

Μηχανική Μάθηση: 1^ο Σετ Ασκήσεων

Πρόβλημα 1: Έστω τυχαίο διάνυσμα $X = [x_1, x_2]$ μήκους 2 για το οποίο έχουμε τις εξής δύο υποθέσεις:

H_0 : x_1, x_2 είναι ανεξάρτητα με πυκνότητα πιθανότητας $f_0(x_1, x_2) = f_0(x_1)f_0(x_2)$ όπου $f_0(x) \sim \mathcal{N}(0, 1)$.

H_1 : x_1, x_2 είναι ανεξάρτητα με πυκνότητα πιθανότητας $f_1(x_1, x_2) = f_1(x_1)f_1(x_2)$ όπου $f_1(x) \sim 0.5\{\mathcal{N}(-1, 1) + \mathcal{N}(1, 1)\}$.

Υποθέστε ότι οι δύο αρχικές πιθανότητες $P(H_0) = P(H_1) = 0.5$. α) Ποιό είναι το βέλτιστο τεστ κατά Bayes που ελαχιστοποιεί την πιθανότητα σφάλματος απόφασης; β) Για να υπολογίσετε την πιθανότητα σφάλματος θα χρησιμοποιήσετε simulations. Δημιουργήστε 10^6 ζευγάρια $[x_1, x_2]$ από την $f_0(x_1, x_2)$ και άλλα 10^6 ζευγάρια από την $f_1(x_1, x_2)$. Όσον αφορά στην $f_1(x)$ με πιθανότητα 0.5 το δείγμα έχει μέσο όρο -1 και με πιθανότητα 0.5 μέσο όρο 1. Αφού δημιουργήσετε τις δύο κατηγορίες δειγμάτων θα εφαρμόσετε τον κανόνα Bayes και θα μετρήσετε τα ποσοστά των λάνθασμένων αποφάσεων στις δύο περιπτώσεις και θα τα αθροίσετε ώστε να βρείτε το τελικό αθροιστικό ποσοστό σφάλματος. Αυτό είναι το βέλτιστο ποσοστό που δεν μπορεί να το ξεπεράσει καμία άλλη μέθοδος.

γ) Τα ζευγάρια που δημιουργήσατε πριν, να τα διατηρήσετε γιατί θα τα χρησιμοποιήσετε στο ερώτημα αυτό. Δημιουργήστε επιπλέον 200 ζευγάρια από κάθε υπόθεση για να τα χρησιμοποιήσετε για την εκπαίδευση νευρωνικών δικτύων. Ενδιαφερόμαστε για fully connected νευρωνικά δίκτυα διαστάσεων $2 \times 20 \times 1$ (είσοδος μεγέθους 2, κρυφό επίπεδο μεγέθους 20 και μία έξοδος) τα οποία εκτιμούν κάποιο γνωστό μετασχηματισμό του λόγου πιθανοφάνειας $r(x_1, x_2) = \frac{f_1(x_1)f_1(x_2)}{f_0(x_1)f_0(x_2)}$. Συγκεκριμένα θα εφαρμόσουμε 1) την cross-entropy μέθοδο με $\phi(z) = -\log(1 - z)$, $\psi(z) = -\log(z)$ η οποία, όπως εξηγήσαμε, εκτιμά την εκ των υστέρων πιθανότητα $\frac{r(x_1, x_2)}{1+r(x_1, x_2)}$ και 2) την exponential με $\phi(z) = e^{0.5z}$, $\psi(z) = e^{-0.5z}$ η οποία εκτιμά το λογάριθμο $\log r(x_1, x_2)$. Χρησιμοποιείτε τα 200+200 δεδομένα εκπαίδευσης (training data) για να εκπαιδεύσετε τα δύο νευρωνικά δίκτυα εφαρμόζοντας τον stochastic gradient descent. Όταν τελειώνουν τα δεδομένα εκπαίδευσης τα επαναχρησιμοποιείτε από τη αρχή έως ότου συγκλίνει ο αλγόριθμος. Τη σύγκλιση μπορείτε να τη διαπιστώσετε παρακολουθώντας τις τιμές του κόστους $\phi(u(x_{1,t}^0, x_{2,t}^0, \theta_t)) + \psi(u(x_{1,t}^1, x_{2,t}^1, \theta_t))$ αφού το εξομαλύνετε υπολογίζοντας τον μέσο όρο των τελευταίων 20 τιμών. Στην περίπτωση του cross-entropy θα πρέπει στην έξοδο του δικτύου να εφαρμόσετε, όπως έχουμε εξηγήσει, μια μη γραμμικότητα ώστε η τιμή της εξόδου να είναι στο διάστημα $[0, 1]$ στο οποίο βρίσκεται η τιμή της εκ των υστέρων πιθανότητας. Στην περίπτωση του exponential κάτι αντίστοιχο δεν είναι απαραίτητο γιατί το exponential εκτιμά το λογάριθμο του λόγου πιθανοφάνειας που είναι οποιοσδήποτε πραγματικός θετικός ή αρνητικός. Όταν θα έχουν συγκλίνει οι δύο αλγόριθμοι τότε θα πάρετε τα δύο νευρωνικά δίκτυα και θα τα εφαρμόσετε στα $10^6 + 10^6$ δεδομένα που έχετε από το ερώτημα β) για να διαπιστώσετε τι ποσοστά λάθους κάνουν στην απόφαση. Θυμόμαστε ότι το τεστ του λόγου πιθανοφάνειας αποφασίζει υπέρ του H_1 όταν $r(x_1, x_2) > 1$ και υπέρ του H_0 όταν $r(x_1, x_2) < 1$. Εάν εφαρμόσετε τον λογάριθμο τότε θα συγκρίνετε το $\log r(x_1, x_2)$ με το 0, ενώ αν εφαρμόσετε την εκ των υστέρων πιθανότητα $\frac{r(x_1, x_2)}{1+r(x_1, x_2)}$ θα την συγκρίνετε με το 0.5. Το $\log r(x_1, x_2)$ το προσεγγίζει το νευρωνικό δίκτυο του exponential και το $\frac{r(x_1, x_2)}{1+r(x_1, x_2)}$ το νευρωνικό δίκτυο του cross-entropy. Πώς συγκρίνονται τα συνολικά σφάλματα των δύο μεθόδων με το βέλτιστο σφάλμα της μεθόδου Bayes;

