Προπτυχιακό μάθημα: **Μηχανική Μάθηση**

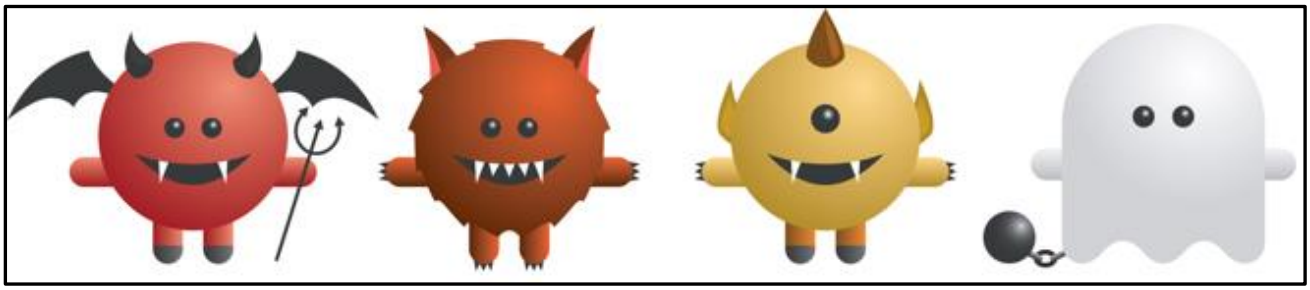
Τμήμα Μηχανικών Η/Υ & Πληροφορικής, Πανεπιστήμιο Ιωαννίνων,
Ακαδημαϊκό έτος 2021-22

1η Σειρά Ασκήσεων

28/3/2022

Ημερομηνία παράδοσης : έως 3/5/2022

Θέμα: Μελέτη μεθόδων ταξινόμησης δεδομένων (*Data Classification algorithms*)



Monster type recognition

Τα τελευταία χρόνια ο ιστότοπος ***Kaggle*** https://www.kaggle.com/ αποτελεί μία πολύτιμη πηγή δεδομένων για την πειραματική μελέτη αλγορίθμων Μηχανικής Μάθησης και συχνά οργανώνει διεθνείς διαγωνισμούς για την επίλυση πολύπλοκων προβλημάτων που σχετίζονται με δεδομένα. Ένα τέτοιο ενδιαφέρον πειραματικό σύνολο δεδομένων αποτελεί και το **Ghouls, Goblins, and Ghosts... Boo**! :

https://www.kaggle.com/competitions/ghouls-goblins-and-ghosts-boo/overview

το οποίο χρησιμοποιήθηκε στα πλαίσια ενός διαγωνισμού πριν από 5 έτη ως ένα πρόβλημα ταξινόμησης (*classification*). Στόχος του προβλήματος είναι η κατασκευή μιας μεθόδου αναγνώρισης του είδους ενός «τέρατος» (*Ghoul, Goblin or Ghost*) ανάλογα με την τιμή πέντε (5) αριθμητικών χαρακτηριστικών του (όπως μήκος μαλλιών, κ.α.). Το 5ο χαρακτηριστικό αποτελεί το χρώμα του τέρατος και έχει 6 διακριτές τιμές (*discrete values*) έναντι των υπολοίπων τα οποία είναι συνεχούς τιμής χαρακτηριστικά (*continuous values features*). Ένας ενδεδειγμένος τρόπος μεταχείρισής του είναι να αντικατασταθεί από μία διακριτή ακέραια τιμή {1, 2, …, 6} ή (διαιρώντας με το 6) σε κανονικοποιημένη μορφή {1/6, 2/6, …, 1} ώστε οι τιμές του χαρακτηριστικού αυτού να είναι στο διάστημα [0, 1] (όπως και των υπολοίπων συνεχών τιμών). Στο πρόβλημα αυτό υπάρχουν δύο σύνολα δεδομένων, ένα που θα χρησιμοποιήσετε για τη διαδικασία εκπαίδευσης (*training*) και ένα για τον έλεγχο και την αξιολόγηση της επίδοσης των αλγορίθμων που θα υλοποιήσετε (*testing*).

