

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΠΜΣ «ΠΛΗΡΟΦΟΡΙΚΗ»**

**ΕΡΓΑΣΙΑ ΜΑΘΗΜΑΤΟΣ**

**«Αναγνώριση Προτύπων»**

**Αυγέρης Χαράλαμπος (ΜΠΠΛ18004) / mppl18004@unipi.gr**

**Καραγιάννη Γεωργία (ΜΠΠΛ18028) / mppl18028@unipi.gr**

**Λοΐζος Εμμανουήλ (ΜΠΠΛ18043) / mppl18043@unipi.gr**

**Οκτώβριος 2020**

**ΕΡΩΤΗΜΑΤΑ 1 ΚΑΙ 2**

Τα δύο πρώτα ζητούμενα της εργασίας περιέχονται στο script με όνομα «ERWTHMATA\_1\_2.m». Ανοίγοντας και εκτελώντας το συγκεκριμένο script αρχικά φορτώνονται από τη βάση sqlite όλα τα απαραίτητα δεδομένα (χαρακτηριστικά ομάδων, προγνωστικά στοιχηματικών εταιρειών κτλ) για τη δημιουργία των ζητούμενων νευρωνικών δικτύων. Κατόπιν, επειδή ο πίνακας Match περιέχει εγγραφές για τις οποίες τα αντίστοιχα διανύσματα προγνωστικών έχουν μηδενικές τιμές, καλείται η συνάρτηση «removeZeros» με την οποία αφαιρούνται από τον Match οι παραπάνω εγγραφές και επιστρέφεται ο πίνακας MatchWithoutZeros.

Εκείνο που μας ζητείται για τα 2 αρχικά ερωτήματα της εργασίας είναι η υλοποίηση ενός γραμμικού και ενός πολυστρωματικού νευρωνικού δικτύου, ώστε ο εκπαιδευόμενος ταξινομητής να υλοποιεί μια συνάρτηση διάκρισης της μορφής  για κάθε στοιχηματική εταιρεία. Ουσιαστικά λοιπόν καλούμαστε να υλοποιήσουμε οκτώ νευρωνικά δίκτυα, από ένα γραμμικό και ένα πολυστρωματικό για κάθε μία από τις τέσσερις στοιχηματικές εταιρείες που περιέχονται στη βάση δεδομένων.

Αφού λοιπόν έχουμε φορτώσει τα δεδομένα της βάσης μέσα στο workspace της matlab και έχουμε αφαιρέσει τις μηδενικές τιμές από τον πίνακα Match, πρέπει να κατασκευάσουμε τα διανύσματα εκείνα τα οποία θα χρησιμοποιηθούν ως είσοδος και ως επιθυμητή έξοδος για την εκπαίδευση των νευρωνικών δικτύων. Όσον αφορά την είσοδο, φτιάχνουμε τέσσερα Νx3 διανύσματα (B365\_odds, BW\_odds, LB\_odds IW\_odds), κάθε ένα από τα οποία περιέχει τις αποδόσεις όλων των αγώνων για κάθε μία στοιχηματική εταιρεία ξεχωριστά, προσαρμόζουμε κάθε γραμμή στην κλίμακα του 1 για να ταιριάζει το κάθε διάνυσμα εισόδου με το διάνυσμα εξόδου που θα αναφέρουμε στη συνέχεια και τέλος τα κάνουμε transpose (το κάθε Νx3 διάνυσμα δηλαδή το μετατρέπουμε στο αντίστοιχο 3xΝ) για να μπορεί το νευρωνικό δικτυο να τα δεχτεί σαν είσοδο. Όσον αφορά την έξοδο, θα είναι και αυτό ένα Νx3 διάνυσμα (το ονομάζουμε target) το οποίο θα αναπαριστά το αποτέλεσμα του κάθε αγώνα ( για παράδειγμα η νίκη της ομάδας εντός έδρας θα αναπαρίσταται σαν [1 0 0], η ισοπαλία σαν [0 1 0] και η νίκη της ομάδας εκτός έδρας σαν [0 0 1]) και το κάνουμε και αυτό transpose. Τέλος, δημιουργούμε τα ζητούμενα νευρωνικά δίκτυα, καλώντας τις συναρτήσεις LinearOdds και MultiOdds και δίνοντάς τους ως είσοδο τα αντίστοιχα διανύσματα.

