ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ
ΠΡΟΗΓΜΕΝΑ ΘΕΜΑΤΑ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ
Ακ. Έτος 2019-2020

ΣΕΠΤΕΜΒΡΙΟΣ 2020
ΑΠΑΛΛΑΚΤΙΚΗ ΕΡΓΑΣΙΑ ΜΑΘΗΜΑΤΟΣ
(ΠΑΡΑΔΟΣΗ: ΗΜΕΡΟΜΗΝΙΑ ΕΞΕΤΑΣΗΣ ΜΑΘΗΜΑΤΟΣ)

Επιλέξτε ένα από τα παρακάτω θέματα εργασιών

Θέμα 1. Κατηγοριοποίηση εικόνων

Στα πλαίσια αυτής της εργασίας ζητείται να εκπαιδεύσετε ένα νευρωνικό δίκτυο ώστε να κατηγοριοπιεί εικόνες σε μία από τις 10 διαθέσιμες κατηγορίες ρούχων που έχουν οριστεί. Μετά την εκπαίδευση, θα πρέπει να αξιολογήσετε την αποδοτικότητα του νευρωνικού δικτύου (NN) σε σχέση με ένα σύνολο από μετρικές όπως *ακρίβεια κατηγοριοποίησης, χρόνος εκπαίδευσης και τρεξίματος δικτύου*. Στη συνέχεια θα υλοποιήσετε μία εφαρμογή (mobile or web-based) η οποία θα δέχεται μία εικόνα σαν είσοδο και θα επιστρέφει α) την κατηγορία στην οποία ανήκει καθώς και πληροφορίες για την συγκεκριμένη κατηγορία, β) τις top-k εικόνες της κατηγορίας που έχουν επιλέξει οι πελάτες.

Α) Δεδομένα
Χρησιμοποιήστε το Fashion MNIST dataset. Μπορείτε να το κατεβάσετε από
https://github.com/zalandoresearch/fashion-mnist, click folder ”data”, then click folder ”fashion” και κατεβάστε τα δεδομένα.
Αυτό το σύνολο δεδομένων περιέχει εικόνες διαφόρων ενδυμάτων και αξεσουάρ, όπως πουκάμισα, τσάντες, παπούτσια και άλλα είδη μόδας. Υπάρχουν 10 ετικέτες.
Το σύνολο δεδομένων αποτελείται από ένα σύνολο δεδομένων εκπαίδευσης περίπου 60,000 (εικόνες) και ένα σύνολο δεδομένων ελέγχου 10,000 παραδειγμάτων. Κάθε παράδειγμα είναι μια εικόνα σε κλίμακα του γκρι 28x28, που σχετίζεται με μια ετικέτα από 10 πιθανές κατηγορίες ενδυμάτων (πουλόβερ, παντελόνι, κλπ).
Για περισσότερες πληροφορίες: https://www.kaggle.com/zalando-research/fashionmnist .

Β) Αξιολόγηση νευρωνικού δικτύου
Β1. Εκπαιδεύστε και αξιολογήστε την αποδοτικότητα των NNs με διαφορετικές παραμέτρους εισόδου:

- Διαφορετικός αριθμός από hidden layers, (θα χρησιμοποιήσετε τον ίδιο αριθμό από κόμβους σε κάθε επίπεδο).
- .Διαφορετικός αριθμός από κόμβους
- Διαφορετικό αριθμό από παραδείγματα εκπαίδευσης
- Διαφορετικό αριθμό από επαναλήψης για τον gradient-descent algorithm

Β2. Θα αξιολογήσουμε την απόδοση των εκπαιδευμένων NNs σε σχέση με :

- *Ακρίβεια ταξινόμησης* (Αξιολογείται στο σύνολο δεδομένων ελέγχου). Την ακρίβεια την μετράμε ως το ποσοστό των σωστά ταξινομημένων εικόνων. Πάρτε τυχαία 1000 εικόνες από το σύνολο δεδομένων ελέγχου, τις τροφοδοτείτε ως είσοδο στο NN και καταγράφετε το ποσοστό των εικόνων που ταξινομήθηκαν σωστά.
- *Χρόνος που χρειάζεται για την εκπαίδευση του NN*, που αναφέρεται ως NN Time Training. Αυτός είναι ο χρόνος από τη στιγμή που ξεκινάμε την εκπαίδευση μέχρι ο αλγόριθμος gradient descent να τελειώσει όλες τις επαναλήψεις.
- *Χρόνος που χρειάζεται για να εκτελέσετε το NN* και να λάβετε αποτελέσματα ταξινόμησης σε νέες εικόνες (από το σύνολο ελέγχου δεδομένων). Αυτός είναι ο χρόνος που απαιτείται για να ταξινομηθεί μια εικόνα, αφού τροφοδοτήσουμε αυτήν την εικόνα ως είσοδο στο NN. Ο χρόνος εκτέλεσης NN μετράτε λαμβάνοντας το μέσο όρο σε 1000 τυχαίες εικόνες από το σύνολο δεδομένων.

Γ) Ανάπτυξη εφαρμογής

Καλείστε να αναπτύξετε μία απλή εφαρμογή όπου θα ενσωματώσετε το εκπαιδευμένο NN προκειμένου να κατηγοριοποιεί νέες εικόνες και να δίνει κάποιες σχετικές πληροφορίες.
Για το σκοπό αυτό μπορείτε σε κάθε κατηγορία εικόνων να αντιστοιχίσετε κάποιο κείμενο. Για παράδειγμα “This Label costs 100 euros” or “This Label is made of silk” etc.
Επίσης μπορείτε να δώσετε κάποια βαθμολογία στις εικόνες που έχετε στο σύνολο εκπαίδευσης και όταν ένα χρήστης δώσει μία εικόνα να του επιστρέφει την κατηγορία της εικόνας και τις k (παράμετρος εισόδου από χρήστη) με μεγαλύτερο βαθμό εικόνες αυτής της κατηγορίας.

