Περιεχόμενα:

1) Λίγα λόγια για τον αλγόριθμους

Αλγόριθμος 1: Αφελής ταξινομητής Bayes, πολυμεταβλητή μορφή Bernoulli

Ο Naive Bayes είναι ο ταξινομητής που χρησιμοποιεί το θεώρημα Bayes. Προβλέπει τις πιθανότητες για κάθε κλάση, δηλαδή την πιθανότητα ένα συγκεκριμένο δεδομένο να ανήκει σε μία κλάση (0 ή 1 στην περίπτωση του Bernoulli). Η κλάση με την μεγαλύτερη πιθανότητα θεωρείται η πιο πιθανή κλάση.

Αλγόριθμος 2: ID3

Με λίγα λόγια, ένα δέντρο απόφασης είναι μία δομή από κόμβους και ακμές και δομείται από μία βάση δεδομένων. Κάθε κόμβος χρησιμοποιείται για να ληφθεί μία απόφαση είτε για να παραστήσει κάποιο αποτέλεσμα.
Συγκεκριμένα, αυτό που κάνει ο id3 είναι να διχοτομεί διαρκώς ιδιότητες σε μία ή περισσότερες ομάδες σε κάθε βήμα.
Ο id3 χρησιμοποιεί **top-down greedy** προσέγγιση για να χτίσει ένα δένδρο απόφασης. Με λίγα λόγια, η **top-down** προσέγγιση σημαίνει ότι ξεκινάμε να χτίζουμε το δένδρο από την αρχή και **greedy** προσέγγιση σημαίνει ότι σε κάθε επανάληψη επιλέγουμε την καλύτερη ιδιότητα στην δεδομένη στιγμή για να δημιουργήσουμε έναν κόμβο.

Αλγόριθμος 3: Λογιστική Παλινδρόμηση με στοχαστική ανάβαση κλίσης

Ο αλγόριθμος Logistic regression χρησιμοποιεί μία εξίσωση για την αναπαράσταση. Τα input values (x) της βάσης δεδομένων συνδυάζονται γραμμικά χρησιμοποιώντας βάρη με σκοπό την πρόβλεψη μιας output value (y).

2) Αρχιτεκτονική προγράμματος

Το notebook αποτελείτε από τα κελιά Imports, Processing data functions, Learning curve and table functions, Naive Bayes class, ID3 class, Logistic Regression class, Main και Comparison with sklearn built in algorithms .Processing data functions:

Proccessing data functions: Περιέχει τις μεθόδους επεξεργασίας των δεδομένων. Αυτές είναι:

Η μέθοδος fetch( ) με την οποία φορτώνονται όλα τα δεδομένα του imdb dataset σε λίστες x_train, y_train για τα δεδομένα εκπαίδευσης και x_test, y_test για τα δεδομένα ελέγχου.

Η μέθοδος vocabulary( ) η οποία δέχεται την λίστα x_train και εξάγει το

λεξιλόγιο των δεδομένων εκπαίδευσης.

Η μέθοδος binary_vectors( ) η οποία δέχεται την λίστα x_train/x_test και το λεξιλόγιο και εξάγει το αντίστοιχο διάνυσμα εκπαίδευσης/ ελέγχου. Το διάνυσμα αποτελείτε από λίστες για κάθε παράδειγμα εκπαίδευσης/ελέγχου με 1 και 0 που ενδεικνύουν την ύπαρξη ή όχι της εκάστοτε λέξης του λεξιλογίου μέσα στο αντίστοιχο παράδειγμα εκπαίδευσης/ελέγχου.

Learning curve and table functions:

Περιέχει τις μεθόδους δημιουργίας των μετρήσεων accuracy, precision, recall, f1 score. Αυτές καλούνται μία προς μία μέσω της μεθόδου calculate( ). Οι μετρήσεις αυτές παράγονται συγκρίνωντας τις εκτιμήσεις που προκύπτουν από τον αλγόριθμο μάθησης (λίστες y_pred_train/ y_pred_test) με τις πραγματικές τιμές των λιστών y_train/y_test.

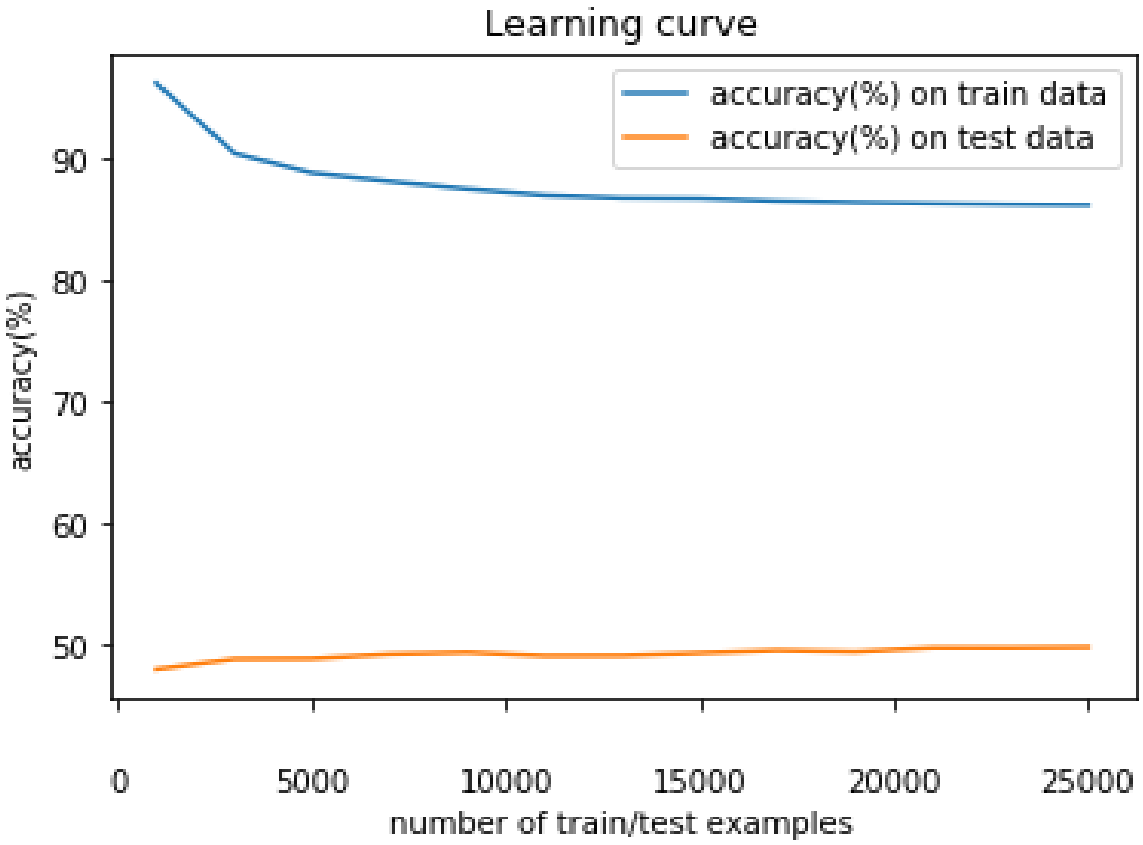
Περιέχει επίσης τις μεθόδους learningcurve( ) και table( ) reform( ). Η reform( ) δεχεται τις μετρήσεις των προαναφερθέντων μεθόδων και τις μορφοποιεί σε μορφή πίνακα. Έπειτα η learningcurve( ) δέχεται την δοσμένη ιδιότητα (accuracy, precision, recall, f1) και για τα δεδομένα εκπαίδευσης και για τα δεδομένα ελέγχου καθώς και τον αντιστοιχο αριθμό παραδειγμάτων εκπαίδευσης/ελέγχου και εξάγει τα δεδομένα αυτά σε μορφή καμπυλών μάθησης συναρτήσει του αριθμού παραδειγμάτων εκπαίδευσης. Αντιστοιχα, η μέθοδος table( ) εξάγει τα ίδια δεδομένα σε μορφή πίνακα.

