2η Εργασία Τεχνητή Νοημοσύνη

Μιχαήλ Πρωτονοτάριος 3200164

Μάριος Γεωργοπετρέας 3200028

Για την επεξεργασία των δεδομένων (reviews) φτιάξαμε μια κλάση tokenizer που ουσιαστικά κάνει το tokenization των reviews σε «λέξεις κλειδία» (features), δηλαδή μετατρέπει κάθε review σε feature vector, όπου έχει 1 στην θέση ενός feature αν αυτό υπάρχει στο κείμενο, αλλιώς 0 αν δεν υπάρχει.

Και για τους 2 αλγορίθμους που είχαμε αναπτύξει, για το φόρτωμα των δεδομένων (πριν κάνουμε initialize τον tokenizer), το πρόγραμμα ζητάει από τον χρήστη να δώσει ως input τα paths των positive και negative training και test data αντίστοιχα.

Χρησιμοποιήσαμε 27 features (λέξεις του λεξιλογίου), φαίνονται κάποιες από αυτές στο παρακάτω pandas dataframe. Καταλήξαμε σε αυτές τις 27 λέξεις φιλτράροντας τα κείμενα από σημεία στίξης, κενά και άλλα άχρηστα strings, έπειτα πήραμε τις 700 πιο συχνές λέξεις από όλα τα κείμενα εκπαίδευσης (αρνητικά και θετικά) και κρατήσαμε αυτές που είχαν information gain μεγαλύτερο ή ίσο του 0.008.

Έχουμε φτιάξει επίσης μια συνάρτηση classification\_report που τυπώνει recall, precision, f1-score και accuracy για κάθε κατηγορία, καθώς και το general accuracy του αλγορίθμου (τα τυπώνει όλα ως dataframe).

Τέλος, χρησιμοποιούμε και μία συνάρτηση draw\_diagram για να σχεδιάσουμε τα διαγράμματα για ότι χρειαζόμαστε. Η συνάρτηση αυτή δέχεται ως παραμέτρους το object που χρησιμοποιήσαμε για να υλοποιήσουμε τον κάθε αλγόριθμο του οποίου τα στατιστικά θέλουμε να σχεδιάσουμε (πχ nb για naïve bayes ή rf για random forest) και ένα string με το όνομα του στατιστικού που θέλουμε να σχεδιάσει (πχ accuracy, recall, precision κλπ.).

A screenshot of a computer

Description automatically generated

**Naïve Bayes algorithm:**

Για την υλοποίηση του αλγορίθμου αυτού χρησιμοποιήσαμε μια κλάση Probability, η οποία με βάση το παραπάνω dataframe για τα δεδομένα μας δημιουργεί 2 πίνακες 2Χ2 τους pX\_1 και pX\_0, όπου για παράδειγμα το pX\_1[0][2] περιέχει την πιθανότητα η 3η λέξη να είναι 1 όταν η κατηγορία είναι 0 και αντίστοιχα το pX\_0[0][2] περιέχει την πιθανότητα η 3η λέξη να είναι 0 όταν η κατηγορία είναι 0. Αυτό το κάνουμε έτσι ώστε να μην χρειάζεται να υπολογίζουμε κάθε φορά για κάθε νέα πρόβλεψη τις πιθανότητες της κάθε λέξης.

(Λέγοντας η λέξη να είναι 0 εννοούμε ότι δεν υπάρχει στο συγκεκριμένο παράδειγμα και αντίστοιχα για το 1, όπως επίσης κατηγορία 0 είναι η αρνητική κριτική και 1 η θετική)

**Stats of our implementation**

* Training data report

A screenshot of a computer

Description automatically generated

* Test data report

A screenshot of a computer

Description automatically generated

|  |  |
| --- | --- |
| Training Curve (Accuracy)  Αυτή η καμπύλη εκπαίδευσης δείχνει το συνολικό ποσοστό ορθότητας (accuracy) και στις δύο κατηγορίες μαζί (δηλαδή τη συνολική ορθότητα και σε θετικά και σε αρνητικά reviews) | Recall Curve (Negative category)  Τα recall των test data και training data ταυτίζονται και είναι σταθερά στο 1 |
| A graph on a white background  Description automatically generated |  |

|  |  |
| --- | --- |
| Precision curve (Negative category)  H καμπύλη έχει αυτήν την μορφή γιατί αρχικά τα πρώτα παραδείγματα που διαβάζουμε είναι positive και στα 12500 όπου τα positive τελειώνουν και ξεκινούν τα negative αρχίζει και ανεβαίνει το precision | F1- Curve (Negative Category) |
| A graph of a training curve  Description automatically generated with medium confidence |  |

**b)**

Σύγκριση των στατιστικών μας με τα στατιστικά της υλοποίησης του naïve bayes algorithm από την βιβλιοθήκη sklearn

