Κωνσταντίνος Τσιάρας AM: 1115201500165

Αναγνώριση Προτύπων – Μηχανική Μάθηση

Προαιρετική Εργασία

Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν για την εργασία αυτή πάρθηκαν, όπως ζητείται από την εκφώνηση, από εδώ και συγκεκριμένα χρησιμοποιήθηκε το αρχείο '*breast-cancer-wisconsin.data*' το οποίο περιέχει 699 γραμμές δεδομένων τα οποία περιγράφονται στο '*breast-cancer-wisconsin.*names' και το οποίο απλά μετέτρεψα στο '*breast-cancer-wisconsin-data.xls*'. Το τελευταίο φορτώνεται στο MATLAB στο αρχείο '*loadCleanData.m*' μέσω *xlsread()*, όπου παραβλέπεται η στήλη "id" και επίσης γίνεται η αντικατάσταση των τιμών που λείπουν με την επικρατέστερη για το χαρακτηριστικό εκείνο. Τέλος, γίνεται κανονικοποίηση όλων των τιμών στο διάστημα [0, 1] με τη χρήση *normalize()*.

Ταξινομητές

Έχοντας φορτώσει τα δεδομένα όπως περιέγραψα παραπάνω, στο αρχείο '*ClassifierPerformance.*m' δημιουργώ τα ζητούμενα 10 splits μέσω της *crossvalind()* και στη συνέχεια για κάθε ταξινομητή πραγματοποιώ 10-fold cross validation, τυπώνοντας Accuracy, Sensitivity και Specificity για τον καθένα. Τα αποτελέσματα μιας ενδεικτικής εκτέλεσης παρατίθενται στον παρακάτω πίνακα (σε %):

Ταξινομητής	**KNN**			**Naive Bayes**		**SVM**		**Decision Tree**	
Παράμετρος Σχεδίασης	*NumNeighbors*			*DistributionNames*		*KernelFunction*		*PredictorSelection*	
	3	*5*	*7*	*normal*	*kernel*	*linear*	*rbf*	*allsplits*	*curvature*
Accuracy	96.85	97.14	97.00	95.99	96.57	96.70	96.85	93.99	94.85
Sensitivity	97.16	97.60	97.60	95.20	97.16	97.16	96.50	94.98	96.23
Specificity	96.27	96.27	95.85	97.51	95.85	95.85	97.51	92.12	92.12

Όπως έχει αναφερθεί και από συναδέλφους, ο ταξινομητής Multilayer Perceptron δεν είναι διαθέσιμος στο πανεπιστημιακό license του MATLAB και για τον λόγο αυτό δεν τον έχω χρησιμοποιήσει.

Δεδομένου ότι μελετάμε περιπτώσεις επικίνδυνης ασθένειας, μας ενδιαφέρει πολύ περισσότερο να ελαχιστοποιήσουμε τα false negatives – να μην έχουμε δηλαδή περιπτώσεις ατόμων που έχουν την ασθένεια και αποτύχαμε να την εντοπίσουμε – παρά το αντίθετο (false positives) – αφού αν λανθασμένα χαρακτηρίζαμε κάποιον ασθενή ενώ δεν ήταν, αυτό θα μπορούσε να ανακαλυφθεί με μεταγενέστερες εξετάσεις. Συνεπώς, **μας ενδιαφέρει να διαλέξουμε τον ταξινομητή με το υψηλότερο Sensitivity**.

Για το λόγο, με βάση τα παραπάνω αποτελέσματα για το πρόβλημα αυτό θα επέλεγα **KNN** και συγκεκριμένα με 5 γείτονες.