Main:

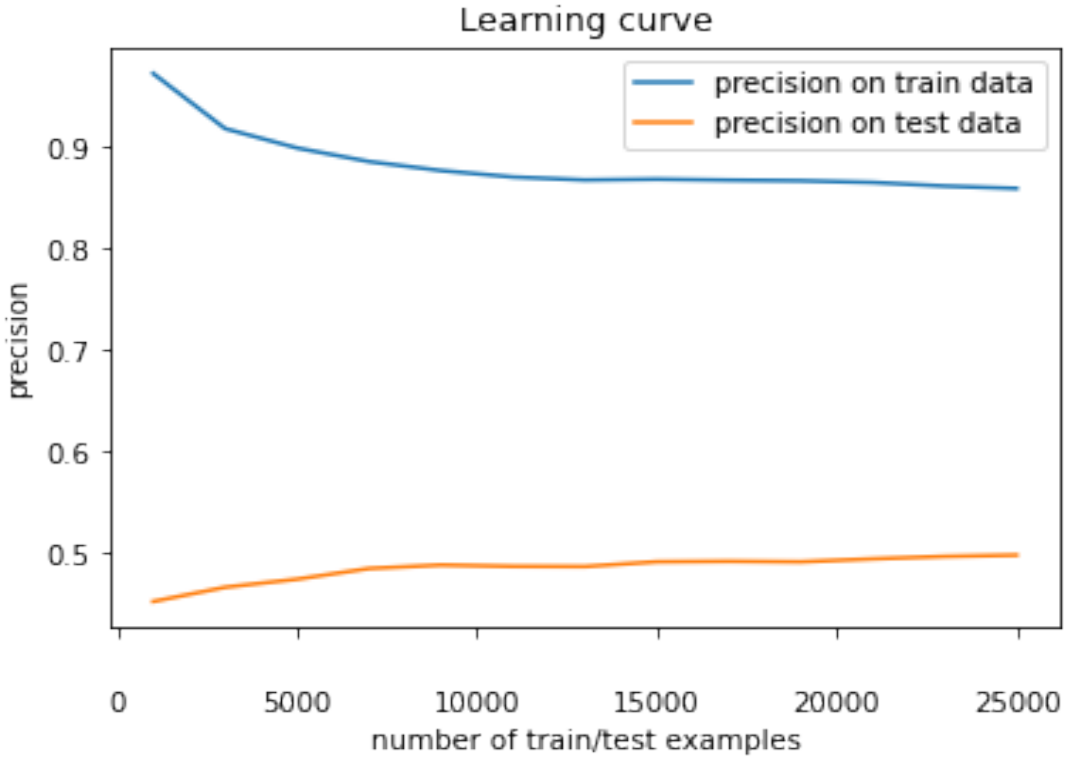
Στην main καλούνται όλες οι μέθοδοι επεξεργασίας δεδομένων (fetch( ), vocabulary( ), binary_vectors( )), γίνεται επιλογή του αλγορίθμου μάθησης που επιθυμεί να χρησιμοποιήσει ο χρήστης, καλείται η αντίστοιχη κλάση αλγορίθμου, γίνεται fitting των δεδομένων εκπαίδευσης στον επιλεγμένο αλγόριθμο και δημιουργούνται οι λίστες εκτιμήσεων για τα δεδομένα εκπαίδευσης και ελέγχου (y_pred_train, y_pred_test) μέσω της μεθόδου predict( ) του εν χρήση αλγορίθμου . Η παραπάνω διαδικασίες εκτελούνται τόσες φορές όσος είναι και ο αριθμός των παραδειγμάτων εκπαίδευσης/ελέγχου. Παράλληλα καλείται η calculate( ) για τον υπολογισμό τον διαφόρων μετρήσεων για κάθε παράδειγμα εκπαίδευσης/ελέγχου. Τέλος όταν ολοκληρωθούν όλα τα παραδείγματα εκπαίδευσης, καλείται η results( ) για την εκτύπωση των μετρήσεων σε μορφή πινάκων και καμπυλών μάθησης.

