Οικονομικό Πανεπιστήμιο Αθηνών, Τμήμα Πληροφορικής

Μάθημα: Τεχνητή Νοημοσύνη

Ακαδημαϊκό έτος: 2020–21

2η Προγραμματιστική εργασία

Παράδοση ως 31-01-2021

Μέλη:

Κυρμπάτσος Παναγιώτης -Χρήστος (ΑΜ:3180226)

Μεϊδάνης Γεώργιος- Στέφανος (ΑΜ:3170107)

**Περιγραφή Προβλήματος**

Το πρόβλημά μας είναι η δημιουργία ενός αλγορίθμου που θα επιβλέπει την αυτόματη κατάταξη δεδομένων εισόδου σε μία από δύο ξένες μεταξύ τους κατηγορίες γνωστό και ως binary classification. Αποτελεί ένα εξαιρετικά συχνό πρόβλημα σήμερα, του οποίου οι εφαρμογές περιλαμβάνουν αλλά δεν περιορίζονται σε: αναγνώριση ανεπιθύμητων/ σημαντικών μηνυμάτων στο email (spam/ham) , έλεγχο και επισήμανση υβριστικών σχολίων ή αναγνώριση μιας ασθένειας από κάποιο διαγνωστικά δεδομένα.Πιο συγκεκριμένα, σκοπός της εν λόγω εργασίας είναι η κατάρτιση αλγορίθμου κατάλληλου για binary classification στην βάση θετικών και αρνητικών κριτικών του IMDB.

**Στοχοθεσία**

Στόχος μας είναι να δημιουργήσουμε 2 αλγορίθμους μηχανικής μάθησης οι οποίοι θα μπορούν να προβλέπουν με την μεγαλύτερη δυνατή ακρίβεια το σκορ των αξιολογήσεων(<=4 αρνητικό, >=7 θετικό). Θα παρέχουμε δεδομένα εκμάθησης ( που θα εκπαιδεύουν τον αλγόριθμο) και θα εξετάσουμε την ευστοχία του σε δεδομένα ελέγχου. Επιλέξαμε τους αλγορίθμους Logistic Regression και Naive Bayes.

Ως δεδομένα εκπαίδευσης χρησιμοποιούνται τα ήδη ταξινομημένα σε αρνητικά και θετικά σχόλια που παρέχονται από το αρχείο της βάσης σχολίων.Ως δεδομένα ελέγχου χρησιμοποιούμε το 20% των εκάστοτε δεδομένων εκπαίδευσης με τυχαία επιλογή.

**Υπερ Παράμετροι**

Παράλειψη n πρώτων λέξεων:Από στατιστικό έλεγχο που συμφωνεί με την διαίσθησή μας αποφασίζουμε να αγνοήσουμε τις πρώτες 48 λέξεις εξαρχής καθώς περιλαμβάνουν μέρη του λόγου που δεν συνεισφέρουν θετικά ή αρνητικά στην εξαγωγή γνώμης για τις κριτικές.

m συχνότερες λέξεις: Σαν αρχική τιμή με την οποία έτρεξαν τα παρακάτω πειράματα έχουμε υιοθετήσει το άνω όριο των 1000 λέξεων του λεξιλογίου.Δίνεται η δυνατότητα στον χρήστη να ορίσει και τις δύο τιμές

**Αλγόριθμος Logistic Regression**

Με δοκιμές στο πρώτο δεκαδικό του ορίου πρόβλεψης σε Logistic Regression καταλήξαμε πως η τιμή 0.5 οδηγεί σε βέλτιστα αποτελέσματα. Ο αλγόριθμος δοκιμάστηκε και με διαφορετικά ποσοστά για τις προβλέψεις του και τα αποτελέσματα ορθότητας ηταν 0.802439024390244 (0.6) και 0.7975609756097561 (0.4)