Υπολογισμός ταξινομητικής ακρίβειας (accuracy)

Αφού εκπαιδευτεί το νευρωνικό δίκτυο, του δίνουμε ξανά ως είσοδο το αντίστοιχο διάνυσμα που χρησιμοποιήσαμε για την εκπαίδευσή του και παίρνουμε σαν απάντηση την πρόβλεψη του νευρωνικού για τα αποτελέσματα των αγώνων την οποία αποθηκεύουμε στον πίνακα outputs. Στη συνέχεια κρατάμε το index με τη μεγαλύτερη τιμή κάθε γραμμής του outputs(το οποίο είναι ουσιαστικά η πρόβλεψη του εκπαιδευμένου νευρωνικού για το αποτέλεσμα του αγώνα), όπως και το αντίστοιχο index της κάθε γραμμής του διανύσματος εισόδου (το οποίο είναι το πραγματικό αποτέλεσμα του αντίστοιχου αγώνα). Τέλος, υπολογίζουμε το πλήθος όλων των indexes του πίνακα outputs που ισούνται με τα αντίστοιχα του διανύσματος εισόδου(όλες οι φορές δηλαδή που το εκπαιδευμένο μας νευρωνικό δίκτυο έχει δώσει πρόβλεψη η οποία συμφωνεί με το πραγματικό αποτέλεσμα του αγώνα) και το πλήθος αυτό το διαιρούμε με το πλήθος των στοιχείων του διανύσματος εισόδου (δηλαδή το πλήθος των αγώνων).

**ΕΡΩΤΗΜΑ 3**

Το τρίτο ερώτημα περιέχεται στο script με όνομα «ERWTHMA\_3.m». Εκείνο που μας ζητείται για το συγκεκριμένο ερώτημα είναι η υλοποίηση ενός πολυστρωματικού νευρωνικού δικτύου, ώστε ο εκπαιδευόμενος ταξινομητής να υλοποιεί μια συνάρτηση διάκρισης της μορφής  **,** όπου το αντιστοιχεί στο πλήρες διάνυσμα χαρακτηριστικών του κάθε αγώνα που δίνεται από την σχέση:

Το παραπάνω νευρωνικό δίκτυο θα παίρνει σαν είσοδο ένα Νx28 διάνυσμα το οποίο θα περιέχει τα προγνωστικά αγώνων και των τεσσάρων στοιχηματικών εταιρειών μαζί με τα 8 χαρακτηριστικά των ομάδων από τις οποίες θα αποτελείται ο κάθε αγώνας και τα οποία ειναι τα εξής: buildUpPlaySpeed, buildUpPlayPassing, chanceCreationPassing, chanceCreationCrossing, chanceCreationShooting, defencePressure, defenceAggregation, defenceTeamWidth και τα οποία περιλαμβάνονται στον πίνακα TeamAttributes (το νευρωνικό δίκτυο που θα δημιουργήσουμε δηλαδή γι αυτό το ερώτημα θα παίρνει σαν είσοδο ένα διάνυσμα το οποίο θα περιέχει όλη την πληροφορία). Αν κοιτάξουμε όμως τον TeamAttributes παρατηρούμε ότι για αγώνες κάποιων σεζόν, δεν έχουμε τα χαρακτηριστικά των ομάδων που συμμετέχουν. Αν πάρουμε για παράδειγμα τον 1ο αγώνα του πίνακα Match βλέπουμε  
ότι παίζει η ομάδα με team\_id 9987 εναντίον της ομάδας με team\_id 9993 και ο αγώνας αναφέρεται στη σεζόν 2008-2009. Για αυτές τις ομάδες όμως δεν υπάρχουν χαρακτηριστικά στον πίνακα Team\_Attributes που να αφορούν εκείνη τη σεζόν. Γι αυτό το λόγο αποφασίσαμε να αφαιρέσουμε και να μη λάβουμε υπόψη στο διάνυσμα εισόδου όλους εκεινους τους αγώνες για τους οποίους δεν έχουμε χαρακτηριστικά ομάδων για τις αντίστοιχες χρονικές περιόδους. Το διάνυσμα εξόδου θα είναι στην ίδια λογική με αυτό των ερωτημάτων 1 και 2 χωρίς βέβαια τους αγώνες που αναφέραμε παραπάνω ότι θα αφαιρέσουμε. Με βάση τα παραπάνω λοιπόν κατασκευάζουμε τα διανύσματα εισόδου και εξόδου και εν τέλη είμαστε σε θέση να προχωρήσουμε στην υλοποίηση του ζητούμενου νευρωνικού δικτύου.

