2/4/2024

Προπτυχιακό μάθημα: **Μηχανική Μάθηση**
Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων,
Ακαδημαϊκό έτος 2023-24

1η Σειρά Ασκήσεων
Ημερομηνία παράδοσης : έως 14/5/2024

Θέμα: Ταξινόμηση δεδομένων (*Data Classification*)

Χρησιμοποιήστε το σύνολο δεδομένων **fashion MNIST** (https://www.kaggle.com/datasets/zalando-research/fashionmnist) το οποίο αποτελείται από δεδομένα από δέκα (10) είδη ενδυμάτων (κλάσεις). Στην αρχική τους μορφή τα δεδομένα αυτά είναι ασπρόμαυρες εικόνες 28x28 pixels, αλλά μπορούν να αναπαρασταθούν σε διανυσματική μορφή μεγέθους 784 (28*28) χαρακτηριστικών (διανύσματα δυαδικών τιμών). Συνολικά υπάρχουν 60,000 δεδομένα για εκπαίδευση (*training set*) και άλλα 10,000 για έλεγχο (*testing set*).[1]

Να κανονικοποιήσετε τα αριθμητικά δεδομένα αρχικά και στη συνέχεια να μελετήσετε και να συγκρίνετε τα παρακάτω μοντέλα ταξινόμησης:

[A]. (διανυσματική αναπαράσταση)

- Χρησιμοποιώντας την τεχνική **max polling** με παράθυρο διάστασης 4x4 μειώστε αρχικά την διάσταση των εικόνων σε όλα τα δεδομένα, και στη συνέχεια χρησιμοποιήστε την διανυσματική τους μορφή (διανύσματα διάστασης 49).
- Κατασκευάστε τον ταξινομητή των κοντινότερων γειτόνων (***Nearest Neighbor Classifier***) χρησιμοποιώντας K=1 ή 3 ή 5 γείτονες και Ευκλείδια απόσταση.
- Κατασκευάστε ένα ***Decision Tree*** (δέντρο απόφασης) με *maximum depth* ίσο με 10 (να κάνετε plot το παραγόμενο δέντρο με τους κανόνες) και ένα ***Random Forest*** με 100 estimators.
- Κατασκευάστε ταξινομητές SVMs χρησιμοποιώντας 3 διαφορετικές τιμές της παραμέτρου C {1, 10, 100} – (3 διαφορετικά μοντέλα για κάθε τύπο) και μέγιστο αριθμό 500 επαναλήψεων:
    - ένα ***linear SVM classifier***,

[1] Σε περίπτωση που αντιμετωπίσετε δυσκολίες στην διαδικασία μάθησης λόγω υψηλής πολυπλοκότητας και ανεπάρκειας σε υπολογιστικούς πόρους, χρησιμοποιήστε ένα υποσύνολο του training set παίρνοντας ένα ποσοστό των δεδομένων ανά κατηγορία. Καθώς υπάρχουν 6,000 δεδομένα εκπαίδευσης σε κάθε κατηγορία, επιλέξτε τυχαία ένα ποσοστό ρ<1 αυτών (π.χ. 0.8 ή 0.5).

- ένα ***SVM με RBF kernel*** ρυθμίζοντας την τιμή της παραμέτρου της συνάρτησης ίση με 0.02, 0.1 ή 1 (βρείτε την καλύτερη τιμή της).

- Κατασκευάστε ένα νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης (***feed-forward neural network***) που να περιλαμβάνει **3 κρυμμένα επίπεδα** με πλήθος νευρώνων 100, 100 και 50, αντίστοιχα, και ένα επίπεδο εξόδου στο τέλος με πλήθος νευρώνων όσες και οι κατηγορίες χρησιμοποιώντας την συνάρτηση ενεργοποίησης softmax. Στους νευρώνες των κρυμμένων επιπέδων χρησιμοποιήστε την συνάρτηση ενεργοποίησης *Leaky ReLu*, ενώ για την εκπαίδευση του δικτύου τον *Adam optimizer*, το τετραγωνικό σφάλμα ως συνάρτηση απώλειας και το accuracy ως μέτρο υπολογισμού. Ποιο είναι το πλήθος παραμέτρων του δικτύου αυτού (μπορείτε να το βρίσκετε αυτόματα καλώντας μια έτοιμη συνάρτηση); Εκπαιδεύστε το δίκτυο για 100 εποχές με *batch size* 50 και κάντε ένα *plot* της τιμής της συνάρτησης απώλειας *loss function* και του *accuracy* (και για τα δύο σύνολα *training* και *testing*).

[Β]. (δεδομένα με την μορφή εικόνας)

- Κατασκευάστε **συνελικτικά νευρωνικά δίκτυα** (Convolutional Neural Networks - CNNs) με τρία διαδοχικά 2D συνελικτικά επίπεδα με 3x3 kernels. Μετά το τέλος των δύο πρώτων επιπέδων θα υπάρχει ένα επίπεδο με *max polling* χρησιμοποιώντας παράθυρο 2x2. Στη συνέχεια χρησιμοποιήστε ένα κρυμμένο επίπεδο (*fully connected layer*) 100 νευρώνων και τον μηχανισμό dropout με τιμή α=0.3. Τέλος χρησιμοποιήστε ένα επίπεδο εξόδου με 10 (softmax) νευρώνες όσες και οι κατηγορίες. Ως συνάρτηση απώλειας δοκιμάστε την *categorical-cross-entropy*, και ακολουθήστε την προηγούμενη στρατηγική μάθησης (100 εποχές, batch size 50). Δοκιμάστε **2 διαφορετικές τιμές** για το πλήθος των φίλτρων ανά επίπεδο και σχολιάστε πως ο αριθμός επιδρά στην επίδοση του δικτύου πάνω στα δεδομένα του συνόλου testing set. Όπως και σε όλες τις παραπάνω περιπτώσεις, κάντε ένα *plot* της τιμής της συνάρτησης απώλειας *loss function* και του *accuracy* (και για τα δύο σύνολα *training* και *testing*).

Για κάθε από τα παραπάνω μοντέλα να σχολιάσετε τόσο την επίδοσή τους, όσο και την συμπεριφορά τους. Χρησιμοποιήστε το περιβάλλον Jupyter Notebook για την παρουσίαση και εκτέλεση του **report** δίνοντας παράλληλα **ικανοποιητικά σχόλια** πάνω στον κώδικα. Δεν θα βαθμολογηθεί κώδικας χωρίς ικανοποιητικά σχόλια. Το παραδοτέο θα είναι αρχείο pdf που μπορεί να παραχθεί απ' ευθείας από το Notebook, όπως επίσης τον πηγαίο κώδικα (φτιάξτε ένα αρχείο zip που θα περιλαμβάνει τα 2 αυτά αρχεία). Επίσης μην ξεχάσετε να γράψετε το όνομα των μελών της ομάδας (σε σχόλιο) στην αρχή του κειμένου. Ανώνυμο report δεν θα ληφθεί υπόψιν.