Θέμα 2. Αναγνώριση συναισθήματος από κείμενο

Ο στόχος της εργασίας είναι η ανάπτυξη ενός συστήματος το οποίο θα αναγνωρίζει το συναίσθημα του χρήστη με βάση την κριτική που δίνεται για μία ταινία. Στα πλαίσια της εργασίας αυτή απαιτείται η εκπαίδευση ενός μοντέλου ανάλυσης συναισθήματος (sentiment analysis) σε κριτικές που δίνονται σε ταινίες. Μετά την εκπαίδευση, θα πρέπει να αξιολογήσετε την αποδοτικότητα του μοντέλου σας σε σχέση με ένα σύνολο από μετρικές όπως *ακρίβεια κατηγοριοποίησης, χρόνος εκπαίδευσης και κατηγοριοποίησης*. Στη συνέχεια θα υλοποιήσετε μία εφαρμογή η οποία θα δέχεται σαν είσοδο ένα κείμενο κριτικής και θα επιστρέφει την κατηγορία στην οποία ανήκει (θετική ή αρνητική.

Α) Δεδομένα

Χρησιμοποιήστε το σύνολο δεδομένων IMDB review dataset. Μπορείτε να το κατεβάσετε από:
https://www.kaggle.com/utathya/imdb-review-dataset
επίσης μπορείτε να δείτε πληροφορίες και να κατεβάσετε δεδομένα από
http://ai.stanford.edu/~amaas/data/sentiment/

Το σύνολο δεδομένων περιέχει κριτικές ταινιών και την αντίστοιχη κατηγοριοποίηση τους σε θετικές/αρνητικές.

Β) Εκπαίδευση μοντέλου

Μπορείτε να επιλέξετε την εκπαίδευση μοντέλου κατηγοριοποίησης των κειμένων κριτικής με βάση είτε SVM ή νευρωνικό δίκτυο.

Η είσοδος στο αλγόριθμο εκπαίδευσης θα είναι η αναπαράσταση των κειμένων σας σε μορφή διανύσματος. Θα πρέπει επομένως να χρησιμοποιήσετε ένα μοντέλο αναπαράστασης κειμένου όπως, TFIDF, word embedding (https://www.tensorflow.org/tutorials/representation/word2vec).

Γ) Αξιολόγηση

Θα αξιολογήσουμε την απόδοση του εκπαιδευμένου μοντέλου σε σχέση με :

- *Ακρίβεια ταξινόμησης* (Αξιολογείται στο σύνολο δεδομένων ελέγχου). Την ακρίβεια την μετράμε ως το ποσοστό των σωστά ταξινομημένων κριτικών. Πάρτε τυχαία 1000 κείμενα από το σύνολο δεδομένων ελέγχου τα δίνεται ως είσοδο στο SVM ή NN μοντέλο και καταγράφετε το ποσοστό των κειμένων που ταξινομήθηκαν σωστά.
- *Χρόνος που χρειάζεται για την εκπαίδευση του μοντέλου*. Αυτή είναι η ώρα από τη στιγμή που ξεκινάμε την εκπαίδευση μέχρι να τελειώσει η διαδικασία εκπαίδευσης.
- *Χρόνος που χρειάζεται για να εκτελέσετε το NN*/SVM και να λάβετε αποτελέσματα ταξινόμησης σε νέα κείμενα (από το σύνολο ελέγχου δεδομένων). Αυτός είναι ο χρόνος που απαιτείται για να ταξινομηθεί ένα κείμενο, αφού τροφοδοτήσουμε το κείμενο ως είσοδο στο μοντέλο. Ο χρόνος εκτέλεσης μπορεί να εκτιμηθεί λαμβάνοντας το μέσο όρο σε 1000 τυχαία κείμενα από το σύνολο δεδομένων.

Δ) Ανάπτυξη εφαρμογής

- Καλείστε να αναπτύξετε μία εφαρμογή όπου θα ενσωματώσετε το εκπαιδευμένο NN ή SVM μοντέλο προκειμένου να κατηγοριοποιεί νέες κριτικές. Η εφαρμογή θα δίνει τη δυνατότητα σε κάποιον να γράφει την κριτική του και το σύστημα θα αναγνωρίζει και θα καταχωρεί στο σύστημα το συναίσθημα του χρήστη (θετικό, αρνητικό).

Παρατηρήσεις

1. Η εργασία είναι ατομική
2. Δεν υπάρχει περιορισμός σε γλώσσα υλοποίησης.

ΠΑΡΑΔΟΤΕΑ

Η εργασία θα υποβληθεί μέσω e-class (http://evdoxos.ds.unipi.gr/).

Μέχρι την ημερομηνία εξέτασης μαθήματος. Θα πρέπει να παραδώσετε ένα αρχείο ΑΜ.zip (ΑΜ είναι ο αριθμός μητρώου σας) το οποίο θα περιλαμβάνει:

- τον πηγαίο κώδικα και
- το κείμενο της εργασίας σε μορφή pdf. Παρουσίαση όλων των βημάτων της εργασίας και των αποτελεσμάτων
- Μία παρουσία (power point or pdf) 15 λεπτών στην οποία θα παρουσιάζεται τα βασικά στοιχεία της εργασίας σας.

Ανεβάστε το .zip αρχείο στην περιοχή «Εργασίες» στον evdoxo.