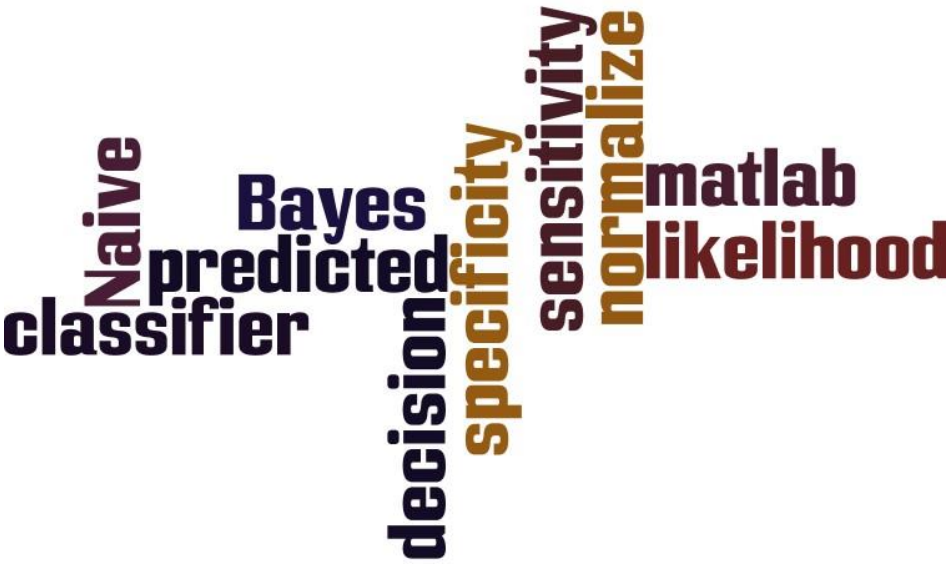

PROJECT ΘΕΩΡΙΑ ΑΠΟΦΑΣΕΩΝ

Pattern Analysis

Παναγιώτης Τριανταφυλλόπουλος
1054367
Ανδρέας Κάλλιστρος
1054351

ΜΕΡΟΣ 2

Ερώτημα 4ο

Αρχικά χρησιμοποιήσαμε την εντολή fitcsvm() και την crossval() του Matlab για την εκπαίδευση του support vector machine ταξινομητή με 5-fold cross validation. Με την χρήση βρόχου επανάληψης και τα κατάλληλα ορίσματα στην συνάρτηση (BoxConstraint, kernel function linear(default), cost) βρήκαμε την τιμή του C για την οποία ο ταξινομητής είναι βέλτιστος. Με έναν ακόμη βρόχο επανάληψης και τα ορίσματα BoxConstraint με την παραπάνω τιμή, cost, kernel function rbf και KernelScale υπολογίσαμε το τελευταίο (δηλαδή το γ) για το οποίο ο ταξινομητής είναι βέλτιστος. Ως συνάρτηση κόστους επιλέξαμε την [0 ,5;1, 0] που σημαίνει ότι μια λάθος ταξινόμηση για κάποιον που έχει πράγματι την ασθένεια έχει κόστος 5. Ως μετρική θα χρησιμοποιήσουμε τον γεωμετρικό μέσο ο οποίος είναι ίσος με **0.66**. Η τιμή αυτή δεν είναι και πολύ καλή για ταξινομητή ειδικά σε αυτή την περίπτωση που έχουμε ιατρικά δεδομένα. Το sensitivity και το specificity είναι 0,60 και 0,74 αντίστοιχα. Επίσης χρησιμοποιώντας την συνάρτηση kfoldloss() είδαμε ότι το ποσοστό λάθους είναι **0,30**.

Στην συνέχεια με χρήση της συνάρτησης fitcknn() και της crossval() του Matlab βρίσκουμε την τιμή του κ για την οποία βελτιστοποιείται ο ταξινομητής. Ο γεωμετρικός μέσος σε αυτή την

περίπτωση είναι ίσος περίπου με **0,53** αρκετά χαμηλή τιμή για ταξινομητή. Το sensitivity και το specificity είναι ίσα με 0,80 και 0,35. Με την χρήση της kfoldloss() βρήκαμε ότι η πιθανότητα λάθους του ταξινομητή είναι ίση με **0.33**.

Κατά την ρύθμιση των παραμέτρων και συγκεκριμένα του C (BoxConstraint) τα αποτελέσματα ήταν τα εξής:

- Γεωμετρικός μέσος: 0,62
- Sensitivity: 0,40
- Specificity: 0,97
- Πιθανότητα λάθους: 0,21

Παρόλο που η πιθανότητα λάθους είναι μικρή, ο γεωμετρικός μέσος είναι χαμηλότερος από τον αντίστοιχο γεωμετρικό μέσο αφού προσδιορίσουμε την παράμετρο γ.

