

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

**ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ**

|  |
| --- |
|  |
|  |

**ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ**

Ακαδημαϊκό Έτος 2021-2022

**Εργαστηριακή Άσκηση**

**Μέρος Α’**

**Γεώργιος Κοντογιάννης**

**1070908 – Δ’ έτος**

**α)** 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.

**β)** 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*.**

Πόρισμα 1: Η **κανονικοποίηση** στο συγκεκριμένο πρόβλημα δεν προσδίδει ιδιαίτερη βελτίωση. Το συγκεκριμένο πόρισμα φαίνεται λογικό εαν σκεφτόμαι οτι τα έγγραφα του κειμένου έχουν υποστεί σημαντική προεπεξεργασία.

Πόρισμα 2. Η **τυποποίηση** και το **κεντράρισμα** εισάγουν αρντηικές τιμές εισόδων. Αυτό δεν είναι επιθυμητό για τον αλγόριθμο για δύο λόγους. Πρώτον οι αρνητικές τιμές συχνότητας επαναλήψεων δεν έχουν ουσιαστικό νόημα. Δεύτερον, καθώς τα έγγραφα έχουν υποστεί αφαίρεση των stopwords και stemming, οι μεγάλες συχνότητες των λέξεων προσδίνουν μεγάλη βαρύτητα το οποίο είναι λογικό.

**α)** Κάθε μια απο τις μετρικές αυτές, είναι μια μέθοδος αξιολόγησης του πόσο καλά μοντελοποιεί ο αλγόριθμός το σύνολο δεδομένων. Εάν οι προβλέψεις είναι εντελώς «εκτός», η συνάρτηση κόστους θα παράγει μεγαλύτερο αριθμό. Αν είναι αρκετά καλά, θα βγάζει μικρότερο αριθμό.Αυτός ο αριθμός είναι πολύ σημαντικός γιατί μας βοηθάει να εντοπίσουμε τοπικά ελάχιστα. Στο δικό μας πρόβλημα, αρχικά να πούμε οτι οι απόλυτες διαφορές ανάμεσα στην πραγματική έξοδο με την επιθυμητή είναι πάντα ανάμεσα στο [0, 1]. Άρα, στη περίπτωση του **Cross-Entropy**, μικρές πιθανότητες πρόβλεψης τιμωρούνται περισσότερο. Στο **MSE**, τα μεγαλύτερα σφάλματα είναι αυτά που «τιμωρούνται» περισσότερο λόγω του τετραγώνου. Τέλος, το **Acc** ελέγχει πόσες απο τις ετικέτες είναι ίδιες και διαιρεί με το πλήθος των πραγματικών και ψευδών ετικετών. Άρα, όσο περισσότερες λανθασμένες ετικέτες προβλέψει οτι υπάρχουν ο αλγόριθμος, τοσο μιρκότερο σφάλμα θα δώσει.

**β)** Η ταξινόμηση πολλαπλών ετικετών μπορεί να υποστηριχθεί απευθείας από νευρωνικά δίκτυα, προσδιορίζοντας τον αριθμό των ετικετών-στόχων που υπάρχουν στο πρόβλημα ως τον αριθμό των κόμβων στο επίπεδο εξόδου. Στη περίπτωση μας, το πρόβλημα έχει είκοσι (20) πιθανές ετικέτες εξόδου (κλάσεις), άρα για το επίπεδο εξόδου του νευρωνικού δικτύου απαιτούνται 20 κόμβοι/νευρώνες εξόδου.

**γ)** Κάθε κόμβος στο κρυφό επίπεδο θα χρησιμοποιεί την **ReLU** συνάρτηση ενεργοποίησης. Τα χαρακτρηριστικά της μας φαίνοται ιδιαίτερα χρήσιμα στο συγκεκριμένο πρόβλημα καθώς οι τιμές των διανύσματων εισόδου αποτελούνται απο διακριτές στο εύρος [0, k], με k , δίνοντας έτσι στην εκάστοτε τιμή το βάρος που της αναλογεί, καθώς λέξη με μεγάλη συχνότητα έχει νόημα να εισαχθεί στο δίκτυο με μεγαλύτερο βάρος.

**δ)** Κάθε κόμβος στο επίπεδο εξόδου πρέπει να χρησιμοποιεί την **σιγμοειδή** συνάρτηση ενεργοποίησης. Ο βασικός λόγος είναι ότι οι πιθανότητες που παράγονται από τη σιγμοειδή είναι ανεξάρτητες και δεν περιορίζονται στο να αθροιστούν στο ‘1’.

**ε)**

|  |  |  |  |
| --- | --- | --- | --- |
| **#Νευρώνων στο κρυφό επίπεδο** | **CE Loss** | **MSE** | **Acc** |
| H1 = O = 20 | **0.5180/0.8540** | **0.2104/** **0.8314** |  |
| H1=(I+O)/2=4270 | **0.5003** |  |  |
| H1 = (I+O) = 8540 |  |  |  |