MELANOMA MACHINE LEARNING

Σαλτερης Γεωργιος 2136
Μουλόπουλος Αντώνιος 2104

Αποτελεσματα απο την εκπαίδευση διαφορων clasifiers που προβλεπουν αν καποιος εχει μελανομα η όχι. Οι classifier που χρησιμοποιουμε είναι οι Random Forest, Linear Regression,SVM,Linear Logistic Regression,OLS και GaussianNB.

Για το παρακατω πειραμα κρατήσαμε ολα τα snp με **pvalue <= p = 0.0001.**
Οι ασθενεις μας ειναι στο συνολο 4980. Απο αυτούς οι 3962 δεν εχουν την ασθενεια**(control = 0)** και οι υπόλοιποι 1018 εχουν την ασθενεια**(case = 1)**.

Τα snps με **pvalue <= p = 0.0001** είναι 5415 για να αποφύγουμε το **overfitting** μειώνουμε τα snps με τον εξής τρόπο, τον οποίο σκεφτήκαμε μόνοι μας. Αφού στήσουμε τον πίνακα με του ασθενείς και τα χαρακτηριστικά τους, στη συνέχεια βρίσκουμε τον πίνακα με το **cosine similarity** μεταξύ των snps χρησιμοποιώντας την εντολή **metrics.pairwise.cosine_similarity.** Ορίζουμε έναν counter για κάθε snp. Στη συνέχεια εξετάζουμε την ομοιότητα μεταξύ των snps και αυξάνουμε κατά ένα τον counter του snp που έχει τιμή στον πίνακα ίση η μικρότερη από 0.3(οσο πιο κοντά στο 1 είναι η τιμή τόσο πιο όμοια είναι μεταξύ τους τα χαρακτηριστικά.) κατα 1. Τέλος κρατήσαμε τα snps των οποίων ο counter είναι μεγαλύτερος ή ίσος από το 60% των συνολικών snps.

Καναμε ενα τυχαίο sample του παραπανω δειγματος όπου το 90% ειναι το trainData και το 10% ειναι το testData. Μετα το sample στο 90% του δειγματος οι 3572 ασθενεις δεν εχουνε την ασθενεια(και 910 ασθενεις εχουν την ασθενεια. Στο 10% του δειγματος οι 390 ασθενεις δεν εχουν την ασθενεια και οι 108 εχουν την ασθενεια.

	Cross validiation	**AUC**	**recall**
Random Forest	**0,84**	**0,64**	**0,28**
Linear Regression	**0,89**	**0,81**	**0,63**
SVM	**0,78**	**0,5**	**0,0**
SVM(kernel = linear)	**0,92**	**0,85**	**0,71**
Linear Logistic Regression	**0,92**	**0,87**	**0,75**
OLS	**0,91**	**0,82**	**0,73**
Naive Bayes Gaussian	**0,96**	**0,93**	**0,87**

Στη συνέχεια θα κανουμε ένα balanced του Data Train. Το αρχικό Data Train έχει 3572 ασθενεις που δεν εχουνε την ασθενεια και 910 ασθενεις που εχουν την ασθενεια. Θα μειώσουμε τους ασθενεις που δεν εχουν την ασθενεια στους 1300 με τυχαίο τροπο ώστε να υπάρχει μια ισορροπια μεταξυ των ασθενών που εχου την ασθενεία και αυτών που δεν την έχουν.

	Cross validiation	**AUC**	**recall**
Random Forest	**0,7**	**0,63**	**0,34**
Linear Regression	**0,81**	**0,77**	**0,61**
SVM	**0,65**	**0,6**	**0,3**
SVM(kernel = linear)	**0,88**	**0,88**	**0,80**
Linear Logistic Regression	**0,87**	**0,9**	**0,85**
OLS	**0,81**	**0,77**	**0,70**
Naive Bayes Gaussian	**0,94**	**0,94**	**0,92**