Ονοματεπώνυμο: Θεοφανόπουλος Μιχαήλ

Αριθμός Μητρώου: 111 520 18 00 053

1η Εργασία "Τεχνητή Νοημοσύνη 2"

Για το καθάρισμα των δεδομένων χρησιμοποίησα την εξής προσέγγιση:

1. Αρχικά κρατάω μόνο τις στήλες των rating και review, καθώς θα χρειαστούμε τη πρώτη για το training του μοντέλου και τη δεύτερη για να γίνει το prediction.

2. Στη συνέχεια, αντικαθιστούμε όλα τα reviews με είτε 0.0 είτε 1.0 εάν είναι <5 ή >5 αντίστοιχα, ώστε να μπορεί να γίνει το prediction.

3. Ύστερα, αφαιρούμε από τα κείμενα των reviews όλα τα emoticons και τα σύμβολα που μπορεί να περιέχονται ώστε να υπάρχει σκέτο κείμενο. Πέραν αυτών, αφαιρούμε τα links, url's, σημεία στίξης και τους αριθμούς από το κείμενο.

4. Εκτελούμε τη διαδικασία του Tokenization στα reviews.

5. Εκτελούμε τη διαδικασία του Stemming στα reviews.

6. Εκτελούμε τη διαδικασία του Lemmatization στα reviews.

Αφού έχουμε τελειώσει με το καθάρισμα των δεδομένων, κρατάμε σε δυο ξεχωριστούς πίνακες τη στήλη review και τη στήλη rating.

Η μορφή στην οποία βρίσκονται αυτή τη στιγμή τα δεδομένα μας είναι η εξής:

	rating	review
0	1.0	thought quiet good movie fun watch liked best ...
1	1.0	wagon master unique film amongst john fords wo...
2	1.0	film near perfect film john ford made film mag...
3	0.0	gave 4 stars lot interesting themes many alrea...
4	1.0	movie really genuine random really hard find m...
...	...	...
45003	0.0	dont even know begin br br worth typing review...
45004	0.0	one worst movies saw 90s id often use benchmar...
45005	0.0	baldwin really stooped low make movies script ...
45006	0.0	liked watching mel gibson million dollar hotel...
45007	1.0	easily best cinematic version william faulkner...

45008 rows × 2 columns

Παρατηρούμε ότι τα reviews είναι πλήρως καθαρισμένα και έχουμε κρατήσει μόνο τις σημαντικές για την επεξεργασία μας λέξεις.

Το επόμενο βήμα είναι να χωρίσουμε τα δεδομένα μας σε training και test set, με σκοπό να προχωρήσουμε στη δημιουργία του μοντέλου μας.

Εάν δωθεί graders_data (**TODO**: store your data set in this variable), τότε ως training set κρατάμε ολόκληρο το data set που έχουμε και ως test set κρατάμε το data set του grader. Διαφορετικά, κάνουμε split το set σε 80-20.

Θα χρειαστούμε έναν vectorizer ώστε να μπορούμε να μετατρέψουμε τα δεδομένα μας τα οποία είναι σε μορφή κειμένου σε πραγματικούς επεξεργάσιμους αριθμούς.

Κάνουμε fit τον vectorizer μας χρησιμοποιώντας ολόκληρο το training set. Οπουδήποτε στο εξής χρειαστούμε να χρησιμοποιήσουμε τα training data μας θα τα περνάμε πρώτα από την transform του vectorizer μας.

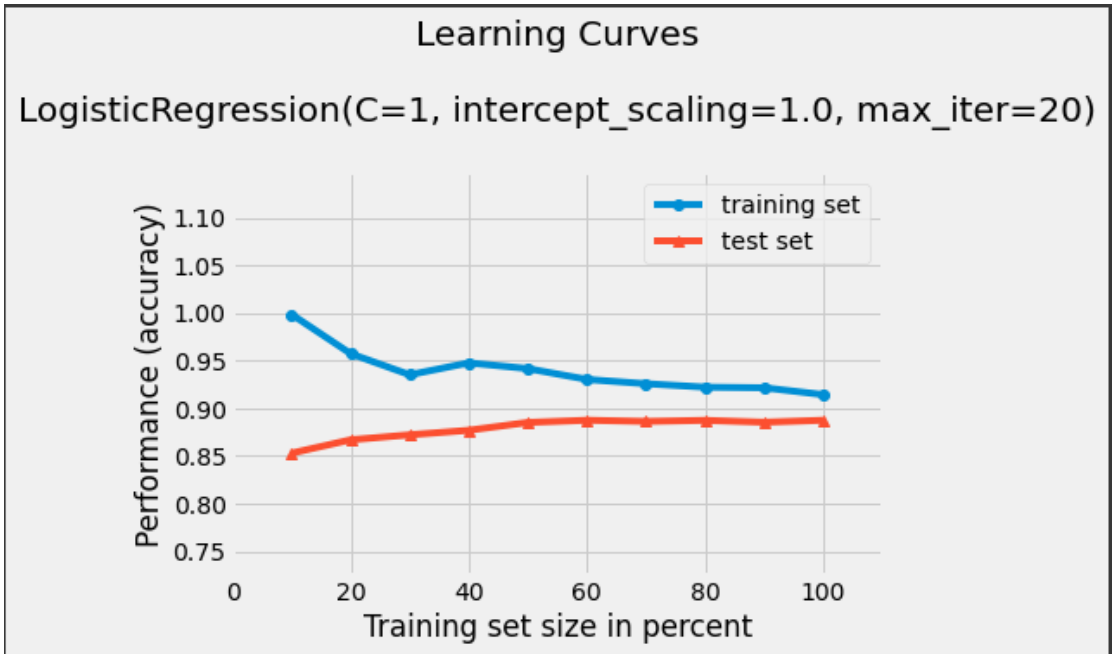
Στη συνέχεια κάνουμε initialize το μοντέλο μας. Οι παράμετροι που χρησιμοποιούμε είναι οι εξής:

