ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
ΕΡΓΑΣΙΑ

ΑΝΔΡΙΝΟΠΟΥΛΟΥ ΧΡΙΣΤΙΝΑ **1115201500006**

Αθήνα
Ιούνιος 2019

Εισαγωγή

Στην παρούσα εργασία είναι υλοποιημένα όλα τα ζητούμενα. Στον φάκελο που παραδόθηκε περιέχεται ο κώδικας σε Matlab, ο οποίος περιέχει τον απαραίτητο σχολιασμό. Στη συνέχεια παρατίθενται πληροφορίες και διευκρινήσεις σχετικά με την υλοποίηση της εργασίας, με σκοπό την καλύτερη κατανόηση του κώδικα.

Υλοποίηση

Το πρόγραμμα εκκινεί και διαβάζεται το dataset από κατάλληλο αρχείο με κατάληξη .xlsx, με τη βοήθεια της xlsread συνάρτησης που προσφέρει το Matlab. Έπειτα, αφαιρείται η στήλη των id από το dataset καθώς δεν παρέχει καμία χρήσιμη πληροφορία για τη μελέτη. Τα ελλειπή δεδομένα αντικαθίστανται από την εκάστοτε επικρατέστερη τιμή με mode. Το preprocessing ολοκληρώνεται με κατάλληλο normalization στο εύρος [0,1].

Στη συνέχεια, εφαρμόζεται k-fold cross validation με k=10 κι έτσι καθένας από τους ταξινομητές που λαμβάνει χώρα εκπαιδεύται σε κατάλληλο train set και πραγματοποιεί έλεγχο στο αντοίστοιχο test set.

Οι ταξινομητές που χρησιμοποιήθηκαν είναι:

- Πολυεπίπεδα Νευρωνικά Δίκτυα Πρόσθιας Τροφοδότησης (Multilayer Perceptron):

Έγινε χρήση των εργαλείων: fitnet, train, net. Σημαντικό είναι στο σημείο αυτό να αναφερθεί ότι χρειάστηκε να χρησιμοποιήθει ανάστροφος, ώστε τα εργαλεία να παράγουν ορθά αποτελέσματα. Ακόμα, πρέπει να επισημανθεί ότι υπολογίστηκε η απόσταση από το 0 και το 1 και αντικαταστάθηκε αναλόγως με 0 ή 1.

- Ταξινομητής Κ-Κοντινότερων Γειτόνων (K-Nearest Neighbor):

Έγινε χρήση του fitcknn και predict.

- Ταξινομητής Bayes:

Έγινε χρήση του fitcnb και predict

- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines):

Έγινε χρήση του fitcsvm και predict.

- Δένδρα Απόφασης (Decision Tree):

Έγινε χρήση του fitctree και predict.

Για καθέναν ταξινομητή υπολογίζονται και τυπώνονται η ακρίβεια (accuracy), η ευαισθησία (sensitivity) και η ειδικότητα (specificity).

Παραμετροποίηση, αποτελέσματα και τελικά συμπεράσματα

Όλοι οι ταξινομητές παραμετροποιήθηκαν κατάλληλα, ώστε να παραχθούν όσο το δυνατό καλύτερα αποτελέσματα. Οι παραμετροποιήσεις που έγιναν παρουσιάζονται ανά ταξινομήτη παρακάτω μαζί με τα αντίστοιχα αποτελέσματα. Έχουν επισημανθεί οι καλύτερες επιδόσεις. Στον παραδοταίο κώδικα θα εντοπίσετε την καλύτερη, με βάση τα αποτελέσματα που προέκυψαν, παραμετροποιημένη εκδοχή κάθε ταξινομητή.

- Πολυεπίπεδα Νευρωνικά Δίκτυα Πρόσθιας Τροφοδότησης (Multilayer Perceptron):

	2 layers (4 1)	5 layers (10 5 3 2 1)
Bayesian Regularization	Accuracy = 0.9470 Sensitivity = 0.9350 Specificity = 0.9558	Accuracy = 0.9470 Sensitivity = 0.9350 Specificity = 0.9558
Gradient Descent	Accuracy = 0.9484 Sensitivity = 0.9473 Specificity = 0.9520	Accuracy = 0.9484 Sensitivity = 0.9473 Specificity = 0.9520

- Ταξινομητής Κ-Κοντινότερων Γειτόνων (K-Nearest Neighbor)

	Euclidean distance	Cosine distance
5 γείτονες	Accuracy = 0,9656 Sensitivity = 0.9546 Specificity = 0.9722	Accuracy = 0.9685 Sensitivity = 0.9335 Specificity = 0.9890
10 γείτονες	Accuracy = 0.9642 Sensitivity = 0.9578 Specificity = 0.9682	Accuracy = 0.9713 Sensitivity = 0.9351 Specificity = 0.9934
20 γείτονες	Accuracy = 0.9656 Sensitivity = 0.9552 Specificity = 0.9722	Accuracy = 0.9713 Sensitivity = 0.9352 Specificity = 0.9933
50 γείτονες	Accuracy = 0.9570 Sensitivity = 0.9533 Specificity = 0.9598	Accuracy = 0.9656 Sensitivity = 0.9342 Specificity = 0.9845

- Ταξινομητής Bayes:

