Logistic Regression

Αρχικά, με την InputToHashMap( ) δημιουργούμε το Vocabulary το οποίο αποτελέιται από όλες μοναδικές λέξεις που περιέχονται στα training mails. Στη συνέχεια, μειώνουμε το πλήθος λέξεων του Vocabulary κρατώντας όσες λέξεις εμφανίζονται απο 30 μέχρι 2000 φορές, με την μέθοδο Prunning( ).

Αφού δημιουργήσουμε το Vocabulary,με την μέθοδο Read( ) φτιάχνουμε τον πίνακα MainTable[ ], ο οποίος έχει ως γραμμές τα training mails και ως στήλες το vocabulary. Αν στο mail (i) περιέχεται η λέξη (j) τότε βάζουμε 1 στο κατάλληλο κελί. Στην τελευταία στήλη του MainTable[ ] αποθηκεύεται αν το συγκεκριμένο μήνυμα είνα spam ή ham.

Παράλληλα , φτιάχνουμε και τον δυσδιάστατο πίνακα words[ ], του οποίου οι γραμμές είναι οι ιδιότητες-λέξεις. Στην πρώτη στήλη αποθηκεύται η λέξη, στη δεύτερη στήλη αποθηκεύεται θέση στην οποία βρίσκεται στον MainTable[] και στην τρίτη το βάρος της κάθε ιδιότητας-λέξης.

Η μέθοδος regression( ), αποτελείται από μια for( ) η οποία τρέχει τόσες επαναλήψεις όσες η μεταβλητή epoxes. Ακόμα, μια εμφωλευμένη for( ) η οποία τρέχει για όλες τις λέξεις του πίνακα words και ενημερώνει για κάθε λέξη το αντίστοιχο βάρος της στην τρίτη στήλη του πίνακα. Κατά την ενημέρωση των βαρών καλείται η μέθοδος MailClassify( ) η οποία υπολογίζει την f(x)=w*x για όλες τις ιδιότητες με τα αντίστοιχα βάρη του πίνακα words[ ]. Η MailClassify() καλεί την μέθοδο sigmoid() στην οποία περνάει ως όρισμα την f(x) που έχει υπολογίσει και η sigmoid με τη σειρά της επιστρέφει το απότελεσμα της σιγμοειδούς σύναρτησης με όρισμα την f(x).

Αφού έχει υπολογίσει τα βάρη για κάθε ιδιότητα, με την μέθοδο calculateTest(), παίρνει ένα καινούργιο mail από τα test δεδομένα, καλεί την MailClassify() η οποία επιστρέφει μέσω της sigmoid() μια πιθανότητα. Αν η τιμή αυτής της πιθανότητας είναι μεγαλύτερη απο την 1-πιθανότητα, τότε το mail είναι ham, στην αντίθετη περίπτωση είναι spam.