Θέμα 1 Ερωτήσεις Θεωρίας (20)

1. Περιγράψτε εν συντομία το φαινόμενο του overfitting.
2. Σε ποιές περιπτώσεις χρησιμοποιείται ο αλγόριθμος EM (Expecation Maximization) και σε ποιές ο αλγόριθμος Maximum Likelihood (ML);
3. Ποιές είναι οι διαφορές μεταξύ των HMM (hidden markov models) και των MM (markov models);
4. Περιγράψτε το πρόβλημα vanishing/exploding gradient. Σε ποιές περιπτώσεις/αρχιτεκτονικές εμφανίζεται και γιατί; Μπορείτε να προτείνετε κάποιες λύσεις;
5. Περιγράψτε ορισμένες ομοιότητες και διαφορές των αλγορίθμων LDA και PCA.
6. Εξηγλησε γιατί ο ταξινομητής Bayes ελαχιστοποιεί το σφάλμα ταξινόμησης;

Θέμα 2 Μοντέλα Μάθησης (30)
Δίνονται τα ακόλουθα δείγματα εκπαίδευσης $D_1 = \{0, 1, 2, -3, 7, 8.2, 9.6, 7.2\}$ και $D_2 = \{2.1, 3, 4, 6.1\}$ από τις κατηγορίες ω_1 και ω_2 αντίστοιχα.
(a) Χρησημοποιήστε τον αλγόριθμο μεγιστοποίησης της πιθανότητας (maximum likelihood) για να υπολογίσετε τις κατανομές $p(x|\omega_1)$ και $p(x|\omega_2)$ και τις a priori πιθανότητες $p(\omega_1)$, $p(\omega_2)$. Υποθέστε ότι η κατανομή ω_1 είναι της μορφής $p(x|\omega_1) = w_{11}\mathcal{N}(\mu_{11}, 1) + w_{12}\mathcal{N}(\mu_{12}, 1)$ και η κατανομή ω_2 είναι της μορφής $\mathcal{N}(\mu_2, 1)$ όπου $\mathcal{N}(\mu, \sigma^2)$ είναι η Γκαουσιανή κατανομή

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

(b) Υπολογίστε τα διαστήματα απόφασης για τις κατηγορίες ω_1 και ω_2 σύμφωνα με τον κανόνα απόφασης του Bayes και τις κατανομές που υπολογίσατε στο (a).
(c) Υπολογίστε τα διαστήματα απόφασης για τις κατηγορίες ω_1 και ω_2 σύμφωνα με τον κανόνα απόφασης των 3 κοντινότερων γειτόνων NNR-3.
(d) Συγκλίνει (σε λύση) ο αλγόριθμος συναρτήσεων διαχωρισμού perceptron για τις κατηγορίες ω_1 και ω_2;
(e) Δεδομένου ότι οι πραγματικές κατανομές των δύο κατηγοριών είναι

$$p(x|\omega_1) = 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(8, 1) \text{ και } p(\omega_1) = \frac{2}{3}$$

$$p(x|\omega_2) = \mathcal{N}(4, 1) \text{ και } p(\omega_2) = \frac{1}{3}$$

ποιός από τους αλγόριθμους Maximum Likelihood, NNR-3 ελαχιστοποιεί το λάθος ταξινόμησης για το συγκεκριμένο παράδειγμα; Συγκρίνετε το λάθος ταξινόμησης Bayes με το λάθος ταξινόμησης για κάθε έναν από τους αλγορίθμους (συγκρίνοντας τα σημεία απόφασης).
(f) Έστω ένα σύνολο δειγμάτων D_t, το οποίο προκύπτει από την ένωση των συνόλων δειγμάτων D_1 και D_2. Συγκεκριμένα $D_t = D_1 \cup D_2$. Θεωρείστε ένα πρόβλημα ταξινόμησης δύο κατηγοριών. Εφαρμόστε τον αλγόριθμο k-means χρησιμοποιώντας ως αρχικά κέντρα τα σημεία 1 και 5

