Μεταπτυχιακό στην Επιστήμη Δεδομένων και Μηχανική Μάθηση
Εθνικό Μετσόβιο Πολυτεχνείο

Αναγνώριση Προτύπων
$2^{\text{ο}}$ Σύνολο Αναλυτικών Ασκήσεων

Νίκος Σταμάτης
nstam84b@gmail.com
ΑΜ: 03400115

10 Ιανουαρίου 2021

Άσκηση 1

Γενικά, δύο κλάσεις σημείων $A_1 = \{x_1, \ldots, x_k\}$ και $A_2 = \{y_1, \ldots, y_m\}$ σε έναν διανυσματικό χώρο X είναι γραμμικά διαχωρίσιμες, αν και μόνο αν οι κυρτές τους θήκες είναι ξένες, δηλαδή $\operatorname{co} A_1 \cap \operatorname{co} A_2 = \emptyset$ ([AB06, Theorem 7.30]). Όπως βλέπουμε από το Σχήμα 1, τα στοιχεία της μπλε κλάσης περιέχονται εξ' ολοκλήρου μέσα στην κυρτή θήκη των σημείων της κόκκινης. Επομένως οι δύο κλάσεις δε διαχωρίζονται γραμμικά στον $\mathbb{R}^2$. Χρησιμοποιώντας την απεικόνιση $\Phi : \mathbb{R}^2 \to \mathbb{R}^4$,

$$\Phi(x_1, x_2) = \left(1, x_1, x_2, \frac{x_1^2 + x_2^2 - 5}{3}\right),$$

τα μετασχηματισμένα δεδομένα μας παίρνουν τη μορφή που βλέπουμε στον διπλανό Πίνακα και πλέον διαχωρίζονται γραμμικά. Αυτό διαπιστώνεται εύκολα σχεδιάζοντας τις 3 τελευταίες συνιστώσες τους στον $\mathbb{R}^3$.

Πίνακας 1: Τα μετασχηματισμένα δεδομένα.

x_1	$(1, 2, 2, 1)$
x_2	$(1, 3, -3, \frac{13}{3})$
x_3	$(1, -2, 2, 1)$
x_4	$(1, -2, 2, 1)$
x_5	$(1, \frac{1}{2}, \frac{1}{2}, -\frac{3}{2})$
x_6	$(1, 1, -1, -1)$
x_7	$(1, -1, -1, -1)$
x_8	$(1, -\frac{1}{2}, \frac{1}{2}, -\frac{3}{2})$

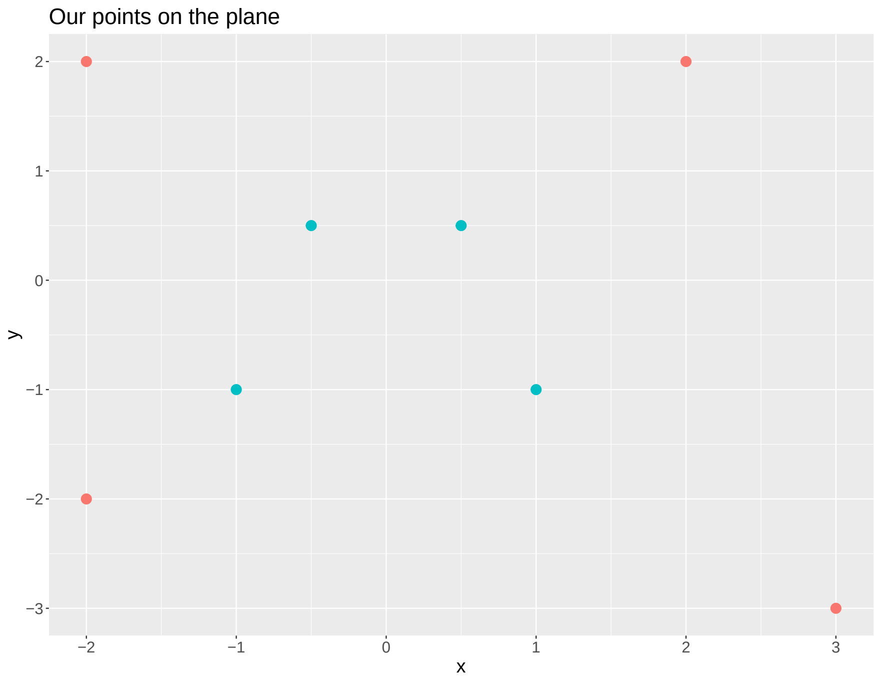
Για την εύρεση του υπερεπιπέδου με το μεγαλύτερο περιθώριο, χρησιμοποιώντας το δυϊκό πρόβλημα, έχουμε να λύσουμε το εξής πρόβλημα βελτιστοποίησης:

$$\begin{aligned} &\max_a f(a) = e'a - a'Aa, \\ &\text{s.t. } a \geq 0, \ y'a = 0, \end{aligned}$$

όπου $e \in \mathbb{R}^8$ είναι το διάνυσμα που αποτελείται από άσσους, $a \in \mathbb{R}^8$, $y = (-1, -1, -1, -1, 1, 1, 1, 1)$ το διάνυσμα που αναθέτει κάθε παρατήρηση στο πρότυπό της, και A είναι ο πίνακας με στοιχεία $A = (y_i y_j \langle x_i, x_j \rangle)_{i,j}$ ([MRT18]).

Θα αγνοήσουμε προς το παρόν τον περιορισμό $a \geq 0$. Η Λαγκραντζιάνη του τροποποιημένου προβλήματος είναι η

$$L(a, \lambda) = e'a - \frac{1}{2}a'Aa + \lambda y'a,$$



Σχήμα 1: Απεικόνιση των σημείων μας στο επίπεδο.

με κρίσιμα σημεία που δίνονται από τις εξισώσεις

$$\nabla_a L(a,\lambda) = e - Aa + \lambda y = 0,$$
$$\frac{\partial L(a,\lambda)}{\partial \lambda} = y' a = 0.$$

Στην περίπτωση που ο A αντιστρέφεται, πρώτη εξίσωση δίνει ότι $a = A^{-1}(e+\lambda y)$. Πολλαπλασιάζοντας επί y', και σε συνδυασμό με τη δεύτερη σχέση, προκύπτει ότι $y'A^{-1}e + \lambda y' A^{-1} y = 0$, από την οποία μπορούμε να εκτιμήσουμε το λ. Τελικά,*

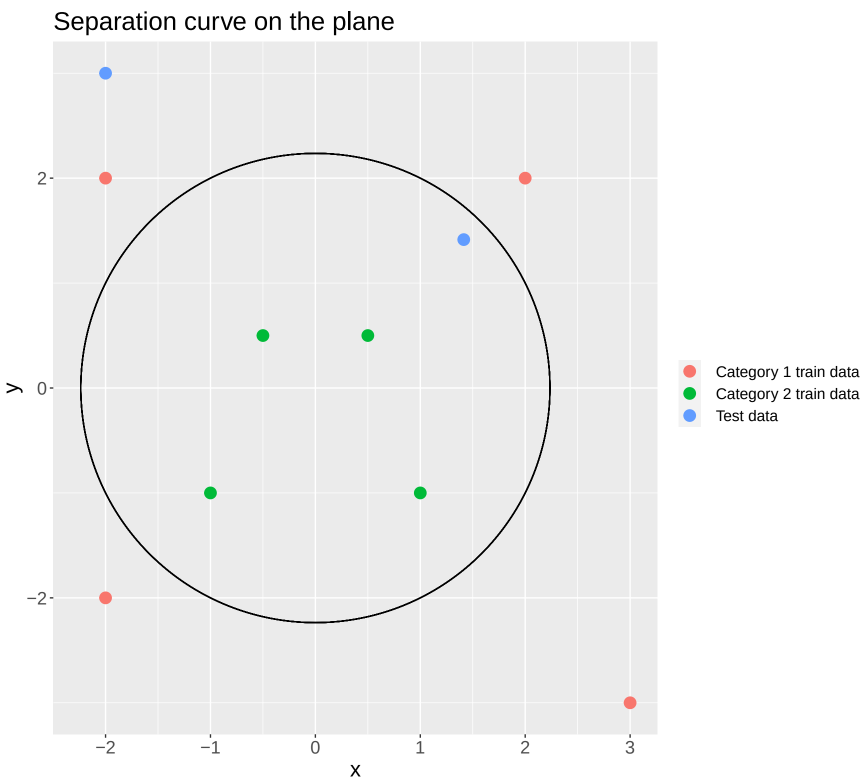
$$\lambda = -\frac{y'A^{-1}e}{y'A^{-1}y} \quad \text{και} \quad a = A^{-1}(e+\lambda y). \tag{1}$$