3) Παραδείγματα χρήσης προγράμματος

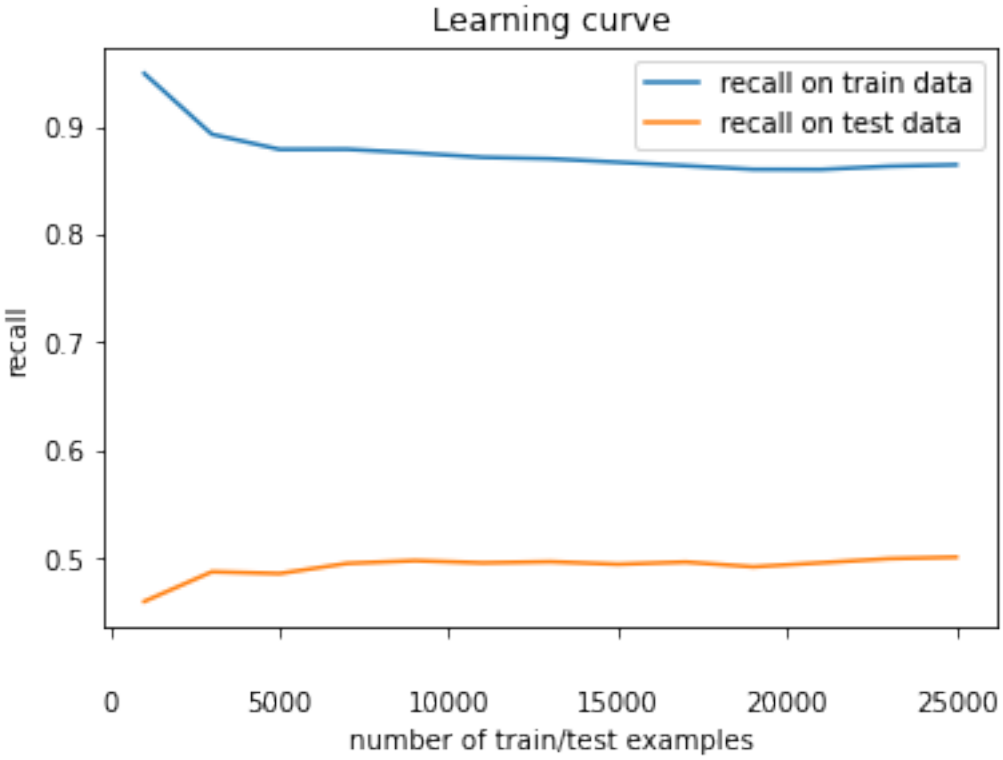
Καμπύλες μάθησης και πίνακες του Naive Bayes: num_words = 1000



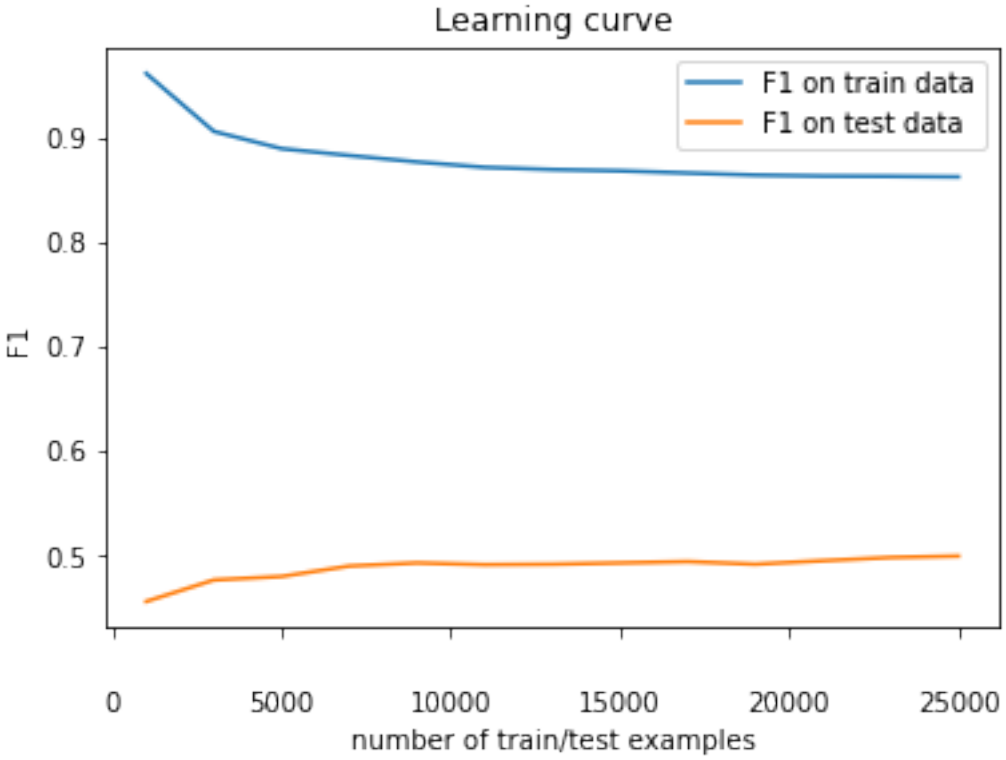
number of examples	accuracy(%) on train data	accuracy(%) on test data
1000	96.2	48.0
3000	90.43333333333334	48.833333333333336
5000	88.84	48.86
7000	88.14285714285714	49.22857142857143
9000	87.53333333333333	49.36666666666667
11000	87.00909090909092	49.09090909090909
13000	86.78461538461538	49.10769230769231
15000	86.74666666666667	49.32
17000	86.52352941176471	49.54117647058823
19000	86.38947368421053	49.42631578947368
21000	86.3	49.74761904761905
23000	86.22173913043478	49.75217391304348
25000	86.176	49.8



number of examples	precision on train data	precision on test data
1000	0.9730290456431535	0.45228215767634855
3000	0.9186124082721815	0.46631087391594395
5000	0.8995176848874598	0.4742765273311897
7000	0.886111900028401	0.48480545299630784
9000	0.8772316508706194	0.4882080670046286
11000	0.8708806050783361	0.4869439942373492
13000	0.867964709461515	0.4867660480681473
15000	0.8687325442213061	0.4914217316132464
17000	0.8678093449166471	0.4918995069265086
19000	0.867216407654086	0.49138386721640764
21000	0.8655518394648829	0.494314381270903
23000	0.8619674402493939	0.4967093869068237
25000	0.8596309845713377	0.49801177031970734

