Μηχανική Μάθηση: 3^{ο} Σετ Ασκήσεων

Πρόβλημα 3.1: Θέλουμε να αξιολογήσουμε τις δυνατότητες της πρώτης μεθόδου kernel στο να προσεγγίζει πυκνότητες πιθανότητας. Για τον σκοπό αυτό δημιουργούμε 1000 υλοποιήσεις μιας τυχαίας μεταβλητής ομοιόμορφα κατανεμημένης στο διάστημα $[0, 1]$. Προσεγγίστε την πυκνότητα πιθανότητας χρησιμοποιώντας το Gaussian kernel

$$\mathsf{K}(x, h) = \frac{1}{\sqrt{2\pi h}} e^{-\frac{1}{2h}x^2}.$$

Ζωγραφίστε την αντίστοιχη προσέγγιση για διαφορετικές τιμές του h. Μην περιορίσετε το γράφημά σας στο $[0, 1]$ αλλά να το επεκτείνετε πέραν των δύο αυτών ορίων. Τι παρατηρείτε σχετικά με τον τρόπο που προσεγγίζετε το διάστημα $[0, 1]$ όπου "ζει" η τυχαία μεταβλητή καθώς και τον τρόπο που προσεγγίζετε την τιμή 1 που είναι το σταθερό πλάτος της συνάρτησης πυκνότητας πιθανότητας.

Πρόβλημα 3.2: Το Matlab αρχείο δεδομένων `data32.mat` περιέχει δύο μήτρες: `stars` και `circles` όπου κάθε μια είναι μια λίστα από δι-διάστατα (2-D) διανύσματα. Κάθε 2-D διάνυσμα αντιστοιχεί σε ένα σημείο στον 2-D επίπεδο και έχει σαν ετικέτα (label) "star" ή "circle". Ενδιαφερόμαστε να δημιουργήσουμε ένα classifier ο οποίος να διακρίνει μεταξύ των δύο συνόλων. Θα χρησιμοποιήσουμε τη μέθοδο με τα kernel ώστε να βρούμε ένα διαχωριστικό σύνορο. Για το σκοπό αυτό αντιστοιχίζουμε την αριθμητική ετικέτα "1" στο `stars` και την "-1" στο `circles`. Καλούμε $\phi(X), X = [x_1, x_2]^\intercal$ το μετασχηματισμό που επιθυμούμε να εφαρμόσουμε στα δεδομένα ο οποίος θα υλοποιεί την εν λόγω αντιστοίχηση. Προκειμένου να προσδιορίσουμε τον $\phi(X)$ επιλύουμε το παρακάτω πρόβλημα βελτιστοποίησης

$$\min_{\phi\in\mathcal{V}} \Big\{ \sum_{X_i\in\text{stars}} \big(1 - \phi(X_i)\big)^2 + \sum_{X_j\in\text{circles}} \big(1 + \phi(X_j)\big)^2 + \lambda\|\phi(X)\|^2 \Big\}, \tag{1}$$

όπου $\mathcal{V}$ είναι ο διανυσματικός χώρος των συναρτήσεων που ορίζονται με τη βοήθεια του Gaussian kernel

$$\mathsf{K}(X, Y) = e^{-\frac{1}{h}\|X-Y\|^2} = e^{-\frac{1}{h}\{(x_1-y_1)^2+(x_2-y_2)^2\}}.$$

α) Χρησιμοποιείστε το Representer Theorem για να αποδείξετε ότι στα πρώτα δύο αθροίσματα μπορούμε να αντικαταστήσουμε το $\phi(X)$ με την ορθογώνια προβολή του $\hat{\phi}(X)$ πάνω στον γραμμικό υπόχωρο ο οποίος δημιουργείται από τις συναρτήσεις $\mathsf{K}(X, X_i), X_i \in$ stars and $\mathsf{K}(X, X_j), X_j \in$ circles δηλαδή

$$\hat{\phi}(X) = \sum_{X_i\in\text{stars}} \alpha_i \mathsf{K}(X, X_i) + \sum_{X_j\in\text{circles}} \beta_j \mathsf{K}(X, X_j). \tag{2}$$

β) Για τον όρο $\|\phi(X)\|^2$ χρησιμοποιείστε την Αρχή της Ορθογωνιότητας για να δείξετε επίσης ότι

$$\|\phi(X)\|^2 = \|\hat{\phi}(X)\|^2 + \|\phi(X) - \hat{\phi}(X)\|^2 \geq \|\hat{\phi}(X)\|^2.$$

