ΔΔΠΜΣ “Προηγμένα Συστήματα Υπολογιστών και Επικοινωνιών”, 2019-2020
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

2ο Σύνολο ασκήσεων για το μάθημα “Τεχνικές Μηχανικής Μάθησης”

Έκδοση 1.0

Εισαγωγικά

Οι παρακάτω ασκήσεις είναι προαιρετικές και λειτουργούν προσθετικά στη βαθμολογία του μαθήματος, μέχρι 1.5 βαθμό.

Παραδοτέα: Συνοπτική αναφορά με απαντήσεις στα ζητούμενα της εργασίας, καθώς και ο σχετικός κώδικας σε γλώσσα python (με επαρκή σχόλια παρακαλώ).

Άσκηση 1: Στην άσκηση αυτή θα χρησιμοποιήσετε το σύνολο δεδομένων “Cover type” που χρησιμοποιήσατε και στο πρώτο σύνολο ασκήσεων, με το ίδιο σύνολο εκπαίδευσης και δοκιμής (δηλ οι πρώτες 15120 εγγραφές είναι το σύνολο εκπαίδευσης και οι υπόλοιπες το σύνολο δοκιμής).

Ζητούνται τα παρακάτω:

1. Χρησιμοποιώντας μόνο το σύνολο εκπαίδευσης, εφαρμόστε διαδικασία διασταυρούμενης επικύρωσης k τμημάτων (k-fold cross-validation) ώστε να επιλέξετε το βέλτιστο συνδυασμό παραμέτρων για ταξινομητές SVM με πυρήνα ’RBF’. Χρησιμοποιήστε $k = 10$ και αναφέρετε τους συνδυασμούς που αξιολογήσατε, καθώς και την επίδοσή τους

2. Εφαρμόστε το βέλτιστο συνδυασμό που επιλέξατε στο σύνολο δοκιμής και αναφέρετε την ορθότητα (accuracy) που πετύχατε. Πως συγκρίνεται αυτή με την αντίστοιχη επίδοση των γραμμικών μοντέλων της πρώτης εργασίας, για το ίδιο σύνολο δεδομένων;

3. Πόσα δυαδικά μοντέλα περιέχει το μοντέλο που εκπαιδεύσατε και γιατί; Πόσα είναι τα διανύσματα υποστήριξης του συνολικού μοντέλου;

Άσκηση 2: Για το σύνολο δεδομένων “diabetes” που σας δίνεται έτοιμο από το python module `sklearn.datasets`, θεωρήστε τα πρώτα 300 δείγματα ως σύνολο εκπαίδευσης, και τα υπόλοιπα ως σύνολο δοκιμής.

1. Βρείτε το καλύτερο μοντέλο που μπορείτε με βάση το δεδομένο σύνολο εκπαίδευσης. Περιοριστείτε σε μοντέλα που συζητήσαμε στο μάθημα (γραμμικά μοντέλα, SVM, random forests). Εξηγήστε τη διαδικασία που εφαρμόσατε για τη διερεύνηση και τους λόγους που σας οδήγησαν στο μοντέλο που επιλέξατε.

2. Εξετάστε την επίδοση του μοντέλου σας στο σύνολο δοκιμής

3. Με ποια διαδικασία θα κατασκευάζατε ένα μοντέλο για χρήση σε άγνωστα δείγματα, και πως θα εκτιμούσατε την επίδοση του μοντέλου σας; Εφαρμόστε αυτή τη διαδικασία στο σύνολο δεδομένων "diabetes".

Άσκηση 3: Σ' αυτό το ερώτημα θα υλοποιήσετε και το στάδιο εξαγωγής χαρακτηριστικών. Σας δίνεται ένα σύνολο προτάσεων που έχουν εξαχθεί από κριτικές χρηστών του Amazon, του IMDB και του Yelp. Αυτές περιέχονται στα αρχεία:
`amazon_cells_labelled.txt`
`imdb_labelled.txt`
`yelp_labelled.txt`

Η μορφή των αρχείων περιγράφεται στο συνοδευτικό αρχείο `readme.txt`. Το dataset χρησιμοποιήθηκε στην εργασία 'From Group to Individual Labels using Deep Features', Kotzias et. al,. KDD 2015.

Στόχος μας είναι να κατασκευάσουμε ένα μοντέλο το οποίο δεδομένης μίας πρότασης, θα αποφασίζει αν αυτή είναι μέρος μίας θετικής ή μίας αρνητικής κριτικής. Για να το πετύχουμε αυτό, θα πρέπει αρχικά να αναπαραστήσουμε την κάθε πρόταση ως ένα διάνυσμα χαρακτηριστικών. Ζητούνται τα παρακάτω:

1. Συγκεντρώστε όλες τις λέξεις που υπάρχουν στο σύνολο δεδομένων και εφαρμόστε μορφολογική επεξεργασία (stemming) χρησιμοποιώντας τον Porter Stemmer που σας δίνεται έτοιμος από το πακέτο NLTK
   `nltk.stem.PorterStemmer`
   Επίσης αφαιρέστε τις λεγόμενες "Stop words" (πολύ συχνές λέξεις, σύνδεσμοι κλπ). Οι stop words για την Αγγλική γλώσσα δίνονται από το
   `nltk.corpus.stopwords`
   Τέλος αφαιρέστε και τα σημεία στίξης, και κατασκευάστε μία ταξινομημένη λίστα όπου η κάθε λέξη να εμφανίζεται μία φορά

2. Αντικαταστήστε την κάθε πρόταση με ένα διάνυσμα όπου το i-οστό στοιχείο είναι ο αριθμός εμφανίσεων της i-οστής λέξης της λίστας που κατασκευάσατε στο προηγούμενο ερώτημα. Πλέον έχετε ένα σύνολο δεδομένων όπου η κάθε πρόταση των αρχικών δεδομένων αναπαρίσταται με ένα διάνυσμα. Επίσης έχετε και μία τιμή στόχο για κάθε πρόταση (σας δίνεται από την αρχή στο σύνολο δεδομένων)

3. Κατασκευάστε ένα μοντέλο το οποίο θα ταξινομεί τις προτάσεις ως θετικές ή αρνητικές, χρησιμοποιώντας ταξινομητές SVM και αναφέρετε την εκτίμησή σας για την επίδοσή του