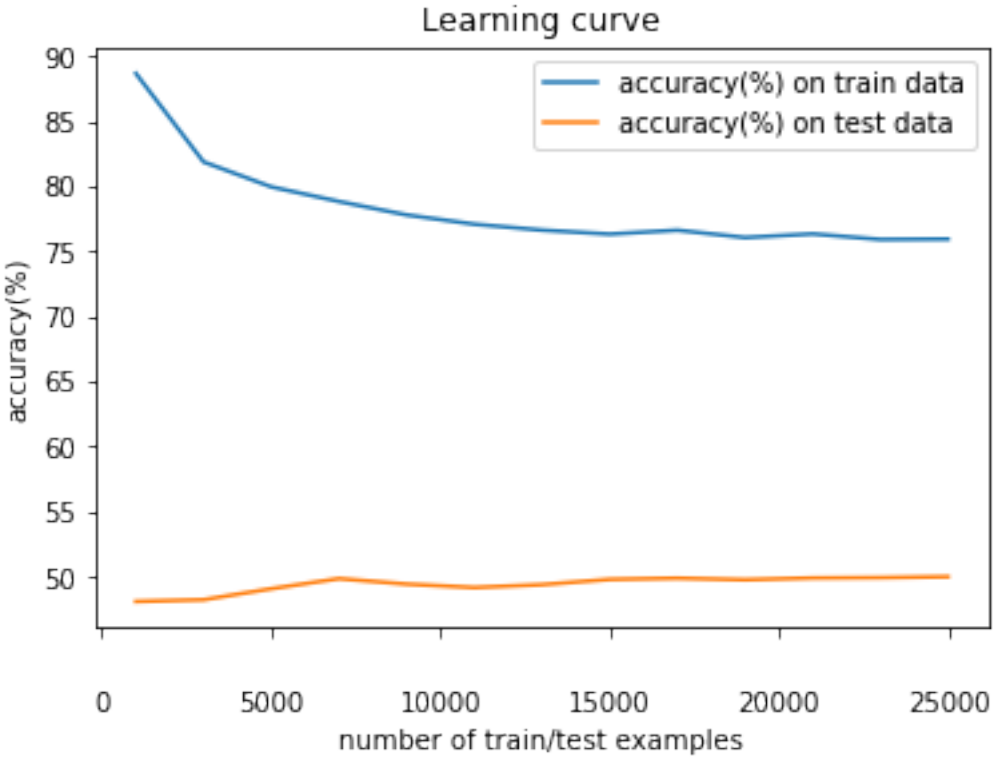


number of examples	recall on train data	recall on test data
1000	0.9493927125506073	0.459915611814346
3000	0.8929961089494164	0.4874476987447699
5000	0.8790259230164965	0.4857966241251544
7000	0.8791208791208791	0.49521322889469105
9000	0.8756875687568757	0.49775280898876406
11000	0.8716654650324441	0.49569202566452797
13000	0.8703477730323368	0.49674014281279105
15000	0.8671180140714191	0.49444667469557074
17000	0.8638541544933972	0.49644549763033174
19000	0.8604846323298018	0.4920601312724963
21000	0.8603723404255319	0.4957830170596128
23000	0.8633879781420765	0.4996080480794356
25000	0.86472	0.50096

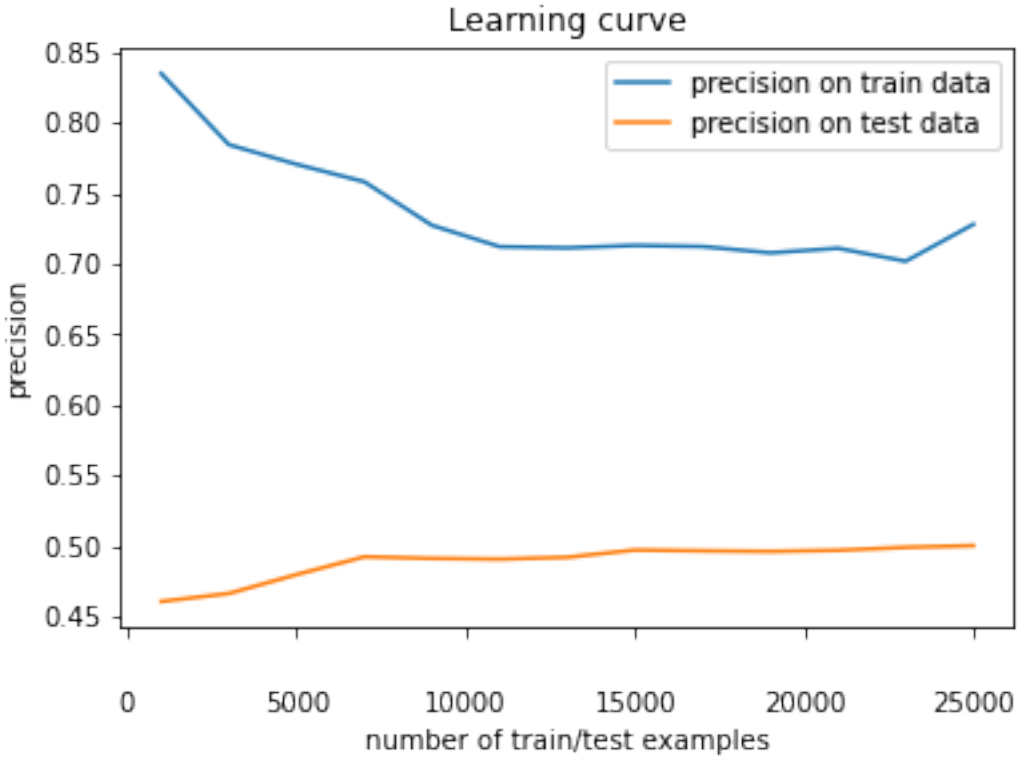


number of examples	F1 on train data	F1 on test data
1000	0.9610655737704918	0.4560669456066946
3000	0.9056231502795133	0.47664507330378453
5000	0.8891537544696066	0.47996745983323164
7000	0.8826025459688825	0.4899540757749713
9000	0.876458929751156	0.4929342383442751
11000	0.8712728583010539	0.4912790697674419
13000	0.8691546077684691	0.49170251997541486
15000	0.8679245283018869	0.4929295624332977
17000	0.8658272327964861	0.49416204741125136
19000	0.8638374052232518	0.4917217667283787
21000	0.8629543181060354	0.4950476099334897
23000	0.8626771244095853	0.49815450084675844
25000	0.8621679827709979	0.4994815346574141

