**ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ**

Διδάσκοντες: Σ. Λυκοθανάσης, Δ. Κουτσομητρόπουλος

Ακαδημαϊκό Έτος 2022-2023

**Εργαστηριακή Άσκηση**

**Μέρος Α’**

Σούρλας Ζήσης

Αρ. Μητρώου: 1072477

sourlas.zisis@upnet.gr

**Α1. Προεπεξεργασία και Προετοιμασία δεδομένων**

1. Για την προεπεξεργασία των δεδομένων χρησιμοποιήθηκε η βιβλιοθήκη **sklearn**.

**Μετατροπή κατηγορικών σε αριθμητικές τιμές**

Για να μετατρέψουμε τις κατηγορικές τιμές της εισόδου σε αριθμητικές χρησιμοποιήθηκε ο **OrdinalEncoder** της **sklearn** ο οποίος κωδικοποιεί τις κατηγορικές τιμές σε ακέραιες τιμές στο διάστημα [0, πλήθος\_κατηγοριών-1]. Για την κωδικοποίηση των κλάσεων χρησιμοποιήθηκε ο **LabelEncoder** της **sklearn** ο οποίος κωδικοποιεί τις κλάσεις σε ακέραιες τιμές στο διάστημα [0, πλήθος\_κλάσεων-1] και προτείνεται από το documentation της scikit learn για κωδικοποίηση κλάσεων.

**Κεντράρισμα (Centering)**

Το κεντράρισμα των δεδομένων είναι μια μεθοδολογία που χρησιμοποιείται πολύ συχνά στην προπεξεργασία δεδομένων. Αφαιρώντας το μέσο όρο από τα δεδομένα δημιουργούμε δεδομένα με μηδενική μέση τιμή. Αυτό σημαίνει ότι κατά τη διαδικασία μάθησης μειώνεται η επίδραση των ακραίων τιμών που μπορεί να οδηγήσουν σε καθυστέρηση στη σύγκλιση και σε μικρότερη ικανότητα γενίκευσης. Συνεπώς κρίθηκε σκόπιμο να χρησιμοποιηθεί στην παρούσα εργασία και υλοποιήθηκε με χρήση του **StandardScaler** της **sklearn**.

**Κανονικοποίηση (Min-Max Scaling)**

Το min-max scaling είναι μια πολύ χρήσιμη τεχνική όταν τα δεδομένα βρίσκονται σε διαφορετικές κλίμακες. Με αυτή την τεχνική μεταφέρονται όλες οι τιμές σε μια συγκεκριμένη κλίμακα ([0,1]) έτσι ώστε να διατηρούνται μεν οι διαφορές (αναλογικά) μεταξύ των δειγμάτων ενός τύπου εισόδου αλλά μεταξύ δειγμάτων διαφορετικών τύπων εισόδων που βρίσκονται αρχικά σε διαφορετικά εύρη τιμών (π.χ. φύλο και βάρος) να υπάρξει εν τέλει μία ενιαία κλίμακα τιμών ώστε να μην δίνεται αρχικά περισσότερο βάρος σε κάποιες εισόδους από άλλες. Στη συγκεκριμένη εργασία το dataset που δίνεται περιλαμβάνει τιμές που βρίσκονται σε πολύ διαφορετικές κλίμακες και κατά συνέπεια κρίθηκε απαραίτητο να χρησιμοποιηθεί min-max scaling (υλοποιήθηκε με τον **MinMaxScaler** της **sklearn**). Ωστόσο η τεχνική αυτή είναι ευαίσθητη στην παρουσία ακραίων τιμών (outliers) καθώς τείνει να συμπιέσει τις συνηθισμένες τιμές (inliers) σε μικρό κομμάτι του εύρους τιμών. Αυτός είναι ακόμα ένας λόγος για τον οποίο πρέπει να εφαρμοστεί κεντράρισμα στα δεδομένα.

