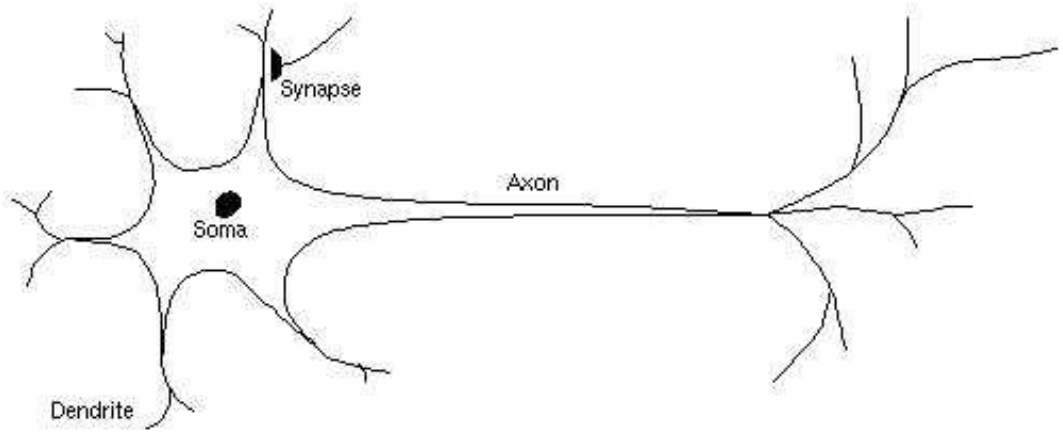
Είναι σημαντικό να σημειωθεί ότι ένας νευρώνας πυροδοτεί μόνο εάν το συνολικό σήμα που λαμβάνεται στο κυτταρικό σώμα υπερβαίνει ένα ορισμένο επίπεδο. Ο νευρώνας είτε πυροδοτεί είτε δεν το κάνει, δεν υπάρχουν διαφορετικές ποιότητες πυροδότησης.

Έτσι, ολόκληρος ο εγκέφαλος αποτελείται από αυτούς τους διασυνδεδεμένους ηλεκτροχημικούς νευρώνες που μεταδίδουν. Από ένα πολύ μεγάλο αριθμό εξαιρετικά απλών μονάδων επεξεργασίας (το καθένα εκτελώντας ένα σταθμισμένο άθροισμα των εισροών του, και στη συνέχεια πυροδοτώντας ένα δυαδικό σήμα αν η συνολική είσοδος ξεπεράσει ένα ορισμένο επίπεδο), ο εγκέφαλος καταφέρνει να εκτελεί εξαιρετικά πολύπλοκες εργασίες.
(https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/index.html, 2019)

Τεχνητοί Νευρώνες

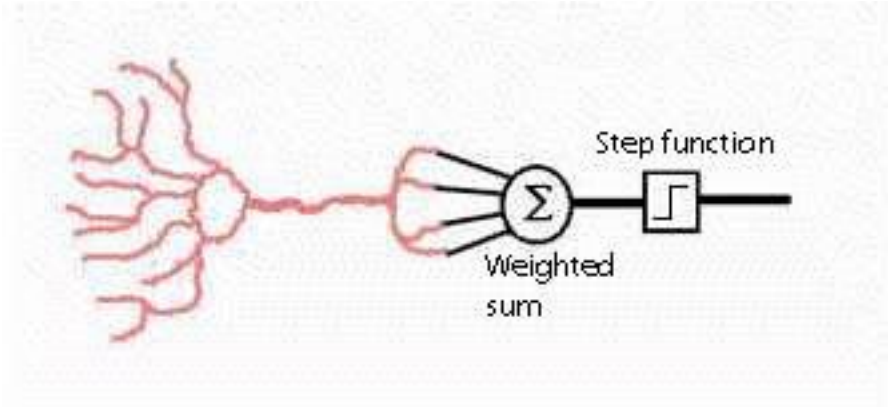
Το Αντίληπτρο(Perceptron). Το Αντίληπτρο είναι ένα μαθηματικό μοντέλο ενός βιολογικού νευρώνα. Ενώ στους πραγματικούς νευρώνες ο δενδρίτης λαμβάνει ηλεκτρικά σήματα από τους άξονες άλλων νευρώνων, στο Αντίληπτρο αυτά τα ηλεκτρικά σήματα αντιπροσωπεύονται ως αριθμητικές τιμές. Στις συνάψεις μεταξύ του δενδρίτη και των αξόνων, τα ηλεκτρικά σήματα διαμορφώνονται σε διάφορες

ποσότητες. Αυτό διαμορφώνεται επίσης στο Αντίληπτρο πολλαπλασιάζοντας κάθε τιμή εισόδου με μια τιμή που ονομάζεται βάρος. Ένας πραγματικός νευρώνας πυροδοτεί ένα σήμα εξόδου μόνο όταν η συνολική ισχύς των σημάτων εισόδου υπερβαίνει ένα συγκεκριμένο όριο. Μοντελοποιήσαμε αυτό το φαινόμενο σε ένα Αντίληπτρο υπολογίζοντας το σταθμισμένο άθροισμα των εισροών για να αντιπροσωπεύει τη συνολική ισχύ των σημάτων εισόδου και εφαρμόζοντας μια συνάρτηση βημάτων στο άθροισμα για να καθορίσουμε την έξοδο του. Όπως και στα βιολογικά νευρωνικά δίκτυα, αυτή η έξοδος τροφοδοτείται σε άλλα Αντίληπτρα.



(https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/bioneuron.jpg, 2019)

Ένας βιολογικός νευρώνας.



(https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/artificial.jpg, 2019)

Ένας τεχνητός νευρώνας (Perceptron /Αντίληπτρο).
(https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html, 2019)

Ο Perceptron(Αντίληπτρο) είναι ένας δυαδικός ταξινομητής, δηλαδή μία συνάρτηση η οποία απεικονίζει την είσοδο x (ένα διάνυσμα με πραγματικές τιμές) σε μία τιμή εξόδου f(x) (μία και μοναδική δυαδική τιμή).

$$f(x) = \begin{cases} 1, if\ w * x + b > 0 \\ 0, else \end{cases}$$

Όπου w είναι ένα διάνυσμα από βάρη με πραγματικές τιμές και w*x είναι το εσωτερικό γινόμενο μεταξύ των διανυσμάτων w και x (Υπολογίζεται δηλαδή ένα βεβαρημένο άθροισμα). Το b είναι το 'bias', ένας σταθερός όρος ο οποίος δεν εξαρτάται από καμία τιμή εισόδου.

Η τιμή της f(x) (0 ή 1) χρησιμοποιείται για να ταξινομήσει το x είτε ως θετικό ή αρνητικό στιγμιότυπο, αν το πρόβλημα ταξινόμησης είναι δυαδικό. Το bias χρησιμοποιείται για την μετατόπιση της συνάρτησης ενεργοποίησης ή για να δώσει στον νευρώνα εξόδου ένα βασικό επίπεδο δραστηριότητας. Αν το b είναι αρνητικό τότε ο βεβαρημένος συνδυασμός των εισόδων πρέπει να παραγάγει μία θετική τιμή μεγαλύτερη του -b έτσι ώστε να αναγκάσει τον νευρώνα που ταξινομεί να έχει τιμή άνω του κατωφλίου 0. Χωρικά, το bias μεταβάλει την θέση (αλλά όχι τον προσανατολισμό) του συνόρου απόφασης.

Εφόσον οι είσοδοι τροφοδοτούνται στο δίκτυο άμεσα μέσω των βεβαρημένων συνδέσεων, ο νευρώνας μπορεί να θεωρηθεί ως ένα απλό είδος νευρωνικού δικτύου εμπρός τροφοδότησης. (https://el.wikipedia.org/wiki/Perceptron#%CE%9F%CF%81%CE%B9%CF%83%CE%BC%CF%8C%CF%82, 2019)

Διαφορετικοί τύποι νευρωνικών δικτύων χρησιμοποιούν διαφορετικές αρχές για τον καθορισμό των δικών τους κανόνων. Υπάρχουν πολλά είδη τεχνητών νευρωνικών δικτύων, το καθένα με μοναδικά πλεονεκτήματα. Εδώ είναι μερικά από τα πιο σημαντικά είδη νευρωνικών δικτύων και οι εφαρμογές τους.

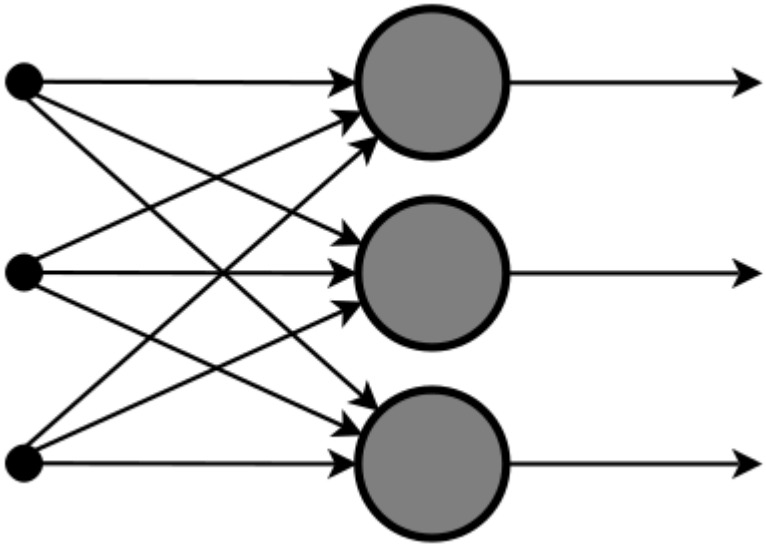
1. Feedforward Neural Network – Artificial Neuron

   Αυτός είναι ένας από τους απλούστερους τύπους τεχνητών νευρωνικών δικτύων. Σε ένα νευρωνικό δίκτυο feedforward, τα δεδομένα περνούν από τους διάφορους κόμβους εισόδου μέχρι να φτάσουν στον κόμβο εξόδου.

