Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Τεχνητή Νοημοσύνη
Ακαδημαϊκό έτος: 2020–21
Διδάσκων: Ι. Ανδρουτσόπουλος

2η Εργασία

Υλοποιήστε σε Java ή C++ ή Python (ή άλλη γλώσσα που θα σας επιτρέψει ο υπεύθυνος των φροντιστηρίων) δύο ή τρεις (ανάλογα με το αν η ομάδα σας έχει δύο ή τρία μέλη) από τους ακόλουθους αλγορίθμους μάθησης, ώστε να μπορούν να χρησιμοποιηθούν για την κατάταξη κειμένων σε δύο (ξένες μεταξύ τους) κατηγορίες (π.χ. θετική/αρνητική γνώμη).

- **Αφελής ταξινομητής Bayes** (πολυμεταβλητή μορφή Bernoulli ή πολυωνυμική μορφή),
- **ID3** (προαιρετικά με πριόνισμα ή πρόωρο τερματισμό της επέκτασης κάθε δέντρου),
- **Random Forest** (προαιρετικά με πριόνισμα ή πρόωρο τερματισμό της επέκτασης κάθε δέντρου, αν και δεν συνηθίζεται),
- **AdaBoost** (με δέντρα απόφασης βάθους 1 ως βασικό ταξινομητή),
- **Logistic Regression** (με στοχαστική ανάβαση κλίσης, προσθέτοντας όρο κανονικοποίησης στην αντικειμενική συνάρτηση),

Κάθε κείμενο θα πρέπει να παριστάνεται από ένα διάνυσμα ιδιοτήτων με τιμές 0 ή 1, οι οποίες θα δείχνουν ποιες λέξεις ενός λεξιλογίου περιέχει το κείμενο. Το λεξιλόγιο θα πρέπει να περιλαμβάνει τις *m* συχνότερες λέξεις των δεδομένων εκπαίδευσης (ή ολόκληρου του συνόλου δεδομένων), ενδεχομένως παραλείποντας πρώτα τις *n* πιο συχνές λέξεις, όπου τα *m* και *n* θα είναι υπερ-παράμετροι. Προαιρετικά μπορείτε να προσθέσετε και επιλογή ιδιοτήτων μέσω υπολογισμού κέρδους πληροφορίας (ή μέσω άλλου τρόπου) στον αφελή ταξινομητή Bayes και στον Logistic Regression. Οι υπόλοιποι αλγόριθμοι ενσωματώνουν ήδη μεθόδους επιλογής ιδιοτήτων.

Επιδείξτε τις δυνατότητες μάθησης των υλοποιήσεών σας χρησιμοποιώντας το σύνολο δεδομένων «Large Movie Review Dataset», το οποίο είναι γνωστό και ως «IMDB dataset» (βλ. https://ai.stanford.edu/~amaas/data/sentiment/, https://keras.io/api/datasets/imdb/). Θα πρέπει να περιλάβετε στην αναφορά σας αποτελέσματα των πειραμάτων που θα εκτελέσετε με τις υλοποιήσεις σας σε αυτό το σύνολο δεδομένων, δείχνοντας (τουλάχιστον):

- **καμπύλες μάθησης και αντίστοιχους πίνακες** που να δείχνουν το ποσοστό **ορθότητας** (accuracy) στα **δεδομένα εκπαίδευσης** (training data, όσα έχουν χρησιμοποιηθεί κάθε φορά) και **ελέγχου** (test data) συναρτήσει του πλήθους των παραδειγμάτων εκπαίδευσης που χρησιμοποιούνται σε κάθε επανάληψη του πειράματος,
- αντίστοιχες καμπύλες και πίνακες με αποτελέσματα **ακρίβειας** (precision), **ανάκλησης** (recall), **F1** συναρτήσει του πλήθους των παραδειγμάτων εκπαίδευσης.

Θα πρέπει να αναφέρετε επίσης στην αναφορά σας τις **τιμές** των **υπερ-παραμέτρων** που χρησιμοποιήσατε (π.χ. τιμή *λ* του όρου κανονικοποίησης στον αλγόριθμο Logistic Regression, πλήθος δέντρων στον Random Forest) και **πώς τις επιλέξατε** (π.χ. με δοκιμές σε ξεχωριστά δεδομένα ανάπτυξης).

Δεν επιτρέπεται να χρησιμοποιήσετε έτοιμες υλοποιήσεις αλγορίθμων μηχανικής μάθησης. Μπορείτε, όμως, να συγκρίνετε προαιρετικά τις επιδόσεις των υλοποιήσεών σας με τις επιδόσεις άλλων διαθέσιμων υλοποιήσεων (π.χ. του Weka ή του Scikit-learn) των ίδιων ή άλλων αλγορίθμων (π.χ. υλοποιήσεις νευρωνικών δικτύων σε Keras) ή υλοποιήσεων άλλων ομάδων, κάτι που θα προσμετρηθεί θετικά στον βαθμό σας. Επιτρέπεται, επίσης, να χρησιμοποιήσετε έτοιμες βιβλιοθήκες για την κατασκευή διαγραμμάτων με καμπύλες. Περαιτέρω διευκρινίσεις θα δοθούν στα φροντιστήρια.

Η προθεσμία παράδοσης της εργασίας θα ανακοινωθεί στο e-class. **Διαβάστε προσεκτικά και το έγγραφο με τις γενικές οδηγίες των εργασιών του μαθήματος** (βλ. e-class). Αν οι κανόνες εκείνου του εγγράφου σας επιτρέπουν να υποβάλετε την εργασία ατομικά, αρκεί να υλοποιήσετε έναν από τους παραπάνω αλγορίθμους.