Παρατηρούμε ότι τα στατιστικά της υλοποίησης μας σε σχέση με την υλοποίηση της sklearn είναι σχεδόν ίδια και στις δύο κατηγορίες (negative positive) σε όλα τα metrics (precision, recall, f1-score και accuracy)

A screenshot of a computer program

Description automatically generated

|  |  |
| --- | --- |
| Training curve (Accuracy): | Recall curve: |
| A graph of a graph  Description automatically generated with medium confidence | A white rectangular object with green lines  Description automatically generated |

|  |  |
| --- | --- |
| Precision curve: | F1 curve: |
| A graph of a training  Description automatically generated | A green line graph on a white background  Description automatically generated |

**Random Forest algorithm:**

Για την υλοποίηση του random forest αρχικά φτιάξαμε μία κλάση decision\_tree όπου κατασκευάζει τα δέντρα του δάσους. Έχει μία μέθοδο build\_tree η οποία κατασκευάζει το δέντρο λαμβάνοντας υπόψην της το max depth που δίνεται ως είσοδος από τον χρήστη. Η build\_tree καλεί την get\_best\_split η οποία διαλέγει κάθει φορά από την λίστα με τα διαθέσιμα features (τα οποία επιλέγονται τυχαία για κάθε δέντρο), με βάση ποιο feature θα συνεχιστεί η κατασκευή του δέντρου υπολογίζοντας για κάθε διαθέσιμο feature το ig του.

To random forest λοιπόν κατασκευάζει τόσα δέντρα όσα και η είσοδος που βάζει ο χρήστης και με όσα features θέλει επίσης ο χρήστης για το κάθε δέντρο.

**a)**

Χρησιμοποιήσαμε τα ίδια 27 features για τα δεδομένα μας (όπως στον naïve bayes παραπάων). Ως υπερπαραμέτρους χρησιμοποιήσαμε 150 δέντρα με 10 features για το κάθε ένα.

**Stats of our implementation:**

* **Training data report:**

**A screenshot of a computer

Description automatically generated**

* **Test data report:**

A screenshot of a computer

Description automatically generated

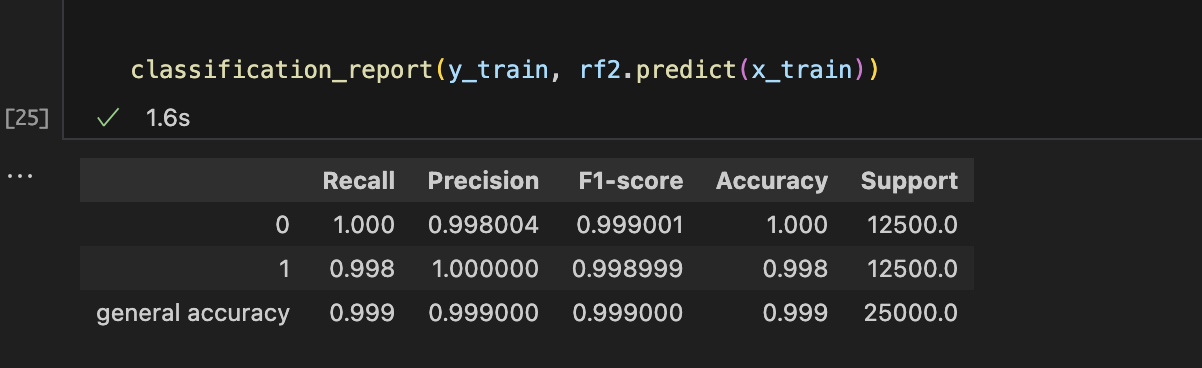
|  |  |
| --- | --- |
| Training curve: | Recall Curve: |
| A graph of a graph  Description automatically generated | A screen shot of a computer  Description automatically generated |

|  |  |
| --- | --- |
| Precision Curve: | F1-score Curve: |
| A graph of a training curve  Description automatically generated | A graph on a white background  Description automatically generated |

**b)**

Σύγκριση των στατιστικών μας με τα στατιστικά της υλοποίησης του random forest algorithm από την βιβλιοθήκη Sklearn

Παρατηρούμε ότι οι διαφορές στα στατιστικά της βιβλιοθήκης Sklearn σε σχέση με την δική μας υλοποίηση είναι πολύ μικρές, με την υλοποίηση της Sklearn να είναι περίπου κατά 2% πιο ακριβής, και στα test data και στα training data



A screen shot of a computer

Description automatically generated

|  |  |
| --- | --- |
| Training Curve (Accuracy): | Recall Curve: |
| A graph of a graph  Description automatically generated with medium confidence | A screen shot of a computer  Description automatically generated |

|  |  |
| --- | --- |
| Precision Curve: | F1-score Curve: |
| A graph of a training  Description automatically generated | A graph of a training  Description automatically generated |