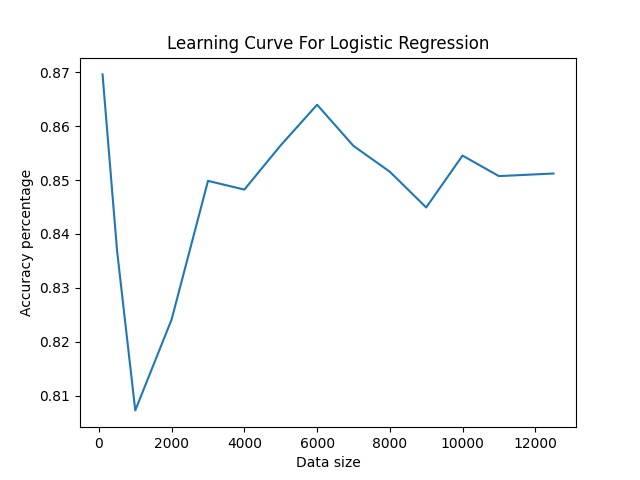
Για τις υπόλοιπες παραμέτρους του logistic Regression τρέξαμε δοκιμές σε συγκεκριμένο αριθμό δεδομένων και καταλήξαμε ότι τα καλύτερα αποτελέσματα(με προτεραιότητα το accuracy) προκύπτουν με lr=0.03 και iterations= 5000. Ακολουθεί πίνακας με τα accuracy από τις δοκιμές μας, και πλήρη σκορ απο τις καλύτερες.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| lr/ iters | 100 | 1000 | 5000 | 10000 |
| 0.0001 | 0.7731707317073171 | 0.775609756097561 | 0.7829268292682927 | 0.7853658536585366 |
| 0.001 | 0.775609756097561 | 0.7853658536585366 | 0.7878048780487805 | 0.7804878048780488 |
| 0.01 | 0.7853658536585366 | 0.7804878048780488 | 0.7975609756097561 | **0.8073170731707318** |
| 0.02 | 0.7829268292682927 | 0.7878048780487805 | **0.8073170731707318** | 0.8048780487804879 |
| 0.03 | 0.7829268292682927 | 0.7829268292682927 | **0.8073170731707318** | 0.8048780487804879 |
| 0.05 | 0.7853658536585366 | 0.7975609756097561 | 0.802439024390244 | 0.8048780487804879 |
| 0.8 | 0.802439024390244 | 0.802439024390244 | 0.7829268292682927 | 0.7780487804878049 |

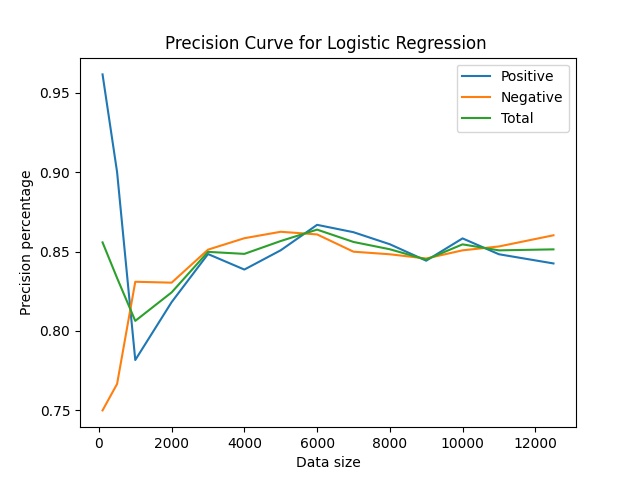
|  |  |  |  |
| --- | --- | --- | --- |
| score/ lr | 0.02(5000 iterns) | 0.03(5000 iters) | 0.01(10000 iters) |
| accuracy | 0.8073170731707318 | 0.8073170731707318 | 0.8073170731707318 |
| pos. Precision | 0.7761194029850746 | 0.7817258883248731 | 0.7761194029850746 |
| Pos Recall | 0.8210526315789474 | 0.8105263157894737 | 0.8210526315789474 |
| Neg Precision | 0.8373205741626795 | 0.8309859154929577 | 0.8373205741626795 |
| Neg Recall | 0.7954545454545454 | 0.8045454545454546 | 0.7954545454545454 |
| F1 | 0.8074860603820718 | 0.8069454621708263 | 0.8074860603820718 |

**Καμπύλες και πίνακες**

* **Μάθησης(ορθότητας)**



|  |  |
| --- | --- |
| Data size\Algorithm | Logistic Regression accuracy |
| 100 | 0.4782608695652174 |
| 500 | 0.8105263157894737 |
| 1000 | 0.8121951219512196 |
| 2000 | 0.8241758241758241 |
| 3000 | 0.8311688311688312 |
| 4000 | 0.8317013463892289 |
| 5000 | 0.8220338983050848 |
| 6000 | 0.8344566133108677 |
| 7000 | 0.8205776173285199 |
| 8000 | 0.8236032642812304 |
| 9000 | 0.8193440800444691 |
| 10000 | 0.8234261349385503 |
| 11000 | 0.8245815180004586 |
| 12500 | 0.8232931726907631 |

