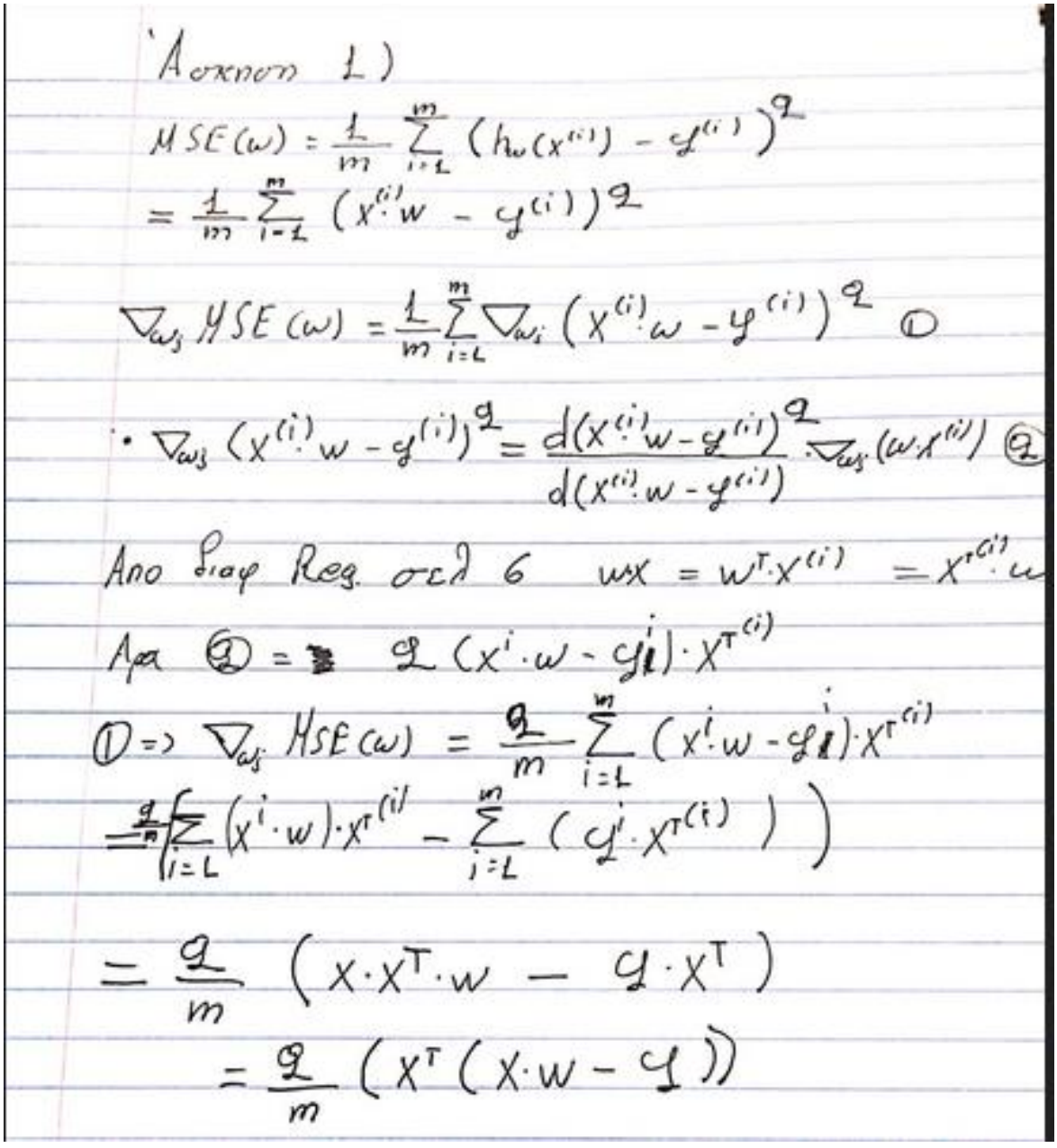
Καρακολης Ιωαννης sdi1800065

Ασκ1:



Άσκηση 1)

$$MSE(w) = \frac{1}{m}\sum_{i=1}^{m}(h_w(x^{(i)}) - y^{(i)})^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}\cdot w - y^{(i)})^2$$

$$\nabla_{w_j} MSE(w) = \frac{1}{m}\sum_{i=1}^{m}\nabla_{w_j}(x^{(i)}\cdot w - y^{(i)})^2 \quad (1)$$

$$\cdot \nabla_{w_j}(x^{(i)}\cdot w - y^{(i)})^2 = \frac{d(x^{(i)}\cdot w - y^{(i)})^2}{d(x^{(i)}\cdot w - y^{(i)})} \cdot \nabla_{w_j}(w\cdot x^{(i)}) \quad (2)$$