1. penalty = ‘l2’, καθώς θα χρειαστούμε να διαχειριστούμε ένα αρκετά μεγάλο data set
2. dual = ‘false’
3. tol = 0.0001
4. C = 1,
5. fit_intercept = True
6. intercept_scaling = 1.0
7. class_weight = None
8. random_state = None
9. max_iter = 20

Κάνουμε fit το μοντέλο μας δίνοντας του το training set (review και rating). Χρησιμοποιώντας το classification report της sklean, train και test set το dataset της άσκησης splitted, παράγουμε τα εξής αποτελέσματα:

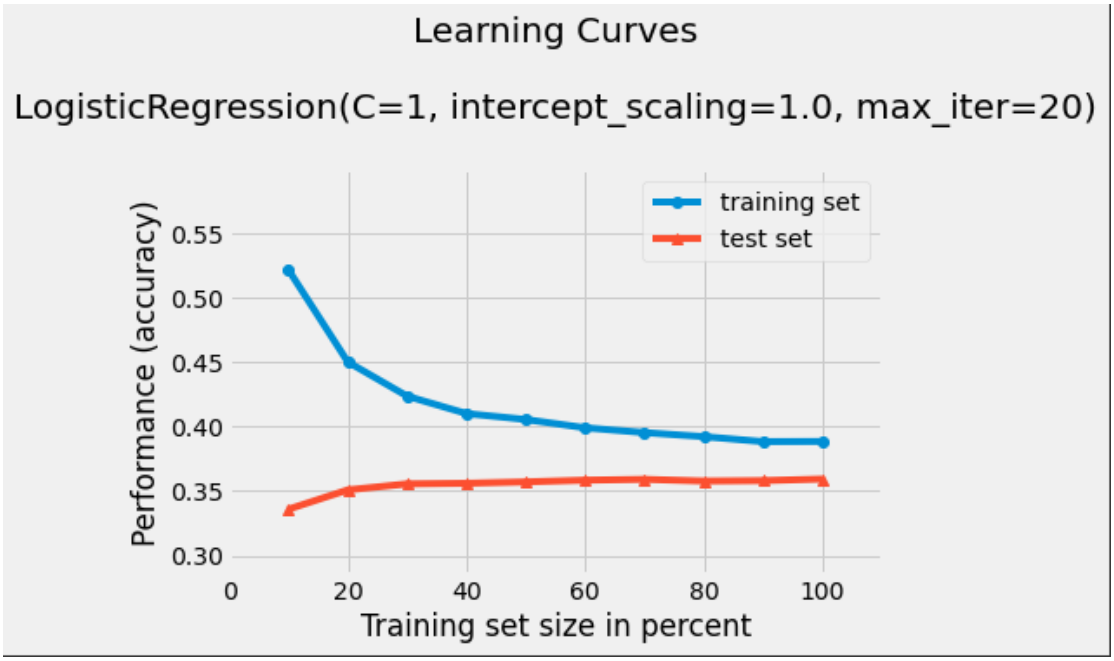
	precision	recall	f1-score	support
0.0	0.89	0.89	0.89	4514
1.0	0.88	0.88	0.88	4488
accuracy			0.89	9002
macro avg	0.89	0.89	0.89	9002
weighted avg	0.89	0.89	0.89	9002

Για τα ίδια sets φτιάχνουμε learning curve κάνοντας πάλι split, το training set αυτή τη φορά, 80-20 αξιοποιώντας την plot_learning_curves της mlxtend. Έχουμε τα εξής αποτελέσματα:



Παρατηρούμε ότι το μοντέλο μας συγκλίνει και δεν έχουμε ούτε overfit, ούτε underfit. Στη συνέχεια θα εμφανίσουμε ένα learning curve για το ίδιο μοντέλο χωρίς καμία προεπεξεργασία στα δεδομένα μας:

	precision	recall	f1-score	support
1.0	0.48	0.74	0.58	1872
2.0	0.13	0.08	0.10	818
3.0	0.21	0.06	0.09	877
4.0	0.23	0.22	0.22	882
7.0	0.25	0.21	0.23	836
8.0	0.25	0.17	0.20	1019
9.0	0.17	0.05	0.08	873
10.0	0.44	0.70	0.54	1825
accuracy			0.37	9002
macro avg	0.27	0.28	0.26	9002
weighted avg	0.31	0.37	0.32	9002



Παρατηρούμε ότι το μοντέλο μας συγκλίνει και έχει αρκετά καλύτερα αποτελέσματα με την κανονικοποίηση των δεδομένων μας.

Οπότε το μοντέλο μας διαχειρίζεται σωστά και παράγει καλά αποτελέσματα για το δεδομένο data set.

Σας εύχομαι καλή διόρθωση 😊