Στόχος της εργασίας είναι να μελετήσετε πειραματικά την επίδοση γνωστών αλγορίθμων Μηχανικής Μάθησης στο πρόβλημα της ταξινόμησης χρησιμοποιώντας αυτό το σύνολο δεδομένων. Η αξιολόγηση της επίδοσης των μεθόδων θα γίνει σύμφωνα με τα παρακάτω μέτρα αξιολόγησης στο σύνολο ελέγχου:

- ***Accuracy*** – **ποσοστό επιτυχίας** (ακρίβεια) **του ταξινομητή**, $ACC = \frac{TP+TN}{P+N}$
- ***F1 score*** $F1 = 2\frac{Precision*Recall}{Precision+Recall}$ $Precision = \frac{TP}{TP+FP}$ $Recall = \frac{TP}{TP+FN}$

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

Να σημειωθεί ότι στον ιστότοπο του *Kaggle* υπάρχουν διαθέσιμα αποτελέσματα (τιμές *accuracy*) από την συμμετοχή άλλων ερευνητικών ομάδων στον διαγωνισμό στο πρόσφατο παρελθόν (επιλογή ***Leaderboard***). Όπως μπορείτε να δείτε υπάρχει μόνο μία ομάδα με accuracy 1 , ενώ οι υπόλοιπες έχουν τιμή κάτω του 0.8. Έτσι, το να βρείτε ένα ποσοστό πάνω από 0.75 θα είναι ένα πολύ αξιόλογο αποτέλεσμα! Να σημειωθεί ότι το Kaggle σας επιτρέπει να κατασκευάσετε έναν κωδικό και να δηλωθείτε ως ερευνητική ομάδα στον διαγωνισμό (έστω και αν έχει ημερολογιακά λήξει) ώστε να καταγραφείτε στην κατάταξη (μπορείτε να χρησιμοποιήσετε τον υπάρχον φοιτητικό λογαριασμός σας στο uoi.gr καθώς είναι google mail account).

Για τις ανάγκες τις παρούσας άσκησης θα πρέπει να μελετήσετε τις παρακάτω **μεθόδους ταξινόμησης**:

[Method 1]. ***k-NN Nearest Neighbor με Ευκλείδια απόσταση*** (δοκιμάστε τιμές *k=1, 3, 5, 10*) – μπορείτε εναλλακτικά να διαχωρίσετε από το μέτρο της απόστασης την 5η διάσταση καθώς έχει διακριτή τιμή και η αντίστοιχη τιμή να είναι binary (0 σε περίπτωση ταύτισης χρώματος ή 1).

[Method 2]. ***Neural Networks*** με σιγμοειδή συνάρτηση ενεργοποίησης (*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 3 νευρώνες (όσες και οι κατηγορίες των δεδομένων) όπου, χρησιμοποιώντας τη συνάρτηση ενεργοποίησης *softmax*, θα υπολογίζεται η πιθανότητα να ανήκει ένα δεδομένο (χαρακτηριστικά τέρατος) σε κάθε μια κατηγορία. Για την εκπαίδευσή τους χρησιμοποιήστε τη μέθοδο βελτιστοποίησης *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 3]. ***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.
Σε κάθε περίπτωση να χρησιμοποιήσετε την στρατηγική ***one-versus-all*** καθώς έχετε να αντιμετωπίσουμε ένα πρόβλημα ταξινόμησης πολλών κατηγοριών (*multi-class problem*).

[Method 4]. ***Naïve Bayes classifier*** υποθέτοντας (ανεξάρτητη) κανονική κατανομή (*normal distribution*) για κάθε ένα από τα 4 πρώτα χαρακτηριστικά συνεχούς τιμής και πολυωνυμική (*multinomial distribution*) κατανομή για το 5ο χαρακτηριστικό διακριτής τιμής.

Δώστε ένα **σύντομο *report* (*pdf*** μορφή αρχείου***)*** το οποίο θα στείλετε ηλεκτρονικά περιγράφοντας εν συντομία την διαδικασία 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 ως παράρτημα.

Καλή επιτυχία!