Εξηγείστε γιατί η προηγούμενη ανισότητα υπονοεί ότι μπορούμε να αντικαταστήσουμε το $\phi(X)$ με το $\hat{\phi}(X)$ στην αρχική συνάρτηση κόστους στο (1). γ) Χρησιμοποιώντας τη μορφή του $\hat{\phi}(X)$ που ορίστηκε στην (2), υπολογίστε τους βέλτιστους συντελεστές α_i, β_j. δ) Όταν προσδιορίσετε το βέλτιστο $\hat{\phi}(X)$ εξηγείστε πως θα το χρησιμοποιήσετε για να κατηγοριοποιήσετε ένα νέο σημείο X_{new} σε "star" ή "circle" λαμβάνοντας υπόψη ότι φυσικά η τιμή του $\hat{\phi}(X_{\text{new}})$ δεν θα είναι ακριβώς 1 ή -1. ε) Αφού καταλήξετε στον τελικό σας κανόνα κατηγοριοποίησης στο δ) βρείτε (αριθμητικά) το διαχωριστικό σύνορο για τις δύο κλάσεις στον 2-D χώρο. Επίσης τοποθετείστε τα δεδομένα εκπαίδευσης στο 2-D επίπεδο ώστε να αξιολογείσετε την ποιότητα του συνόρου σας. Επαναλάβετε τη διαδικασία για διαφορετικές τιμές των παραμέτρων h και λ.

Πρόβλημα 3.3: Το αρχείο data33.mat περιέχει $N = 200$ διανύσματα μήκους 2, τα οποία επιθυμούμε να ομαδοποιήσουμε σε δύο ομάδες. Σας δίνεται η **ΜΥΣΤΙΚΗ πληροφορία την οποία ΑΠΑΓΟΡΕΥΕΤΑΙ να χρησιμοποιήσετε στη μέθοδο ομαδοποίησης που θα δημιουργήσετε:** ότι τα πρώτα 100 διανύσματα ανήκουν σε μια ομάδα και τα δεύτερα 100 σε άλλη.

α) Εφαρμόστε την K-means ώστε να κάνετε την ομαδοποίηση. Αφού συγκλίνει η μέθοδος μετρήστε το ποσοστό των σφαλμάτων που κάνατε κάνοντας χρήση της μυστικής πληροφορίας. *Προσοχή!!!: Το τι είναι πρώτη ομάδα και τι δεύτερη δεν είναι καθορισμένο. Οπότε από τις δύο δυνατές επιλογές θα διαλέξετε εκείνη που κάνει τα λιγότερα λάθη! Επίσης έχετε υπόψη σας ότι εάν επιλέξετε εντελώς τυχαία την ομάδα σε κάθε διάνυσμα η πιθανότητα να κάνετε λάθος είναι 0.5. Επομένως εάν η μέθοδός σας αποδίδει ποσοστό σφάλματος κοντά στο 0.5 τότε φυσικά κάτι δεν έχετε κάνει σωστά!*

β) Ένας τρόπος να βελτιώσετε την απόδοση της K-means είναι να παρατηρήσετε ότι έχει την τάση να δημιουργεί γραμμικά διαχωριστικά όρια. Επομένως εάν αυξήσετε με τεχνητό τρόπο τη διάσταση των δεδομένων σας αυτό μπορεί να βελτιώσει τα αποτελέσματα. Δημιουργείστε μια τρίτη συντεταγμένη όπου κάθε δισδιάστατο X_i θα αντικατασταθεί από το τρισδιάστατο $\{X_i, \|X_i\|^2\}$. Με άλλα λόγια, σημεία X_i που είναι κοντά στην αρχή των αξόνων στις τρεις διαστάσεις θα είναι πιο κοντά στο οριζόντιο επίπεδο από ό,τι σημεία που βρίσκονται μακρύτερα. Φυσικά εδώ κάνουμε μια ελαφρά χρήση της μυστικής πληροφορίας μιας και παρατηρούμε ότι τα σημεία της μιας ομάδας βρίσκονται συγκεντρωμένα γύρω από την αρχή των αξόνων, αλλά δεν πειράζει. Επαναλάβατε την K-means με τα σημεία στον τρισδιάστατο χώρο και υπολογίστε πάλι το σφάλμα ομαδοποίησης.

Για να δείτε τα κατορθώματά σας, τοποθετείστε τα δεδομένα στο επίπεδο σαν τελείες με διαφορετικό χρώμα κάθε ομάδα, χρησιμοποιώντας τη μυστική πληροφορία που έχετε. Κατόπιν σε κάθε σημείο βάλτε ένα κύκλο με το χρώμα της ομάδας που επιλέγετε για το σημείο αυτό. Το κάνετε αυτό για κάθε μία από τις δύο μεθόδους σε χωριστές εικόνες. Σε όσο πιο πολλά σημεία οι τελείες και οι κύκλοι έχουν το ίδιο χρώμα, τόσο καλύτερη είναι η αντίστοιχη μέθοδος.

