Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής
Μάθημα: Τεχνητή Νοημοσύνη
Ακαδημαϊκό έτος: 2024–25
Διδάσκων: Ι. Ανδρουτσόπουλος

2η Εργασία

Μέρος Α' (50%): 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),
- **Τυχαίο Δάσος (**Random Forest, 16η διάλεξη) χρησιμοποιώντας τον ID3 ή παραλλαγή του (π.χ. που θα παράγει δέντρα τα οποία δεν θα υπερβαίνουν ένα μέγιστο βάθος, το οποίο θα δίνεται ως υπερ-παράμετρος) για την παραγωγή των δέντρων,
- **AdaBoost** (17η διάλεξη) με δέντρα απόφασης βάθους 1 (decision stumps), δηλαδή κάθε «δέντρο» θα ρωτά την τιμή μόνο μίας ιδιότητας, εκείνης που οδηγεί στο μεγαλύτερο κέρδος πληροφορίας στα δεδομένα εκπαίδευσης του «δέντρου»,[2]
- **Λογιστική Παλινδρόμηση** (Logistic Regression) με στοχαστική ανάβαση κλίσης (stochastic gradient ascent), 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 κατασκευάζεται παραλείποντας πρώτα τις n πιο συχνές και τις k πιο σπάνιες λέξεις των κειμένων εκπαίδευσης, θεωρώντας ότι η συχνότητα μια λέξης ισούται με το πλήθος των κειμένων εκπαίδευσης στα οποία εμφανίζεται. Από τις λέξεις των δεδομένων εκπαίδευσης που θα απομένουν, θα πρέπει να επιλέγονται ως λέξεις του λεξιλογίου οι m λέξεις με το υψηλότερο πληροφοριακό κέρδος (βλ. διαφάνειες 15ης διάλεξης).[3]

Επιδείξτε τις δυνατότητες μάθησης των υλοποιήσεών σας εκτελώντας με αυτές πειράματα στο σύνολο δεδομένων «Large Movie Review Dataset», το οποίο είναι γνωστό και ως «IMDB dataset»[4]. Χρησιμοποιήστε ένα υποσύνολο

[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).
[3] Οι αλγόριθμοι Τυχαίου Δάσους και AdaBoost εκτελούν κατόπιν εσωτερικά και τη δική τους πρόσθετη επιλογή ιδιοτήτων (μεταξύ των m με το υψηλότερο πληροφοριακό κέρδος).
[4] Βλ. https://ai.stanford.edu/~amaas/data/sentiment/, https://keras.io/api/datasets/imdb/, https://pytorch.org/text/stable/datasets.html#imdb).

των δεδομένων εκπαίδευσης ως δεδομένα ανάπτυξης (development data). Θα πρέπει να περιλάβετε στο έγγραφο της εργασία σας:

- **καμπύλες μάθησης** (18η διάλεξη) που να δείχνουν αποτελέσματα **ακρίβειας** (precision), **ανάκλησης** (recall), **F1**, για **μία από τις δύο κατηγορίες** (όποια προτιμάτε), στα **δεδομένα εκπαίδευσης** (training data, όσα έχουν χρησιμοποιηθεί σε κάθε επανάληψη) και **ανάπτυξης** (development data, πάντα όλα τα δεδομένα ανάπτυξης) συναρτήσει του πλήθους των παραδειγμάτων εκπαίδευσης που χρησιμοποιούνται σε κάθε επανάληψη του πειράματος,
- **πίνακες** με αποτελέσματα **ακρίβειας** (precision), **ανάκλησης** (recall), **F1** για **κάθε μία από τις δύο κατηγορίες** και **μέσους όρους (micro- και macro-averaged**), στα **δεδομένα αξιολόγησης** (test data), όταν χρησιμοποιούνται όλα τα δεδομένα εκπαίδευσης.