* **Ακρίβειας**

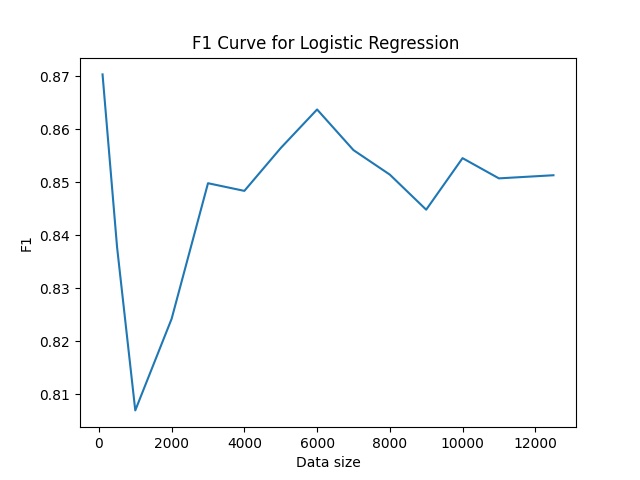
|  |  |  |  |
| --- | --- | --- | --- |
| Data size\Algorithm | Θετική | Αρνητική | Συνολική |
| 100 | 0.9615384615384616 | 0.75 | 0.8557692308 |
| 500 | 0.9 | 0.7666666666666667 | 0.8333333333 |
| 1000 | 0.7817258883248731 | 0.8309859154929577 | 0.8063559019 |
| 2000 | 0.8181818181818182 | 0.830423940149626 | 0.8243028792 |
| 3000 | 0.8484349258649094 | 0.8512 | 0.8498174629 |
| 4000 | 0.8386714116251482 | 0.8584070796460177 | 0.8485392456 |
| 5000 | 0.8508180943214629 | 0.8624612202688728 | 0.8566396573 |
| 6000 | 0.8668280871670703 | 0.8607929515418502 | 0.8638105194 |
| 7000 | 0.8621883656509696 | 0.8499245852187028 | 0.8560564754 |
| 8000 | 0.8546120952962737 | 0.8482892188508715 | 0.8514506571 |
| 9000 | 0.8442534908700322 | 0.8456221198156681 | 0.8449378053 |
| 10000 | 0.8582914572864322 | 0.8507761642463696 | 0.8545338108 |
| 11000 | 0.8483069977426636 | 0.853215284249767 | 0.850761141 |
| 12500 | 0.84251968503937 | 0.8602459016393442 | 0.8513827933 |

* **Ανάκλησης**

****

|  |  |  |  |
| --- | --- | --- | --- |
| Data size\Algorithm | Θετική | Αρνητική | Συνολική |
| 100 | 0.8333333333333334 | 0.9375 | 0.8854166667 |
| 500 | 0.8108108108108109 | 0.8734177215189873 | 0.8421142662 |
| 1000 | 0.8105263157894737 | 0.8045454545454546 | 0.8075358852 |
| 2000 | 0.8341463414634146 | 0.8141809290953546 | 0.8241636353 |
| 3000 | 0.8470394736842105 | 0.8525641025641025 | 0.8498017881 |
| 4000 | 0.8632478632478633 | 0.8331288343558282 | 0.8481883488 |
| 5000 | 0.8692232055063913 | 0.8432760364004045 | 0.856249621 |
| 6000 | 0.8717532467532467 | 0.8555166374781086 | 0.8636349421 |
| 7000 | 0.8621883656509696 | 0.8499245852187028 | 0.8560564754 |
| 8000 | 0.8561811505507956 | 0.8466494845360825 | 0.8514153175 |
| 9000 | 0.8543478260869565 | 0.8350398179749715 | 0.844693822 |
| 10000 | 0.8514456630109671 | 0.8576476527006562 | 0.8545466579 |
| 11000 | 0.8564266180492252 | 0.8449469312413475 | 0.8506867746 |
| 12500 | 0.8625554212011286 | 0.8399359743897559 | 0.8512456978 |