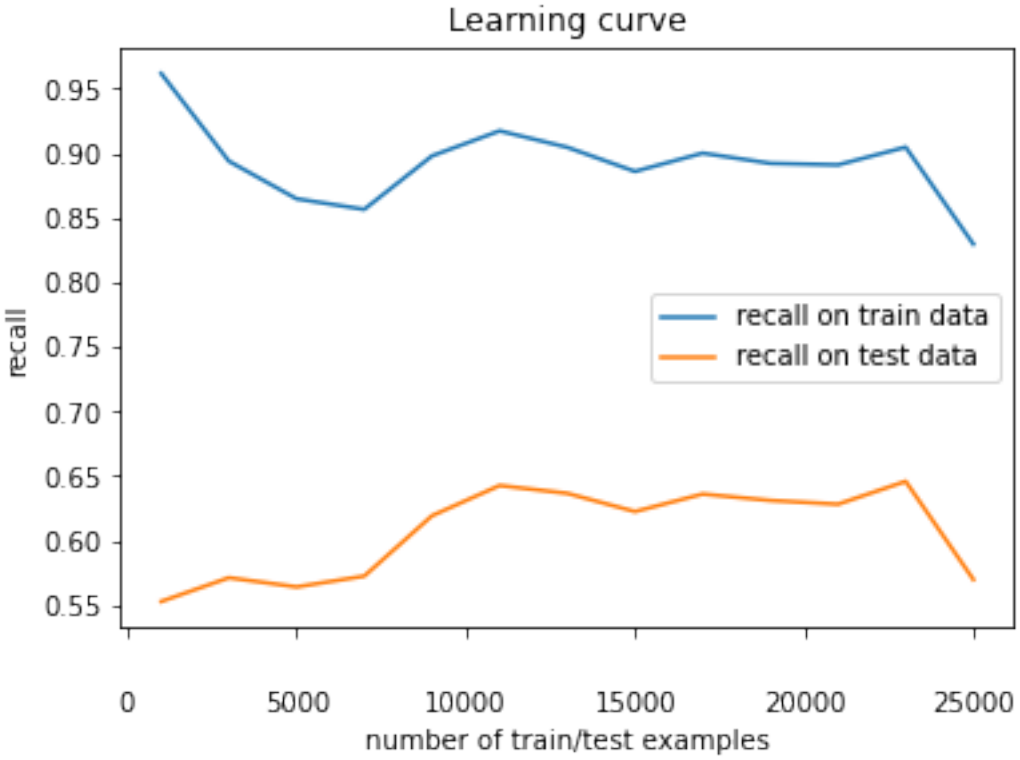
Καμπύλες μάθησης και πίνακες του ID3: num_words = 400



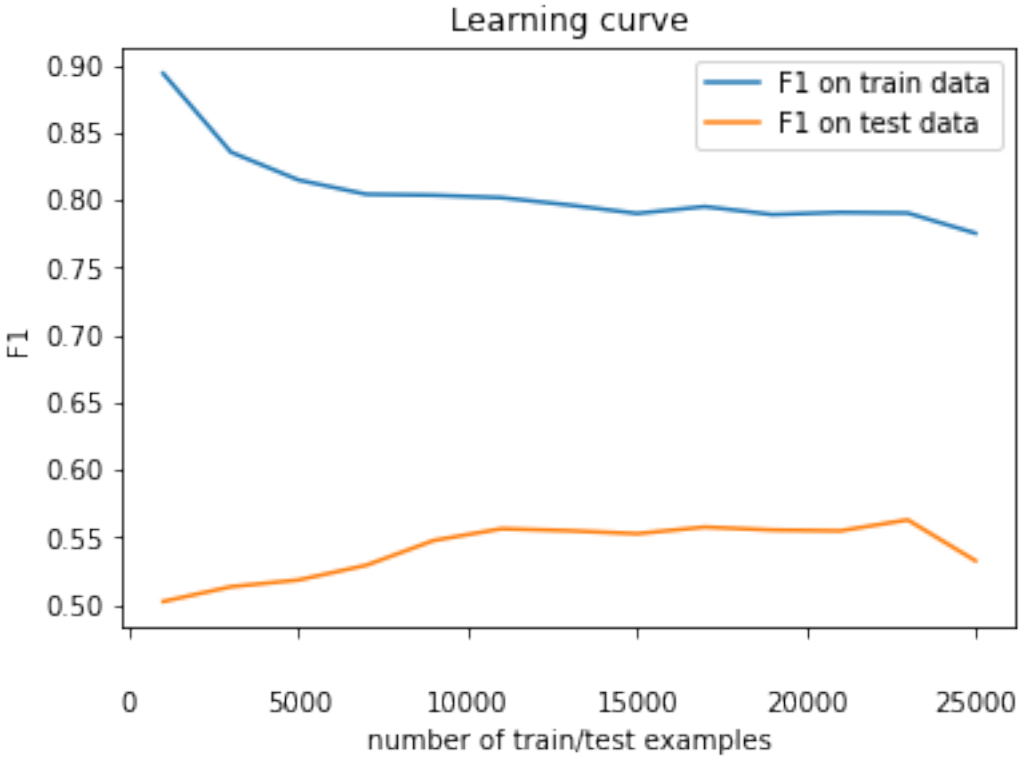
number of examples	accuracy(%) on train data	accuracy(%) on test data
1000	88.7	48.1
3000	81.9	48.233333333333334
5000	79.98	49.08
7000	78.85714285714286	49.857142857142854
9000	77.82222222222222	49.43333333333333
11000	77.10909090909091	49.17272727272727
13000	76.65384615384616	49.39230769230769
15000	76.34666666666666	49.8
17000	76.6470588235294	49.88823529411765
19000	76.08421052631579	49.77368421052632
21000	76.36190476190477	49.904761904761905
23000	75.93478260869566	49.93478260869565
25000	75.964	50.004



number of examples	precision on train data	precision on test data
1000	0.8347978910369068	0.46045694200351495
3000	0.7842914058053501	0.46613545816733065
5000	0.7703885194259713	0.47952397619880993
7000	0.7580444000997755	0.4921426789723123
9000	0.7271431117447871	0.4909998217786491
11000	0.7120067170445005	0.4904841869577386
13000	0.7111670864819479	0.4917836152093079
15000	0.7128966769954055	0.4969548028635538
17000	0.7121057985757884	0.49634698973457875
19000	0.7075809270200549	0.49596405092785223
21000	0.7108853850818678	0.49674044875682233
23000	0.7016552280985062	0.4988561431839591
25000	0.72786632029769	0.5000351049638418