   Με άλλα λόγια, τα δεδομένα μετακινούνται σε μία μόνο κατεύθυνση από την πρώτη βαθμίδα μέχρι να φτάσουν στον κόμβο εξόδου. Αυτό είναι επίσης γνωστό ως ένα εμπρόσθιο πολλαπλασιασμένο κύμα το οποίο συνήθως επιτυγχάνεται με τη χρήση μιας συνάρτησης ενεργοποίησης κατηγοριοποίησης.

   Σε αντίθεση με τους πιο σύνθετους τύπους νευρωνικών δικτύων, δεν υπάρχει κανένας μετασχηματισμός και τα δεδομένα μετακινούνται προς μία μόνο κατεύθυνση. Ένα νευρωνικό δίκτυο feedforward μπορεί να έχει ένα μόνο στρώμα ή μπορεί να έχει κρυμμένα στρώματα.

   Σε ένα νευρωνικό δίκτυο feedforward, υπολογίζεται το άθροισμα των προϊόντων των εισροών και των βαρών τους. Αυτό στη συνέχεια τροφοδοτείται στην έξοδο. Ακολουθεί ένα παράδειγμα ενός μονοστρωματικού νευρωνικού δικτύου feedforward.



(https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-2-3.png, 2019)

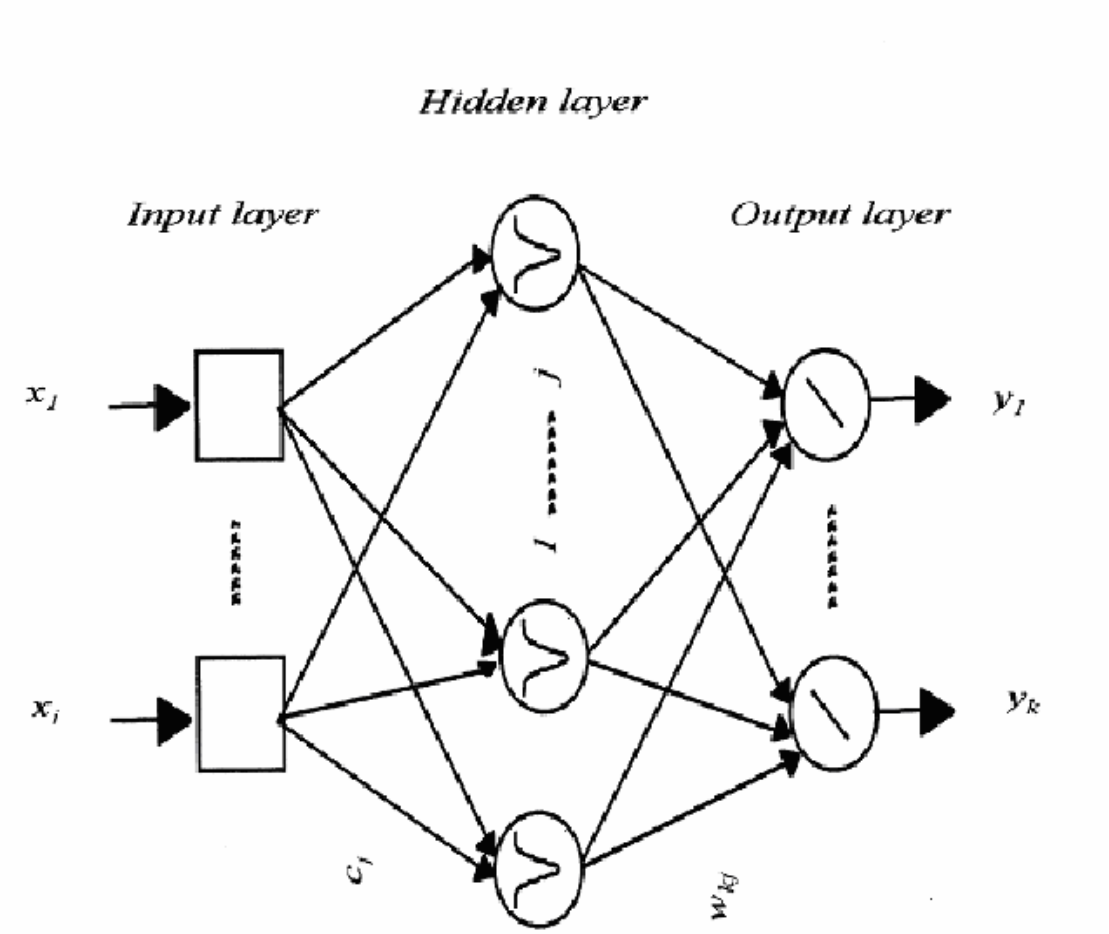
Τα νευρωνικά δίκτυα feedforward χρησιμοποιούνται σε τεχνολογίες όπως η αναγνώριση προσώπου και η όραση στον υπολογιστή. Αυτό οφείλεται στο γεγονός ότι οι κατηγορίες στόχοι σε αυτές τις εφαρμογές είναι δύσκολο να ταξινομηθούν.

Ένα απλό feedforward νευρωνικό δίκτυο είναι εξοπλισμένο για να ασχολείται με δεδομένα που περιέχει πολύ θόρυβο. Τα νευρωνικά δίκτυα feedforward είναι επίσης σχετικά απλά να διατηρηθούν.

2. Radial Basis Function Neural Network

Μια λειτουργία ακτινικής βάσης λαμβάνει υπόψη την απόσταση οποιουδήποτε σημείου σε σχέση με το κέντρο. Τέτοια νευρικά δίκτυα έχουν δύο στρώματα. Στο εσωτερικό στρώμα, τα χαρακτηριστικά συνδυάζονται με τη λειτουργία ακτινικής βάσης.

Στη συνέχεια, η έξοδος αυτών των χαρακτηριστικών λαμβάνεται υπόψη κατά τον υπολογισμό της ίδιας εξόδου στο επόμενο βήμα. Εδώ υπάρχει ένα διάγραμμα που αντιπροσωπεύει ένα νευρωνικό δίκτυο με ακτινική βάση.



(https://www.researchgate.net/profile/Michael_Friswell/publication/240829694/figure/fig1/AS:298579240472582@1448198246661/Radial-basis-function-RBF-neural-network-structure.png, 2019)

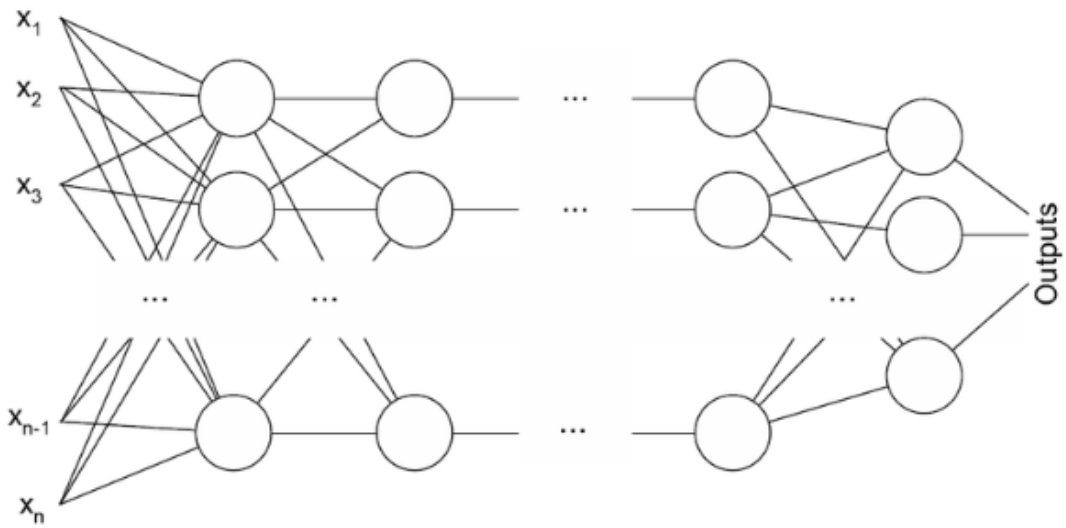
Το νευρωνικό δίκτυο της ακτινικής βάσης χρησιμοποιείται εκτενώς στα συστήματα αποκατάστασης ισχύος. Τις τελευταίες δεκαετίες, τα συστήματα ισχύος έχουν γίνει μεγαλύτερα και πιο περίπλοκα.

Αυτό αυξάνει τον κίνδυνο συστολής. Αυτό το νευρικό δίκτυο χρησιμοποιείται στα συστήματα αποκατάστασης ισχύος για να αποκατασταθεί η ισχύς στο συντομότερο δυνατό χρονικό διάστημα.

