Linear Least Squares Regression

Οι τάξεις που αφορούν το συγκεκριμένο ερώτημα βρίσκονται στον φάκελο LinearRegression. Η τάξη Separator χωρίζει ένα σύνολο δεδομένων σε δεδομένα εκπαίδευσης (60%), δεδομένα επικύρωσης (20%) και δεδομένα αξιολόγησης (20%). Η τάξη TrainingExample συμβολίζει δεδομένα εκπαίδευσης και αξιολόγησης. Περιέχει ένα ArrayList με τις ιδιότητες X και μία double μεταβλητή y η οποία συμβολίζει την ορθή απόκριση των συγκεκριμένων ιδιοτήτων. Επίσης περιέχει μεθόδους που υπολογίζουν το σφάλμα E(w) και την απόκλιση f(x)-y. Η τάξη LinearRegression περιέχει μεθόδους με τις οποίες διαβάζουμε τα δεδομένα (εκπαίδευσης, επικύρωσης και αξιολόγησης) και υλοποιούμε τον αλγόριθμο στοχαστικής κατάβασης με τον οποίο βρίσκουμε τα βάρη w ώστε να υπολογίσουμε την f(x) για νέα δεδομένα (αξιολόγησης ή επικύρωσης). Αφού υπολογίσουμε τις αποκρίσεις για όλα τα παραδείγματα, εκτυπώνουμε την τιμή του αθροίσματος των σφαλμάτων και τον μέσο όρο σφάλματος. Διαιρούμε δηλαδή το άθροισμα των σφαλμάτων που κάναμε κατά την αξιολόγηση με το πλήθος των παραδειγμάτων που χρησιμοποιήσαμε κατά την αξιολόγηση.

Λεπτομέριες Υλοποίησης :

Στην τάξη LinearRegression υπάρχει η μεταβλητή "position" στην οποία δίνουμε την θέση στην οποία βρίσκεται η ορθή απόκριση στα δεδομένα που θα επεξεργαστούμε (η πρώτη θέση είναι 0, η δεύτερη 1 κτλ..). Η μεταβλητή "attrno" περιέχει το πλήθος των ιδιοτήτων κάθε παραδείγματος (συμπεριλαμβανομένης και της ορθής απόκρισης). Η μεταβλητή n είναι η σταθερά "η" που θα χρησιμοποιήσουμε στον αλγόριθμο στοχαστικής κατάβασης. Θα εξηγήσουμε παρακάτω γιατί την αρχικοποιούμε με την τιμή 10^{-6}. Η μεταβλητή "split_element" περιέχει το στοιχείο το οποίο χώρίζει τις ιδιότητες των δεδομένων και το χρειαζόμαστε ώστε να κάνουμε τις κατάλληλες επεξεργασίες πάνω σε αυτά. Η μέθοδος "findW()" υλοποιεί τον αλγόριθμο στοχαστικής κατάβασης. Η μεταβλητή "trainpath" περιέχει το path των δεδομένων εκπαίδευσης, ενώ η μεταβλητή "path" το path των δεδομένων προς αξιολόγηση. Η τιμή της μεταβλητής "epoxes" δείχνει το πόσες φορές θα σαρώσουμε όλα τα παραδείγματα εκπαίδευσης στον αλγόριθμο στοχαστικής κατάβασης, σε περίπτωση που το σφάλμα E(w) δεν έχει συγκλίνει.

Σημαντικοί μέθοδοι :

Τάξη LinearRegression :

- public void findW(): Υλοποιεί τον αλγόριθμο στοχαστικής κατάβασης. Μέσω αυτής της μεθόδου δηλαδή βρίσκουμε τα βάρη.