Θα πρέπει να αναφέρετε στο έγγραφο της εργασίας σας τις τιμές των υπερ-παραμέτρων που χρησιμοποιήσατε (π.χ. κατώφλια συχνοτήτων λέξεων *k* και *n*, μέγεθος λεξιλογίου *m*, τιμή *λ* του όρου ομαλοποίησης στον αλγόριθμο Λογιστικής Παλινδρόμησης, πλήθος δέντρων στο Τυχαίο Δάσος) και πώς τις επιλέξατε (π.χ. με δοκιμές στα δεδομένα ανάπτυξης, χρήση προτεινόμενων τιμών της βιβλιογραφίας). Μπορείτε να χρησιμοποιήσετε την υλοποίηση του ID3 των φροντιστηρίων ή άλλη έτοιμη υλοποίηση του ID3 (π.χ. του Scikit-learn).[5] Δεν επιτρέπεται να χρησιμοποιήσετε έτοιμες υλοποιήσεις άλλων 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.[6]

Μέρος Β' (20%): Συγκρίνετε τις επιδόσεις των υλοποιήσεών σας με τις επιδόσεις άλλων διαθέσιμων υλοποιήσεων (π.χ. του Scikit-learn) των ίδιων αλγορίθμων μάθησης που υλοποιήσατε στο Μέρος Α' ή άλλων αλγορίθμων μάθησης (π.χ. MLP 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).[7] 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). Σε κάθε σύγκριση μεταξύ δικής σας υλοποίησης και άλλης διαθέσιμης υλοποίησης του *ιδίου* αλγορίθμου, το λεξιλόγιο και οι τιμές των υπερ-παραμέτρων θα πρέπει να είναι κατά το δυνατόν ίδια. Σε συγκρίσεις μεταξύ υλοποιήσεων *διαφορετικών* αλγορίθμων, δεν τίθεται τέτοιος περιορισμός αλλά θα πρέπει να εξηγείτε στο έγγραφό σας πώς επιλέξατε τις τιμές των υπερ-παραμέτρων των δύο αλγορίθμων. Μπορείτε κι εδώ, να χρησιμοποιήσετε έτοιμες υλοποιήσεις προ-επεξεργασίας κειμένων, επιλογής ιδιοτήτων και κατασκευής διαγραμμάτων με καμπύλες.

---

[5] Βλ. https://scikit-learn.org/.

[6] Βλ. π.χ. https://matplotlib.org/stable/tutorials/pyplot.html.

[7] 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.

Μέρος Γ' (30%): Συγκρίνετε τα αποτελέσματα των Μερών Α' και Β' με τα αποτελέσματα ενός στοιβαγμένου διπλής κατεύθυνσης RNN (stacked bidirectional RNN) με κελιά LSTM ή GRU και global max pooling (22η διάλεξη), που θα υλοποιήσετε σε PyTorch.[8] Χρησιμοποιήστε τον Adam optimizer ή άλλον, αντί της απλής στοχαστικής κατάβασης κλίσης.[9] Πρέπει να χρησιμοποιήσετε έτοιμες ενθέσεις λέξεων (word embeddings, διάλεξη 21).[10] Χρησιμοποιήστε τα δεδομένα ανάπτυξης (development) για να επιλέξετε την καλύτερη εποχή της εκπαίδευσης. Πρέπει να αναφέρετε στο έγγραφό σας τις τιμές των υπερ-παραμέτρων που χρησιμοποιήσατε (π.χ. πλήθος στοιβαγμένων επιπέδων του RNN) και πώς τις επιλέξατε (π.χ. με δοκιμές στα δεδομένα ανάπτυξης). Θα πρέπει να περιλάβετε, επίσης, στο έγγραφό σας:

- **καμπύλες που να δείχνουν το σφάλμα** (loss) στα παραδείγματα **εκπαίδευσης** (πάντα όλα τα δεδομένα εκπαίδευσης) και **ανάπτυξης** (πάντα όλα τα δεδομένα ανάπτυξης), **συναρτήσει του αριθμού των εποχών** ή του αριθμού βημάτων ενημέρωσης βαρών (19η διάλεξη).
- πίνακες με αποτελέσματα **ακρίβειας** (precision), **ανάκλησης** (recall), **F1** για **κάθε μία από τις δύο κατηγορίες** και **μέσους όρους (micro- και macro-averaged**), στα **δεδομένα αξιολόγησης** (test data), όταν χρησιμοποιούνται όλα τα δεδομένα εκπαίδευσης.

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).

[8] Βλ. https://pytorch.org/. Θα καλυφθεί και στα φροντιστήρια του μαθήματος.
[9] Βλ. https://pytorch.org/docs/stable/optim.html και https://aclanthology.org/2024.eacl-long.157/.
[10] Βλ. π.χ. https://radimrehurek.com/gensim/models/word2vec.html.