ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Χειμερινό εξάμηνο 2021 -2022

2η Εργασία

|  |  |
| --- | --- |
| Καθηγητής | Ίων Ανδρουτσόπουλος |
| Συμμετέχον 1ος | Δημήτρης Καββαδάς |
| Αριθμός Μητρώου | 3190064 |
| Συμμετέχον 2ος | Αλέξης Φιλιππακόπουλος |
| Αριθμός Μητρώου | 3190212 |

# Contents

[Naive Bayes 2](#_bookmark0)

[Logistic Regression 5](#_bookmark2)

Naive Bayes

**Ζητούμενο**

Για την κατάταξη κειμένων στις 2 κατηγορίες(θετικές, αρνητικές) υλοποιήσαμε τον Naive Bayes αλγόριθμο σε πολυωνυμική μορφή με εκτιμήτρια Laplace κατά τον υπολογισμό τον prior πιθανοτήτων. Τις class conditional πιθανότητες τις υπολογίσαμε με την Gauss formula.

**Objective**

To rank the text files into 2 categories (positive, negative) we implemented the Naive Bayes algorithm from scratch with the Laplace criterion for the computation of the prior probabilities. The class conditional probabilities were computed according to the Gauss Formula.

**Επιλογή δεδομένων**

Καθώς τα δεδομένα έχουν γίνει random shuffle, η επιλογή των train δεδομένων είναι τυχαία, καθώς από τα 25.000 για την αναπαράσταση των καμπύλων έχω πάρει τα 12.500 train και όλα τα test δεδομένα.

**Data Selection**

Due to the data being randomly shuffled, the selection of the train data is random, so from the 25.000 train data available we selected 12.500 and used all the test data to generate the graphical representation.

**Υπερπαράμετροι**

Οι υπερπαράμετροι χρησιμοποιούνται καθώς αναπαριστώ όλα τα δεδομένα σε 0,1 (encrypt\_reviews(file\_path,dict,m,n)). Μετά από πολλές δοκιμές του αλγορίθμου.

Για το m που είναι η επιλογή των συχνότερων λέξεων δοκιμάσαμε τις τιμές: 500,1000,1500,3000

Και για το n που είναι η παράλειψη των n συχνότερων λέξεων δοκιμάσαμε τις τιμές: 1,10,15,20,30

Μεγιστοποίηση του accuracy των test δεδομένων γίνεται με:

m=1000 και n=15

**Super-Parameters**

The super parameters are being used since we depict all data in a 0,1 format (encrypt\_reviews(file\_path,dict,m,n)). Also the super parameter m (selection of the (m) most used words in a category) has the value of : 500,1000,1500,3000 and the super parameter n (ignorance of (n) most used words in a category)

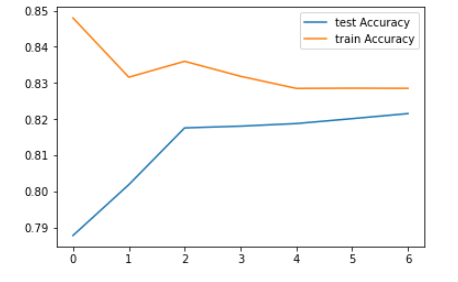
has the value of 1,10,15,20,30.The maximization of the test data accuracy occurs when m=1000 and n=15.

**Καμπύλες Μάθησης**

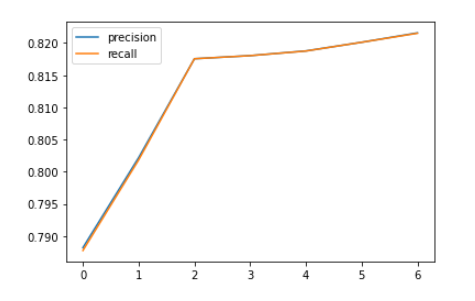
Παρατηρούμε ότι αυξάνοντας το πλήθος των train δεδομένων το accuracy για αυτά που χρησιμοποιήθηκαν μειώνεται , καθώς το αντίθετο γίνεται στα test δεδομένα.

**Learning Curves**

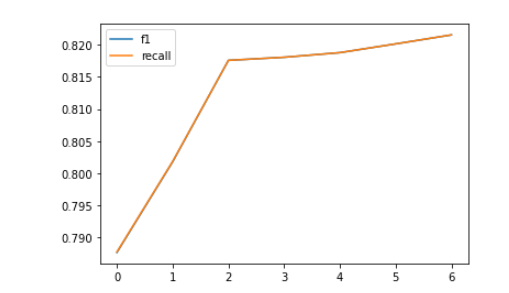
We concluded that increasing the number of the train data used, the accuracy of them decreases while exactly the opposite happens for the test data.