- public void newExample(): Επεξεργάζεται δεδομένα προς αξιολόγηση και επικύρωση. Υπολογίζει την f(x) με βάση τις τα βάρη που βρήκαμε με την μέθοδο "findW()". Αφού υπολογίσει την παραπάνω απόκριση, αφαιρεί την τιμή που βρήκε από την ορθή απόκριση ώστε να βρούμε την τιμή της απόκλισης, δηλαδή το πόσο διαφέρει ο υπολογισμός μας από την ορθή τιμή "y".

Τάξη TrainingExample :

- public double apoklisi(ArrayList<Double> w) : Υπολογίζει την απόκλιση f(x)-y.

Υπολογίζει δηλαδή την τιμή της f(x) με βάση τα βάρη που δέχεται σαν όρισμα και αφαιρεί την ορθή απόκριση.

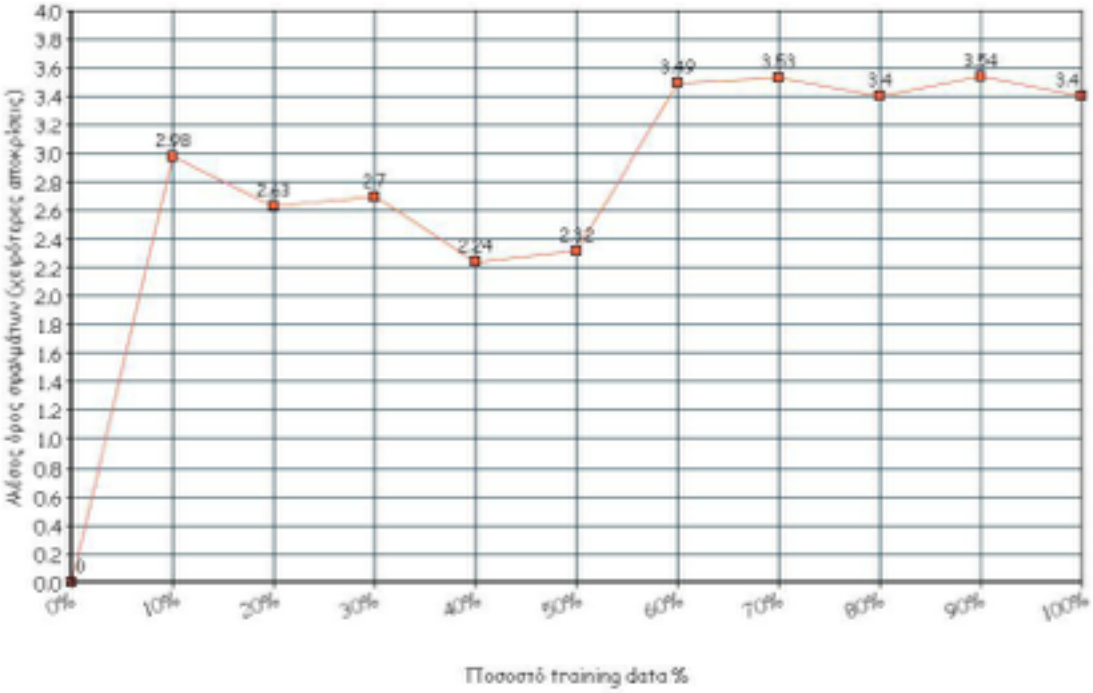
- public double error(ArrayList<Double> w) : Υπολογίζει το σφάλμα E(w) με βάση τα βάρη που δέχεται σαν όρισμα.