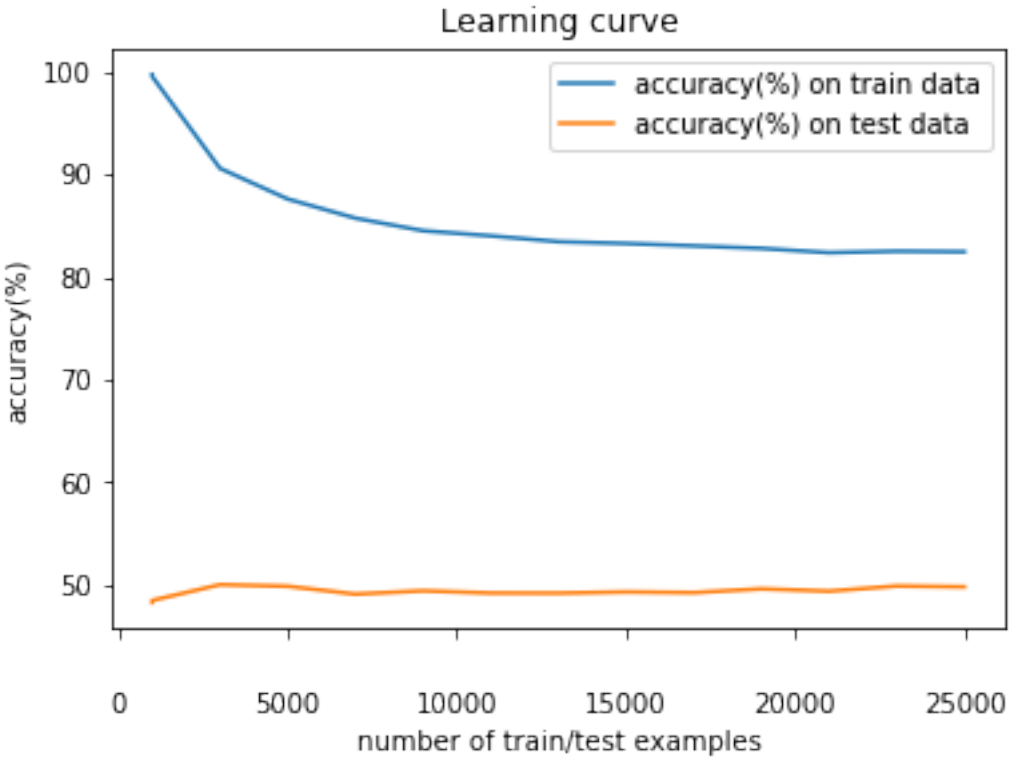


number of examples	recall on train data	recall on test data
1000	0.9615384615384616	0.5527426160337553
3000	0.893644617380026	0.5711297071129707
5000	0.8644933228593873	0.5640181144503911
7000	0.8562975486052409	0.5723817812590658
9000	0.8976897689768977	0.6191011235955056
11000	0.9170872386445565	0.6425297891842346
13000	0.9043624161073825	0.6364483079788885
15000	0.8857029072082836	0.622373879298809
17000	0.8998480775972888	0.6359004739336492
19000	0.8919542641351096	0.6309549015456278
21000	0.8907674772036475	0.6280429365535749
23000	0.9045016913869373	0.6457625642365648
25000	0.82936	0.56976