Παρατηρήσεις

- Παράδοση αναφοράς έως την Παρασκευή 10 Ιουνίου, 12:00 το μεσημέρι. Η παράδοση θα γίνει ηλεκτρονικά στο eclass. Το όνομα του αρχείου θα πρέπει να έχει την μορφή:

  AM-3.pdf

  Το ΑΜ είναι ο αριθμός μητρώου σας ΜΟΝΟ με τα νούμερα ΔΙΧΩΣ το UP. Το "-3" υποδηλώνει ότι είναι η τρίτη σειρά ασκήσεων. ΟΧΙ ΚΕΝΑ ΕΚΑΤΕΡΩΘΕΝ ΤΗΣ ΠΑΥΛΑΣ.

- Στην πρώτη σελίδα της αναφοράς να γράψετε όνομα-επώνυμο, τμήμα και έτος σπουδών. Αν κάνετε μεταπτυχιακό ή διδακτορικό τότε το μεταπτυχιακό/διδακτορικό πρόγραμμα.

- Η αναφορά σας θα είναι σε μορφή PDF και θα έχει ονομασία AM-3.pdf. ΚΑΝΕΝΑ ΑΛΛΟ format (π.χ. AM-2.doc) ή καμία άλλη ονομασία ΔΕΝ ΘΑ ΓΙΝΟΝΤΑΙ ΔΕΚΤΑ. ΘΑ ΣΑΣ ΕΠΙΣΤΡΕΦΕΤΑΙ ΤΟ ΑΡΧΕΙΟ και θα πρέπει να το υποβάλετε εκ νέου με τον σωστό τύπο ή/και ονομασία.

- Μη στέλνετε κώδικα Python ή Matlab σε χωριστά αρχεία. Ο κώδικάς σας σε μορφή text να ενσωματωθεί στο PDF αρχείο της αναφοράς σας **ΜΕΤΑ το τέλος της παρουσίασης των αποτελεσμάτων**. **ΜΗ** βάζετε κώδικα ανάμεσα στο κείμενο της παρουσίασης γιατί δυσκολεύει στην κατανόηση του τι κάνατε και πέρα από την ταλαιπωρία που δημιουργεί υπάρχει κίνδυνος να μας διαφύγουν τα ουσιαστικά και να πάρετε κακό βαθμό. Ο κώδικας **ΔΕΝ αποτελεί αναφορά**. Τον επισυνάπτετε για την περίπτωση που θα θελήσουμε να δούμε με μεγαλύτερη λεπτομέρεια πως βγάλατε κάποιο αποτέλεσμα. **Αναφορά μόνο με κώδικα θα βαθμολογηθεί με 0.**

- Η ΗΜΕΡΑ ΚΑΙ ΩΡΑ υποβολής ΕΙΝΑΙ ΑΝΕΛΑΣΤΙΚΕΣ. Αυτό περιλαμβάνει και την περίπτωση οποιουδήποτε λάθους κάνετε. Αν υποβάλετε λάθος, η διόρθωση ΠΡΕΠΕΙ ΝΑ ΥΠΟΒΛΗΘΕΙ ΠΡΙΝ ΤΙΣ 12 το μεσημέρι της Παρασκευής!!!! **Ωρα 12:01 είναι ήδη ΑΡΓΑ!!! Απλά στις 12:00 κλείνει το σύστημα (eclass) και δεν δέχεται πλέον αναφορές. ΟΥΤΕ ΠΡΟΚΕΙΤΑΙ ΝΑ ΔΕΧΘΟΥΜΕ ΑΝΑΦΟΡΕΣ ΑΡΓΟΠΟΡΗΜΕΝΕΣ ΕΚΤΟΣ ECLASS ΜΕ ΑΠ' ΕΥΘΕΙΑΣ EMAIL.**

- Φροντίστε το αρχείο σας να μην είναι μεγάλο σε όγκο. Ο λόγος είναι ότι φυλάσσονται τα αρχεία όλων των ετών και δεν επιθυμούμε να γεμίσει ο δίσκος με τα γραπτά σας!!! Σε περίπτωση που έχετε χειρόγραφη αναφορά *να "σκανάρετε" με μαύρο/άσπρο (2 επίπεδα χρωμάτων) οπότε για να είναι ευανάγνωστο το αποτέλεσμα φροντίστε να γράψετε με **μαύρο** στυλό (αλλά όχι μολύβι).*
- **Μη στέλνετε ερωτήσεις με emails. Είναι ΑΔΥΝΑΤΟ να βρισκόμαστε επί ώρες μπροστά στον υπολογιστή και να απαντάμε στα ίδια ερωτήματα στον καθένα σας!!!!**
- Βαθμολογούνται αναφορές μόνον όσων δηλώσουν το μάθημα στο Progress.

Ε. Ψαράκης
Γ. Μουστακίδης