**Τυποποίηση (Standarization)**

Η τυποποίηση χρησιμοποιείται για να μετατρέψει τα δεδομένα ώστε αυτά να έχουν μέση τιμή μηδέν και διακύμανση ένα. Ωστόσο η τυποποίηση προϋποθέτει ότι τα δεδομένα έχουν Gaussian κατανομή ή κάτι κοντά σε αυτήν. Από την στιγμή που δεν γνωρίζουμε τίποτα για την κατανομή των δεδομένων μας δεν κρίνεται σκόπιμο να χρησιμοποιήσουμε αυτή την τεχνική ειδικά καθώς και αυτή είναι ευαίσθητη στην παρουσία outliers.

1. Για την δημιουργία των 5 folds που απαιτούνται χρησιμοποιήθηκε η **StratifiedKFold** της **sklearn.** Η συγκεκριμένη συνάρτηση σε αντίθεση με την **KFold** διατηρεί σε κάθε fold την αρχική αναλογία δειγμάτων ανά κλάση.

**Α2. Επιλογή αρχιτεκτονικής**

1. **Επιλογή συνάρτησης κόστους**

**Διεντροπία (Cross Entropy)**

Η διεντροπία μετράει την διαφορά μεταξύ της κατανομής πιθανότητας των παραγόμενων αποτελεσμάτων του νευρωνικού δικτύου και της κατανομής πιθανότητας των επιθυμητών αποτελεσμάτων. Ελαχιστοποιώντας την διεντροπία το μοντέλο αυξάνει την πιθανότητα ανάθεσης σωστής κλάσης στα δεδομένα εισόδου. Αυτό είναι ένα πολύ χρήσιμο χαρακτηριστικό για προβλήματα ταξινόμησης τα οποία απαιτούν ακριβώς αυτό, δηλαδή την ανάθεση ετικετών σε σετ δεδομένων εισόδου. Για αυτόν το λόγο οι διεντροπία χρησιμοποιείται κατά βάση σε προβλήματα ταξινόμησης και για αυτόν τον λόγο χρησιμοποιήθηκε και στην παρούσα εργασία.

**Μέσο Τετραγωνικό Σφάλμα (MSE)**

Το Μέσο Τετραγωνικό Σφάλμα υπολογίζεται παίρνοντας το μέσο όρο των τετραγώνων των διαφορών των πραγματικών με των επιθυμητών τιμών. Δείχνει επί της ουσίας πόσο απέχουν οι προβλεπόμενες τιμές από τις πραγματικές. Χρησιμοποιείται κυρίως σε προβλήματα παλινδρόμησης.

**Ακρίβεια ταξινόμησης (Accuracy)**

Η ακρίβεια ταξινόμησης δείχνει απλώς το ποσοστό σωστών προβλέψεων ενός μοντέλου. Είναι χρήσιμη μετρική για να μετρήσουμε την απόδοση ενός μοντέλου σε προβλήματα ταξινόμησης αφού μας δείχνε ακριβώς πόσο καλά ταξινομήθηκαν τα δεδομένα εισόδου ωστόσο δεν είναι κατάλληλη για συνάρτηση κόστους καθώς δεν είναι διαφορίσιμη πράγμα απαραίτητο για τις συναρτήσεις κόστους. Επιπλέον η ακρίβεια ταξινόμησης δε δίνει κάποια πληροφορία σχετικά με τη “σιγουριά” με την οποία ταξινομούνται τα δεδομένα εισόδου. Π.χ. Ας πάρουμε την περίπτωση που ένα μοντέλο ταξινομεί τη φωτογραφία ενός σκύλου στην κλάση “σκύλος” με πιθανότητα 70% και στην κλάση “γάτα” με πιθανότητα 30% ενώ ένα άλλο μοντέλο ταξινομεί την ίδια φωτογραφία στην κλάση “σκύλος” με 90% και στην κλάση “γάτα” με 10%. Τα δύο μοντέλα θα έχουν την ίδια ακρίβεια ταξινόμησης αφού και τα δύο εν τέλει ταξινομούν σωστά την φωτογραφία όμως το δεύτερο είναι πολύ πιο “σίγουρο” για την ταξινόμηση που έκανε κάτι το οποίο φαίνεται μόνο στην διεντροπία του και όχι στην ακρίβεια ταξινόμησης.