24/3/2023

Προπτυχιακό μάθημα: **Μηχανική Μάθηση**

Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων,
Ακαδημαϊκό έτος 2022-23

1η Σειρά Ασκήσεων

Ημερομηνία παράδοσης : έως 2/5/2023

Θέμα: Μέθοδοι ταξινόμησης ή κατηγοριοποίησης δεδομένων (*Machine Learning classification methods*)

Ο ιστότοπος ***Kaggle*** https://www.kaggle.com/ αποτελεί μία πολύτιμη πηγή δεδομένων για την πειραματική μελέτη αλγορίθμων Μηχανικής Μάθησης, όπου συχνά οργανώνει διεθνείς διαγωνισμούς για την επίλυση πολύπλοκων προβλημάτων που σχετίζονται με δεδομένα. Να σημειωθεί ότι μπορείτε εύκολα να συνδεθείτε στο Kaggle χρησιμοποιώντας τον ιδρυματικό σας λογαριασμό στο uoi.gr καθώς είναι google mail account.

Στην εργασία αυτή θα μελετηθούν δύο τέτοια πειραματικά σύνολο δεδομένων:

- **Mobile Price Classification** – 2,000 δεδομένα 20 διάστασης σε 4 κατηγορίες (0, 1, 2, 3)
  https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification
- **Airlines Delay** – 539,382 δεδομένα 6 διάστασης σε 2 κατηγορίες (δυαδική ταξινόμηση)
  https://www.kaggle.com/datasets/ulrikthygepedersen/airlines-delay

  Παρατηρήσεις: μην λαμβάνεται υπόψιν το 1ο χαρακτηριστικό (*flight*) καθώς αφορά τον κωδικό πτήσης. Τρία από τα υπόλοιπα χαρακτηριστικά είναι διακριτά αλφαριθμητικά τα οποία θα πρέπει να δώσετε μία ακέραια τιμή για κάθε συνδυασμό, π.χ. 1 για ATL, 2 για OO κλπ., που αφορούν αεροπορική εταιρεία (*Airline*), και 1 για ATL, 2 για COS, 3 για BOS, κλπ που αφορούν κωδικούς αεροδρομίων (*Airport From & To*).

Για το πρώτο σύνολο δεδομένων θα χρησιμοποιήσετε **μόνο** το αρχείο για training (*train_mobile.csv*) το οποίο θα χωρίσετε (τυχαία) σε δύο υποσύνολα για μάθηση και έλεγχο σε αναλογία 70-30, ενώ στο δεύτερο σύνολο δεδομένων το υπάρχον αρχείο (*airlines.csv*) το οποίο παρόμοια θα το χωρίσετε τυχαία σε δύο υποσύνολα για μάθηση και αξιολόγηση (70-30).

Στόχος της εργασίας είναι να μελετήσετε πειραματικά την επίδοση γνωστών αλγορίθμων Μηχανικής Μάθησης στο πρόβλημα της ταξινόμησης. Για κάθε μεθοδολογία υϊοθετήστε την γνωστή τεχνική του *10-fold cross validation* για την αντιμετώπιση του *overfitting* και της αύξησης της γενίκευσης του εκάστοτε ταξινομητή.. Η αξιολόγηση της επίδοσης των μεθόδων θα γίνει στο εκάστοτε σύνολο

ελέγχου (*testing set*) με βάση τα γνωστά μέτρα αξιολόγησης ***accuracy*** (ακρίβεια ή ποσοστό επιτυχίας) και ***F1-score***.

Να μελετήσετε τις παρακάτω **μεθόδους ταξινόμησης**:

[Method 1]. ***k-NN Nearest Neighbors*** (δοκιμάστε *k=1, 3, 5 ή 10*). Προσοχή στην περίπτωση χαρακτηριστικών με διακριτές τιμές. Υποθέστε *Ευκλείδια* απόσταση για τις συνεχείς μεταβλητές (*χαρακτηριστικά*) και απόσταση *Hamming* για τις διακριτές μεταβλητές. Η απόσταση θα προκύπτει από το άθροισμα των δύο παραπάνω αποστάσεων.

[Method 2]. ***Naïve Bayes classifier*** υποθέτοντας (ανεξάρτητη) κανονική κατανομή (*normal distribution*) για κάθε ένα από τα χαρακτηριστικά συνεχούς τιμής και πολυωνυμική (*multinomial distribution*) κατανομή για τα διακριτά χαρακτηριστικά (αν υπάρχουν).

[Method 3]. ***Neural Networks*** με σιγμοειδή συνάρτηση ενεργοποίησης στους κρυμμένους νευρώνες (*sigmoid activation function*) *sigmoid* or *tanh*

(α) με 1 κρυμμένο επίπεδο και *K* κρυμμένους νευρώνες, και

(β) με 2 κρυμμένα επίπεδα αποτελούμενο από *K1* και *K2* νευρώνες, αντίστοιχα.

Η έξοδος του δικτύου θα αποτελείται από Κ νευρώνες (όσες και οι κατηγορίες των δεδομένων) όπου, χρησιμοποιώντας τη συνάρτηση ενεργοποίησης *softmax*, θα υπολογίζεται η πιθανότητα να ανήκει ένα δεδομένο σε κάθε κατηγορία. Για την εκπαίδευσή χρησιμοποιήστε τον αλγόριθμο *Stochastic Gradient Descent*. Ενδεικτικές τιμές του αριθμού των νευρώνων είναι*: K = 50 ή 100 ή 200, και (K1, K2)= (50, 25) ή (100, 50) ή (200, 100)*.

[Method 4]. ***Support Vector Machines (SVM)*:** Μηχανές διανυσματικής στήριξης, χρησιμοποιώντας

(α) **Γραμμική** συνάρτηση πυρήνα (*linear kernel*), και

(β) ***Gaussian*** συνάρτηση πυρήνα RBF (*kernel*) δοκιμάζοντας διάφορες τιμές της παραμέτρου της.

Στην περίπτωση ταξινόμησης σε πολλές κατηγορίες (*multi classification problem*) χρησιμοποιήστε την στρατηγική ***one-versus-all.***

Δώστε ένα **σύντομο *report* (*pdf*** μορφή αρχείου***)*** το οποίο θα στείλετε ηλεκτρονικά (μέσω turnin – θα δοθούν οδηγίες) περιγράφοντας εν συντομία α) την διαδικασία κατασκευής των μεθόδων, β) τα αποτελέσματα των δοκιμών ανά μέθοδο με την μορφή πίνακα, και γ) την βέλτιστη μέθοδο που θα προκύψει από την σύγκρισή τους και τα συμπεράσματά σας. Τέλος, στο κείμενο θα πρέπει να ενσωματωθεί και ο κώδικας που κατασκευάσατε ως παράρτημα.

Καλή επιτυχία!