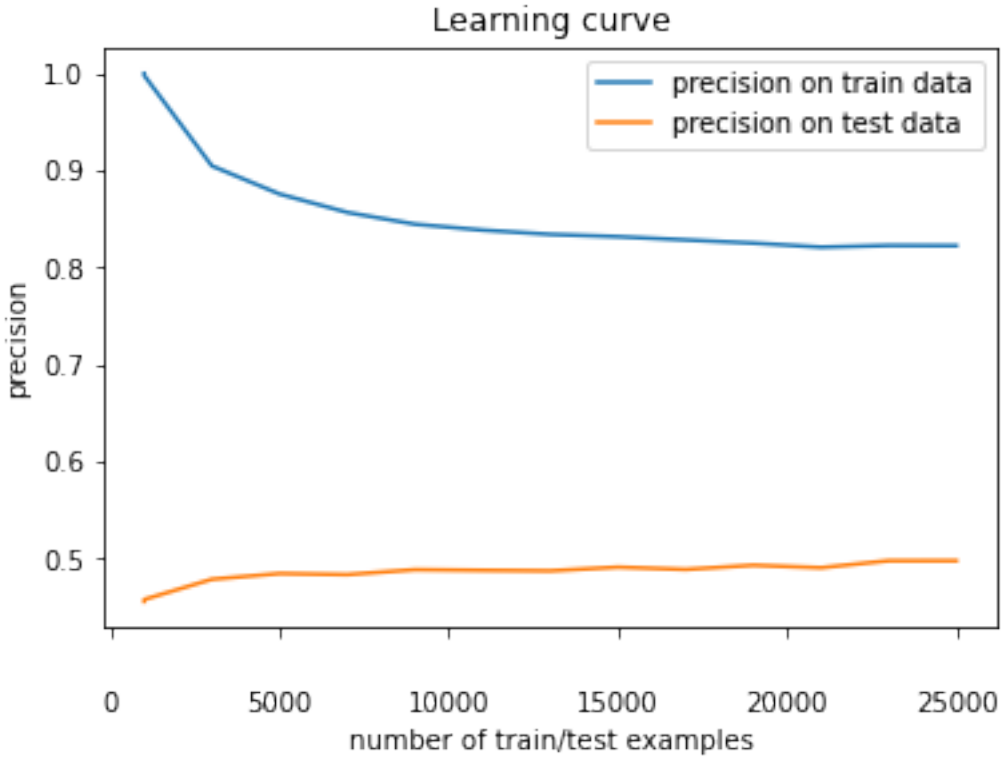


number of examples	F1 on train data	F1 on test data
1000	0.8936970837253058	0.5023969319271333
3000	0.8354046680812367	0.5133187088686932
5000	0.8147325559874145	0.5183503594400303
7000	0.8041810002646201	0.5292381974248926
9000	0.8034659314690823	0.5476592784017494
11000	0.8016385694028675	0.5563050551543528
13000	0.7962129859665614	0.5548413289126463
15000	0.7899597442576368	0.5526378326996197
17000	0.7950438822922045	0.5575235028307276
19000	0.7891415313225058	0.5553743651866003
21000	0.790725126475548	0.5547278422077373
23000	0.7902694100261453	0.5628819800326462
25000	0.7753056874696181	0.5326253599072654

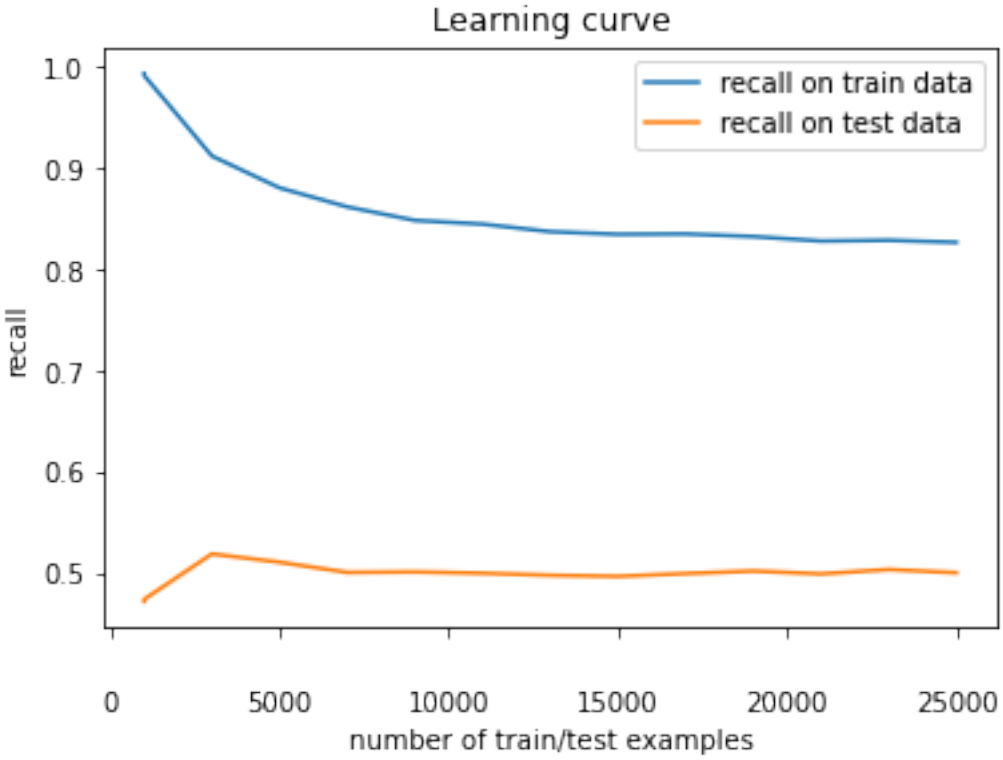
Καμπύλες μάθησης και πίνακες του Logistic Regression: num_words = 250



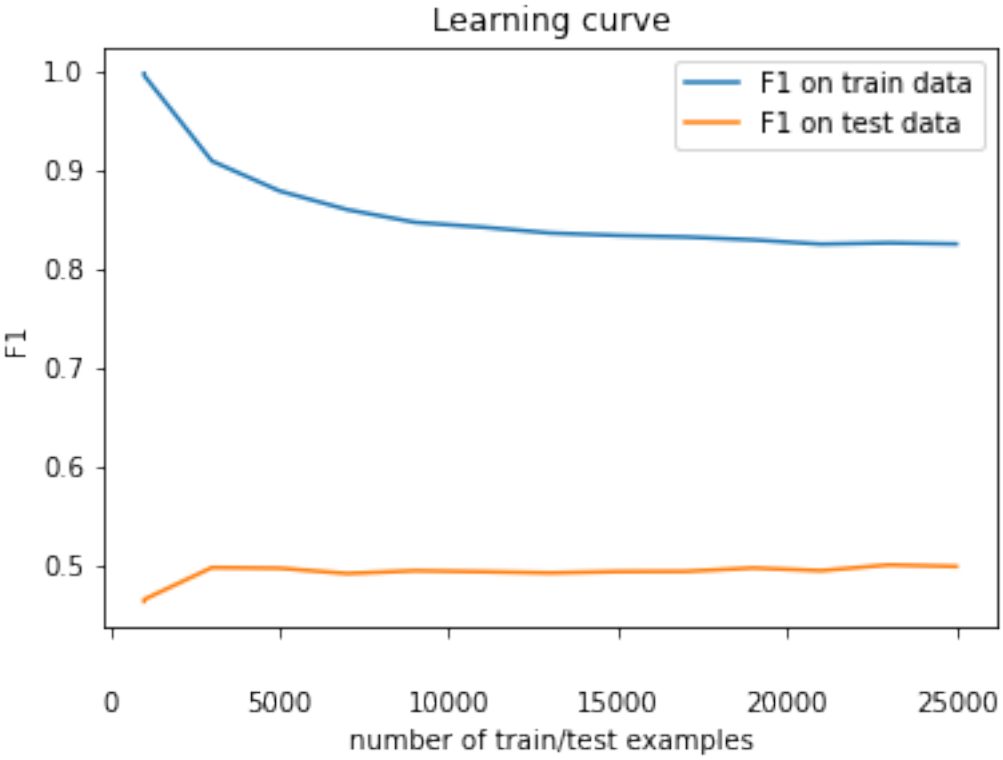
number of examples	accuracy(%) on train data	accuracy(%) on test data
1000	99.7	48.3
1000	99.5	48.5
3000	90.56666666666666	50.03333333333333
5000	87.58	49.88
7000	85.71428571428571	49.114285714285714
9000	84.4888888888889	49.43333333333333
11000	83.99090909090908	49.21818181818182
13000	83.42307692307692	49.207692307692305
15000	83.24666666666667	49.32666666666667
17000	83.01764705882353	49.258823529411764
19000	82.76842105263158	49.626315789473686
21000	82.35238095238095	49.4
23000	82.49130434782609	49.9
25000	82.444	49.82