3. Multilayer Perceptron

Ένας πολλαπλών στρώσεων perceptron έχει τρία ή περισσότερα στρώματα. Χρησιμοποιείται για την ταξινόμηση δεδομένων που δεν μπορούν να διαχωριστούν γραμμικά. Είναι ένα είδος τεχνητού νευρικού δικτύου που είναι πλήρως συνδεδεμένο. Αυτό συμβαίνει επειδή κάθε κόμβος σε ένα στρώμα συνδέεται με κάθε κόμβο στο επόμενο στρώμα.

Ένας πολλαπλών στρώσεων perceptron χρησιμοποιεί μια μη γραμμική συνάρτηση ενεργοποίησης (κυρίως υπερβολική εφαπτομένη ή logistic λειτουργία). Εδώ είναι αυτό που μοιάζει με ένα πολυστρωματικό perceptron.



(https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-5-1.png, 2019)

Αυτός ο τύπος νευρωνικού δικτύου εφαρμόζεται εκτεταμένα στις τεχνολογίες αναγνώρισης ομιλίας και μηχανικής μετάφρασης.\

4. Convolutional Neural Network

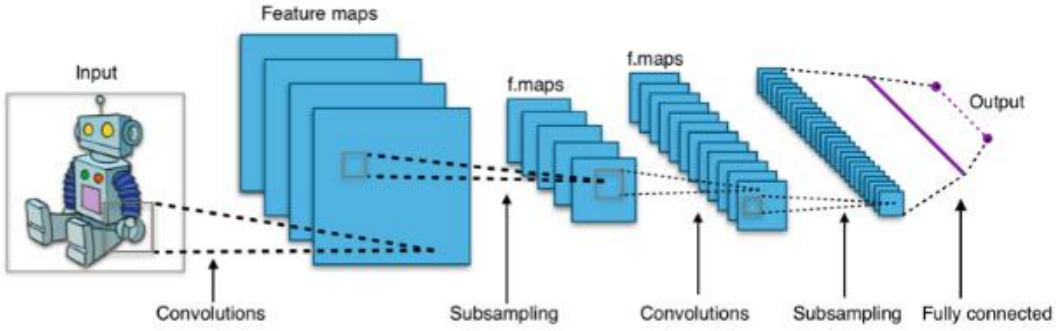
Ένα συνελικτικό νευρωνικό δίκτυο (CNN) χρησιμοποιεί μια παραλλαγή των πολλαπλών στρώσεων perceptrons. Ένα CNN περιέχει ένα ή περισσότερα από ένα συνελικτικά στρώματα. Αυτά τα στρώματα μπορούν είτε να αλληλοσυνδεθούν είτε να συγκεντρωθούν.

Πριν περάσει το αποτέλεσμα στο επόμενο στρώμα, το συνελικτικό στρώμα χρησιμοποιεί μια συνελικτική λειτουργία στην είσοδο. Λόγω αυτής της συνελικτικής λειτουργίας, το δίκτυο μπορεί να είναι πολύ βαθύτερο αλλά με πολύ λιγότερες παραμέτρους.

Λόγω αυτής της ικανότητας, τα συνεργατικά νευρωνικά δίκτυα παρουσιάζουν πολύ αποτελεσματικά αποτελέσματα στην αναγνώριση εικόνων και βίντεο, στην επεξεργασία φυσικής γλώσσας και στα συστήματά τους.

Τα περιελισσόμενα νευρωνικά δίκτυα παρουσιάζουν επίσης εξαιρετικά αποτελέσματα στη σημασιολογική ανάλυση και την παραφρακτική ανίχνευση. Εφαρμόζονται επίσης στην επεξεργασία σήματος και στην ταξινόμηση εικόνων.

Τα CNN χρησιμοποιούνται επίσης στην ανάλυση εικόνας και την αναγνώριση στη γεωργία, όπου τα χαρακτηριστικά του καιρού εξάγονται από δορυφόρους όπως το LSAT για να προβλέψουν την ανάπτυξη και την απόδοση ενός τεμαχίου γης. Ακολουθεί μια εικόνα για το πώς μοιάζει με το Convolutional Neural Network.



(https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-6-1.png, 2019)

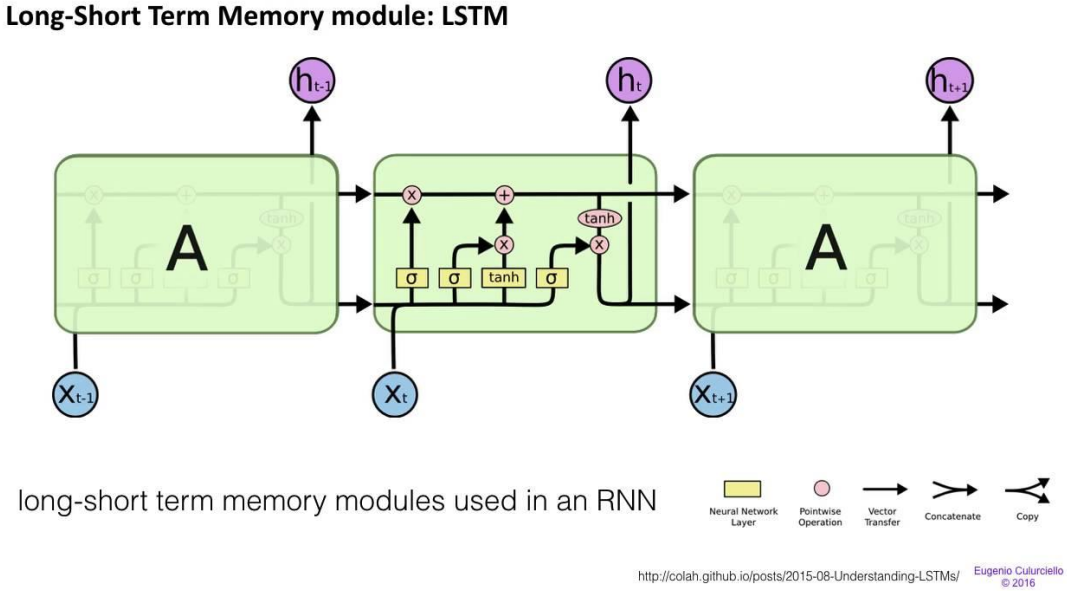
5. Recurrent Neural Network(RNN) – Long Short Term Memory(Επαναλαμβανόμενο νευρωνικό δίκτυο (RNN) - Μακροπρόθεσμη μνήμη)

Ένα επαναλαμβανόμενο νευρωνικό δίκτυο είναι ένα είδος τεχνητού νευρικού δικτύου στο οποίο η έξοδος ενός συγκεκριμένου στρώματος αποθηκεύεται και τροφοδοτείται πίσω στην είσοδο. Αυτό βοηθά στην πρόβλεψη του αποτελέσματος του στρώματος.

Το πρώτο στρώμα σχηματίζεται με τον ίδιο τρόπο όπως στο δίκτυο προώθησης. Δηλαδή, με το προϊόν του αθροίσματος των βαρών και των χαρακτηριστικών. Ωστόσο, σε επόμενα στρώματα, αρχίζει η διαδικασία επαναλαμβανόμενου νευρικού δικτύου.

Από κάθε βήμα-βήμα στο επόμενο, κάθε κόμβος θα θυμάται κάποιες πληροφορίες που είχε στο προηγούμενο βήμα. Με άλλα λόγια, κάθε κόμβος λειτουργεί ως κύτταρο μνήμης ενώ υπολογίζει και εκτελεί πράξεις. Το νευρικό δίκτυο αρχίζει με την εμπρόσθια διάδοση όπως συνήθως, αλλά θυμάται τις πληροφορίες που μπορεί να χρειαστεί να χρησιμοποιήσει αργότερα.

Εάν η πρόβλεψη είναι λάθος, το σύστημα αυτο-μαθαίνει και εργάζεται για να κάνει τη σωστή πρόβλεψη κατά τη διάρκεια του backpropagation. Αυτός ο τύπος νευρωνικού δικτύου είναι πολύ αποτελεσματικός στην τεχνολογία μετατροπής κειμένου σε ομιλία. Εδώ είναι ένα επαναλαμβανόμενο νευρωνικό δίκτυο.

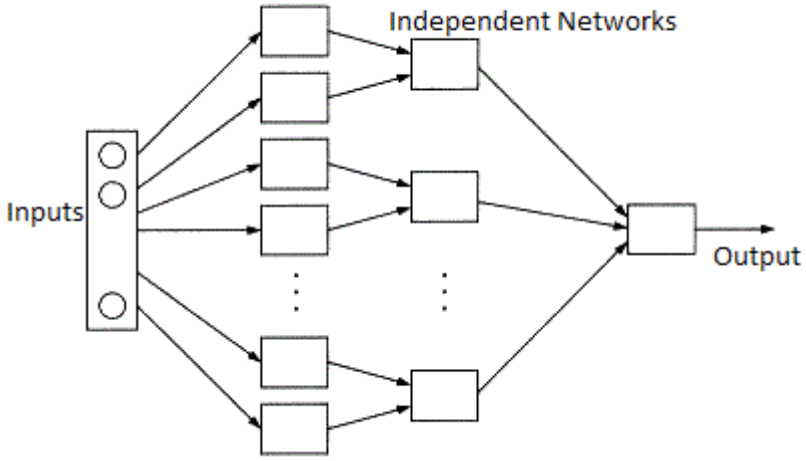


(https://i.ytimg.com/vi/kMLl-TKaEnc/maxresdefault.jpg, 2019)

6. Modular Neural Network

Ένα δομοστοιχειωτό νευρωνικό δίκτυο διαθέτει διάφορα δίκτυα που λειτουργούν ανεξάρτητα και εκτελούν επιμέρους εργασίες. Τα διαφορετικά δίκτυα δεν αλληλεπιδρούν ή αλληλοενημερώνονται κατά τη διάρκεια της διαδικασίας υπολογισμού. Δουλεύουν ανεξάρτητα προς την επίτευξη της απόδοσης.