* **F1**



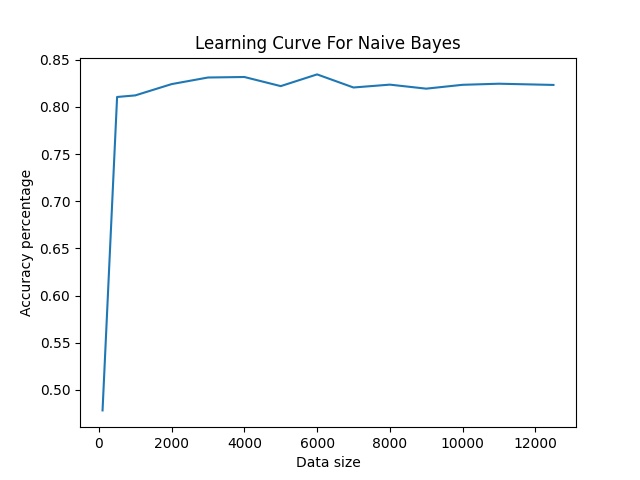
|  |  |
| --- | --- |
| Data size\Algorithm | F1 |
| 100 | 0.8703405430280718 |
| 500 | 0.8377007895453308 |
| 1000 | 0.8069454621708263 |
| 2000 | 0.8242332513416757 |
| 3000 | 0.8498096254560247 |
| 4000 | 0.8483637609345788 |
| 5000 | 0.8564445947173415 |
| 6000 | 0.8637227218122498 |
| 7000 | 0.8560564754348362 |
| 8000 | 0.8514329869418057 |
| 9000 | 0.8448157960712757 |
| 10000 | 0.8545402342628208 |
| 11000 | 0.8507239561955602 |
| 12500 | 0.8513142400479372 |

**Naive Bayes**

Ο αλγόριθμος Bayes παρουσιάζει καλύτερο σκορ αν αγνοήσουμε τις 50 συχνότερες λέξεις (ποσοστό ορθότητας 0.8121951219512196) αντί για τις πρώτες 48 ( 0.8073170731707318) , ενώ κρατάμε το upper Limit σταθερο (1000)

**Καμπύλες και πίνακες**

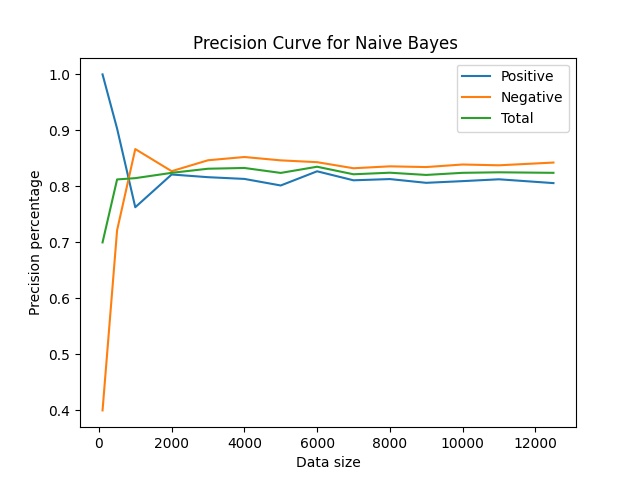
* **Μάθησης**



|  |  |
| --- | --- |
| Data size\Algorithm | Naive Bayes accuracy |
| 100 | 0.8695652173913043 |
| 500 | 0.8368421052631579 |
| 1000 | 0.8073170731707318 |
| 2000 | 0.8241758241758241 |
| 3000 | 0.8498376623376623 |
| 4000 | 0.8482252141982864 |
| 5000 | 0.8564307078763709 |
| 6000 | 0.8639427127211458 |
| 7000 | 0.8563176895306859 |
| 8000 | 0.8515379786566227 |
| 9000 | 0.8449138410227904 |
| 10000 | 0.854527213443692 |
| 11000 | 0.8507223113964687 |
| 12500 | 0.8512048192771084 |

