

2η προαιρετική εργασία για το μάθημα «Αναγνώριση προτύπων»

Σημειώσεις:

1. Η παρούσα εργασία είναι η δεύτερη από 2 συνολικά εργασίες, οι οποίες είναι προαιρετικές και θα βαθμολογηθούν (και οι δύο) με μία μονάδα επιπλέον του 10.
2. Οι απαντήσεις να όσο το δυνατόν συντομότερες.
3. Οι απαιτούμενοι κώδικες επισυνάπτονται.

1^η άσκηση (αλγόριθμος perceptron):

Να παράγετε τέσσερα δισδιάστατα σύνολα δεδομένων X_i , $i = 1, \dots, 4$, καθένα από τα οποία περιέχει δεδομένα από δύο κλάσεις. Σε όλα τα X_i , η πρώτη κλάση (σημειώνεται με -1) περιέχει 150 διανύσματα ομοιόμορφα ταξινομημένα στο τετράγωνο $[0, 2] \times [0, 2]$. Η δεύτερη κλάση (σημειώνεται με $+1$) περιέχει άλλα 150 σημεία, τα οποία είναι ομοιόμορφα ταξινομημένα στα τετράγωνα $[3, 5] \times [3, 5]$, $[2, 4] \times [2, 4]$, $[0, 2] \times [2, 4]$ και $[1, 3] \times [1, 3]$, για τα σύνολα X_1 , X_2 , X_3 και X_4 , αντίστοιχα. Κάθε διάνυσμα επαυξάνεται με μία τρίτη συνιστώσα που ισούται με 1, για όλα τα διανύσματα.

Εκτελέστε τα ακόλουθα βήματα:

1. Απεικονίστε γραφικά τα τέσσερα σύνολα δεδομένων και παρατηρήστε ότι, καθώς κινούμαστε από το X_1 προς το X_3 οι κλάσεις προσεγγίζουν μεταξύ τους αλλά παραμένουν γραμμικώς διαχωρίσιμες. Στο X_4 οι δύο κλάσεις παρουσιάζουν επικάλυψη.
2. Εκτελέστε τον αλγόριθμο του perceptron για κάθε X_i , $i = 1, \dots, 4$, για ρυθμούς εκμάθησης $\rho = 0.01, 0.05$ και αρχική εκτίμηση για το διάνυσμα παραμέτρων την $[1, 1, -0.5]^T$.
3. Εκτελέστε τον αλγόριθμο του perceptron για σύνολο X_3 , για ρυθμό εκμάθησης $\rho = 0.05$ και αρχικές εκτιμήσεις για το διάνυσμα παραμέτρων τις $[1, 1, -0.5]^T$, $[1, 1, 0.5]^T$.
4. Σχολιάστε τα αποτελέσματα που προκύπτουν.

2^η άσκηση (ταξινομητής ελάχιστου αθροίσματος τετραγωνικών σφαλμάτων):

1. Να παράγετε ένα σύνολο X_1 , που περιέχει $N_1 = 300$ διανύσματα δεδομένων, έτσι ώστε τα πρώτα 150 να προέρχονται από την κλάση ω_1 , η οποία μοντελοποιείται από την κανονική κατανομή με μέση τιμή $m_1 = [0, 0, 0, 0, 0]^T$. Τα υπόλοιπα διανύσματα προέρχονται από την κλάση ω_2 , η οποία αποτελείται από την κανονική κατανομή με μέση τιμή $m_2 = [1, 1, 1, 1, 1]^T$. Το (κοινό) μητρώο συνδιασποράς για αμφότερες τις κατανομές είναι

$$S = \begin{bmatrix} 0.9 & 0.3 & 0.2 & 0.05 & 0.02 \\ 0.3 & 0.8 & 0.1 & 0.2 & 0.05 \\ 0.2 & 0.1 & 0.7 & 0.015 & 0.07 \\ 0.05 & 0.2 & 0.015 & 0.8 & 0.01 \\ 0.02 & 0.05 & 0.07 & 0.01 & 0.75 \end{bmatrix}$$

Να παράγετε ένα επιπλέον σύνολο δεδομένων X_2 , που περιέχει $N_2 = 300$ διανύσματα δεδομένων, ακολουθώντας τις οδηγίες για την παραγωγή του X_1 . Εφαρμόστε τον βέλτιστο ταξινομητή Bayes στο σύνολο X_2 και υπολογίστε το λάθος ταξινόμησης.