Ως αποτέλεσμα, μια μεγάλη και πολύπλοκη υπολογιστική διαδικασία μπορεί να γίνει πολύ πιο γρήγορα, χωρίζοντάς την σε ανεξάρτητες συνιστώσες. Η ταχύτητα υπολογισμού αυξάνεται επειδή τα δίκτυα δεν αλληλεπιδρούν ή και συνδέονται μεταξύ τους. Ακολουθεί μια οπτική αναπαράσταση ενός Modular Neural Network.



(https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-8.gif, 2019)

7. Sequence-To-Sequence Models(Μοντέλα ακολουθίας)

Ένα μοντέλο ακολουθίας αποτελείται από δύο επαναλαμβανόμενα νευρωνικά δίκτυα. Υπάρχει ένας κωδικοποιητής που επεξεργάζεται την είσοδο και έναν αποκωδικοποιητή που επεξεργάζεται την έξοδο. Ο κωδικοποιητής και ο αποκωδικοποιητής μπορούν είτε να χρησιμοποιούν τις ίδιες είτε διαφορετικές παραμέτρους. Αυτό το μοντέλο εφαρμόζεται ιδιαίτερα στις περιπτώσεις όπου το μήκος των δεδομένων εισόδου δεν είναι το ίδιο με το μήκος των δεδομένων εξόδου.

Τα μοντέλα ακολουθίας εφαρμόζονται κυρίως σε συστήματα chatbots, μηχανικής μετάφρασης και απαντήσεων σε ερωτήσεις.
(Mehta, 2019)

2..4 Εφαρμογές (γιατί χρειάζεται το classification)

Η κατηγοριοποίηση είναι απαραίτητη και εφαρμόζεται από τον άνθρωπο εδώ και χιλιάδες χρόνια για τους εξής λόγους:

- Είναι σημαντικό για την καθημερινή ζωή καθώς το χρησιμοποιούμε σε όλα όσα κάνουμε.
- Κάνει τα πράγματα ευκολότερα να βρεθούν και να αναγνωριστούν π.χ. Ένα πιρούνι είναι ένα μαχαιροπίρουνο, έτσι θα κοιτάξω στο συρτάρι με τα μαχαιροπίρουνα.
- Η ταξινόμηση μπορεί και πρέπει να χρησιμοποιηθεί για τη διαλογή οτιδήποτε από έγγραφα έως σπουδαστές.
- Εάν πήγαμε σε μια βιβλιοθήκη χωρίς ταξινόμηση, από πού θα ξεκινούσαμε αναζητώντας ένα συγκεκριμένο βιβλίο;
- Η διαφοροποίηση των αντικειμένων μας επιτρέπει να τα κατατάξουμε σε ομάδες. (http://www.sims.monash.edu.au/subjects/ims2603/resources/week6/6.9.pdf, 2019)

3 Εφαρμογή σε R

3..1 Περιγραφή της εφαρμογής

Η εφαρμογή θα είναι γραμμένη σε γλώσσα R. Η R είναι μια γλώσσα και περιβάλλον για τη στατιστική πληροφορική. Πρόκειται για ένα έργο GNU παρόμοιο με τη γλώσσα και το περιβάλλον S που αναπτύχθηκε στα Εργαστήρια Bell (πρώην AT&T, τώρα Lucent Technologies) από τον John Chambers και τους συναδέλφους του. Η R μπορεί να θεωρηθεί ως διαφορετική εφαρμογή του S. Υπάρχουν μερικές σημαντικές διαφορές, αλλά πολύς κώδικας γραμμένος για S τρέχει αναλλοίωτος κάτω από την R.

Η R παρέχει μια μεγάλη ποικιλία στατιστικών (γραμμική και μη γραμμική μοντελοποίηση, κλασικές στατιστικές δοκιμές, ανάλυση χρονοσειρών, ταξινόμηση, ομαδοποίηση, ...) και γραφικών τεχνικών και είναι ιδιαίτερα επεκτάσιμη. Η γλώσσα S είναι συχνά το όχημα επιλογής για έρευνα στη στατιστική μεθοδολογία και το R παρέχει μια οδό ανοιχτού κώδικα για τη συμμετοχή σε αυτή τη δραστηριότητα.

Ένα από τα πλεονεκτήματα της R είναι η ευκολία με την οποία μπορούν να παραχθούν καλά σχεδιασμένα σχέδια, συμπεριλαμβανομένων μαθηματικών συμβόλων και τύπων όπου απαιτείται. Έχει παρθεί μεγάλη προσοχή στις προεπιλογές για τις μικρές επιλογές σχεδίασης στα γραφικά, αλλά ο χρήστης διατηρεί τον πλήρη έλεγχο.

Η R είναι διαθέσιμο ως Ελεύθερο Λογισμικό σύμφωνα με τους όρους του Free Software Foundation's GNU General Public License σε μορφή πηγαίου κώδικα. Συντάσσει και λειτουργεί σε μια μεγάλη ποικιλία πλατφορμών UNIX και παρόμοιων συστημάτων (συμπεριλαμβανομένων του FreeBSD και του Linux), Windows και MacOS.

Το περιβάλλον R

Το R είναι μια ολοκληρωμένη σουίτα εγκαταστάσεων λογισμικού για χειρισμό δεδομένων, υπολογισμό και γραφική απεικόνιση. Περιλαμβάνει

- Αποτελεσματικό χειρισμό και αποθήκευση δεδομένων,
- μια σουίτα χειριστών για υπολογισμούς σε συστοιχίες, ιδίως πίνακες,
- μια μεγάλη, συνεκτική, ολοκληρωμένη συλλογή ενδιάμεσων εργαλείων για την ανάλυση δεδομένων,
- γραφικές διευκολύνσεις για ανάλυση δεδομένων και προβολή είτε στην οθόνη είτε σε έντυπη μορφή, και
- μια καλά αναπτυγμένη, απλή και αποτελεσματική γλώσσα προγραμματισμού που περιλαμβάνει όρους, βρόχους, επαναλαμβανόμενες λειτουργίες καθορισμένες από τον χρήστη και εγκαταστάσεις εισόδου και εξόδου.

Ο όρος «περιβάλλον» προορίζεται να τον χαρακτηρίσει ως ένα πλήρως σχεδιασμένο και συνεκτικό σύστημα, και όχι ως αύξηση πολύ συγκεκριμένων και άκαμπτων εργαλείων, όπως συμβαίνει συχνά με άλλα λογισμικά ανάλυσης δεδομένων.

Η R, όπως η S, έχει σχεδιαστεί γύρω από μια πραγματική γλώσσα υπολογιστή και επιτρέπει στους χρήστες να προσθέτουν πρόσθετες λειτουργίες καθορίζοντας νέες λειτουργίες. Μεγάλο μέρος του συστήματος είναι το ίδιο γραμμένο στη διάλεκτο R του S, γεγονός που διευκολύνει τους χρήστες να ακολουθούν τις αλγοριθμικές επιλογές που έχουν γίνει. Για υπολογιστικές εντατικές εργασίες, οι κωδικοί C, C ++ και Fortran μπορούν να συνδεθούν και να κληθούν κατά το χρόνο εκτέλεσης. Οι προχωρημένοι χρήστες μπορούν να γράψουν κώδικα C για να χειριστούν αντικείμενα R άμεσα.

Πολλοί χρήστες θεωρούν την R ως ένα στατιστικό σύστημα. Προτιμούμε να το θεωρήσουμε ως ένα περιβάλλον εντός του οποίου εφαρμόζονται οι στατιστικές τεχνικές. Το R μπορεί να επεκταθεί (εύκολα) μέσω πακέτων. Υπάρχουν περίπου οκτώ πακέτα που παρέχονται με τη διανομή R και πολλά

άλλα είναι διαθέσιμα μέσω της οικογένειας διαδικτυακών τόπων CRAN που καλύπτουν ένα πολύ ευρύ φάσμα σύγχρονων στατιστικών.

Η R έχει τη δική της μορφή τεκμηρίωσης τύπου LaTeX, η οποία χρησιμοποιείται για την παροχή ολοκληρωμένης τεκμηρίωσης, τόσο on-line σε διάφορες μορφές όσο και σε έντυπη μορφή. (https://www.r-project.org/about.html, 2020)

Στην εφαρμογή θα εισάγουμε ένα σύνολο δεδομένων στο οποίο θα γίνει κατηγοριοποίηση decision tree και θα βγει μια πρόβλεψη σχετικά με το σε πια κλάση υπάγεται η κάθε περίπτωση.