Από διαφ Reg. σελ 6 $wx = w^T\cdot x^{(i)} = x^{T(i)}\cdot w$

Άρα (2) = $2(x^i\cdot w - y^i_1)\cdot x^{T(i)}$

(1) $\Rightarrow \nabla_{w_j} MSE(w) = \frac{2}{m}\sum_{i=1}^{m}(x^i\cdot w - y^i_1)\cdot x^{T(i)}$

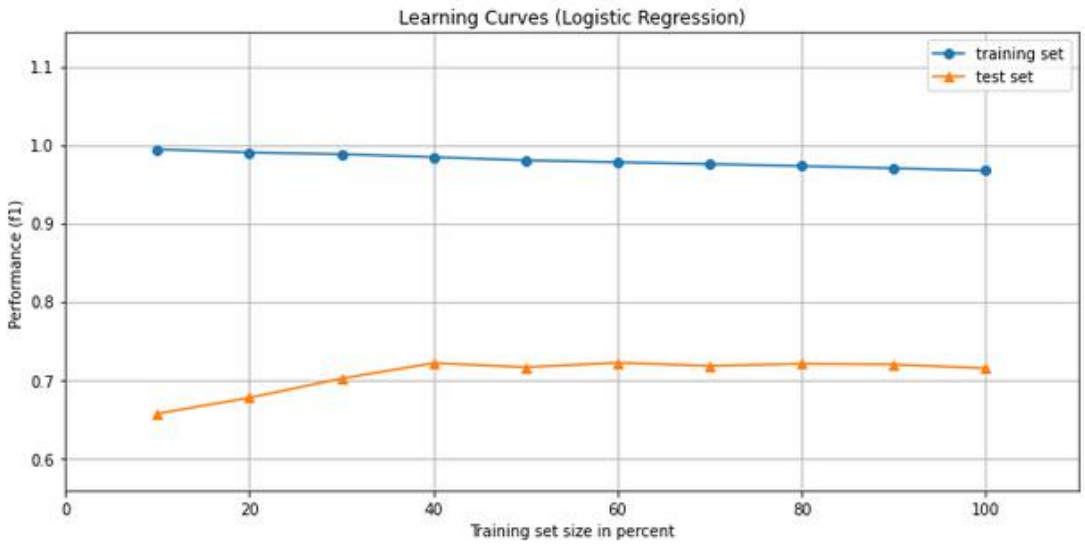
$$= \frac{2}{m}\left(\sum_{i=1}^{m}(x^i\cdot w)\cdot x^{T(i)} - \sum_{i=1}^{m}(y^i\cdot x^{T(i)})\right)$$

$$= \frac{2}{m}(X\cdot X^T\cdot w - y\cdot X^T)$$

$$= \frac{2}{m}(X^T(X\cdot w - y))$$

Report:

Η αρχική υλοποίηση έγινε χρησιμοποιώντας CountVectorizer και τα default parameters σε αυτόν αλλά και στον
classifier (logistic regression). Απεικονίζοντας τα αποτελέσματα αυτής της υλοποίησης μέσω των learning curves έγινε φανερό πως υπήρχε overfitting.



Το πρόβλημα αυτό επιλύθηκε με την παράμετρο min_df όπως εξηγεί και συμφοιτητής μου στο piazza :
("*"You should plot learning curves that show that your models are not overfitting or underfitting"*

Αν κοιτώντας τις καμπύλες διαπιστώσουμε overfit (ανεξαρτήτως vectorizer και χρήσης regularization) τότε ποια θα ήταν μια καλή κατεύθυνσή να κινηθούμε?

Δοκιμάζοντας να μετρήσω το cross entropy , precision και recall ως συνάρτηση του training set size και τα 3 γραφήματα δείχνουν το μοντέλο να τα πηγαίνει καλά στο training set αλλά όχι αναλόγως καλά στο validation set. Οι καμπύλες του training set και του validation set δεν είναι ποτέ πολύ κοντά.