- Επεκτείνετε κάθε διάνυσμα χαρακτηριστικών στα X_1 και X_2 προσθέτοντας μία μονάδα ως τελευταία συνιστώσα. Ορίστε τις ετικέτες κλάσης ως -1 και $+1$ για τις δύο κλάσεις, αντίστοιχα. Χρησιμοποιώντας το X_1 ως σύνολο εκπαίδευσης, εφαρμόστε την $SSErr$ συνάρτηση του *MATLAB* (με $C = 0$) για να πάρετε την εκτίμηση του ελάχιστου αθροίσματος τετραγωνικών σφαλμάτων, \hat{w} . Χρησιμοποιήστε την εκτίμηση αυτή προκειμένου να ταξινομήσετε τα διανύσματα του X_2 , σύμφωνα με την ανισότητα

$$\hat{w}^T x > (<) 0$$

Υπολογίστε την πιθανότητα λάθους. Συγκρίνετε τα αποτελέσματα με αυτά που προέκυψαν από το βήμα 1.

- Επαναλάβετε τα προηγούμενα βήματα, αντικαθιστώντας πρώτα το X_2 με ένα σύνολο X_3 , που περιέχει $N_3 = 10.000$ σημεία και, στη συνέχεια, με ένα σύνολο X_4 , που περιέχει $N_4 = 100.000$ σημεία. Αμφότερα τα σύνολα X_3 και X_4 παράγονται χρησιμοποιώντας τη συνταγή για το X_1 . Σχολιάστε τα αποτελέσματα.

3^η άσκηση (ταξινομητής ελάχιστου αθροίσματος τετραγωνικών σφαλμάτων):

- Θεωρείστε ένα πρόβλημα δύο κλάσεων, $+1$ και -1 , στον δισδιάστατο χώρο, οι οποίες μοντελοποιούνται από κανονικές κατανομές με μέσες τιμές $m_1 = [0, 0]^T$, $m_2 = [7, 7]^T$ και μητρώα συνδιασποράς $S_1 = 2I$ και $S_2 = 0.2I$, αντίστοιχα. Δημιουργήστε ένα σύνολο X_1 (σύνολο εκπαίδευσης), με $N_1 = 300$ σημεία συνολικά, από τα οποία τα 150 προέρχονται από τη μία κλάση ενώ τα υπόλοιπα από την άλλη κλάση. Με τον ίδιο τρόπο δημιουργήστε το σύνολο X_2 (σύνολο δοκιμής).
 - Προσδιορίστε το γραμμικό ταξινομητή που ελαχιστοποιεί το άθροισμα των τετραγωνικών σφαλμάτων, χρησιμοποιώντας το σύνολο X_1 .
 - Απεικονίστε γραφικά το σύνολο X_1 (χρησιμοποιήστε διαφορετικά σύμβολα για σημεία διαφορετικών κλάσεων) και την γραμμή του ταξινομητή.
 - Προσδιορίστε το σφάλμα του ταξινομητή πάνω στο σύνολο X_2 και συγκρίνετέ το με το αντίστοιχο σφάλμα που προκύπτει από τον ταξινομητή Bayes.
 - Εξάγετε τα συμπεράσματά σας.
- Θεωρείστε ένα πρόβλημα δύο κλάσεων, $+1$ και -1 , στον δισδιάστατο χώρο, οι οποίες μοντελοποιούνται από κανονικές κατανομές με μέσες τιμές $m_1 = [0, 0]^T$, $m_2 = [4, 4]^T$ και μητρώα συνδιασποράς $S_1 = S_2 = I$, αντίστοιχα. Δημιουργήστε ένα σύνολο X_1 (σύνολο εκπαίδευσης), με $N_1 = 300$ σημεία συνολικά, από τα οποία

τα 255 προέρχονται από τη μία κλάση ενώ τα υπόλοιπα από την άλλη κλάση. Με τον ίδιο τρόπο δημιουργήστε το σύνολο X_2 (σύνολο δοκιμής).

- Προσδιορίστε το γραμμικό ταξινομητή που ελαχιστοποιεί το άθροισμα των τετραγωνικών σφαλμάτων, με βάση το X_1 .
- Απεικονίστε γραφικά το σύνολο X_1 (χρησιμοποιήστε διαφορετικά σύμβολα για σημεία διαφορετικών κλάσεων) και την γραμμή του ταξινομητή.
- Προσδιορίστε το σφάλμα του ταξινομητή πάνω στο σύνολο X_2 και συγκρίνετέ το με το αντίστοιχο σφάλμα που προκύπτει από τον ταξινομητή Bayes.
- Εξάγετε τα συμπεράσματά σας.