number of examples	precision on train data	precision on test data
1000	1.0	0.456211812627291 27
1000	0.9979633401221996	0.45824847250509165
3000	0.9048231511254019	0.4790996784565916
5000	0.8758297540023429	0.4849668098399063
7000	0.8569028283394007	0.48389806776813216
9000	0.8447558572366981	0.48894241296255747
11000	0.838669289930245	0.4882847433375067
13000	0.8342700896247911	0.48777153273583473
15000	0.8319227717535044	0.4915366305210262
17000	0.8285813630041725	0.489221140472879
19000	0.8252780376260264	0.4935037937844299
21000	0.8211259649783468	0.4909621540199586
23000	0.8228610776381478	0.49819245997589945
25000	0.8226589227464397	0.4982098814543719



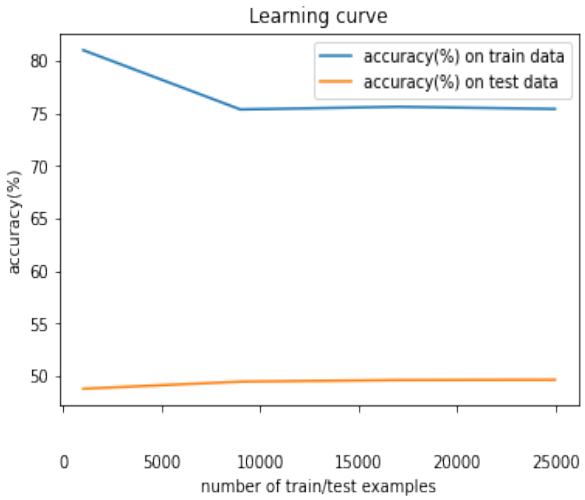
number of examples	recall on train data	recall on test data
1000	0.9939271255060729	0.47257383966244726
1000	0.9919028340080972	0.47468354430379744
3000	0.9124513618677043	0.5195258019525802
5000	0.8809897879025923	0.5113215314944421
7000	0.8622147083685545	0.5013054830287206
9000	0.8488448844884489	0.5017977528089888
11000	0.8451694304253785	0.5004582951420715
13000	0.837705918242831	0.49844768705371
15000	0.8351254480286738	0.4973906061822561
17000	0.8354563515250671	0.5001184834123222
19000	0.8328962551138152	0.502646622909168
21000	0.8284574468085106	0.49980831895725514
23000	0.8292132882296817	0.5041372702726243
25000	0.8272	0.50096



number of examples	F1 on train data	F1 on test data
1000	0.9969543147208122	0.4642487046632125
1000	0.9949238578680204	0.46632124352331605
3000	0.9086212463674525	0.49849447975911676
5000	0.8784021930683376	0.49779559118236477
7000	0.8595505617977528	0.4924479908805927
9000	0.8467954345917471	0.49528668071420645
11000	0.8419068138971182	0.4942965779467681
13000	0.8359844737042393	0.4930518234165067
15000	0.8335210334547865	0.4944462919853675
17000	0.8320046552225778	0.4946097961096789
19000	0.8290696460269396	0.49803325116693764
21000	0.8247754137115839	0.4953457446808511
23000	0.8260249708385536	0.501147235811074
25000	0.8249232119350594	0.49958115600941394

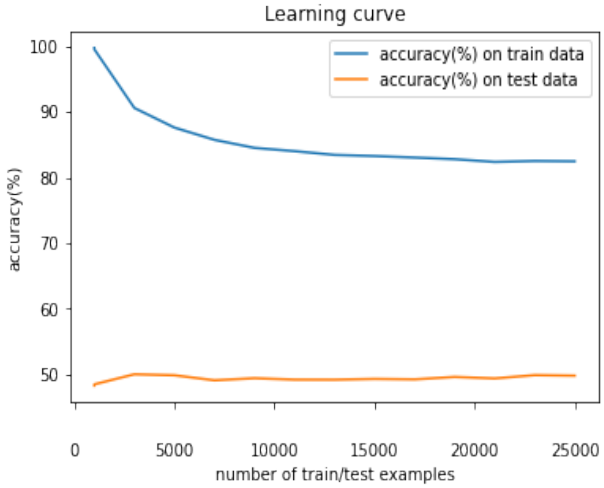
Σύγκριση καμπύλης μάθησης και πίνακα Logistic Regression sklearn με τις καμπύλες και τους πίνακες του δικού μας αλγορίθμου: num_words = 200

Sklearn results:



number of examples	accuracy(%) on train data	accuracy(%) on test data
1000	81.00000023841858	48.80000054836273
9000	75.35555362701416	49.477776885032654
17000	75.61176419258118	49.62941110134125
25000	75.40799975395203	49.65600073337555

Our results:



number of examples	accuracy(%) on train data	accuracy(%) on test data
1000	99.7	48.3
1000	99.5	48.5
3000	90.56666666666666	50.03333333333333
5000	87.58	49.88
7000	85.71428571428571	49.114285714285714
9000	84.4888888888889	49.43333333333333
11000	83.99090909090908	49.21818181818182
13000	83.42307692307692	49.207692307692305
15000	83.24666666666667	49.32666666666667
17000	83.01764705882353	49.258823529411764
19000	82.76842105263158	49.626315789473686
21000	82.35238095238095	49.4
23000	82.49130434782609	49.9
25000	82.444	49.82

Παρατήρηση: Τα αποτελέσματα ορθότητας μεταξύ του built in αλγορίθμου sklearn με τα δικά μας αποτελέσματα παρουσιάζουν μηδαμινή διαφοροποίηση μεταξύ τους.