Προπτυχιακό μάθημα: **Μηχανική Μάθηση,**
Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων,
(ακαδ. έτος 2020-21)

1η Σειρά Ασκήσεων
(Ημερομηνία παράδοσης : έως Τρίτη 20/4/2021)

23/3/2020

Θέμα: Μέθοδοι ταξινόμησης δεδομένων

Classify images of clothing

Μία σημαντική πειραματική βάση δεδομένων είναι η **fashion MNIST** που περιέχει από ασπρόμαυρες εικόνες χαμηλής ευκρίνειας (διάστασης 28 x 28) με ενδυματολογικό περιεχόμενο. Αποτελείται από **60,000 εικόνες 10 κατηγοριών** ρούχων (με απόλυτη συμμετρία ως προς τις κατηγορίες) που χρησιμοποιούνται για εκπαίδευση και ακόμα **10,000 εικόνες για τεστ**. Περισσότερες λεπτομέρειες για το σύνολο **fashion MNIST** μπορείτε να βρείτε στον επόμενο σύνδεσμο:

https://www.tensorflow.org/tutorials/keras/classification

απ' όπου μπορείτε να κατεβάσετε το πειραματικό σύνολο δεδομένων.

Στόχος της εργασίας είναι να μελετήσετε πειραματικά την επίδοση γνωστών αλγορίθμων ταξινόμησης πάνω σ' ένα πραγματικό σύνολο δεδομένων. Η αξιολόγηση της επίδοσης των μεθόδων θα γίνει με τα παρακάτω μέτρα αξιολόγησης πάνω στο σύνολο ελέγχου (*testing*):

- ***Accuracy*** – δηλ. το **ποσοστό επιτυχίας** (%) **των αποφάσεων** του ταξινομητή, και

$$ACC = 100 \times \frac{TP + TN}{P + N}$$

- ***F1 score***

$$F1 = 2\frac{Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

,

όπου *TP: true positives, TN: true negative, FN: false negative, FP: false positives, P: positives, N: negatives.*

Θα μελετήσετε τις εξής **μεθόδους ταξινόμησης**:

[Method 1]. ***Nearest Neighbor k-NN*** *α)* ***με Ευκλείδια απόσταση*** και β) ***με συνημιτονοειδή απόσταση*** (***cosine distance***), υποθέτοντας ***k*** κοντινότερους γείτονες (δοκιμάστε τιμές *k=1, 5, 10*).

[Method 2]. ***Neural Networks*** με σιγμοειδή συνάρτηση ενεργοποίησης (*sigmoid activation function*) σε κάθε νευρώνα (α) με 1 κρυμμένο επίπεδο και *K* κρυμμένους νευρώνες, και (β) με 2 κρυμμένα επίπεδα αποτελούμενο από *K1* και *K2* νευρώνες, αντίστοιχα. Για την εκπαίδευσή τους χρησιμοποιήστε τη μέθοδο βελτιστοποίησης *Stochastic Gradient Descent*. Η έξοδος του δικτύου θα αποτελείται από 10 νευρώνες όπου, χρησιμοποιώντας τη συνάρτηση ενεργοποίησης *softmax,* θα υπολογίζεται η πιθανότητα να ανήκει ένα δεδομένο (εικόνα) σε κάθε μια κατηγορία. Ενδεικτικές τιμές του αριθμού των νευρώνων είναι*: K = 500, K1=500, K2=200*.

[Method 3]. ***Support Vector Machines (SVM)*:** Μηχανές διανυσματικής στήριξης, χρησιμοποιώντας (α) **γραμμική** συνάρτηση πυρήνα (*linear kernel*), (β) ***Gaussian*** συνάρτηση πυρήνα (*kernel*) δοκιμάζοντας διάφορες τιμές της παραμέτρου της, και γ) **συνημιτονοειδή** συνάρτηση πυρήνα (*cosine kernel*). Σε κάθε περίπτωση να χρησιμοποιήσετε την στρατηγική ***one-versus-all*** καθώς έχουμε να αντιμετωπίσουμε ένα πρόβλημα πολλαπλής ταξινόμησης (*multi-class*).

[Method 4]. ***Naïve Bayes classifier*** υποθέτοντας (ανεξάρτητη) κανονική κατανομή (*normal distribution*) για κάθε χαρακτηριστικό.

Δώστε ένα **σύντομο *report* (*pdf αρχείο) το οποίο και θα στείλετε* σχετικά** με τον τρόπο κατασκευής των μεθόδων, τα αποτελέσματα των δοκιμών ανά μέθοδο, όπως επίσης την βέλτιστη μέθοδο που θα προκύψει από την σύγκριση. Στο κείμενο θα πρέπει να ενσωματωθεί επίσης και ο κώδικας που κατασκευάσατε *(Python, Matlab*) ως παράρτημα.