ΠΕΙΡΑΜΑΤΑ

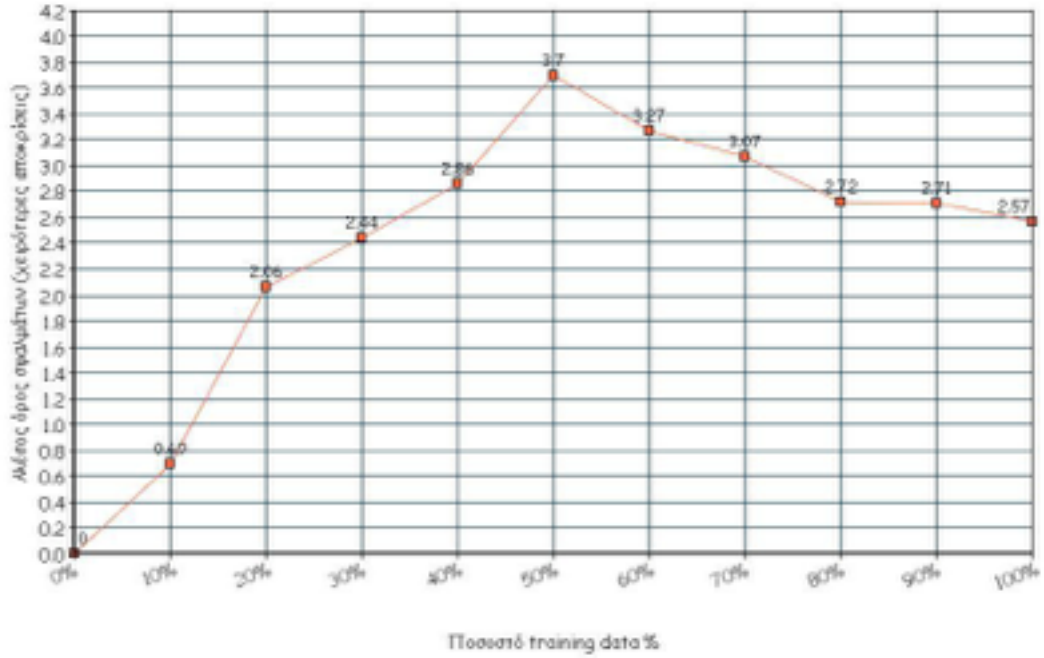
Έχοντας τα δεδομένα επικύρωσης θα προσπαθήσουμε να βρούμε την σταθερά "η" η οποία μας δίνει τα καλύτερα αποτελέσματα. Για κάθε ποσοστο του Training Set δοκιμάζουμε αρκετές φορές τον αλγόριθμο ώστε να βρούμε την χειρότερο μέσο όρο σφαλμάτων. Έχουμε 1174 παραδείγματα επικύρωσης. Για το κάθε παράδειγμα υπολογίζουμε την απόλυτη τιμή της απόκλισης του υπολογισμού μας με την ορθή απόκριση του παραδείγματος. Αφού αθροίσουμε όλες αυτές τις αποκλίσεις, διαιρούμε με τον αριθμό 1174 (πλήθος test data) ώστε να βρούμε τον μέσο όρο. Όπως φένεται και στον πίνακα παρακάτω για η=10-6 έχουμε τις καλύτερες αποκρίσεις.

	10^{-4}	10^{-5}	10^{-6}	10^{-7}
10%	34.8	5.49	17	17
20%	23.2	9.92	17	7
30%	17.67	7.02	14	14
40%	11.75	7.71	14	14
50%	17.08	2.68	14	16
60%	71.12	1.28	14	14
70%	41.4	1.32	11	13
80%	93.19	1.34	11	12
90%	67.81	1.45	9	12
100%	64.76	1.46	10	12

Θα δοκιμάσουμε λοιπόν τον αλγόριθμο με τα δεδομένα αξιολόγησης για η=10^{-6} . Θα δοκιμάσουμε αρκετές φορές τον αλγόριθμο για κάθε ποσοστό του Training Set ώστε να βρούμε τον χειρότερο μέσο όρο σφαλμάτων για κάθε Training Set



Θα δοκιμάσουμε τώρα τον αλγόριθμο με τα δεδομένα εκπαίδευσης.



Στην πρώτη περίπτωση παρατηρούμε πως ο μέσος όρος σφαλμάτων στην αρχή μειώνεται και μετά το 50% των δεδομένων αυξάνεται. Ενώ στην δεύτερη περίπτωση συμβαίνει το αντίθετο.