3..2 ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΘΑ ΑΝΑΛΥΣΟΥΜΕ

Το σύνολο δεδομένων λουλουδιών Iris ή το σύνολο δεδομένων Fisher's Iris είναι ένα σύνολο δεδομένων πολλαπλών παραλλαγών που εισήγαγε ο Βρετανός στατιστικός και βιολόγος Ronald Fisher στο έγγραφο του το 1936. Αυτό είναι ένα πολύ διάσημο και ευρέως χρησιμοποιούμενο σύνολο δεδομένων από όλους όσους προσπαθούν να μάθουν μηχανική μάθηση και στατιστικά στοιχεία. (https://en.wikipedia.org, 2020)

Το σύνολο δεδομένων περιέχει 3 κλάσεις από 50 περιπτώσεις η κάθε μια, όπου κάθε κλάση αναφέρεται σε ένα είδος φυτού Iris. Μια κλάση είναι γραμμικά διαχωρίσιμη από τις άλλες 2, οι τελευταίες ΔΕΝ είναι γραμμικά διαχωρίσιμες η μια από την άλλη.

Το χαρακτηριστικό που θα προβλέψουμε είναι η κλάση του φυτού Iris.

Τα χαρακτηριστικά περιέχουν τις εξής πληροφορίες :

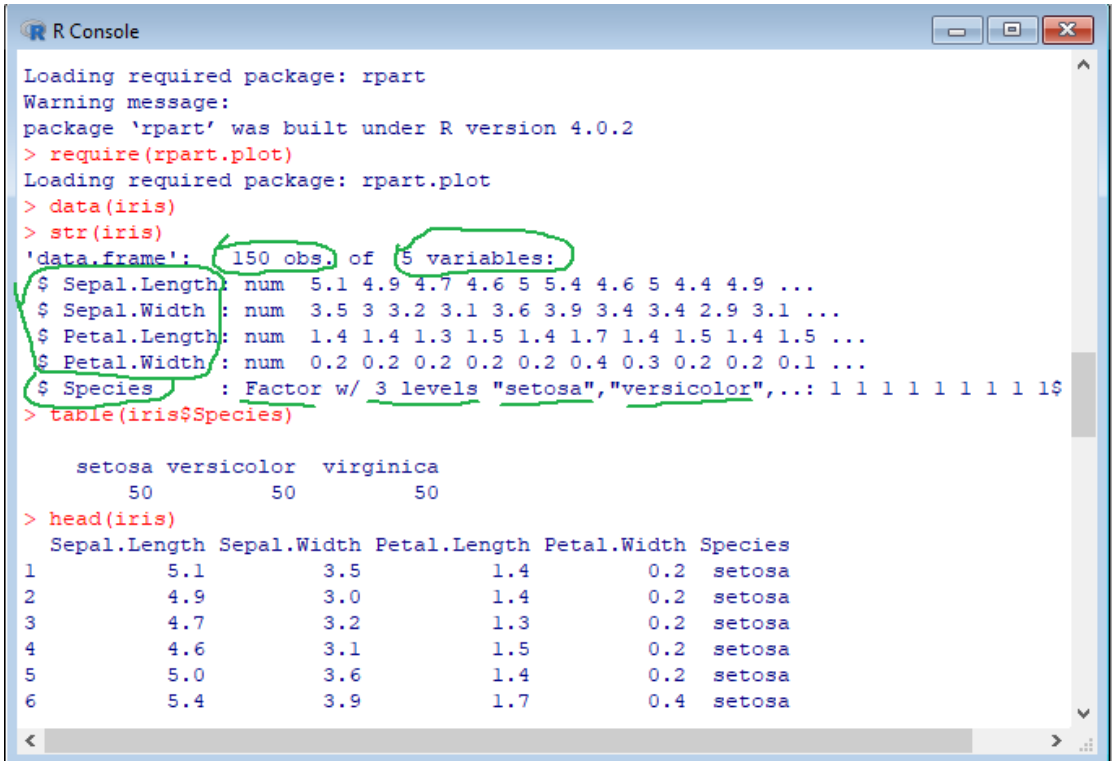
1. Μήκος σέπαλου σε cm
2. Πλάτος σέπαλου σε cm
3. Μήκος πέταλου σε cm
4. Πλάτος πέταλου σε cm
5. Κλάση:
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

Το σύνολο δεδομένων αποκτήθηκε από https://archive.ics.uci.edu/ml/datasets/Iris.

3..3 ΑΠΟΤΕΛΕΣΜΑΤΑ – ΣΥΓΚΡΙΤΙΚΗ ΑΝΑΛΥΣΗ

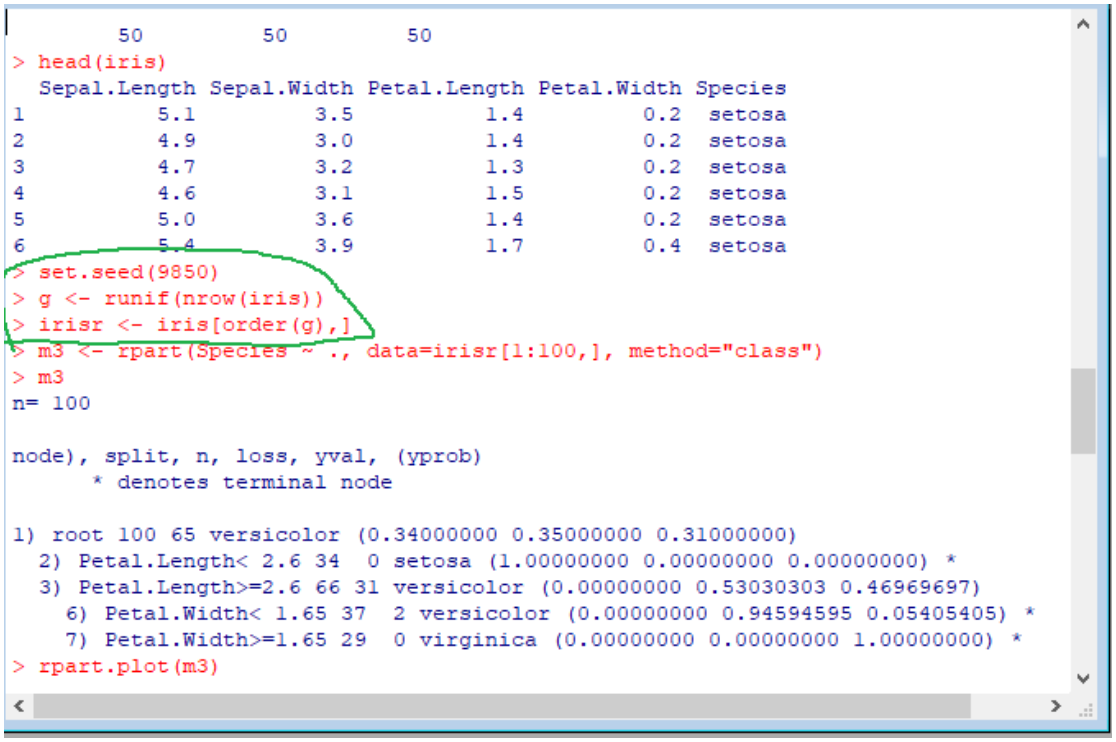
Για αρχή βλέπουμε ότι έχουμε 150 παραδείγματα, 5 χαρακτηριστικά. Τα 4 χαρακτηριστικά είναι αριθμητικά και το 1 είναι κατηγορίας.

Η πρόβλεψη θα γίνει με βάση το χαρακτηριστικό κατηγορίας(setosa, versicolor, virginica).



Πριν κάνουμε οτιδήποτε πάντα πρέπει να αναμιγνύουμε τα δεδομένα αλλιώς δεν θα έχει νόημα να προχωρήσουμε σε στατιστική ανάλυση αν είναι ήδη σε σειρά.

Αποθηκεύουμε τα δεδομένα με τυχαία σειρά σε ένα καινούργιο σετ δεδομένων irisr για να ξεχωρίζουμε τη τυχαία με τη κανονική βάση δεδομένων μας

