Οικονομικό Πανεπιστήμιο Αθηνών – Τμήμα Πληροφορικής

Τεχνητή Νοημοσύνη

Αναφορά 2ης εργασίας(Υλοποίηση Αλγορίθμων Μηχανικής Μάθησης)

Μαριάνθη Μηνδρινού – 3150110

Μιχαήλ Ρούσσος – 3150148

Χρήστος Τασιόπουλος – 3150170

Ακολουθεί συνοπτική περιγραφή της λειτουργίας των κλάσεων μας:

(**Σημ**: Χρησιμοποιούμε το dataset lingspam_public\lemm)

- Κλάση **Reader:** Με την κλάση αυτή διαβάζουμε τα δεδομένα εισόδου από ένα αρχείο (txt) και ελέγχουμε για πιθανά σφάλματα και αστοχίες κατά το διάβασμα.
- Κλάση **TreeNode:** Τα αντικείμενα τύπου TreeNode αντιπροσωπεύουν τους κόμβους του δέντρου απόφασης. Κάθε αντικείμενο TreeNode έχει δύο αναφορές TreeNode (leftNode,rightNode) που ειναι τα "παιδιά'' του καθώς και μια αναφορά τύπου Examples που είναι το "περιεχόμενο" του .Ακόμη έχουμε ορίσει setters και getters για τα πεδία του αντικειμένου TreeNode.
- Κλάση **Example:** Σε αυτήν την κλάση περνάμε τα παραδείγματα (τύπου String) σε μια λίστα. Επίσης έχουμε μια μέθοδο για να παίρνουμε το μέγεθος της λίστας, ολόκληρη την λίστα, ένα παράδειγμα από την λίστα και να τοποθετούμε ένα παράδειγμα στην λίστα.
- Κλάση **Examples:** Σε αυτήν την κλάση περνάμε τα παραδείγματα (τύπου Example) σε μια λίστα. Επίσης έχουμε ένα κενό κατασκευαστή, μια μέθοδο για να προσθέτουμε στοιχεία στη λίστα, να παίρνουμε το μέγεθος της λίστας και να παίρνουμε ένα συγκεκριμένο στοιχείο από την λίστα.
- Κλάση **ID3:Στην κλάση αυτή έχουμε αρχικά την μέθοδο** ID3method(TreeNode type), στην οποία καλούμε την αναδρομική μορφή της μεθόδου ID3recursive με ορίσματα την λίστα των παραδειγμάτων, των ιδιοτήτων και μια συμβολοσειρά για την κατηγορία. Στην μέθοδο ID3recursive στην ουσία ακολουθούμε τον ψευδοκώδικα της 16$_{ης}$ διάλεξης για τον συγκεκριμένο αλγόριθμο, όπου ελέγχουμε αν τα παραδείγματα ανήκουν στην ίδια κατηγορία, στην συνέχεια ελέγχουμε αν η λίστα των ιδιοτήτων είναι κενή, ύστερα κοιτάμε για την καλύτερη ιδιότητα κι ορίζουμε το υποδέντρο.
- Κλάση **main:** Η κλάση main περιέχει την συνάρτηση main η οποία μας δίνει την δυνατότητα να τρέξουμε τους αλγορίθμους ID3 ,αφελείς ταξινομητές Bayes ή τον αλγόριθμο της Λογιστικής Παλινδρόμησης.(Με την προϋπόθεση ότι έχει καθοριστεί σωστά το path προς το dataset). Στην συνέχεια υλοποιούμε την μέθοδο loadFile για το διάβασμα των αρχείων και προσθέτουμε στις αντίστοιχες λίστες ham ,spam τα στοιχεία για να δημιουργήσουμε το λεξικό μας.Έπειτα κοιτάμε τα μέιλ και τοποθετούμε σε ένα vector για κάθε μέιλ που μας δείχνει αν περιέχεται ή όχι κάθε λέξη του λεξικού. Τέλος η μέθοδος readExample χρησιμοποιείται για τον ίδιο λόγο με την προηγούμενη (μόνο στο κομμάτι δημιουργίας vector) και χρησιμοποιείται από την Bayes.

- Κλάση **Bayes:**Σε αυτή την κλάση υλοποιείται η μέθοδος h_method που υπολογίζει τις απαραίτητες πιθανότητες όπως ορίστηκαν και στις διαλέξεις του μαθήματος και επιστρέφει έναν ακέραιο που είναι η κατηγοριοποίηση του αντικειμένου(π.χ. αν είναι spam ή ham).Σημειώνεται πως για τους υπολογισμούς των πιθανοτήτων χρησιμοποιήθηκαν λογάριθμοι για γρηγορότερους και ευκολότερους υπολογισμούς.
- Κλάση **Properties:** Η κλάση Properties είναι η κλάση τα αντικείμενα της οποίας περιέχουν τις ιδιότητες των παραδειγμάτων. Η Properties περιέχει 2 κατσκευαστές και άλλες τρεις μεθόδους τις numofprop(), getprop(int i) ,add(String ex). Κάθε μία
από αυτές αντίστοιχα επιστρέφουν το πλήθος των ιδιοτήτων , επιστρέφουν μια συγκεκριμένη ιδιότητα και προσθέτουν μια ιδιότητα στη δομή αποθήκευσής τους.
- Κλάση **Functions:** Η κλάση Functions περιέχει 6 μεθόδους. Η probability1 υπολογίζει την πιθανότητα που έχει ένα όρισμα ενός παραδείγματος να έχει μια συγκεκριμένη τιμή c. Η probability2 υπολογίζει την πιθανότητα που έχει ένα όρισμα ενός παραδείγματος να έχει μια συγκεκριμένη τιμή c δεδομένου ότι ένα άλλο όρισμα έχει την τιμή x. Η entropy1 υπολογίζει την εντροπία μιας ιδιότητας. Η entropy2 υπολογίζει την εντροπία μιας ιδιότητας δεδομένου ότι ένα άλλο όρισμα έχει μια συγκεκριμένη τιμή . Η IG υπολογίζει το κέρδος πληροφορίας που μας παρέχει μια ιδιότητα. Η choosebest επιλέγει την καλύτερη ιδιότητα με βάση το κέρδος πληροφορίας.
- Κλάση **RegressionInstance:** Η κλάση αυτή αντιπροσωπεύει ένα αντικείμενο μέιλ, έτσι ώστε να μπορεί να χρησιμοποιηθεί από τον αλγόριθμο μας.
- Κλάση **LogisticRegression:** Η κλάση αυτή περιέχει τις μεθόδους που χρησιμοποιούμε για την υλοποιήση του αλγορίθμου Λογιστικής Παλινδρόμησης.Περιέχει μια μέθοδο για την ρύθμιση των βαρών(**train**), μια μέθοδο για την υλοποίηση της σιγμοειδούς συνάρτησης(**sigmoidFunction**),μια μέθοδο για τον έλεγχο ενός συγκεκριμένου παραδείγματος αν είναι ή όχι σπαμ(**test**) και τέλος μια μέθοδο που διαβάζει τα αρχεία μας και τα αποθηκεύει(**read**).