Βιοπληροφορική και προσομοίωση Φυσιολογικών Συστημάτων 2021

Άσκηση 2

(Ατομική)

Στην παρούσα άσκηση καλείστε να δημιουργήστε μοντέλα μηχανικής μάθησης για την πρόβλεψη ασθένειας με τη χρήση δειγμάτων έκφρασης γονιδίων μεταξύ (Lung) Cancer & Normal. Για την υλοποίηση πρέπει να χρησιμοποιήσετε τα παρακάτω αρχεία:

- **Training set**: http://139.91.190.186/tei/bioinformatics/LungTrain.txt
- **Test set**: http://139.91.190.186/tei/bioinformatics/LungTest.txt
- **KEGG human pathways**: http://139.91.190.186/tei/bioinformatics/c2.cp.kegg.v7.4.entrez.gmt

Και να υλοποιήσετε τα ακόλουθα βήματα:

1. **Data pre-processing (10/100) - προ επεξεργασία των δεδομένων**:
    1. Διαβάστε τα δεδομένα εκπαίδευσης (Training set) γονιδιακής έκφρασης από το πείραμα με δείγματα από υγιή και καρκινικό ιστό
    2. Ετοιμάστε τα δεδομένα σας για να δημιουργήσετε μοντέλα μηχανικής μάθησης με τη χρήση της βιβλιοθήκης sklearn.
        - ***Hint****: Για την προετοιμασία θα πρέπει τα δεδομένα σας να είναι array οπου έχουμε τα γονίδια στις κολώνες και τα δείγματα στις γραμμές. Επίσης η κατηγορία των δειγμάτων (labels) θα είναι μια λίστα (προσέξτε η ονομασία των labels να είναι κοινή για όλα τα δείγματα Cancer και αντιστοίχως για τα Normal).*
    3. Ομοίως διαβάστε τα δεδομένα επαλήθευσης (test set).
2. **Data analysis (40/100) – Ανάλυση δεδομένων:**
    1. **Support vector machines model (10/100)**:
        - Εκπαιδεύστε ένα μοντέλο SVM με τα δεδομένα από το training set.
        - Τρέξτε το εκπαιδευμένο μοντέλο στα test set, τυπώστε την πρόβλεψη του μοντέλου για κάθε δείγμα στο test dataset, τυπώστε το confusion matrix και το accuracy.
    2. **Decision tree model (10/100)**:
        - Εκπαιδεύστε ένα μοντέλο decision tree με τα δεδομένα από το training set (δεν χρειάζεται να κάνετε k-fold), τυπώστε το accuracy και το δένδρο.
        - Τρέξτε το εκπαιδευμένο μοντέλο στα test set, τυπώστε την πρόβλεψη του μοντέλου για κάθε δείγμα στο test dataset και τυπώστε το confusion matrix.
        - Σχολιάστε τα δυο μοντέλα (Decision tree, SVM) και αιτιολογείστε ποιο από τα 2 είναι αποδοτικότερο (αν είναι κάποιο).
    3. **Gene selection (20/100) – επιλογή γονιδίων:**
        - Από το μοντέλο του SVM με τη χρήση της συνάρτηση coef_ μπορείτε να πάρετε το επίπεδο σημαντικότητας κάθε μεταβλητής (στην περίπτωση μας γονίδιου). Αν ο SVM classifier σας ονομάζεται clf τότε μπορείτε να πάρετε τα σημαντικότητα κάθε γονιδίου με την εντολή clf.coef_[0].
        - Επιλέξτε τα 20 γονίδια με το μέγιστο coef_ και τα 20 με το ελάχιστο.

- Δημιουργήστε ένα heatmap με τα 40 αυτά γονίδια από τις εκφράσεις γονιδίου αφου πρώτα τα έχετε κανει sorted σύμφωνα με το coef_ value.

  Hint: καντε και ένα .sort_index στις για να εχετε τα δειγματα από το Cancer και το Normal μαζι.

3. **Data annotation (30/100):**
   1. Με τη χρήση της βιβλιοθήκης mygene βρείτε την ονομασία των 40 επιλεγμένων γονιδίων από το προηγούμενο βήμα σε ονοματολογία entrez. Για την υλοποίηση χρησιμοποιήστε την συνάρτηση

      querymany(genes, scopes = 'reporter', fields='entrezgene', species='human')

      οπου *genes* η λίστα με τα ονόματα από τα 40 επιλεγμένα γονίδια.

   2. Διαβάστε το αρχείο KEGG human pathways που περιέχει λίστα με τα γονίδια ανα KEGG Pathway σε ονοματολογία entez. Κάθε γραμμή του αρχείου είναι ένα pathway, όπου στην πρώτη στήλη έχουμε το όνομα του pathway, στη δεύτερη ένα url και ακολουθούν ανά tab όλα τα γονίδια που μετέχουν στο συγκεκριμένο pathway.
   3. Βρείτε τα 5 pathways που περιέχουν τα περισσότερα γονίδια από τη λιστα των επιλεγμένων διακοσίων και τυπώστε το όνομα τους.

      Hint: Η συνάρτηση querymany θα επιστρεψει τα entrezids σαν string και η pandas θα διαβάσει τη λίστα με τα γονίδια ανα pathway σαν Int. Για να κανετε τη σύγκριση πρεπει πρωτα να εχετε τον ιδιο datatype. Μπορείτε τη λιστα με τα αποτελεσματα του mygene να την κανετε int με τη συνάρτηση int(x) (π.χ. προσπελάστε ολη τη λιστα και δημιουργείστε μια νεα τύπου Int)

4. **Pathway viewer (20/100):**
   1. Με τη βοήθεια της IPython.display.HTML και σε iframes οπτικοποιήστε τα 5 pathways από το προηγούμενο βήμα (που περιέχουν τα περισσότερα γονίδια από τη λιστα των επιλεγμένων διακοσίων).

      Hint: Για να βρείτε το pathway name e.g. hsa04010 χρησιμοποιήστε το google με το ονομα του pathway και τη λέξη KEGG. Βρειτε το χάρτη στη KEGG και από το url παρτε το pathway name (string after ?).

Όλα τα αρχεία είναι tab delimited files.

Τρόπος παράδοσης: Eclass

Ανεβάστε ενα αρχείο με όνομα τον αριθμό μητρώου που θα περιέχει το ipynb file <αριθμός_ μητρώου>_exercise2.ipynb (αν το σύστημα δεν σας επιτρέπει να ανεβάσετε το αρχείο λόγο κατάληξης, προσθέστε το .txt ή συμπιέστε το και ανεβάστε το).

Deadline: 18/05/2021 23:55