

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

**ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ**

|  |
| --- |
|  |
|  |

**ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ**

Ακαδημαϊκό Έτος 2021-2022

**Εργαστηριακή Άσκηση**

**Μέρος Α’**

**Γεώργιος Κοντογιάννης**

**1070908**

**a)** Η κωδικοποίηση των λέξεων είναι γωνστή και ανήκει στο εύρος [0, 8519]. Για το λόγο αυτό, ο CountVectorizer παραμετροποιήθηκε με αυτό το ήδη υπάρχον λεξικό. Καλώντας την μέθοδο transform, μετατράπηκαν τα δεδομένα εισόδο σε BoW και κατ’ επέκταση δημιουργήθηκε το Document Term Matrix.

**b)** Η κλιμάκωση(scaling) των δεδομένων, είναι ένα από τα πιο σημαντικά βήματα προεπεξεργασίας δεδομένων στη μηχανική εκμάθηση. Οι αλγόριθμοι που υπολογίζουν την απόσταση μεταξύ των χαρακτηριστικών, ωθούνται προς τις αριθμητικά μεγαλύτερες τιμές (outliers) εάν τα δεδομένα δεν είναι κλιμακωμένα.

1. To **Κεντράρισμα** (centering) αφαιρεί μία σταθερή τιμή απο κάθε μεταβλητή εισόδου. Πρακτικά, επαναπροσδιορίσει το σημείο 0 για τον προγνωστικό παράγοντα, ώστε να είναι οποιαδήποτε τιμή αφαιρέθηκε. Μετατοπίζει την κλίμακα, αλλά διατηρεί τις μονάδες.

2. Η **κανονικοποίηση** (scaling) χρησιμοποιείται για τη μετατροπή των χαρακτηριστικών σε παρόμοια κλίμακα. Το νέο σημείο υπολογίζεται απο το τύπο: . Αυτό κλιμακώνει το εύρος σε [0, 1] ή μερικές φορές [-1, 1]. *Η κανονικοποίηση είναι χρήσιμη όταν δεν υπάρχουν ακραίες τιμές, καθώς δεν μπορεί να τις αντιμετωπίσει*.

3. Η **Τυποποίηση** (standardization) είναι ο μετασχηματισμός χαρακτηριστικών με αφαίρεση τον μέσο όρο και διαίρεση με τυπική απόκλιση. . Η τυποποίηση μπορεί να είναι χρήσιμη σε περιπτώσεις όπου τα δεδομένα ακολουθούν μια κατανομή Gauss. Ωστόσο, αυτό δεν είναι απαραίτητο να ισχύει. *Η τυποποίηση δεν επηρεάζεται από ακραίες τιμές επειδή δεν υπάρχει προκαθορισμένο εύρος μετασχηματισμένων χαρακτηριστικών.*

***\*\*\* Πρακτικά, ο συνδυασμός του κεντραρίσματος και κανονικοποίησης έχει ώς αποτέλεσμα την τυποποίηση*.**

**b) Κώδικας**

Όσον αφορά την τοπολογία των ΤΝΔ για την εκπαίδευση τους με τον Αλγόριθμο Οπισθοδιάδοσης του Σφάλματος (back-propagation), θα χρησιμοποιήσετε ΤΝΔ με *ένα κρυφό* 3

*επίπεδο* και θα πειραματιστείτε με τον αριθμό των κρυφών κόμβων. Για την εκπαίδευση του δικτύου χρησιμοποιήστε αρχικά ρυθμό μάθησης *η =* 0.001.