MEM704 - Μηχανική Μάθηση
3η Εργαστηριακή Άσκηση
Οι ταξινομητές Naive Bayes και SVM
Νευρωνικά Δίκτυα

Παράδοση: Πέμπτη 19/05/2021, 18:00. Εξέταση Παρασκευή 20/05/2021 11:00

Για την υλοποίηση των αλγορίθμων αυτής της άσκησης θα χρησιμοποιηθούν οι συναρτήσεις της βιβλιοθήκης **Scikit-Learn**.

1 Φίλτρο ανεπιθύμητων(spam) ηλεκτρονικών μηνυμάτων(sms, email)

Σε αυτή την άσκηση θα χρησιμοποιήσετε τον απλοικό αλγόριθμο Bayes(Naive Bayes) και τις Μηχανές Διανυσμάτων Υποστήριξης (SVM) για την κατασκεύη ενός ταξινομητή-φίλτρο για ανεπιθύμητα ηλεκτρονικά μηνύματα. Το αρχείο *spam_train.txt* περιέχει περίπου 5500 μηνύματα τύπου SMS με το 87% να είναι κανονικά και τα υπόλοιπα να είναι ανεπιθύμητα. Το άρχειο αποτελείται από δύο στήλες, στη 1η στήλη υπάρχει η ετικέττα του μηνύματος η οποίο είναι είτε ham στη περίπτωση κανονικού μηνύματος ή spam όταν είναι ανεπιθύμητο, ενώ στη 2η στήλη είναι το κείμενο του μηνύματος. Οι δύο στήλες χωρίζονται μεταξύ τους με ενα διάκενο(tab). Το αρχείο *spam_test.txt* περιέχει τα μηνύματα με τα οποία θα δοκιμάσουμε τους αλγοριθμούς.

1. **Προετοιμασία των δεδομένων**. Σε κάθε μήνυμα οι λέξεις χωρίζονται με ένα ή παραπάνω κενά. Η αναπαράσταση του κάθε μηνύματος θα γίνει με έναν διαφορετικό τρόπο από αυτόν που περιγράψαμε στο μάθημα. Φτιάχνουμε ένα λεξικό (V) με όλες τις λέξεις όλων των μηνυμάτων του συνόλου εκπαίδευσης. Με x_j συμβολίζουμε την $j-$ λέξη του μηνύματος. Δηλαδή το x_j είναι ένας ακέραιος με τιμές στο σύνολο $\{1, \ldots, |V|\}$, όπου $|V|$ είναι ο αριθμός των λέξεων στο λεξικό. Ενα μήνυμα με $d-$ λέξεις αναπαραστάται με το διάνυσμα $(x_1, x_2, \ldots, x_d)$ μήκους d. Στο λεξικό θα καταχωρηθούν οι λέξεις που εμφανίζονται σε τουλάχιστον 5 μηνύματα. Επίσης όλα τα κεφαλαία γράμματα θα μετατραπούν σε μικρά (κονονικοποίηση).

2. **Απλοικός(Naive) αλγόριθμος Bayes**. Από το *Scikit-Learn* χρησιμοποίηστε την συνάρτηση που υλοποίει τον απλοικό ταξινομήτη (classifier) Bayes με πολυωνυμική(multinomial) κατανομή και εξομάλυνση Laplace. Χρησιμοποιήστε το αρχείο *spam_train.txt* για την εκπαίδευση του μοντέλου και το *spam_test.txt* για να εκτιμήσετε την αποδοτικότητα του και υπολογίστε το σφάλμα. Διαισθητικά υπάρχουν κάποιες λέξεις κλειδιά που είναι ιδιαίτερα ενδεικτικές σε ποιά κλάση ανήκει ένα μήνυμα. Ένας τρόπος για να δούμε πόσο ενδεικτική είναι η λέξη i είναι να υπολογίσουμε