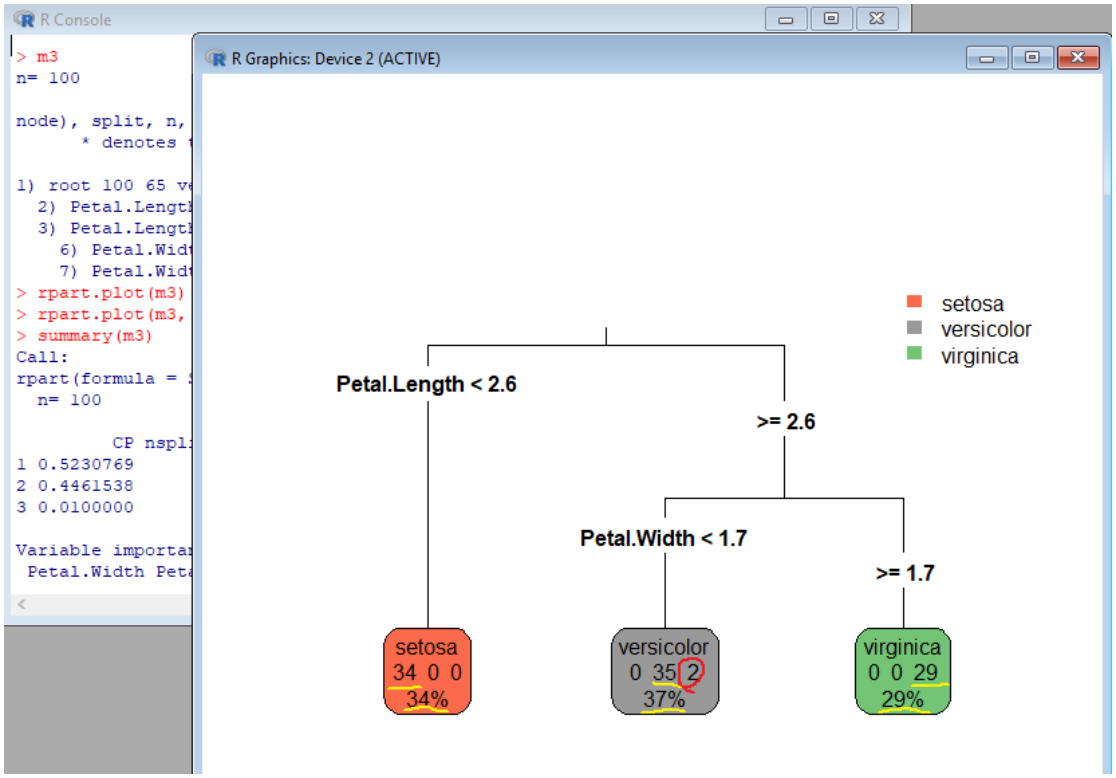


Χρήση rpart

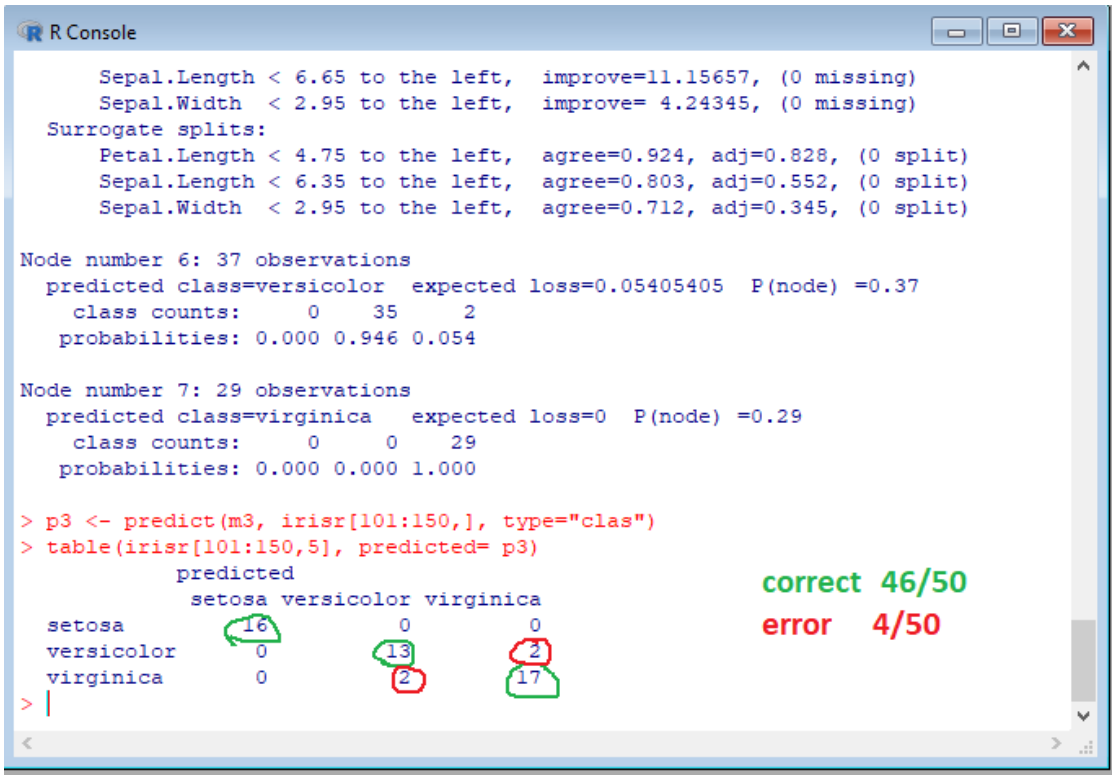
Θα χρησιμοποιήσουμε το πακέτο ανάλυσης rpart το οποίο περιέχει τη μαθηματική φόρμουλα για να γίνει η ανάλυση. Θα αναλύσουμε τα πρώτα 100 δεδομένα από τα 150 γιατί θα είναι το σετ εκπαίδευσης για να τεστάρουμε τα τελευταία 50 δεδομένα και να γίνει η εκτίμηση ποιο ακριβής και να δούμε πόσο καλά μπορεί αυτό το μοντέλο να γενικευτεί. Θα ονομάσουμε το μοντέλο με τα πρώτα 100 δεδομένα m3.

```
        50         50         50
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> set.seed(9850)
> g <- runif(nrow(iris))
> irisr <- iris[order(g),]
> m3 <- rpart(Species ~ ., data=irisr[1:100,], method="class")
> m3
n= 100

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 100 65 versicolor (0.34000000 0.35000000 0.31000000)
  2) Petal.Length< 2.6 34  0 setosa (1.00000000 0.00000000 0.00000000) *
  3) Petal.Length>=2.6 66 31 versicolor (0.00000000 0.53030303 0.46969697)
    6) Petal.Width< 1.65 37  2 versicolor (0.00000000 0.94594595 0.05405405) *
    7) Petal.Width>=1.65 29  0 virginica (0.00000000 0.00000000 1.00000000) *
> rpart.plot(m3)
```

Με τη χρήση της rpart.plot για το μοντέλο m3 βλέπουμε με γραφική απεικόνιση πως κατηγοριοποιούνται τα δεδομένα μας σύμφωνα με τις οδηγίες που του δώσαμε δηλαδή ποιος είναι ο στόχος μας(Species) και ποιοι οι προβλεπτές μας (Sepal.Length,Sepal.Width,Petal.Length,Petal.Width), έχουμε με ποσοστό και με ακριβείς αριθμούς τις κατηγορίες με τα δεδομένα τους και επίσης βλέπουμε ότι έχουμε δύο λάθη.



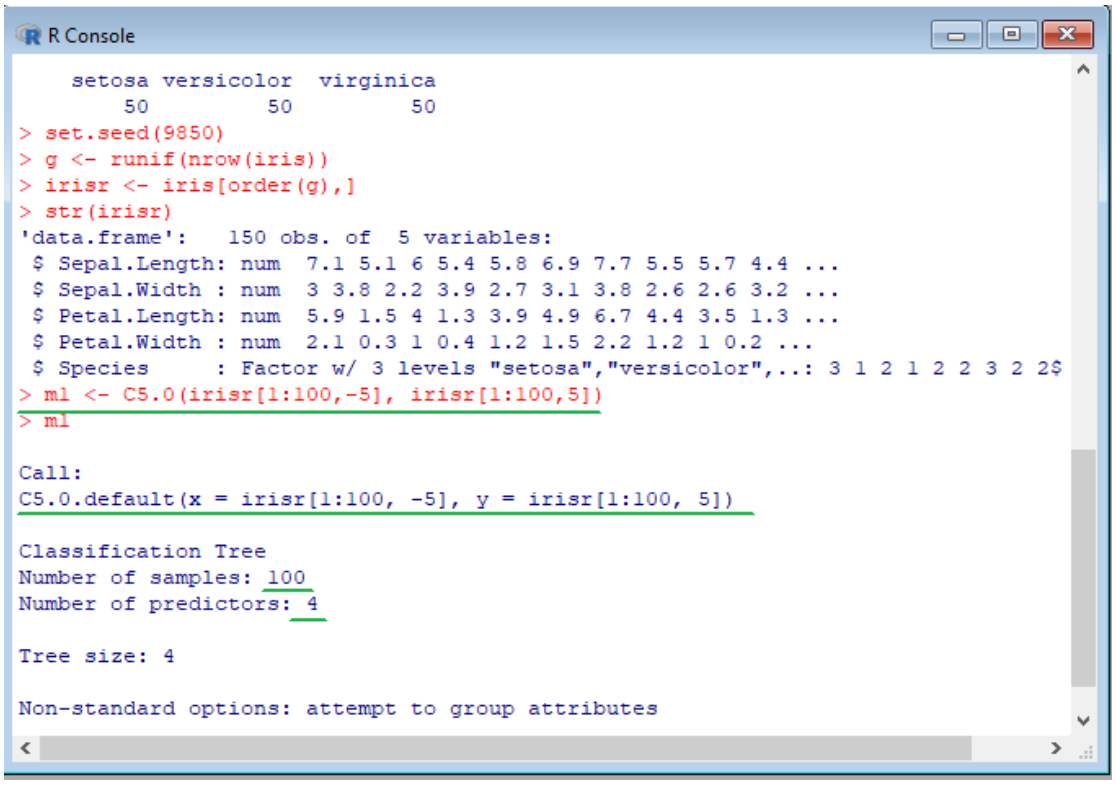
Έπειτα με τη χρήση confusion matrix θα δούμε πόσα θα προβλέψει σωστά και πόσα θα κάνει λάθος.