Πρόβλημα 2: Εφαρμόστε την ιδέα του ερωτήματος γ) του Προβλήματος 1 στα δεδομένα της βιβλιοθήκης MNIST. Απομονώστε τα νούμερα 0 και 8 και δημιουργήστε classifier που διακρίνει μεταξύ των δύο αριθμών χρησιμοποιώντας νευρωνικό δίκτυο που το εκπαιδεύετε με α) Hinge, β) Cross-entropy και γ) Exponential. Η MNIST έχει 5500 εικόνες διαστάσεων 28×28 σε αποχρώσεις του γκρι για training. Χρησιμοποιείτε πλήρες νευρωνικό δίκτυο διαστάσεων $784 \times 300 \times 1$ ($784 = 28 \cdot 28$ εισόδους, δηλαδή οι εικόνες θα γίνονται διανύσματα!, 300 κρυφό επίπεδο και 1 έξοδος). Φροντίστε τα pixel των εικόνων να έχουν τιμές στο διάστημα $[0, 1]$. Αν τα επίπεδα του γκρι είναι μεταξύ $[0, 255]$ διαιρέστε τις τιμές τους με 255. Η MNIST διαθέτει επίσης επιπλέον εικόνες για testing. Για κάθε νούμερο ο αριθμός των εικόνων αυτών είναι διαφορετικός. Όταν σχεδιάσετε τα νευρωνικά δίκτυα με τα training δεδομένα μετά τα εφαρμόζετε στα testing προκειμένου να δείτε τα ποσοστά σφαλμάτων. Κάντε ένα πίνακα που για κάθε μέθοδο θα δίνει τα ποσοστά σφάλματος ανά υπόθεση καθώς και το συνολικό ποσοστό σφάλματος.

Υπόδειξη: Η αρχικοποίηση των παραμέτρων των νευρωνικών δικτύων πρέπει να γίνει ως εξής: Τα offset θα τα ξεκινάτε με 0. Κάθε επίπεδο έχει μια μήτρα βαρών W που έχει διαστάσεις $m \times n$. Τα στοιχεία της μήτρας θα τα επιλέγετε τυχαία, Gaussian με μέση τιμή 0 και διασπορά $\frac{1}{n+m}$. Θα πρέπει να υπολογίσετε την κλίση του δικτύου ως προς τις παραμέτρους του. Σας διατίθεται για το σκοπό αυτό το αρχείο derivatives.pdf που εξηγεί πως να βρίσκεται τις παραγώγους καθώς και πως να τις κανονικοποιείτε (μέθοδος ADAMS) κατά την εφαρμογή του training.

Στην επόμενη σελίδα παρατηρήσεις!

Παρατηρήσεις

- Παράδοση αναφοράς έως την Παρασκευή 6 Μαΐου, 12:00 το μεσημέρι. Η παράδοση θα γίνει ηλεκτρονικά στο eclass. Το όνομα του αρχείου θα πρέπει να έχει την μορφή:

AM-1.pdf

Το AM είναι ο αριθμός μητρώου σας ΜΟΝΟ με τα νούμερα ΔΙΧΩΣ το UP. Το “-1” υποδηλώνει ότι είναι η πρώτη Άσκηση. 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-1.pdf. ΚΑΝΕΝΑ ΑΛΛΟ format (π.χ. AM-1.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 επισυνάπτεται 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.**
- 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 βρίσκομαστε επί ώρες μπροστά στον υπολογιστή και να απαντάμε στα ίδια ερωτήματα στον καθένα σας!!!!** Υπάρχει βίντεο με ερωτήσεις και απαντήσεις.
- Βαθμολογούνται αναφορές μόνον όσων δηλώσουν το μάθημα στο Progress. Οπότε η αξιολόγηση των αναφορών θα ξεκινήσει αφού ολοκληρωθεί η περίοδος δήλωσης μαθημάτων.

Ε. Ψαράκης

Γ. Μουστακίδης