	Distribution name = kernel	Distribution name = normal
Prior = uniform	Accuracy = 0.9670 Sensitivity = 0.9480 Specificity = 0.9780	Accuracy = 0.9613 Sensitivity = 0.9167 Specificity = 0.9890
Prior = empirical	Accuracy = 0.9656 Sensitivity = 0.9476 Specificity = 0.9760	Accuracy = 0.9599 Sensitivity = 0.9161 Specificity = 0.9865

- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines):

	Rbf kernel	Polynomial kernel
Kernel scale = default	Accuracy = 0.9684 Sensitivity = 0.9370 Specificity = 0.9867	Accuracy = 0.9570 Sensitivity = 0.9414 Specificity = 0.9665
Kernel scale = auto	Accuracy = 0.9699 Sensitivity = 0.9375 Specificity = 0.9890	Accuracy = 0.9485 Sensitivity = 0.9336 Specificity = 0.9578

- Δένδρα Απόφασης (Decision Tree):

	Categories = 5	Categories = 10	Categories = 50
Merge leaves = on	Accuracy = 0.9399 Sensitivity = 0.9096 Specificity = 0.9586	Accuracy = 0.9399 Sensitivity = 0.9096 Specificity = 0.9586	Accuracy = 0.9399 Sensitivity = 0.9096 Specificity = 0.9586
Merge leaves = off	Accuracy = 0.9399 Sensitivity = 0.9096 Specificity = 0.9586	Accuracy = 0.9399 Sensitivity = 0.9096 Specificity = 0.9586	Accuracy = 0.9399 Sensitivity = 0.9096 Specificity = 0.9586

Έπειτα από τον παραπάνω πειραματισμό μπορεί κανείς να ξεχωρίσει τους ταξινομητές K-κοντινότερων γειτόνων, Bayes και τις μηχανές διανυσμάτων υποστήριξης ως “αποδοτικότερους” ταξινομητές για την καλύτερη και πιο αποτελεσματική διαχώριση των δεδομένων σε καλοήθεις και κακοήθεις όγκους.

Σε μία πρώτη ανάγνωση, θα μπορούσε κανείς να θεωρήσει αρκετό για την αξιολόγηση των ταξινομητών μόνο το Accuracy, το οποίο αναφέρεται στο ποσοστό των ορθών προβλέψεων. Αν προσπαθήσουμε όμως να σκεφτούμε με την αρμόζουσα προσοχή για το πρόβλημα, καθώς μιλάμε για dataset που αφορά γυναίκες που μπορεί να πάσχουν ή όχι από καρκίνο του μαστού, θα καταλάβουμε ότι ίσως το Accuracy να μην εξυπηρετεί ακριβώς τους σκοπούς μας, τουλάχιστον έτσι όπως θα θέλαμε.

Θα χρειαστεί μόνο ένα παράδειγμα για να συνειδητοποιήσουμε ακριβώς σε τί αναφερόμαστε. Έστω ότι το μοντέλο μας πετυχαίνει Accuracy 0.92, που σημαίνει ότι πετύχαμε 92% σωστές προβλέψεις. Παρατηρώντας κανείς, ωστόσο, τον παρακάτω πίνακα που περιλαμβάνει αναλυτικά τα δεδομένα και εφαρμόζοντας τον τύπο για το Accuracy, δηλαδή Accuracy = (2 + 90) / (2 + 90 + 1 + 7)

True positive: 2 πραγματικότητα: κακοήθεια τεστ: κακοήθεια	*False negative: 7* πραγματικότητα: κακοήθεια τεστ: καλοήθεια
False positive: 1 πραγματικότητα: καλοήθεια τεστ: κακοήθεια	*True negative: 90* πραγματικότητα: καλοήθεια τεστ: καλοήθεια

αντιλαμβάνεται ότι από τα 9 δείγματα κακοήθειας, μόνο τα 2 προβλέφθηκαν σωστά.

Συνεπώς, πρέπει εκτός από το Accuracy που αναμφισβήτητα θα παίξει καθοριστικό ρόλο για την επιλογή του καταλληλότερου αλγορίθμου να λάβουμε υπόψιν και το Sensitivity, δηλαδή το ποσοστό εκείνων που νοσούν στην πραγματικότητα και έχουν θετική την εν λόγω δοκιμασία.

Με βάση όλα τα παραπάνω, αλλά κι από προσωπική μελέτη καταλήγουμε στο συμπέρασμα ότι αν έπρεπε να επιλεχθεί μόνο ένας ταξινομητής για να διαχωρίσει τα δεδομένα με τον καλύτερο δυνατό τρόπο αυτός θα ήταν οι μηχανές διανυσμάτων υποστήριξης.

Σκοπός των SVM είναι να βρεθεί ένα υπερεπίπεδο που λειτουργώντας ως επιφάνεια απόφασης να διαχωρίζει τα δεδομένα με τέτοιον τρόπο ώστε το περιθώριο διαχωρισμού μεταξύ θετικών και αρνητικών στιγμιοτύπων να μεγιστοποιείται. Με άλλα λόγια, στην περίπτωσή μας τα στιγμιότυπα κακοήθειας να διαχωρίζονται από τα στιγμιότυπα καλοήθειας με τέτοιον τρόπο ώστε το μεταξύ τους περιθώριο να μεγιστοποιείται και συνεπώς να λάβουμε έναν πιο αυστηρό διαχωρισμό των δεδομένων μας.