Υπόδειξη:
Αλγόριθμος Perceptron: Έστω $\mathbf{w}(t)$ η εκτίμηση του διανύσματος βάρους και $\mathbf{x}_t$ το αντίστοιχο διάνυσμα χαρακτηριστικών στο t-οστό βήμα επανάληψης. Ο αλγόριθμος έχει ως εξής:

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \rho\mathbf{x}_t \quad \text{αν} \quad \mathbf{x}_t \in \omega_1 \quad \text{και} \quad \mathbf{w}(t)^\top \mathbf{x}_t \leq 0$$

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \rho\mathbf{x}_t \quad \text{αν} \quad \mathbf{x}_t \in \omega_2 \quad \text{και} \quad \mathbf{w}(t)^\top \mathbf{x}_t \geq 0$$

$$\mathbf{w}(t+1) = \mathbf{w}(t) \quad \text{αλλιώς}$$

Θέμα 3 Ταξινομητές και υπολογισμός παραγώγων (20)

Απαντήστε στα παρακάτω ερωτήματα

1. Έστω 3 κόστη/κριτήρια εκπαίδευσης νευρωνικών δικτύων. Συγκεκριμένα a) Cross-Entropy, b) Mean Square Error. Εξηγήστε σε τί είδους πρόβλημα θα χρησιμοποιούσατε το καθένα. Τι σημαίνει το κάθε ένα διαισθητικά;
2. Σας δίνεται ένα σύνολο από N το πλήθος δείγματα ως $\{x_i, y_i\}$ όπου x_i είναι ένα διάνυσμα ($\mathbb{R}^d$) εισόδου και y_i η επισημείωση (label). Γράψτε το συνολικό αναμενόμενο κόστος για όλα τα δείγματα για κριτήρια (1α) και (1β). (Γενίκευση των τύπων για N το πλήθος δεδομένα.)
3. Σε προβλήματα ταξινόμησης χρησιμοποιούμε τη συνάρτηση softmax $s(\cdot)$ ως:
$$\pi_i = s(z)|_i = \frac{e^{z_i}}{\sum_{k=1}^{d} z_k}, z \in \mathbb{R}^d$$
   (a) Δείξτε ότι $\sum_{i=1}^{d} \pi_i = 1$. Τι σημαίνει αυτό το αποτέλεσμα;
   (b) Δείξτε ότι $s(\mathbf{z}) = s(\mathbf{z} + c)$, όπου c μια σταθερά η οποία προστίθεται σε κάθε στοιχείο του διανύσματος $\mathbf{z}$.
   (c) Υπολογίστε την παράγωγο $\frac{\vartheta \pi_i}{\vartheta z_j}$, για όλα τα $j = \{1, \cdots, N\}$.
   Tip: Υπολογίστε χωριστά όταν $i = j$ και $i \neq j$
4. Έστω πρόβλημα ταξινόμησης 3 κλάσεων. Και έστω η αρακάτω τοπολογία με 3 (γραμμικούς) νευρώνες. Θεωρήστε σαν θ_i τα διανύσματα βαρών που αντιστοιχούν σε κάθε νευρώνα (trainable parameters). Γράψτε τη συνολική συνρτηση κόστους $C(\theta)$ της παρακάτω τοπολογίας. (Εαν

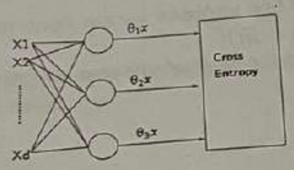

Figure 1: Cross Entropy Classifier

χρειαστεί οποιοσδήποτε επιπλέον υπολογισμός να τον αναφέρετε και θεωρήστε ότι περιλαμβάνετε μέσα στο Cross Entropy module).

5. Έστω τώρα η ακόλουθη τοπολογία. Υπολογίστε και εδώ τη συνάρτηση κόστους $C(\theta)$. Το block Log-Softmax υποδηλώνει πρώτα την εφαρμογή του τελεστή Softmax και έπειτα του Log. Το NLL block υποδηλώνει το Negative Log Likelihood block. Σχολιάστε το αποτέλεσμα.

