ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρονικών Μηχ. Και Μηχ. Υπολογιστών

Τομέας Τηλεπικοινωνιών

Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων

Εργασία 15.2:

Call Routing System

Η εταιρία ΤΟΕ έχει ένα τηλεφωνικό κέντρο για την εξυπηρέτηση των πελατών της. Στο τηλεφωνικό αυτό κέντρο απαντούν τηλεφωνήτριες η κάθε μία από τις οποίες ειδικεύεται σε ένα είδος συναλλαγής, για παράδειγμα "υπόλοιπο λογαριασμού", "πληρωμές δόσεων", "μεταφορά χρημάτων μεταξύ λογαριασμών". Για να βελτιωθεί η εξυπηρέτηση των πελατών θέλουμε το τηλεφωνικό κέντρο της ΤΟΕ να δρομολογεί αυτόματα την κλήση κάθε χρήστη στον κατάλληλο τηλεφωνητή με βάση το αίτημα κάθε χρήστη.

Συνολικά υπάρχουν 15 διαφορετικές επιλογές δρομολόγησης της κλήσης. Κατά τη διάρκεια μιας δοκιμαστικής περιόδου συλλέγουμε και μετεγγράφουμε δείγματα προτάσεων των χρηστών και τα αντιστοιχούμε στις κατάλληλες κατηγορίες δρομολόγησης.

Στο παρακάτω link θα βρείτε τα αρχεία «final.train» και «final.test:
http://www.telecom.tuc.gr/patreco/res/projects/15.2_CallRouting/data/
Τα αρχεία αυτά αντιστοιχούν σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου και σε κάθε γραμμή περιέχουν τον αριθμό της πραγματικής δρομολόγησης (κλάσης) στην οποία αντιστοιχεί η υπόθεση ου αναγνωριστή (πρόταση) που ακολουθεί.

Ο πίνακας term-document

Η βασική δομή που ακολουθείται για την εξαγωγή χαρακτηριστικών από κείμενο είναι ο πίνακες term-document. Σε έναν term-document πίνακα M ∈ Rn×p , όπου n είναι το πλήθος των γραμμών και p το πλήθος των στηλών, η πληροφορία για τα κείμενα αναπαρίσταται ως εξής: κάθε γραμμή $i, 1 \leq i \leq n$, αναπαριστά ένα κείμενο/πρόταση ενώ κάθε στήλη j, 1 ≤ j ≤ p, αναπαριστά μια λέξη του κειμένου. Το στοιχείο m_{ij} του M είναι η συχνότητα εμφάνισης (ακέραιος αριθμός) του όρου j στο κείμενο i την οποία για ευκολία συμβολίζουμε και ως f_{ij} , όπου $f_{ij=}m_{ij}$.

Γενικά, οι πίνακες term-document είναι αραιοί (sparse) γιατί έχουν πολλά στοιχεία ίσα ε το 0. Έχοντας ήδη δημιουργήσει έναν term-document πίνακα, μπορούμε να εφαρμόσουμε κάποιους μετασχηματισμούς στα στοιχεία του.

Μετασχηματισμοί term-document πινάκων

Σε αρκετές περιπτώσεις η εφαρμογή ενός μετασχηματισμού 1-1 πάνω στα στοιχεία ενός term-document πίνακα οδηγεί στη βελτίωση της επίδοσης των αλγορίθμων μάθησης λόγω καλύτερης

αναπαράστασης της πληροφορίας. Τρεις από τους δημοφιλέστερους μετασχηματισμούς είναι ο δυαδικός, ο λογαριθμικός, και η εντροπία. Αν l_{ij} είναι ο μετασχηματισμός του f_{ij} οι παραπάνω μετασχηματισμοί ορίζονται ως εξής:

- **Δυαδικός Μετασχηματισμός**: $l_{ij} = 1$ αν $f_{ij} > 0$
- **Λογαριθμικός Μετασχηματισμός**: $l_{ij} = \log(1 + f_{ij})$
- **Εντροπία**: Αν ορίσουμε την ποσότητα $p_{ij} = f_{ij} / \sum_{i=1}^{n} f_{ij}$ τότε η εντροπία της j-οστής λέξης ορίζεται ως:

$$e_j = 1 + \frac{\sum_{i=1}^{n} p_{ij} \log(p_{ij})}{\log(n)}$$

Η τελική μορφή του μετασχηματισμού σε αυτή την περίπτωση ορίζεται ως:

$$l_{ij} = e_j \log(1 + f_{ij})$$

Βήματα

Γράψτε ένα πρόγραμμα (προτείνεται Python) το οποίο:

1) θα διαβάζει τα δεδομένα εκπαίδευσης και ελέγχου που σας δίνονται.
2) Δημιουργήστε το λεξικό με τις λέξεις που υπάρχουν στα δεδομένα εκπαίδευσης και υπολογίστε την αντίστοιχη συχνότητα εμφάνισης της κάθε λέξης στο dataset.
3) Για κάθε κλάση δρομολόγησης υπολογίστε το πόσο συχνά εμφανίζεται μια λέξη υπολογίζοντας το Term Frequency (TF) κάθε λέξης όπου:

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

4) Μετρήστε το πόσο σημαντική είναι μια λέξη υπολογίζοντας τον όρο Inverse Document Frequency (IDF) κάθε λέξης όπου:

$$IDF(t) = ln\left(\frac{Total\ number\ of\ documents}{Total\ number\ of\ document\ with\ term\ t\ in\ it}\right)$$

5) Υπολογίστε την τιμή $(TF - IDF)(t) = TF(t) \cdot IDF(t)$ και ταξινομήστε τις λέξεις με βάση το πόσο σημαντικές είναι.
6) Φτιάξτε term-document πίνακες για τις προτάσεις εκπαίδευσης που έχετε χρησιμοποιώντας τις τιμές του TF-IDF που υπολογίσατε στο προηγούμενο βήμα για κάθε εμφάνιση των λέξεων. Οι γραμμές του πίνακα term-document θα χρησιμοποιηθούν ως διανύσματα χαρακτηριστικών για να εκπαιδεύσετε το σύστημα σας. Μπορείτε να δοκιμάσετε διαφορετικές μορφές term-document πινάκων χρησιμοποιώντας μετασχηματισμούς ή λαμβάνοντας υπόψη σας το βάρος που έχουν οι λέξεις με βάση της τιμής TFIDF που έχουν.
    a. Επειδή το μήκος των χαρακτηριστικών είναι πολύ μεγάλο επιλέξτε να εφαρμόσετε κάποια μέθοδο για να κάνετε μείωση της διάστασης (π.χ. εφαρμόστε SVD για να μειώσετε τη διάσταση σε 50 ή και λιγότερο).
    b. Προτείνετε και χρησιμοποιήστε εναλλακτικές μεθόδους μείωσης διάστασης και συγκρίνετε τα αποτελέσματα σας.
7) Χρησιμοποιώντας τα παραπάνω διανύσματα χαρακτηριστικών εκπαιδεύστε ένα GMM για κάθε κλάση. Η επιλογή του αριθμού των Gaussian components είναι

θέμα πειραματισμού. Επίσης εφόσον τα δεδομένα σας δεν είναι αρκετά μπορείτε να θεωρήσετε διαγώνιους πίνακες διακύμανσης. Μια άλλη προσέγγιση που θα μπορούσατε να εξετάσετε είναι να θεωρήσετε ότι όλες οι κατηγορίες έχουν GMMs με τα ίδια Gaussian components και διαφοροποιούνται μόνο στο βάρος που έχουν αυτά.

8) Εφαρμόστε την ίδια διαδικασία πάνω στον αντίστοιχο πίνακα των test-προτάσεων και κάντε ταξινόμηση
9) Επαναλάβετε τα πειράματα σας αλλάζοντας τις παραμέτρους του προβλήματος

Επεκτάσεις (Bonus)

10) Χρησιμοποιήστε term-document πίνακες όπου αντί για λέξεις θα συμπεριλάβετε τα σημαντικότερα ζεύγη λέξεων. Για να βρείτε τα σημαντικότερα ζεύγη χρησιμοποιήστε το μέτρο της αμοιβαίας πληροφορίας (mutual information): $I(w_1, w_2) = \log_2\left(\frac{P(w_2|w_1)}{P(w_2)}\right)$
11) Το Cosine Similarity είναι ένα μέτρο της ομοιότητας δύο n-διάστατων διανυσμάτων x και y που ορίζεται ως εξής:

$$\cos(x, y) = \frac{xy^T}{\sqrt{(\sum_{1 \le i \le n} x_i^2) \cdot (\sum_{1 \le i \le n} y_i^2)}}$$

Χρησιμοποιήστε το κατάλληλα για να κάνετε ταξινόμηση

12) Προτείνετε άλλες μεθόδους ταξινόμησης