

2^η Εργασία

Μέρος Α (30%): Χρησιμοποιήστε έτοιμες υλοποιήσεις (π.χ. από τα φροντιστήρια ή/και από το Scikit-learn¹), δύο ή τριών (ανάλογα με το αν η ομάδα σας έχει δύο ή τρία μέλη) από τους παρακάτω αλγορίθμους μηχανικής μάθησης, για να αναπτύξετε ένα σύστημα που θα κατατάσσει κριτικές ταινιών του «Large Movie Review Dataset» (γνωστού και ως «IMDB dataset»²) σε θετικές ή αρνητικές.³

- **Αφελής ταξινομητής Bayes**, πολυμεταβλητή μορφή Bernoulli (όπως στις διαφάνειες της 8^{ης} διάλεξης) ή πολυωνυμική μορφή (βλ. παραπομπές στο τέλος των διαφανειών της 8^{ης} διάλεξης),
- **Τυχαίο Δάσος** (Random Forest, 9η 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** (10^η 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»,⁴
- **Λογιστική Παλινδρόμηση** (Logistic Regression) με στοχαστική ανάβαση κλίσης (stochastic gradient ascent), με ομαλοποίηση (regularization) L1 ή L2 (βλ. διαφάνειες 11^{ης} διάλεξης).

Κάθε κείμενο κριτικής θα πρέπει να παριστάνεται ως 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 7^{ης} διάλεξης).⁵ Το λεξιλόγιο θα πρέπει να είναι το ίδιο για όλους τους αλγορίθμους μάθησης. Χρησιμοποιήστε ένα υποσύνολο των δεδομένων εκπαίδευσης ως δεδομένα ανάπτυξης (development data). Μπορείτε να χρησιμοποιήσετε δικές σας ή έτοιμες υλοποιήσεις προ-επεξεργασίας των

¹ Βλ. <https://scikit-learn.org/>. Μπορείτε να χρησιμοποιήσετε και δικές σας υλοποιήσεις αλλά δεν θα λάβετε πρόσθετες μονάδες για αυτό.

² Βλ. <https://ai.stanford.edu/~amaas/data/sentiment/>, <https://pytorch.org/text/stable/datasets.html#imdb>.

³ Αν οι γενικές οδηγίες για τις εργασίες του μαθήματος σας επιτρέπουν να παραδώσετε την εργασία ατομικά, μπορείτε να χρησιμοποιήσετε μόνο έναν αλγόριθμο μάθησης.

⁴ 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).

⁵ Οι αλγόριθμοι Τυχαίου Δάσους και AdaBoost εκτελούν κατόπιν εσωτερικά και τη δική τους πρόσθετη επιλογή ιδιοτήτων (μεταξύ των m με το υψηλότερο πληροφοριακό κέρδος).

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

Θα πρέπει να περιλάβετε στο έγγραφο της εργασίας σας:

- **καμπύλες μάθησης** που να δείχνουν αποτελέσματα **ακρίβειας** (precision), **ανάκλησης** (recall), **F1** (9^η διάλεξη), 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 (παρόμοια διαγράμματα με τα αντίστοιχα της 11^{ης} διάλεξης),
- **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).

Μέρος Β (35%): Συγκρίνετε τα αποτελέσματα του Μέρους Α με τα αποτελέσματα ενός στοιβαγμένου διπλής κατεύθυνσης RNN (stacked bidirectional RNN) με κελιά LSTM ή GRU και global max pooling ή self-attention MLP (15^η διάλεξη), που θα υλοποιήσετε σε PyTorch.⁷ Χρησιμοποιήστε τον Adam optimizer ή άλλον, αντί της απλής στοχαστικής κατάβασης κλίσης.⁸ Πρέπει να χρησιμοποιήσετε έτοιμες ενθέσεις λέξεων (word embeddings, 14^η διάλεξη).⁹ 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 (12^η διάλεξη).
- **πίνακες** με αποτελέσματα **ακρίβειας** (precision), **ανάκλησης** (recall), **F1** για **κάθε μία από τις δύο κατηγορίες** και **μέσους όρους (micro-**

⁶ Βλ. π.χ. <https://matplotlib.org/stable/tutorials/pyplot.html>.

⁷ Βλ. <https://pytorch.org/>. Θα καλυφθεί και στα φροντιστήρια του μαθήματος.

⁸ Βλ. <https://pytorch.org/docs/stable/optim.html> και <https://aclanthology.org/2024.eacl-long.157/>.

⁹ Βλ. π.χ. <https://radimrehurek.com/gensim/models/word2vec.html>.

και macro-averaged), στα **δεδομένα αξιολόγησης** (test data), όταν χρησιμοποιούνται όλα τα δεδομένα εκπαίδευσης.

Μέρος Γ (35%): Κατασκευάστε έναν ταξινομητή εικόνων, βασισμένο σε κωδικοποιητή CNN συνδεδεμένο με MLP (16^η διάλεξη), που θα υλοποιήσετε σε PyTorch, για το σύνολο δεδομένων FashionMNIST.¹⁰ Μπορείτε να χρησιμοποιήσετε έναν έτοιμο προ-εκπαιδευμένο (π.χ. στο ImageNet) κωδικοποιητή CNN¹¹, τον οποίο θα εκπαιδεύσετε περαιτέρω (fine-tuning) στα δεδομένα εκπαίδευσης του FashionMNIST, εκμεταλλευόμενοι και τεχνικές επαύξησης δεδομένων (16^η διάλεξη). Χρησιμοποιήστε τον Adam optimizer ή άλλον, αντί της απλής στοχαστικής κατάβασης κλίσης. Κρατήστε ένα υποσύνολο των δεδομένων εκπαίδευσης ως δεδομένα ανάπτυξης (development data). Χρησιμοποιήστε τα δεδομένα ανάπτυξης για να επιλέξετε την καλύτερη εποχή της εκπαίδευσης. Πρέπει να αναφέρετε στο έγγραφό σας τις τιμές των υπερ-παραμέτρων που χρησιμοποιήσατε (π.χ. πλήθος κρυφών επιπέδων του MLP) 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), **συναρτήσεως του αριθμού των εποχών** ή του αριθμού βημάτων ενημέρωσης βαρών (12^η διάλεξη).
- πίνακες με αποτελέσματα **ακρίβειας** (precision), **ανάκλησης** (recall), **F1** για **κάθε μία από τις δέκα 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. έγγραφο του μαθήματος στο e-class).

¹⁰ Βλ. <https://arxiv.org/abs/1708.07747>, <https://github.com/zalandoresearch/fashion-mnist>, <https://docs.pytorch.org/vision/stable/generated/torchvision.datasets.FashionMNIST.html>.

¹¹ Βλ. π.χ. <https://docs.pytorch.org/vision/main/models.html>.