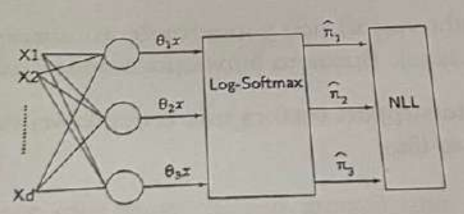

Figure 2: Negative Log Likelihood Topology

6. Υπολογίστε τις παραγώγους $\frac{\vartheta C(\theta)}{\vartheta \theta_2}$ για την τοπολογία του Σχήματος 1 (Cross Entropy Classifier). Κατά αντιστοιχία γράψτε και τους τύπους για τις μερικές παραγώγους ως προς τις μεταβλητές θ_1 και θ_3.

7. Υπολογίστε τις παραγώγους $\frac{\vartheta C(\theta)}{\vartheta \theta_2}$ για την τοπολογία του Σχήματος 2. Γενικεύστε και για τις μερικές παραγώγους ως προς τις παραμέτρους θ_1 και θ_3. Σχολιάστε το αποτέλεσμα.

Χρήσιμες Σχέσεις

Cross Entropy: $L(y, \hat{y}) = -\sum_{k=1}^{C} y_k \log \hat{y}_k$ MSE: $L(y, \hat{y}) = (y - \hat{y})^2$

Negative Log-Likelihood: $L(\hat{y}) = -\hat{\pi}_i$, i η σωστή κλάση

Θέμα 4 LDA (15)

Στο μάθημα είδαμε ότι η Linear Discriminant Analysis (LDA) βασίζεται στην ανάστροφη σχέση των μητρών (πινάκων) S_W και S_B:

$$S_W = \sum_{i=1}^{|Classes|} \mathbb{E}_{x|x\in\omega_i}[(\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^T]$$

$$S_B = \sum_{i=1}^{|Classes|} P(\omega_i)(\vec{\mu_i} - \vec{\mu})(\vec{\mu_i} - \vec{\mu})^T$$

όπου το ω_i αναπαριστά μια κλάση με μέση τιμή $\vec{\mu_i}$, $|Classes|$ είναι το πλήθος των κλάσεων και $\vec{\mu}$ είναι η μέση τιμή όλων των δειγμάτων.

(a) Δείξτε ότι στην περίπτωση διαχωρισμού δύο κλάσεων ω_1 και ω_2, ο πίνακας S_B μπορεί να γραφτεί στη μορφή $S_B = P(\omega_1)P(\omega_2)(\vec{\mu_2} - \vec{\mu_1})(\vec{\mu_2} - \vec{\mu_1})^T$

(b) Βασιζόμενοι στο υποερώτημα (a), να βρείτε το ιδιοδιάνυσμα του πίνακα $S_W^{-1} S_B$ και την ιδιοτιμή του.

Θέμα 5 SVM (15)

Έστω τα ακόλουθα δεδομένα εκπαίδευσης σε δύο διαστάσεις.

class	x_1	x_2
+	1	1
+	2	2
+	2	0
−	0	0
−	1	0
−	0	1

1. Απεικονίστε τα έξι σημεία εκπαίδευσης. Είναι γραμμικά διαχωρίσιμα;

2. Υπολογίστε το βάρος (weight vector) που χαρακτηρίζει το υπερεπίπεδου μέγιστου περιθωρίου (maximum margin hyperplane). Βρείτε τα διανύσματα υποστήριξης (support vectors).

3. Εαν αφαιρέσουμε ένα από τα support vectors τότε τι συμβαίνει στο βέλτιστο περιθώριο; Αυξάνεται, μειώνεται ή παραμένει το ίδιο;

4. (Bonus) Είναι η απάντηση που δώσατε στο 3. καθολική για κάθε σύνολο δεδομένων εκπαίδευσης; Δώστε ένα αντιπαράδειγμα ή μια σύντομη απόδειξη για την απάντησή σας.