Σημείωση: Για την άσκηση αυτή θα πρέπει να δημιουργήσετε μόνοι σας τον κώδικα που απαιτείται, ακολουθώντας τα βασικά βήματα της προηγούμενης άσκησης.

4^η άσκηση (Μηχανές διανυσματικής στήριξης (SVM) – η γραμμική περίπτωση):

Δίνονται δύο ισοπίθανες κλάσεις στο δισδιάστατο χώρο, οι οποίες ακολουθούν κανονικές κατανομές με μέσες τιμές $m_1 = [0, 0]^T$ και $m_2 = [1.2, 1.2]^T$, ενώ τα μητρώα συνδιασποράς τους είναι $S_1 = S_2 = 0.2I$, όπου I είναι ο ταυτοτικός πίνακας τάξης 2×2 .

- Να παράγετε και να απεικονίσετε γραφικά ένα σύνολο δεδομένων X_1 , το οποίο περιέχει 250 σημεία από κάθε κλάση (500 σημεία συνολικά), προκειμένου να χρησιμοποιηθεί για εκπαίδευση (χρησιμοποιήστε την τιμή 5 ως «σπόρο» (seed) για τη συνάρτηση *randn* του *MATLAB*). Να παράγετε και ένα επιπλέον σύνολο δεδομένων, X_2 , το οποίο περιέχει 250 σημεία από κάθε κλάση, προκειμένου να χρησιμοποιηθεί για έλεγχο (χρησιμοποιήστε την τιμή 10 ως «σπόρο» (seed) για τη συνάρτηση *randn* του *MATLAB*).
- Με βάση το X_1 , εκτελέστε τον αλγόριθμο του Platt προκειμένου να παράγετε έξι μηχανές διανυσματικής στήριξης που διαχωρίζουν τις δύο κλάσεις, με $C = 0.1, 0.2, 0.5, 1, 2, 20$. Χρησιμοποιήστε $tol = 0.001$.
 - Υπολογίστε το λάθος ταξινόμησης για τα σύνολα εκπαίδευσης και δοκιμής.
 - Μετρήστε τα διανύσματα στήριξης.
 - Υπολογίστε το περιθώριο ($2/||w||$).
 - Απεικονίστε γραφικά τον ταξινομητή και τις γραμμές περιθωρίου.

Άσκηση 5^η (Νευρωνικά δίκτυα):

Θεωρείστε ένα πρόβλημα ταξινόμησης δύο κλάσεων στο δισδιάστατο χώρο. Τα σημεία της πρώτης (δεύτερης) κλάσης σημειώνονται με $+1$ (-1), και προκύπτουν από τρεις (τέσσερις) κανονικές κατανομές με μέσες τιμές $[-5, 5]^T$, $[5, -5]^T$, $[10, 0]^T$ ($[-5, -5]^T$, $[0, 0]^T$, $[5, 5]^T$, $[15, -5]^T$), με ίσες πιθανότητες. Το μητρώο συνδιασποράς για κάθε κατανομή είναι $\sigma^2 I$, όπου $\sigma^2 = 1$ και I το ταυτοτικό μητρώο τάξεως 2×2 .

- Να παράγετε και να απεικονίσετε γραφικά ένα σύνολο δεδομένων X_1 (σύνολο εκπαίδευσης), το οποίο περιέχει 75 σημεία από την κλάση $+1$ (περίπου 25 από κάθε κατανομή) και 100 σημεία από την κλάση -1 (περίπου 25 σημεία από κάθε κατανομή). Με χρήση της ίδιας συνταγής, να παράγετε και ένα ακόμη σύνολο, το X_2 (σύνολο ελέγχου).

