Υπολογιστική Εργασία του Μαθήματος Αναγνώριση Προτύπων
Ακαδημαϊκό Έτος 2019 – 2020
Γ.Τσιχριντζής – Δ. Σωτηρόπουλος

Ανάλυση και Πρόγνωση Αθλητικών Γεγονότων με χρήση Αλγορίθμων Μηχανικής Μάθησης

Στόχος της συγκεκριμένης εργασίας είναι η ανάπτυξη αλγορίθμων μηχανικής μάθησης για την πρόβλεψη του αποτελέσματος ενός ποδοσφαιρικού αγώνα. Το σύνολο των δεδομένων που θα χρησιμοποιήσετε βρίσκεται στην παρακάτω δικτυακή τοποθεσία, https://www.kaggle.com/hugomathien/soccer, υπό την μορφή μιας βάσης δεδομένων SQLite.
Μάλιστα, η ωφέλιμη πληροφορία για την υλοποίηση των ζητούμενων μηχανισμών μάθησης είναι αποθηκευμένη στους πίνακες **Match** και **Team_Attributes**.

Θεωρώντας το σύνολο των διαφορετικών αγώνων της βάσης ως $\boldsymbol{M} = \{\boldsymbol{m}_1, \cdots \boldsymbol{m}_N\}$ και το αντίστοιχο σύνολο των διαφορετικών ομάδων ως $\boldsymbol{T} = \{\boldsymbol{t}_1, \cdots \boldsymbol{t}_K\}$, τότε η πιο αφηρημένη αναπαράσταση του κάθε αγώνα μπορεί να πραγματοποιηθεί ως μια διατεταγμένη τριάδα της μορφής $\boldsymbol{m} = \langle \boldsymbol{h}, \boldsymbol{a}, \boldsymbol{r} \rangle$, με $\boldsymbol{h} \neq \boldsymbol{a}$, όπου η $\boldsymbol{h} \in \boldsymbol{T}$ είναι η ομάδα που αγωνίζεται εντός έδρας (home team), $\boldsymbol{a} \in \boldsymbol{T}$ είναι η ομάδα που αγωνίζεται εκτός έδρας (away team) και $\boldsymbol{r} \in \{\boldsymbol{H}, \boldsymbol{D}, \boldsymbol{A}\}$ το αποτέλεσμα του αγώνα. Συγκεκριμένα, το **H** (home win) υποδηλώνει νίκη της ομάδας που αγωνίζεται εντός έδρας, το **D** (Draw) υποδηλώνει την ισόπαλη έκβαση του αγώνα και το **A** (away win) υποδηλώνει την νίκη της ομάδας που αγωνίζεται εκτός έδρας. Αν με $\boldsymbol{G_m}(\boldsymbol{h}) \geq 0$ και $\boldsymbol{G_m}(\boldsymbol{a}) \geq 0$ συμβολίσουμε το πλήθος των τερμάτων που επιτυγχάνονται από τις ομάδες εντός και εκτός έδρας κατά την διεξαγωγή του αγώνα $\boldsymbol{m}$, τότε η μεταβλητή έκβασης του αγώνα $\boldsymbol{r}$ μπορεί να ορισθεί συναρτήσει της διαφοράς των συνολικών τερμάτων $\boldsymbol{\Delta G_m}(\boldsymbol{h}, \boldsymbol{a}) = \boldsymbol{G_m}(\boldsymbol{h}) - \boldsymbol{G_a}(\boldsymbol{h})$ ως:

$$\boldsymbol{r} = \begin{cases} \boldsymbol{H}, \boldsymbol{\Delta G_m}(\boldsymbol{h}, \boldsymbol{a}) > 0; \\ \boldsymbol{D}, \boldsymbol{\Delta G_m}(\boldsymbol{h}, \boldsymbol{a}) = 0; \\ \boldsymbol{A}, \boldsymbol{\Delta G_m}(\boldsymbol{h}, \boldsymbol{a}) < 0\,. \end{cases}$$

Η διαδικασία εκπαίδευσης των εμπλεκόμενων ταξινομητών θα πρέπει να βασιστεί σε ένα σύνολο χαρακτηριστικών γνωρισμάτων της κάθε ομάδας καθώς και σε ένα σύνολο προγνωστικών (odds) για την πιθανή έκβαση του κάθε αγώνα από έναν αριθμό στοιχηματικών εταιρειών. Συγκεκριμένα, κάθε ομάδα $\boldsymbol{t} \in \boldsymbol{T}$ είναι συσχετισμένη με ένα διάνυσμα χαρακτηριστικών $\boldsymbol{\varphi}(\boldsymbol{t}) \in \mathbb{R}^8$ τα οποία αντιστοιχούν στις παρακάτω στήλες του πίνακα **Team_Attributes** {**buildUpPlaySpeed**, **buildUpPlayPassing**, **chanceCreationPassing**, **chanceCreationCrossing**, **chanceCreationShooting**, **defencePressure**, **defenceAggregation**, **defenceTeamWidth**}. Επιπλέον, κάθε αγώνας $\boldsymbol{m} \in \boldsymbol{M}$ είναι συσχετισμένος με τέσσερα διανύσματα προγνωστικών $\boldsymbol{\psi_\kappa}(\boldsymbol{m}) \in \mathbb{R}^3$ με $\boldsymbol{k} \in \{\boldsymbol{B}365, \boldsymbol{BW}, \boldsymbol{IW}, \boldsymbol{LB}\}$ τα αντιστοιχούν στις παρακάτω στήλες του πίνακα **Match** {**B365H**, **B365D**, **B365A**,**BWH**, **BWD**, **BWA**, **IWH**, **IWD**, **IWA**, **LBH**, **LBD**, **LBA**}. Δηλαδή, το κάθε διάνυσμα

