Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Τεχνητή Νοημοσύνη
Ακαδημαϊκό έτος: 2023–24
Διδάσκων: Ι. Ανδρουτσόπουλος

2η Εργασία

Μέρος Α' (60%): Υλοποιήστε σε Java ή C++ ή Python (ή άλλη γλώσσα που θα σας επιτρέψουν οι υπεύθυνοι 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).[1]

- **Αφελής ταξινομητής Bayes**, πολυμεταβλητή μορφή Bernoulli (όπως στις διαφάνειες της 16ης διάλεξης) ή πολυωνυμική μορφή (βλ. παραπομπές στο τέλος των διαφανειών της 16ης διάλεξης),
- **Random Forest** χρησιμοποιώντας τον ID3 ή παραλλαγή του (π.χ. που θα παράγει δέντρα τα οποία δεν θα υπερβαίνουν ένα μέγιστο βάθος, που θα δίνεται ως υπερ-παράμετρος) για την παραγωγή των δέντρων,
- **AdaBoost** με δέντρα απόφασης βάθους 1, δηλαδή κάθε «δέντρο» θα ρωτά την τιμή μόνο μίας ιδιότητας, εκείνης που οδηγεί στο μεγαλύτερο κέρδος πληροφορίας στα δεδομένα εκπαίδευσης του «δέντρου»,[2]
- **Logistic Regression** με στοχαστική ανάβαση κλίσης, προσθέτοντας ομαλοποίηση (regularization, βλ. διαφάνειες 18ης διάλεξης).

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, παραλείποντας πρώτα τις *n* πιο συχνές και τις *k* πιο σπάνιες λέξεις των δεδομένων εκπαίδευσης, όπου τα *m, n, k* 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)

[1] Αν οι γενικές οδηγίες για τις εργασίες του μαθήματος σας επιτρέπουν να παραδώσετε την εργασία ατομικά, μπορείτε να υλοποιήσετε μόνο έναν αλγόριθμο μάθησης.

[2] Στον AdaBoost, κατά τους υπολογισμούς πιθανοτήτων από τα παραδείγματα εκπαίδευσης, μπορείτε να θεωρείτε ότι ένα παράδειγμα με βάρος β εμφανίζεται β φορές στα παραδείγματα εκπαίδευσης (ακόμα και αν το β δεν είναι ακέραιος).

συναρτήσει του πλήθους των παραδειγμάτων εκπαίδευσης που χρησιμοποιούνται σε κάθε επανάληψη του πειράματος,
- αντίστοιχες καμπύλες και πίνακες με αποτελέσματα **ακρίβειας** (precision), **ανάκλησης** (recall), **F1** για μία από τις δύο κατηγορίες, συναρτήσει του πλήθους των παραδειγμάτων εκπαίδευσης.

Θα πρέπει να αναφέρετε στο έγγραφο της εργασίας σας τις **τιμές** των **υπερ-παραμέτρων** που χρησιμοποιήσατε (π.χ. τιμή *λ* του όρου ομαλοποίησης στον αλγόριθμο Logistic Regression, πλήθος δέντρων στον Random Forest) και **πώς τις επιλέξατε** (π.χ. με δοκιμές σε ξεχωριστά δεδομένα ανάπτυξης, development data). Δεν επιτρέπεται να χρησιμοποιήσετε έτοιμες υλοποιήσεις αλγορίθμων μηχανικής μάθησης σε αυτό το μέρος της εργασίας. Μπορείτε, όμως, να χρησιμοποιήσετε έτοιμες υλοποιήσεις προ-επεξεργασίας των κειμένων (π.χ. χωρισμού των κειμένων σε λέξεις) και επιλογής ιδιοτήτων (π.χ. κέρδος πληροφορίας). Επιτρέπεται, επίσης, να χρησιμοποιήσετε έτοιμες βιβλιοθήκες για την κατασκευή διαγραμμάτων με καμπύλες.

Μέρος Β' (20%): Συγκρίνετε τις επιδόσεις των υλοποιήσεών σας με τις επιδόσεις άλλων διαθέσιμων υλοποιήσεων (π.χ. του Weka ή του Scikit-learn) των ίδιων αλγορίθμων μάθησης που υλοποιήσατε στο μέρος Α' ή άλλων αλγορίθμων μάθησης (π.χ. υλοποιήσεις MLPs του Scikit-learn), κατασκευάζοντας τις ίδιες καμπύλες και πίνακες όπως στο Μέρος Α'. Θα πρέπει να συγκρίνετε με διαθέσιμες υλοποιήσεις τουλάχιστον δύο ή τριών αλγορίθμων μάθησης (ανάλογα με το αν η ομάδα σας έχει δύο ή τρία μέλη).[3] Θα πρέπει να χρησιμοποιήσετε τις ίδιες παραστάσεις των κειμένων όπως στο Μέρος Α' (διανύσματα ιδιοτήτων με τιμές 0 ή 1 που θα δείχνουν ποιες λέξεις του λεξιλογίου εμφανίζονται ή όχι στο κείμενο, με το ίδιο λεξιλόγιο όπως στο Μέρος Α'). Μπορείτε κι εδώ, να χρησιμοποιήσετε έτοιμες υλοποιήσεις προ-επεξεργασίας των κειμένων και επιλογής ιδιοτήτων, καθώς και έτοιμες βιβλιοθήκες για την κατασκευή διαγραμμάτων με καμπύλες.

Μέρος Γ' (20%): Συγκρίνετε τα αποτελέσματα των Μερών Α' και Β' με τα αποτελέσματα ενός MLP ή/και ενός RNN, που θα υλοποιήσετε σε Tensorflow/Keras, παριστάνοντας τις λέξεις με ενθέσεις λέξεων (word embeddings). Στην περίπτωση του MLP, μπορείτε να χρησιμοποιήσετε κεντροειδή ενθέσεων λέξεων (διάλεξη 21). Στην περίπτωση του RNN, μπορείτε να χρησιμοποιήσετε την τελευταία κατάσταση του RNN ως παράσταση του κειμένου ή να προσθέσετε έναν μηχανισμό αυτο-προσοχής (self-attention, διάλεξη 22). Κατασκευάστε κι εδώ τις ίδιες καμπύλες και πίνακες, όπως στα μέρη Α' και Β'. Κατασκευάστε επίσης καμπύλες που να δείχνουν τη μεταβολή στου σφάλματος (loss) στα παραδείγματα εκπαίδευσης και ανάπτυξης, συναρτήσει του αριθμού των εποχών.

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. έγγραφα του μαθήματος στο e-class).

[3] Αν οι γενικές οδηγίες για τις εργασίες του μαθήματος σας επιτρέπουν να παραδώσετε την εργασία ατομικά, μπορείτε να συγκρίνετε με διαθέσιμη υλοποίηση ενός μόνο αλγορίθμου μάθησης.