2. Χρησιμοποιώντας το X_1 , εκπαιδεύστε δύο νευρωνικά δίκτυα εμπρόσθιας διάδοσης δύο επιπέδων (two-layer feedforward neural networks – 2LFNN) με δύο και τέσσερεις νευρώνες στο κρυμμένο επίπεδο¹. Όλοι οι νευρώνες του κρυμμένου επιπέδου χρησιμοποιούν την υπερβολική εφαπτομένη (\tanh) ως συνάρτηση ενεργοποίησης, ενώ ο νευρώνας εξόδου χρησιμοποιεί τη γραμμική συνάρτηση ενεργοποίησης². Εκτελέστε τον καθιερωμένο αλγόριθμο οπίσθιας διάδοσης (standard backpropagation algorithm - BP) για 9000 επαναλήψεις με ρυθμό μάθησης ίσο με 0.01. Υπολογίστε τα σφάλματα για τα σύνολα εκπαίδευσης και ελέγχου (χρησιμοποιώντας τα X_1 και X_2 , αντίστοιχα) και απεικονίστε γραφικά τα σημεία του X_1 καθώς επίσης και τις περιοχές απόφασης που δημιουργούνται από κάθε δίκτυο. Επίσης, απεικονίστε γραφικά το σφάλμα εκπαίδευσης ως προς τον αριθμό των επαναλήψεων.
3. Επαναλάβετε το βήμα 2 χρησιμοποιώντας ρυθμό εκμάθησης ίσο με 0.0001 στον αλγόριθμο BP.
4. Επαναλάβετε το βήμα 2, χρησιμοποιώντας τώρα τον προσαρμοστικό (adaptive) BP αλγόριθμο για 6000 επαναλήψεις και $r_i = 1.05$, $r_d = 0.7$, $c = 1.04$.
5. Σχολιάστε τα αποτελέσματα που προκύπτουν από την εκτέλεση των βημάτων 2, 3 και 4.

6^η άσκηση (Μηχανές διανυσματικής στήριξης (SVM) – η μη γραμμική περίπτωση):

1. Να παράγετε ένα δισδιάστατο σύνολο δεδομένων X_1 (σύνολο εκπαίδευσης) ως ακολούθως. Θεωρείστε τα εννέα τετράγωνα $[i, i + 1] \times [j, j + 1]$, $i = 0, 1, 2$, $j = 0, 1, 2$ και επέλεξε τυχαία από το καθένα 35 σημεία, σύμφωνα με την ομοιόμορφη κατανομή. Τα σημεία, που προέρχονται από τα τετράγωνα για τα οποία ο αριθμός $i + j$ είναι άρτιος (περιττός), καταχωρούνται στην κλάση +1 (−1) (το πρόβλημα θυμίζει τα λευκά και μαύρα τετράγωνα μίας σκακιέρας). Απεικονίστε γραφικά τα δεδομένα και παράγετε ένα επιπλέον σύνολο X_2 (σύνολο ελέγχου) ακολουθώντας τη συνταγή για το X_1 (θέσατε την τιμή «σπόρου» (seed) για την rand στο 5 για το X_1 και στο 10 για το X_2).
2. (α) Σχεδιάστε έναν γραμμικό SVM ταξινομητή, χρησιμοποιώντας την πρώτη παραλλαγή του αλγορίθμου του Platt, με $C = 200$ και $\text{tol} = 0.001$. Υπολογίστε τα σφάλματα για τα σύνολα εκπαίδευσης και ελέγχου και μετρήστε τον αριθμό των διανυσμάτων στήριξης.
(β) Χρησιμοποιήστε τον προηγούμενο αλγόριθμο προκειμένου να σχεδιάσετε μη γραμμικούς SVM ταξινομητές, χρησιμοποιώντας συναρτήσεις ακτινωτής βάσης ως συναρτήσεις πυρήνων, για $C = 0.2, 2, 20, 200, 2000, 20000$. Χρησιμοποιήστε

¹ Ο αριθμός των νευρώνων στο επίπεδο εισόδου ισούται με τη διάσταση του χώρου των δεδομένων (δύο στην προκειμένη περίπτωση), ενώ ο αριθμός των νευρώνων στο επίπεδο εξόδου ισούται με τον αριθμό των κλάσεων μειωμένο κατά μία μονάδα (ένas στην περίπτωσή μας).

² Σημειώνουμε ότι και η τελευταία συνάρτηση θα μπορούσε να ήταν η υπερβολική εφαπτομένη. Ωστόσο, γενικά, μπορεί κάποιος να χρησιμοποιήσει διαφορετικές συναρτήσεις ενεργοποίησης στα διάφορα επίπεδα του δικτύου. Αυτή είναι και η φιλοσοφία που υιοθετείται εδώ.

$\sigma = 1, 1.5, 2, 5$. Υπολογίστε τα σφάλματα για τα σύνολα εκπαίδευσης και ελέγχου και μετρήστε τον αριθμό των διανυσμάτων στήριξης.

(γ) Επαναλάβετε για πολυωνυμικές συναρτήσεις πυρήνων, χρησιμοποιώντας $n = 3.5$ και $\beta = 1$.

3. Εξάγετε τα συμπεράσματά σας.