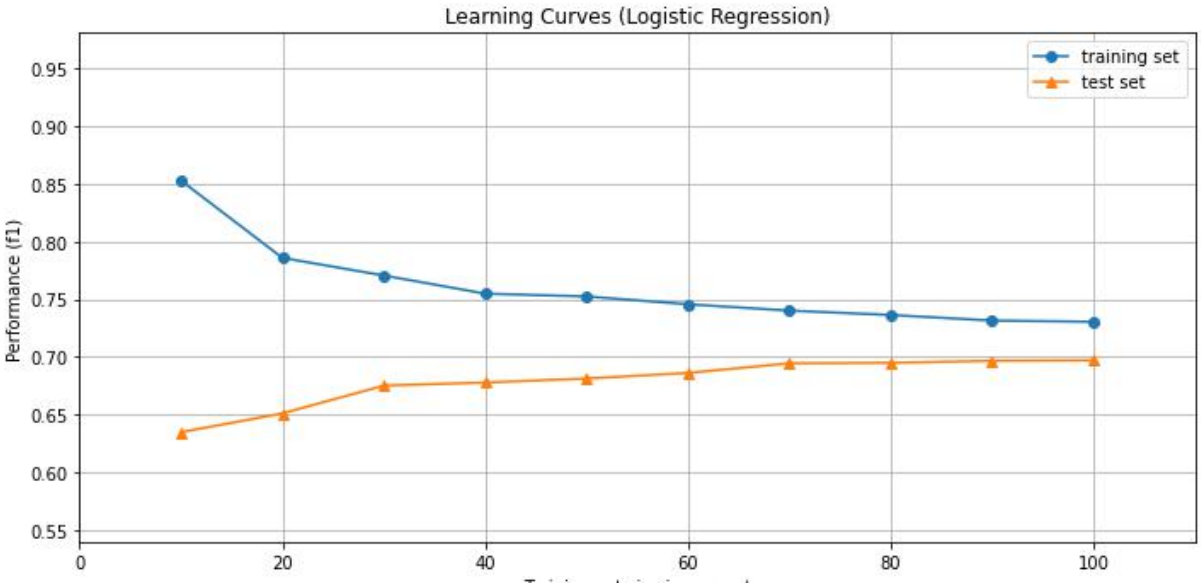
Ενδεικτικά το precision στο validation set φτάνει (το πολύ) κοντά στο 65-68% (αντίστοιχα για το recall). Για το cross entropy αν αφήσω τον classifier να κάνει πολλά iterations , χειροτερεύει λίγο όσο αυξάνεται το training set size. Χρησιμοποιώ το sklearn.linear_model.LogisticRegression του sklearn που είναι εύκολο να παραμετροποιηθεί για multinomial (softmax) regression.

EDIT (σε περίπτωση που κάποιος πέσει σε ανάλογο πρόβλημα):
Δεν είχα προσέξει το ποσό πολλά features προέκυπταν από default από τους vectorizer (ανεξαρτήτως του ngram_range). Είχαμε δει και στο μάθημα οτι πολλά

features σε συνδυασμό με όχι πολλά δεδομένα μπορούν να δώσουν εικόνες overfitting. Καλώντας την get_feature_names() φάνηκε οτι μεναν tags σε λογαριασμούς, αριθμοί κλπ.

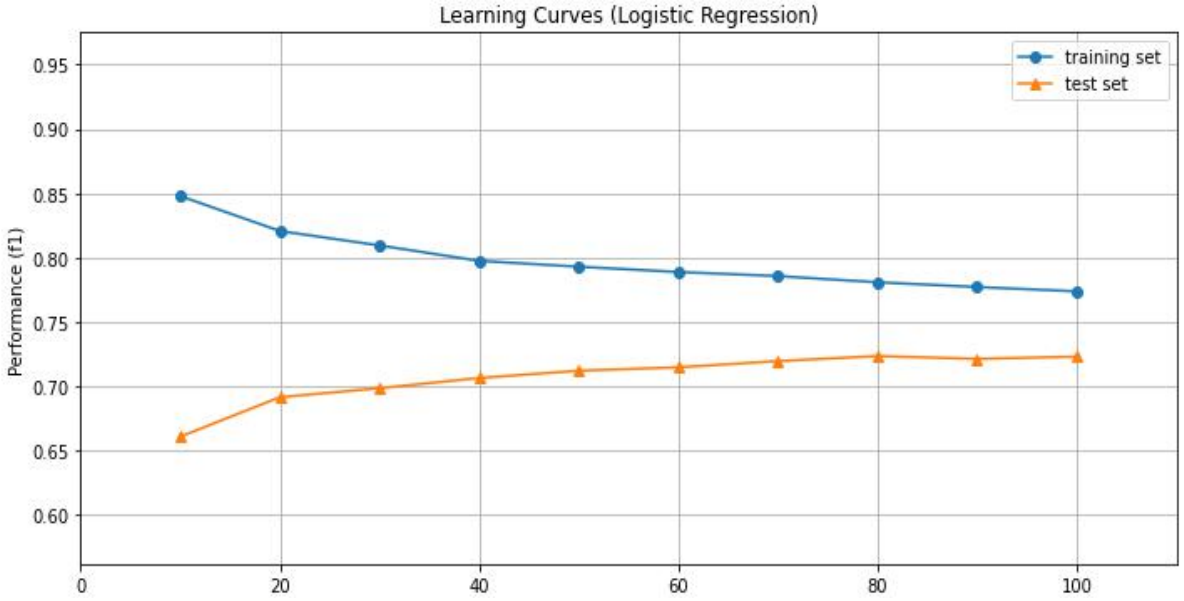
Χρησιμοποιώντας τις παραμέτρους max_df και min_df στους vectorizer trimαρεται ένα καλό μέρος του θορύβου και τα features είναι λιγότερα. Αν και η απόδοση στο training set πέφτει, οι καμπύλες για το training και το validation είναι πλέον κοντά και όλες οι μετρικές δείχνουν να βελτιώνονται όσο αυξάνεται το training set size (μικρο cross entropy loss , ~70% average precision). ")

