MELANOMA MACHINE LEARNING

Σαλτερής Γεώργιος 2136
Μουλόπουλος Αντώνιος 2104

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

Για το παρακατω πειραμα κρατήσαμε ολα τα snp με **pvalue <= 0.0001** από το αρχείο **assoc** που πήραμε απο το **plink.** Το assoc αρχείο είναι της μορφής:

CHR	**Chromosome code**
SNP	**Variant identifier**
BP	**Base-pair coordinate**
A1	**Allele 1 (usually minor)**
F_A	**Allele 1 frequency among cases**
F_U	**Allele 1 frequency among controls**
A2	**Allele 2**
CHISQ	**Allelic test chi-square statistic. *Not present with 'fisher'/'fisher-midp' modifier.***
P	**Allelic test p-value**
OR	**odds(allele 1 | case) / odds(allele 1 | control)**

Για να στήσουμε τον πινακα **ασθενείς-features,** όπου feature είναι ένα snp, κωδικοποιούμε το κάθε snp σε 0 ή 1 ή 2. Για να ξεκινήσουμε την κωδικοποίηση των snps όπως είπαμε και πριν πρέπει πρώτα να εξάγουμε το lgen αρχείο των ασθενών με τη χρήση του plink. Το lgen αρχείο είναι ξεχωριστό για κάθε χρωμόσωμα και περίεχει το snp του ασθενή με τα allele του. Είναι της μορφής:

1. **Family ID**
2. **Within-family ID**
3. **Variant identifier**
4. **Allele 1**
5. **Allele 2**

Το snp κάθε ασθένη παίρνει τον κωδικό 0 αν το allele1 **και** το allele2 απο το lgen αρχείο είναι διαφορετικο από το allele1(A1) του assoc αρχείου, παίρνει τον κωδικό 1 αν **μόνο** ένα από τα allele1 και allele2 απο το lgen αρχείο είναι ίδια με το allele1(A1) του assoc αρχείου και τέλος παίρνει τον κωδικό 2 αν το allele1 **και** το allele2 απο το lgen αρχείο είναι ίδια με το allele1(A1) του assoc αρχείου.

Τα snps με **pvalue <= 0.0001** είναι 5415 για να αποφύγουμε το **overfitting** μειώνουμε τα snps με τον εξής τρόπο, τον οποίο σκεφτήκαμε μόνοι μας. Για να επιλέξουμε τα snps που θα χρησιμοποιήσουμε για την εκπαίδευση, θα βρούμε το correlation μεταξύ των snps ανάλογα με την κωδικοποίηση που είπαμε και πιο πάνω. Το correlation μεταξύ ενός snps X και ενός snp Y ορίζεται ως εξής **COR(X,Y) = COV(X,Y) / VAR(X)*VAR(Y).**
Ισχύει

- $-1 \leq COR(X,Y) \leq 1$
- COR(X,Y) = COR(Y,X)
- COR(X,X) = 1

Αν είναι ίσο με 1 τότε τα snp δεν είναι ανεξάρτητα μεταξύ τους, δηλαδή η κωδικοποίηση αλλάζει προς την ίδια κατεύθυνση. Αν είναι κοντά στο 0 είναι ανεξάρτητα μεταξύ τους και αν είναι κοντά στο -1 η κωδικοποίηση αλλάζει προς την αντίθετη κατεύθυνση. Αρχικά βρίσκουμε snp που έχουνε πολύ μεγάλο correlation. Αυτό σημαίνει ότι βρίσκουμε snps που η κωδικοποίηση αλλάζει με ίδιο τρόπο όταν κάποιος έχει την ασθένεια ή όταν κάποιος δεν έχει την ασθένεια. Εμείς εκπαιδεύουμε τους classifiers με τα υπόλοιπα snps. Αυτό το κάνουμε επειδή τα snp με υψηλό correlation έχουν μεταβάλονται κατα τον ίδιο τρόπο ανάλογα με το αν έχουμε case ή control. Ετσι χρησιμοποιούμε τα υπόλοιπα snps που είναι ανεξαρτητα μεταξύ τους με βάση το correlation για να βγάλουμε την αποφασή μας

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)**.

Για κάθε classifier χωρίζουμε τα δεδομένα που έχουμε σε 10 σύνολα που είναι ξένα μεταξύ τους. Άρα τρέχουμε κάθε classifier δέκα φορές και κάθε φορά χρησιμοποιούμε τα 9 σύνολα ως δεδομένα εκπαίδευσης και το 1 σύνολο ως test για να δούμε την ικανότητα γενίκευση του classifier. Τα αποτελέσματα που φαίνονται στους παρακάτω πίνακες είναι ο μέσος όρος των αποτελεσμάτων που πήραμε από τις 10 φορές που τρέξαμε τον clasifier. Το Cross validation είναι το ποσοστό επιτυχιας της πρόβλεψης. Το recall υπολογίζεται από τον τύπο **true positives / (true positives + false negatives)** και το precision από τον τύπο **true positives / (true positives + false positives). True positives** είναι η πρόβλεψη του classifier για κάποιον ασθενή ότι έχει την ασθενεια και πράγματι την έχει. **False positives** είναι η πρόβλεψη του classifier για κάποιον ασθενή ότι έχει την ασθενεια και στην πραγματικότητα δεν την έχει. **False negatives** είναι η πρόβλεψη του classifier για κάποιον ασθενή ότι δεν έχει την ασθενεια και στην πραγματικότητα την έχει.

	ACCURACY	**AUC**	**RECALL**	**PRECISION**	**F_MEASURE**
Low Correlation	**0,96**	**0,97**	**0,99**	**0,87**	**0,92**
High Correlation	**0,93**	**0,86**	**0,75**	**0,91**	**0,82**

GNB

	ACCURACY	AUC	RECALL	PRECISION	F_MEASURE
Low Correlation	**0,96**	**0,91**	**0,83**	**0,98**	**0,90**
High Correlation	**0,87**	**0,78**	**0,63**	**0,72**	**0,67**

SVM(KERNEL)

	ACCURACY	AUC	RECALL	PRECISION	F_MEASURE
Low Correlation	**0,93**	**0,84**	**0,70**	**0,95**	**0,80**
High Correlation	**0,87**	**0,76**	**0,57**	**0,75**	**0,65**

Linear Logistic Regression

Στη συνέχεια με τη βοήθεια του plink κάναμε εξαγωγή ενος αρχείου assoc μεβάση τον κανόνα fisher. Κρατήσαμε **snps** με **pvalue <= 0.001** και στη συνέχεια επιλέξαμε χαρακτηριστικά με τον παραπάνω τρόπο που είπαμε.

	ACCURACY	AUC	RECALL	PRECISION	F_MEASURE
Low Correlation	**0,74**	**0,78**	**0,85**	**0,43**	**0,57**
High Correlation	**0,49**	**0,63**	**0,87**	**0,27**	**0,41**

GNB

	ACCURACY	AUC	RECALL	PRECISION	F_MEASURE
Low Correlation	**0,96**	**0,93**	**0,88**	**0,94**	**0,91**
High Correlation	**0,93**	**0,88**	**0,81**	**0,84**	**0,82**

SVM(KERNEL)

	ACCURACY	AUC	RECALL	PRECISION	F_MEASURE
Low Correlation	**0,96**	**0,93**	**0,87**	**0,96**	**0,92**
High Correlation	**0,93**	**0,88**	**0,79**	**0,87**	**0,83**

Linear Logistic Regression