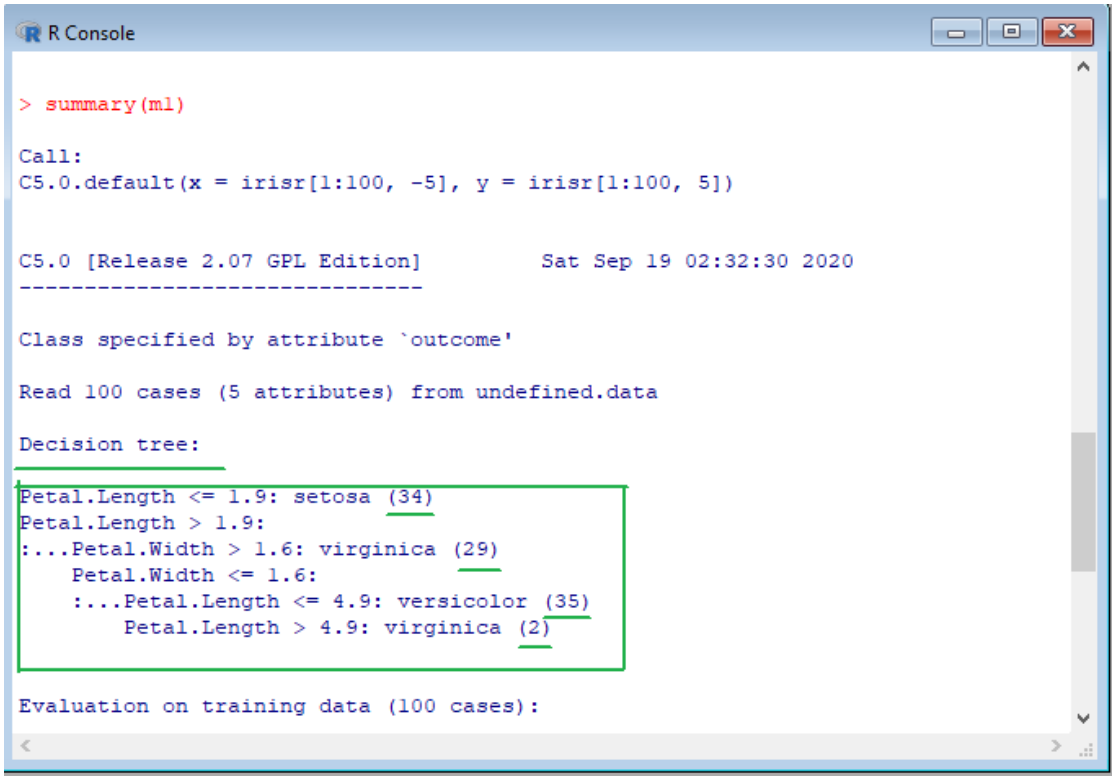
Χρήση C5.0

Τα δεδομένα μας είναι τα ίδια, θα γίνει πρόβλεψη και τυχαία κατανομή όπως και πριν αλλά με χρήση κατηγοριοποίησης C5.0.

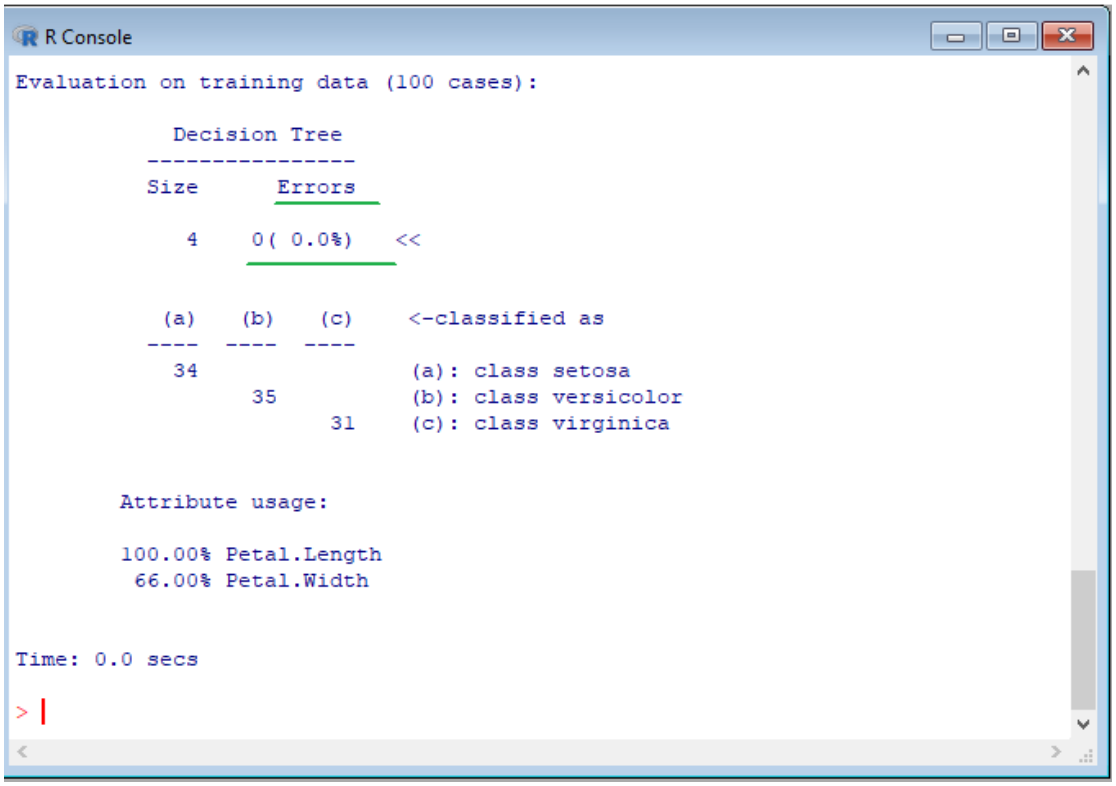
Θα χρησιμοποιήσουμε το πακέτο ανάλυσης C5.0 το οποίο περιέχει τη μαθηματική φόρμουλα για να γίνει η ανάλυση. Θα αναλύσουμε τα πρώτα 100 δεδομένα από τα 150 γιατί θα είναι το σετ εκπαίδευσης για να τεστάρουμε τα τελευταία 50 δεδομένα και να γίνει η εκτίμηση ποιο ακριβής και να δούμε πόσο καλά μπορεί αυτό το μοντέλο να γενικευτεί. Θα ονομάσουμε το μοντέλο με τα πρώτα 100 δεδομένα m1.



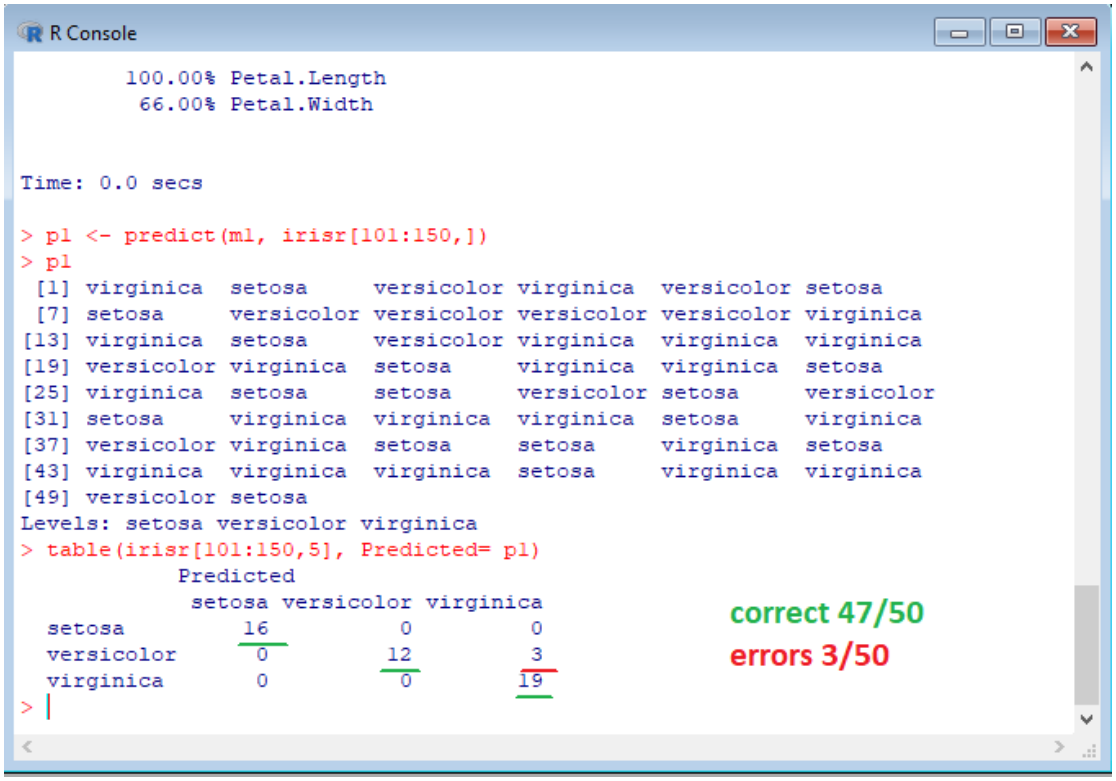
Με τη χρήση της εντολής summary(m1) βλέπουμε το δέντρο απόφασης.