Χρησιμοποιώντας λοιπόν το min_df ,με value πειραματικά καλύτερο το 0.004 , οι καμπύλες διαμορφώθηκαν ως εξής :



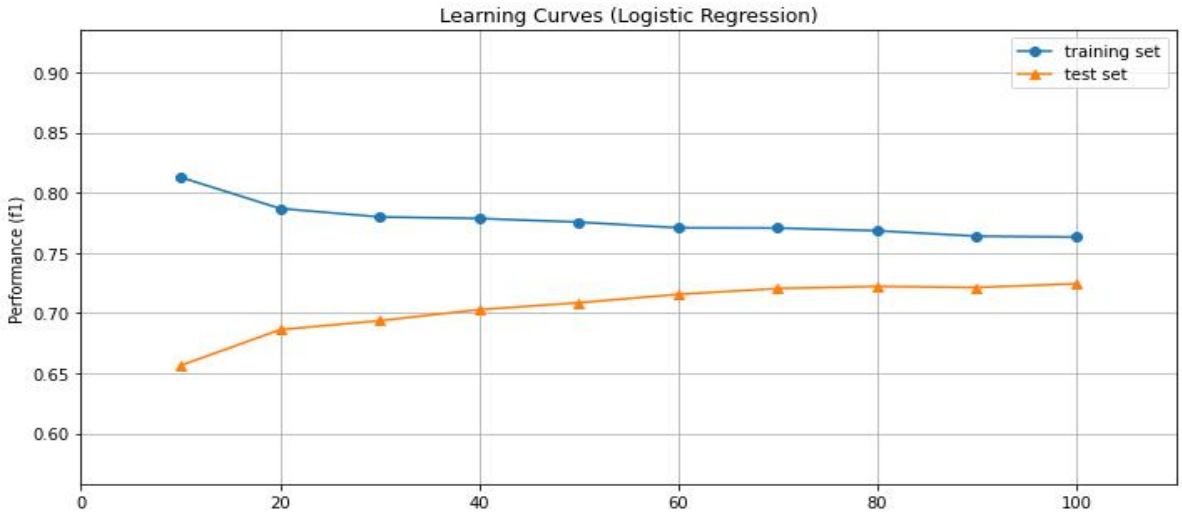
Πέρα από την λύση στο πρόβλημα του overfit η χρήση του min_df μείωσε πολύ και το χρόνο που χρειάζεται για το training.Φυσικά αυτό μείωσε ελάχιστα τα scores.

Μέσα από περισσότερο πειραματισμό τελικά ο vectorizer που φαίνεται να δίνει τα καλύτερα αποτελέσματα είναι ο TfidfVectorizer με min_df =0.001.Ο δεύτερος αυτός vectorizer δεν μετράει απλά πόσες φορές υπάρχει η λέξη αλλά προσθέτει και ένα βάρος που είναι σχετικό με το πόσο σημαντική είναι αυτή η λέξη.



Για την επιλογή των hyperparameters αρχικά χρησιμοποιήθηκε η GridSearchCV η οποία όμως αν και βρήκε μεγαλύτερο score οδηγούσε σε εικόνες overfitting η underfitting .

Μέσω εκτεταμένου πειραματισμού κατέληξα σε μια πολύ μικρή αύξηση στα scores χρησιμοποιώντας ως παραμέτρους :
C=1,penalty='l2',solver='liblinear'



Τέλος θα ήθελα να αναφέρω συνοπτικά την λειτουργία και την δομή του προγράμματος. Αρχικά γίνεται το read των αρχείων καθως και το κατάλληλο pre-processing .Στην συνέχεια γίνεται train το μοντέλο καθως και εκτυπώνονται ενδεικτικά οι τρεις τύποι score που ζητούνται με prediction πάνω στο validation set.Τέλος εκτυπώνονται τα learning curves και για τους τρεις τύπους scoring, χρησιμοποιώντας 10 σημεία για την αναπαράσταση.