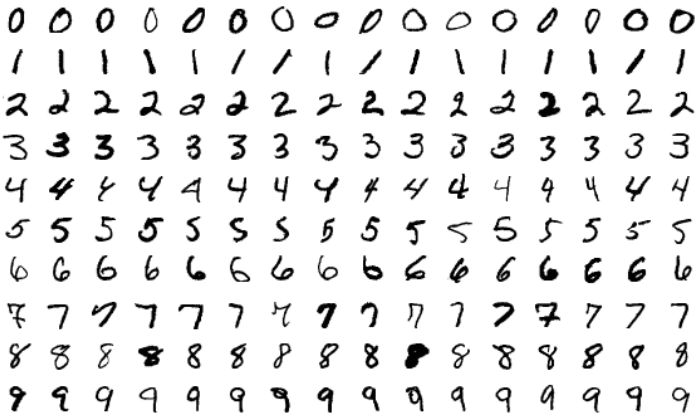
$$\log \frac{p(x_j = i \mid y = 1)}{p(x_j = i \mid y = 0)} = \log \left(\frac{P(word\, i \mid email\ is\ SPAM)}{P(word\, i \mid email\ is\ NoSPAM)} \right)$$

Χρησιμοποιήστε τον παραπάνω τύπο και υπολογίστε τις 5 πιο ενδεικτικές λέξεις.

3. **Μηχανές Διανυσμάτων Υποστήριξης**. Χρησιμοποιήστε τις συναρτήσεις που παρέχει το *Scikit-Learn* για τους ταξινομητές SVM για τον χαρακτηρισμό των μηνυμάτων σε κανονικά η ανεπιθύμητα. Χρησιμοποιήστε τον πυρήνα τύπου Gauss (Radial Basis Function) και πειραματιστείτε με τις παραμέτρους του πυρήνα. Βρείτε την ακτίνα του πυρήνα που δίνει το μικρότερο σφάλμα ως προς το σύνολο δοκιμής.

2 Νευρωνικά Δίκτυα: Ταξινόμηση Εικόνων

Με την βοήθεια της *Scikit-Learn*, θα υλοποιήσετε ένα απλό νευρωνικό δίκτυο για την ταξινόμηση ασπρόμαυρων εικόνων από χειρόγραφα ψηφία $0-9$ από την βάση δεδομένων MINST (https://www.nist.gov). Η βάση έχει 60000 εικόνες χειρόγραφων ψηφίων $0-9$ και 10000 εικόνες δοκιμής. Κάθε εικόνα, π.χ. Σχήμα 1, έχει διάσταση 28×28 και αναπαραστάται με ένα διάνυσμα από $784 = 28^2$ αριθμούς. Για κάθε εικόνα υπάρχει επίσης και μία ετικέττα πού δείχνει τον πραγματικό αριθμό που βρίσκεται σε αυτή. Τα σχετικά αρχεία είναι τα εξής: *images_train.csv, labels_train.csv* και *images_test.csv, labels_test.csv*. Το νευρωνικό δίκτυο που θα υλοποιήσετε θα



Σχήμα 1: Χειρόγραφα ψηφία

έχει ένα κρυφό επίπεδο(layer) που θα εκπαιδευτεί με το σύνολο εκπαίδευσης χρησιμοποιώντας την εντροπία ως σύναρτηση κόστους. Θα χρησιμοποιηθεί η σιγμοειδής-συνάρτηση ενεργοποίησης και την συνάρτηση $softmax$ στο επίπεδο εξόδου. Θυμίζουμε ότι για ένα δείγμα (x, y) η συνάρτηση κόστους εντροπίας είναι

$$J_y(y, \hat{y}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k, \quad \text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{k=1}^{K} \exp(z_k)}$$