Βλέπουμε ότι δεν υπάρχουν λάθη.



Έπειτα με τη χρήση confusion matrix θα δούμε πόσα θα προβλέψει σωστά και πόσα θα κάνει λάθος.



4 Βιβλιογραφία

Bigelow, S. J. (2005, 11 02). *techtarget*. Ανάκτηση από http://searchstorage.techtarget.com/news/1139254/Data-classification-An-overview

Brown, M. (2012, 12 11). *IBM*. Ανάκτηση από https://www.ibm.com/developerworks/library/ba-data-mining-techniques/

Brownlee, J. (2013, 12 25). *machine learning mastery*. Ανάκτηση από http://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/

CHEN, J. (2019, 10 13). *https://www.investopedia.com/terms/n/neuralnetwork.asp*. Ανάκτηση από https://www.investopedia.com/terms/n/neuralnetwork.asp: https://www.investopedia.com/terms/n/neuralnetwork.asp

Cichosz, P. (2015). *Data Mining Algorithms: Explained Using R.* John Wiley & Sons, Ltd.

Eckerson, W. (2007, 5 10). *The Data Warehouse Institute*. Ανάκτηση από https://tdwi.org/articles/2007/05/10/predictive-analytics.aspx?sc_lang=en: https://tdwi.org/articles/2007/05/10/predictive-analytics.aspx?sc_lang=en

Everitt, B. S., & Hothorn, T. (2010). *A handbook of statistical analyses using R, 2nd ed.* Chapman & Hall/CRC.

Fabricio Voznika, L. V. (χ.χ.). *https://courses.cs.washington.edu*. Ανάκτηση 11 11, 2016, από https://courses.cs.washington.edu: https://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf

Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data,Myths, Misconceptions and Methods.* Palgrave Macmillan UK.

Foster Provost, T. F. (2013). *Data Science for Business, What You Need to Know About Data Mining and Data-Analytic Thinking.* O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Glenn J. Myatt, W. P. (2014). *MAKING SENSE OF DATA I, A Practical Guide to Exploratory Data Analysis and Data Mining.* John Wiley & Sons, Inc., Hoboken, New Jersey .

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers.

http://www.sims.monash.edu.au/subjects/ims2603/resources/week6/6.9.pdf. (2019, 12 16). *http://www.sims.monash.edu.au/subjects/ims2603/resources/week6/6.9.pdf*. Ανάκτηση από http://www.sims.monash.edu.au/subjects/ims2603/resources/week6/6.9.pdf: http://www.sims.monash.edu.au/subjects/ims2603/resources/week6/6.9.pdf

http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html. (2019, 12 16). *http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html*. Ανάκτηση από http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html: http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/images/neuron.jpg. (2019, 12 16). *https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/images/neuron.jpg*. Ανάκτηση από https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/images/neuron.jpg: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/images/neuron.jpg

https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/index.html. (2019, 12 16). *https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/index.html*. Ανάκτηση από

https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/index.html: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Biology/index.html

https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/artificial.jpg. (2019, 12 16). *https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/artificial.jpg*. Ανάκτηση από https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/artificial.jpg: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/artificial.jpg

https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/bioneuron.jpg. (2019, 12 16). *https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/bioneuron.jpg*. Ανάκτηση από https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/bioneuron.jpg: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/images/bioneuron.jpg

https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html. (2019, 12 16). *https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html*. Ανάκτηση από https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html

https://el.wikipedia.org/wiki/%CE%91%CE%BB%CE%B3%CF%8C%CF%81%CE%B9%CE%B8%CE%BC%CE%BF%CF%82_ID3. (2019, 11 30). *https://el.wikipedia.org/wiki/%CE%91%CE%BB%CE%B3%CF%8C%CF%81%CE%B9%CE%B8%CE%BC%CE%BF%CF%82_ID3*. Ανάκτηση από https://el.wikipedia.org/wiki/%CE%91%CE%BB%CE%B3%CF%8C%CF%81%CE%B9%CE%B8%CE%BC%CE%BF%CF%82_ID3: https://el.wikipedia.org/wiki/%CE%91%CE%BB%CE%B3%CF%8C%CF%81%CE%B9%CE%B8%CE%BC%CE%BF%CF%82_ID3

https://el.wikipedia.org/wiki/Perceptron#%CE%9F%CF%81%CE%B9%CF%83%CE%BC%CF%8C%CF%82. (2019, 11 30). *https://el.wikipedia.org/wiki/Perceptron#%CE%9F%CF%81%CE%B9%CF%83%CE%BC%CF%8C%CF%82*. Ανάκτηση από https://el.wikipedia.org/wiki/Perceptron#%CE%9F%CF%81%CE%B9%CF%83%CE%BC%CF%8C%CF%82: https://el.wikipedia.org/wiki/Perceptron#%CE%9F%CF%81%CE%B9%CF%83%CE%BC%CF%8C%CF%82

https://en.wikipedia.org. (2020, 5 30). Ανάκτηση από https://en.wikipedia.org/wiki/Iris_flower_data_set: https://en.wikipedia.org/wiki/Iris_flower_data_set

https://en.wikipedia.org/wiki/ID3_algorithm. (2019, 11 30). *https://en.wikipedia.org/wiki/ID3_algorithm*. Ανάκτηση από https://en.wikipedia.org/wiki/ID3_algorithm: https://en.wikipedia.org/wiki/ID3_algorithm

https://i.ytimg.com/vi/kMLl-TKaEnc/maxresdefault.jpg. (2019, 12 16). *https://i.ytimg.com/vi/kMLl-TKaEnc/maxresdefault.jpg*. Ανάκτηση από https://i.ytimg.com/vi/kMLl-TKaEnc/maxresdefault.jpg: https://i.ytimg.com/vi/kMLl-TKaEnc/maxresdefault.jpg

https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm. (2019, 11 15). *https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm*. Ανάκτηση από https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm: https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm

https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-2-3.png. (2019, 12 16). *https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-2-3.png*. Ανάκτηση από https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-2-3.png: https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-2-3.png

https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-5-1.png. (2019, 12 16). *https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-5-1.png*. Ανάκτηση από https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-5-1.png: https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-5-1.png

https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-6-1.png. (2019, 12 16). *https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-6-1.png*. Ανάκτηση από https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-6-1.png: https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-6-1.png

https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-8.gif. (2019, 12 16). *https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-8.gif*. Ανάκτηση από https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-8.gif: https://www.digitalvidya.com/wp-content/uploads/2019/01/Image-8.gif

https://www.researchgate.net/profile/Michael_Friswell/publication/240829694/figure/fig1/AS:298579240472582@1448198246661/Radial-basis-function-RBF-neural-network-structure.png. (2019, 12 16). *https://www.researchgate.net/profile/Michael_Friswell/publication/240829694/figure/fig1/AS:298579240472582@1448198246661/Radial-basis-function-RBF-neural-network-structure.png*. Ανάκτηση από https://www.researchgate.net/profile/Michael_Friswell/publication/240829694/figure/fig1/AS:298579240472582@1448198246661/Radial-basis-function-RBF-neural-network-structure.png: https://www.researchgate.net/profile/Michael_Friswell/publication/240829694/figure/fig1/AS:298579240472582@1448198246661/Radial-basis-function-RBF-neural-network-structure.png

https://www.r-project.org/about.html. (2020, 5 30). Ανάκτηση από https://www.r-project.org/about.html: https://www.r-project.org/about.html

https://www.sas.com. (2019, 12 16). *https://www.sas.com/en_us/insights/analytics/neural-networks.html*. Ανάκτηση από https://www.sas.com/en_us/insights/analytics/neural-networks.html: https://www.sas.com/en_us/insights/analytics/neural-networks.html

Lord, N. (2016, 10 11). *digitalguardian*. Ανάκτηση από https://digitalguardian.com/blog/what-data-classification-data-classification-definition

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook, 2nd ed.* Springer.

Margaret Taft, R. K. (2005, 6). *Oracle*. Ανάκτηση από https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/3predictive.htm

Mehta, A. (2019, 1 25). *https://www.digitalvidya.com/blog/types-of-neural-networks/*. Ανάκτηση από https://www.digitalvidya.com/blog/types-of-neural-networks/: https://www.digitalvidya.com/blog/types-of-neural-networks/

Mohammed J. Zaki, W. M. (2014). *Data Mining And Analysis, Fundamental Concepts and Algorithms.* Cambridge University Press.

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications.* Elsevier Inc.

Nyce, C. (2007). *Predictive Analytics White Paper.* American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America.

Serengil, S. I. (2018, 05 13). *https://sefiks.com/*. Ανάκτηση από https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/: https://sefiks.com/2018/05/13/a-step-by-step-c4-5-decision-tree-example/

Shah, J. (2017, 11 16). *https://blog.statsbot.co/neural-networks-for-beginners-d99f2235efca*. Ανάκτηση από https://blog.statsbot.co/neural-networks-for-beginners-d99f2235efca:

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Addison-Wesley.

wikipedia. (2019, 11 14). *https://en.wikipedia.org*. Ανάκτηση από https://en.wikipedia.org/wiki/Decision_tree_learning:

wikipedia. (χ.χ.). *Wikipedia*. Ανάκτηση 2016, από https://en.wikipedia.org/wiki/Predictive_analytics