$\boldsymbol{\psi_\kappa(m) = \left[d_k^H(m), d_k^D(m), d_k^A(m)\right]}$ συγκεντρώνει τις στοιχηματικές αποδόσεις για κάθε πιθανή έκβαση του αγώνα $\boldsymbol{m}$ για κάθε στοιχηματική εταιρεία $\boldsymbol{k \in B = \{B365, BW, IW, LB\}}$. Λάβετε υπόψιν πως υπάρχουν εγγραφές στον πίνακα **Match** για τις οποίες τα αντίστοιχα διανύσματα προγνωστικών έχουν μηδενικές τιμές. Οι συγκεκριμένες εγγραφές θα πρέπει να αφαιρεθούν.

Ερωτήματα:

I. Να υλοποιήσετε ένα γραμμικό νευρωνικό δίκτυο, ώστε ο εκπαιδευόμενος ταξινομητής να υλοποιεί μια συνάρτηση διάκρισης της μορφής $\boldsymbol{g_k(\psi_\kappa(m)): \mathbb{R}^3 \to \{H, D, A\}}$ για κάθε στοιχηματική εταιρεία. Να αναγνωρίσετε την στοιχηματική εταιρεία τα προγνωστικά της οποίας οδηγούν σε μεγαλύτερη ακρίβεια ταξινόμησης.

II. Να υλοποιήσετε ένα πολυστρωματικό νευρωνικό δίκτυο, ώστε ο εκπαιδευόμενος ταξινομητής να υλοποιεί μια συνάρτηση διάκρισης της μορφής $\boldsymbol{g_k(\psi_\kappa(m)): \mathbb{R}^3 \to \{H, D, A\}}$ για κάθε στοιχηματική εταιρεία. Να αναγνωρίσετε την στοιχηματική εταιρεία τα προγνωστικά της οποίας οδηγούν σε μεγαλύτερη ακρίβεια ταξινόμησης.

III. Να υλοποιήσετε ένα πολυστρωματικό νευρωνικό δίκτυο, ώστε ο εκπαιδευμένος ταξινομητής να υλοποιεί μια συνάρτηση διάκρισης της μορφής $\boldsymbol{g(\Phi(m)): \mathbb{R}^{28} \to \{H, D, A\}}$, όπου το $\boldsymbol{\Phi(m) \in \mathbb{R}^{28}}$ αντιστοιχεί στο πλήρες διάνυσμα χαρακτηριστικών του κάθε αγώνα που δίνεται από την σχέση:

$$\boldsymbol{\Phi(m) = \left[\varphi(h),\ \varphi(\alpha), \psi_{B365}(m), \psi_{BW}(m), \psi_{IW}(m), \psi_{LW}(m)\right]}$$

IV. Να εφαρμόσετε τον αλγόριθμο ομαδοποίησης c – means επάνω στο σύνολο των διανυσμάτων προγνωστικών $\boldsymbol{\Psi_k = \{\psi_\kappa(m) \in \mathbb{R}^3 : m \in M\}}$ για κάθε στοιχηματική εταιρεία $\boldsymbol{k \in B}$, θέτοντας την τιμή του c ίση με 3. Με το τρόπο αυτό, θα παράξετε μια διαφορετική διαμέριση του συνόλου των αγώνων $\boldsymbol{M}$ σε τρείς συστάδες για κάθε στοιχηματική εταιρεία. Λαμβάνοντας υπόψιν το αποτέλεσμα του κάθε αγώνα να υπολογίσετε την κατανομή των τριών αποτελεσμάτων εντός της κάθε συστάδας για κάθε στοιχηματική εταιρεία. Υπάρχει κάποιο αποτέλεσμα που να επικρατεί σε συχνότητα εντός της κάθε συστάδας;

Παρατηρήσεις:

I. Για κάθε ταξινομητή που θα υλοποιήσετε θα πρέπει να αναφέρετε την ταξινομητική του ακρίβεια τόσο κατά την φάση της εκπαίδευσης όσο και κατά την φάση του ελέγχου σύμφωνα με την μέθοδο της 10-πλής διεπικύρωσης (**10 fold cross validation).**

II. Στο αρχείο **EuropeanSoccerDatabaseRetriever.m** σας παρέχετε κώδικας για την άντληση των δεδομένων από την βάση SQLite.

III. Παραδοτέα της εργασίας αποτελούν ο **κώδικας** της υλοποίησής σας σε MATLAB ή Python καθώς και ένα συνοδευτικό **κείμενο τεκμηρίωσης**.

IV. Μπορείτε να εργασθείτε σε ομάδες των **δύο** ή **τριών ατόμων**.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!