όπου $\hat{y} \in \mathbb{R}^K$ είναι το διάνυσμα εξόδου από το μοντέλο για το δείγμα εκπαίδευσης x και $y \in \mathbb{R}^K$ το διάνυσμα που αντιστοιχεί στο δείγμα x της μορφής $y = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ με το 1 στη θέση της κατάλληλης κλάσης. Για τα δεδομένα $(x^{(i)}, y^{(i)})_{i=1}^n$ έχουμε $x^i \in \mathbb{R}^d$ και $y^{(i)} \in \mathbb{R}^K$ όπως παραπάνω. Έστω m ο αριθμός των κρυφών μονάδων του νευρωνικού δικτύου τότε $W^{(1)} \in$

$\mathbb{R}^{d\times m}, W^{(2)} \in \mathbb{R}^{d\times K}$ και $b^{(1)} \in \mathbb{R}^m,\ b^{(2)} \in \mathbb{R}^K$. Για το δείγμα $x^{(i)}$ έχουμε

$$\begin{aligned} a^{(i)} &= \sigma\left((W^{(1)})^T x^{(i)} + b^{(1)}\right) \in \mathbb{R}^m \quad \sigma = \text{σιγμοειδή} \\ z^{(i)} &= (W^{(2)})^T x^{(i)} + b^{(2)} \in \mathbb{R}^K \\ \hat{y}^{(i)} &= softmax(z^{(i)}) \in \mathbb{R}^K \end{aligned}$$

Για τα n-δείγματα εκπαίδευσης η εντροπική συνάρτηση κόστους είναι

$$J\left(W^{(1)},\ W^{(2)},\ b^{(1)},\ b^{(2)}\right) = \frac{1}{n}\sum_{i=1}^{n} J_{y^{(i)}}\left(y^{(i)}, \hat{y}^{(i)}\right) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k^{(i)}$$

Το δίκτυο θα εκπαιδευτεί με το σύνολο εκπαίδευσης και υπολογίσετε το σφάλμα του συνόλου δοκιμής. Να φτιαχτεί το γράφημα μείωσης της συνάρτησης κόστους ως πρός τον αριθμό επαναλήψεων. Επαναλάβετε τα ίδια προσθέτοντας ένα όρο εξομάλυνσης στη συνάρτηση κόστους της μορφής

$$J\left(W^{(1)},\ W^{(2)},\ b^{(1)},\ b^{(2)}\right) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k^{(i)} + \alpha\left(\|W^{(1)}\|_2^2 + \|W^{(2)}\|_2^2\right).$$

3 Παρατηρήσεις

1. Σκόπος της ασκήσης είναι να δούμε της συμπεριφορά των ταξινομητών και των νευρωνικών δικτύων και την χρήση τους σε πραγματικά προβλήματα.
2. Εξοικίωση με τις δυνατότητες της βιβλιοθήκες *Scikit-Learn* στο χώρο της Μηχανικής Μάθησης. Διαβάστε προσεκτικά της οδηγίες και δυνατότητες της κάθε συνάρτησης που θα χρησιμοποιήσετε.
3. Μην ξεχάσετε να κανονικοποιήσετε τα δεδομένα σας. Η βιβλιοθήκη *Scikit-Learn* έχει κατάλληλες συναρτήσεις για αυτό το σκοπό.

4 Παράδοση - Εξέταση

Για κάθε μέρος της άσκησης να φτιάξετε έναν διαφορετικό κώδικα Python με όνομα π.χ: $\{math, tem, ph\}XXXX\_Lab3\{a, b, c\}.py$ όπου $XXXX$ είναι ο αριθμός μητρώου σας. Επίσης στις πρώτες γραμμές του κάθε προγράμματος θα υπάρχουν σαν σχόλιο τα στοιχεία σας: όνομα, επώνυμο και ΑΜ. Θα στείλετε τους κώδικες (ως συνημμένα αρχεία) με email από τον **ιδρυματικό σας λογαριασμό** στη διεύθυνση $mem704labs@gmail.com$ το αργότερο μέχρι 18:00, Πέμπτη 19 Μαίου. Εκπρόθεσμες ασκήσεις δεν θα βαθμολογηθούν. Εργαστείται ατομικά. Κώδικες που είναι προιόν αντιγραφής θα μηδενίζονται.