Και ο υπολογισμός της ταξινομητικής ακρίβειας (accuracy) είναι στην ίδια λογική με αυτή που ακολουθήσαμε στα 2 προηγούμενα ερωτήματα.

**ΕΡΩΤΗΜΑ 4**

Το τέταρτο ερώτημα περιέχεται στο script με όνομα «ERWTHMA\_4.m». Ζητούμενο του ερωτήματος αυτού είναι η εφαρμογή του αλγόριθμου ομαδοποίησης c – means επάνω στο σύνολο των διανυσμάτων προγνωστικών για κάθε στοιχηματική εταιρεία , θέτοντας την τιμή του c ίση με 3.

Καλούμαστε δηλαδή να εφαρμόσουμε τον παραπάνω αλγόριθμο 4 φορές(από μία για κάθε στοιχηματική εταιρεία). Φτιάχνουμε λοιπόν αρχικά τα διανύσματα που θα δινουμε σαν είσοδο στον αλγόριθμο τα οποία είναι τα ίδια με αυτά που κατασκευάσαμε για τα 2 πρώτα ερωτήματα (τέσσερα Νx3 διανύσματα, ένα για κάθε στοιχηματική εταιρεία με τις αποδόσεις της). Κατόπιν, τρέχουμε τον αλγόριθμο fuzzy c-means ([[centers](https://www.mathworks.com/help/fuzzy/fcm.html?fbclid=IwAR1PmhrqUJrgKfkLiBZGjuNUuG3OZjels2Z6XJklivyNSwXgpiTE4mzaBFM#buv_wo_-centers),[U](https://www.mathworks.com/help/fuzzy/fcm.html?fbclid=IwAR1PmhrqUJrgKfkLiBZGjuNUuG3OZjels2Z6XJklivyNSwXgpiTE4mzaBFM#buv_wo_-U)] = fcm([data](https://www.mathworks.com/help/fuzzy/fcm.html?fbclid=IwAR1PmhrqUJrgKfkLiBZGjuNUuG3OZjels2Z6XJklivyNSwXgpiTE4mzaBFM#buv_wo_-data),[Nc](https://www.mathworks.com/help/fuzzy/fcm.html?fbclid=IwAR1PmhrqUJrgKfkLiBZGjuNUuG3OZjels2Z6XJklivyNSwXgpiTE4mzaBFM#buv_wo_-Nc)) ), όπου για data θα είναι το αντίστοιχο διάνυσμα της εκάστοτε στοιχηματικής και για Nc = 3. Από τη συνάρτηση αυτή θα μας επιστραφεί μια διαμέριση του συνόλου των αγώνων σε τρία clusters μαζί με τις συντεταγμένες του κέντρου του καθε cluster. Υπολογίζοντας το μέγιστο στοιχείο κάθε στήλης και λαμβάνοντας υπόψιν το αποτέλεσμα κάθε αγώνα μπορούμε να δούμε την κατανομή των τριών αποτελεσμάτων εντός του κάθε cluster για κάθε στοιχηματική εταιρεία και ποιο αποτέλεσμα ανήκει περισσότερο σε ένα συγκεκριμένο cluster.