**Καμπύλη Precision-Recall**

****

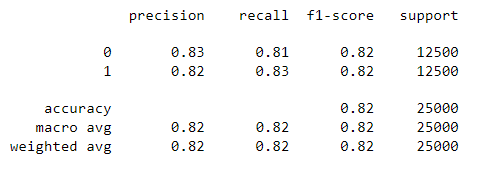
**Καμπύλη F1-Recall**

****

**Classification Report**

Εφαρμόζουμε το καλύτερο μοντέλο στο οποίο το accuracy μεγιστοποιείται με τις κατάλληλες υπερπαραμέτρους.

We use the best model where the overall data accuracy is maximized with the appropriate super parameters.



Logistic Regression

**Ζητούμενο**

Για την κατάταξη κειμένων στις 2 κατηγορίες(θετικές, αρνητικές) υλοποιήσαμε τον Logistic Regression αλγόριθμο

προσθέτοντας όρο κανονικοποίησης (λ) στην αντικειμενική συνάρτηση.

**Objective**

To rank the text files into 2 categories(positive, negative)

We implemented the Logistic Regression algorithm from scratch adding a normalization term (λ) into the objective function.

**Επιλογή δεδομένων**

Καθώς τα δεδομένα έχουν γίνει random shuffle, η επιλογή των train δεδομένων είναι τυχαία, καθώς από τα 25000 για την αναπαράσταση των καμπύλων έχω πάρει τα 12500 train και όλα τα test δεδομένα.

**Data Selection**

Due to the data being randomly shuffled, the selection of the train data is random, so from the 25.000 train data available we selectεδ 12.500 and used all the test data to generate the graphical representation.

**Υπερπαράμετροι**

Οι υπερπαράμετροι χρησιμοποιούνται καθώς αναπαριστώ όλα τα δεδομένα σε 0,1 (encrypt\_reviews(file\_path,dict,m,n)). Μετά από πολλές δοκιμές του αλγορίθμου.

Για το m που είναι η επιλογή των συχνότερων λέξεων δοκιμάσαμε τις τιμές: 500,1000,1500,3000

Και για το n που είναι η παράλειψη των n συχνότερων λέξεων δοκιμάσαμε τις τιμές: 1,10,15,20,30

Μεγιστοποίηση του accuracy των test δεδομένων γίνεται με:

m=1000 και n=15

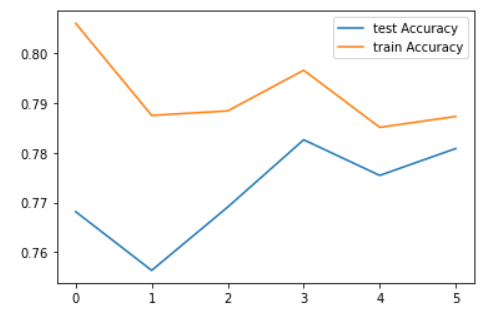
Tέλος ο όρος κανονικοποιήσης (λ) παίρνει την τιμή 0,001.

**Super-Parameters**

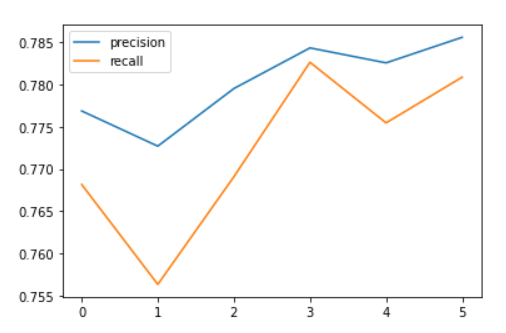
The super parameters are being used since we depict all data in a 0,1 format (encrypt\_reviews(file\_path,dict,m,n)). Also the super parameter m (selection of the (m) most used words in a category) has the value of : 500,1000,1500,3000 and the super parameter n (ignorance of (n) most used words in a category)

has the value of 1,10,15,20,30.The maximization of the test data accuracy occurs when m=1000 and n=15. Finally the normalization term (λ) takes the value of 0.001.

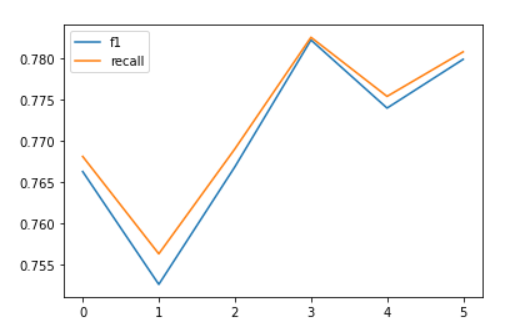
**Καμπύλες Μάθησης**



**Καμπύλη Precision-Recall**



**Καμπύλη F1-Recall**



**Classification Report**

Εφαρμόζουμε το μοντέλο με τις ίδιες παραμέτρους που βάλαμε και στον πρώτο αλγόριθμο

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Όλες οι δοκιμές των αλγορίθμων γίνανε με τις ίδιες τιμές στις υπερπαραμέτρους καθώς κάθε φορά για τις αναπαραστάσεις των καμπύλων χρησιμοποιήσαμε και πάλι των ίδιο αριθμό train δεδομένων και όλα τα test δεδομένα.