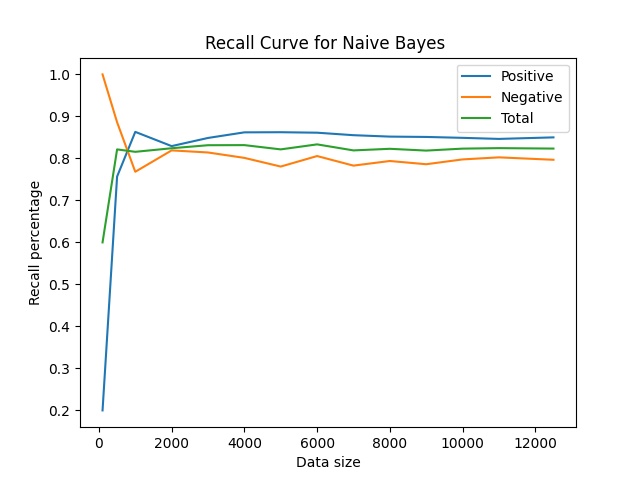
* **Ακρίβειας**

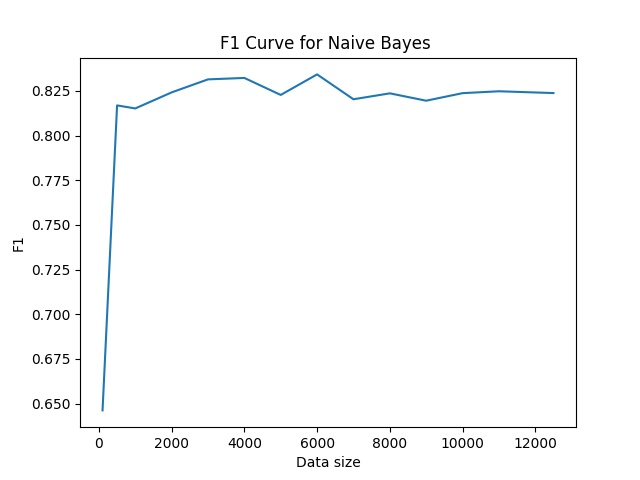
|  |  |  |  |
| --- | --- | --- | --- |
| Data size\Precision | Θετική | Αρνητική | Συνολική |
| 100 | 1.0 | 0.4 | 0.7 |
| 500 | 0.9032258064516129 | 0.7216494845360825 | 0.8124376455 |
| 1000 | 0.7627906976744186 | 0.8666666666666667 | 0.8147286822 |
| 2000 | 0.821256038647343 | 0.8271604938271605 | 0.8242082662 |
| 3000 | 0.8164556962025317 | 0.8466666666666667 | 0.8315611814 |
| 4000 | 0.8133640552995391 | 0.8524804177545692 | 0.8329222365 |
| 5000 | 0.8016453382084096 | 0.8464912280701754 | 0.8240682831 |
| 6000 | 0.8269680436477007 | 0.843263061411549 | 0.8351155525 |
| 7000 | 0.8108995403808273 | 0.8323977546110666 | 0.8216486475 |
| 8000 | 0.8130841121495327 | 0.835820895522388 | 0.8244525038 |
| 9000 | 0.8063851699279093 | 0.8345410628019324 | 0.8204631164 |
| 10000 | 0.8094106463878327 | 0.8390865639936272 | 0.8242486052 |
| 11000 | 0.812691466083151 | 0.8376685934489403 | 0.8251800298 |
| 12500 | 0.8058846006878104 | 0.8425730004231908 | 0.8242288006 |

****

* **Ανάκλησης**

|  |  |  |  |
| --- | --- | --- | --- |
| Data size\Precision | Θετική | Αρνητική | Συνολική |
| 100 | 0.2 | 1.0 | 0.6 |
| 500 | 0.7567567567567568 | 0.8860759493670886 | 0.8214163531 |
| 1000 | 0.8631578947368421 | 0.7681818181818182 | 0.8156698565 |
| 2000 | 0.8292682926829268 | 0.8190709046454768 | 0.8241695987 |
| 3000 | 0.8486842105263158 | 0.8141025641025641 | 0.8313933873 |
| 4000 | 0.8620268620268621 | 0.8012269938650307 | 0.8316269279 |
| 5000 | 0.8623402163225172 | 0.7805864509605662 | 0.8214633336 |
| 6000 | 0.8612012987012987 | 0.8056042031523643 | 0.8334027509 |
| 7000 | 0.8552631578947368 | 0.7828054298642534 | 0.8190342939 |
| 8000 | 0.8518971848225214 | 0.7938144329896907 | 0.8228558089 |
| 9000 | 0.8510869565217392 | 0.7861205915813424 | 0.8186037741 |
| 10000 | 0.8489531405782652 | 0.7975769813225644 | 0.823265061 |
| 11000 | 0.8463992707383774 | 0.8024919243193355 | 0.8244455975 |
| 12500 | 0.8500604594921403 | 0.79671868747499 | 0.8233895735 |



* **F1**

|  |  |
| --- | --- |
| Data size\Algorithm | F1 |
| 100 | 0.6461538461538462 |
| 500 | 0.8169023284108236 |
| 1000 | 0.8151989976607914 |
| 2000 | 0.8241889319971957 |
| 3000 | 0.8314772759092051 |
| 4000 | 0.8322740782488872 |
| 5000 | 0.8227637465151494 |
| 6000 | 0.8342582725981268 |
| 7000 | 0.8203393877609534 |
| 8000 | 0.8236533825527143 |
| 9000 | 0.8195323905978735 |
| 10000 | 0.8237565394890098 |
| 11000 | 0.824812650158626 |
| 12500 | 0.8238089732861558 |

Τέλος, Ένα διάγραμμα με τις καμπύλες ακρίβειας των 2 αλγορίθμων.