Η κάθε μέθοδος δίνει διαφορετικά αποτελέσματα γιατί κάθε μια από αυτές χρησιμοποιεί διαφορετικό αλγόριθμο για να ταξινομήσει τα δεδομένα. Οι μηχανές διανυσματικής στήριξης (svm) έχουν ως σκοπό να επιτύχουν την μέγιστη απόσταση από το όριο απόφασης μέχρι το κοντινότερο σημείο κάθε κλάσης. Από την άλλη ο ταξινομητής κ-κοντινότερων γειτόνων για κάθε στοιχείο που θέλει να ταξινομήσει κοιτάει τα κ κοντινότερα στοιχεία (γείτονες) και το ταξινομεί στην κλάση στην οποία ανήκει η πλειοψηφία των γειτόνων.

Από τις δύο μεθόδους πιο αποδοτική κρίνουμε την **SVM** καθώς έχει μικρότερη πιθανότητα λάθους και μεγαλύτερο γεωμετρικό μέσο από τον Knn. Έτσι παρόλο που η μέθοδος ταξινόμησης Knn έχει μεγαλύτερο sensitivity(το οποίο προτιμάμε όταν έχουμε ιατρικά δεδομένα) θεωρούμε ότι ο γεωμετρικός μέσος και η πιθανότητα λάθους είναι πιο σημαντικά.

Ερώτημα 5ο

Θεωρούμε ως βέλτιστο ταξινομητή τον SVM. Χρησιμοποιούμε την συνάρτηση corr() για να υπολογίσουμε τους δείκτες συσχέτισης Pearson. Τα αποτελέσματα που παίρνουμε είναι τα εξής:

-0.1374

0.0824

-0.2202

-0.2460

-0.1849

-0.1634

-0.1519

0.0350

0.1614

0.1623

Κρατάμε τους 4 θετικούς δείκτες δηλαδή τις στήλες 2,8,9,10 και με αυτές επαναλαμβάνουμε την εκπαίδευση του svm ταξινομητή

βρίσκοντας τις νέες βέλτιστες τιμές για το C και το γ. Ο γεωμετρικός μέσος είναι 0,53 η πιθανότητα λάθους είναι 0,41 και τα sensitivity και specificity είναι 0,43 και 0,65 αντίστοιχα. Η απόδοση σε αυτόν τον ταξινομητή είναι ακόμα χαμηλότερη από αυτόν του ερωτήματος 4 το οποίο είναι λογικό καθώς παρόλο που στην εκπαίδευση λαμβάνουμε υπόψιν μας τα 4 σημαντικότερα ως προς την πρόβλεψη χαρακτηριστικά, οι συντελεστές συσχέτισής τους είναι αρκετά χαμηλοί (μέγιστη τιμή 1).

Ερώτημα 6ο

Χρησιμοποιήσαμε τον ταξινομητή svm του 4ου ερωτήματος με κόστη ίδια με το ερώτημα 4. Για τους άνδρες ο γεωμετρικός μέσος είναι 0,62 με sensitivity 0,40 και specificity 0,95 και πιθανότητα λάθους 0,24. Τα αποτελέσματα δεν είναι για ακόμη μια φορά ικανοποιητικά. Αντίστοιχα για τις γυναίκες ο γεωμετρικός μέσος είναι 0,48 με sensitivity και specificity 0,25 και 0,96 αντίστοιχα. Η πιθανότητα λάθους είναι 0,20.

Όσον αφορά τις γυναίκες, από τα δείγματα που έχουμε, βγάζουμε το συμπέρασμα ότι το 65% πάσχει από την ασθένεια ενώ όσον αφορά τους άντρες το 73% είναι ασθενής. Παρ ‘όλα αυτά, δεν μπορούμε να εξάγουμε ασφαλή συμπεράσματα για το ποιο από τα δυο φύλλα έχει μεγαλύτερη πιθανότητα να πάσχει από την ασθένεια καθώς δεν έχουμε εξαρχής επαρκή αριθμό δειγμάτων. Επίσης η διαφορά μεταξύ αρσενικών και θηλυκών δειγμάτων είναι πολύ μεγάλη (441 άνδρες και 142 γυναίκες) το

οποίο κάνει ακόμη πιο απίθανη την πιθανότητα τα συμπεράσματα που βγάζουμε να είναι ασφαλή.