Εδώ ο πίνακας A είναι συμμετρικός και θετικά ημιορισμένος, επομένως εν γένει μπορεί να μην αντιστρέφεται. Αυτό συμβαίνει όντως αν προσπαθήσουμε να επιλύσουμε το πρόβλημα θεωρώντας και τα οκτώ διανύσματα $x_1, \ldots, x_8$. Στην περίπτωση αυτή, ο 8×8 πίνακας A έχει rank ίσο με 4, άρα δεν αντιστρέφεται.

Για αυτό το λόγο επιλέγουμε μόνο τα σημεία x_1, x_3, x_6, x_7. Η επιλογή έγινε ζωγραφίζοντάς τα στο χώρο και ελέγχοντας εποπτικά ποια από αυτά είναι κοντά στο σύνορο των υπερεπιπέδων που διαχωρίζουν τις δύο κλάσεις. Υπολογίζοντας τα εσωτερικά γινόμενα, βρίσκουμε ότι ο πίνακας A αντιστρέφεται, με

$$A = \begin{bmatrix} 10 & -6 & 0 & 4 \\ -6 & 10 & 0 & -4 \\ 0 & 0 & 4 & 2 \\ 4 & -4 & 2 & 4 \end{bmatrix} \quad \text{και} \quad A^{-1} = \begin{bmatrix} 0.22 & 0.03 & 0.12 & -0.25 \\ 0.03 & 0.22 & -0.12 & 0.25 \\ 0.12 & -0.12 & 0.50 & -0.50 \\ -0.25 & 0.25 & -0.50 & 1.00 \end{bmatrix}. \tag{2}$$

*Γενικά το πρόβλημα της μεγιστοποίησης της συνάρτησης $f(x) = c'x - \frac{1}{2}x'Dx$, υπό τον περιορισμό ότι $Ax = b$, όπου ο D είναι συμμετρικός και θετικά ορισμένος πίνακας, έχει ως μοναδική λύση την $x = -D^{-1}[A'(AD^{-1}A')^{-1}(AD^{-1}c - b) + c]$ [Sun96, Exercise 5.6].



Σχήμα 2: Η καμπύλη διαχωρισμού στον αρχικό μας χώρο.

Αντικαθιστώντας στην (1) βρίσκουμε ότι $\lambda = 0$ και $a = (0.125, 0.375, 0, 0.5)$. Καθώς όλες οι συνιστώσες του a είναι μη αρνητικές, η λύση αυτή ικανοποιεί και τον επιπλέον περιορισμό ότι $a_i \geq 0$ για κάθε i, δηλαδή αποτελεί λύση όχι μόνο του τροποποιημένου προβλήματος, αλλά και του αρχικού μας. Παρατηρούμε επιπλέον ότι ο συντελεστής του x_6 μηδενίζεται, άρα μόνο τα διανύσματα x_1, x_3 και x_7 είναι τα support vectors μας.

Το αντίστοιχο διάνυσμα w ισούται με

$$w = \sum_i a_i y_i x_i = -0.125x_1 - 0.375x_3 + 0.5x_7 = (0,0,0,-1)'.$$

Το εσωτερικό γινόμενο του w με τα διάφορα x_i θα ισούται απλά με την τελευταία συντεταγμένη του x_i πολλαπλασιασμένη με το -1. Έτσι τα 4 πρώτα διανύσματα θα δίνουν τις τιμές $-1, -\frac{13}{3}$, -1, -1 ενώ τα υπόλοιπα 4 τις τιμές $\frac{3}{2}, 1, 1, \frac{3}{2}$. Οι πρώτοι 4 αριθμοί είναι μικρότεροι ή ίσοι του -1 ενώ οι υπόλοιποι 4 μεγαλύτεροι ή ίσοι του 1, όπως αναμενόταν. Αντίστοιχα, η συνάρτηση διαχωρισμού θα ισούται με

$$g(x_1, x_2) = w'\left(1, x_1, x_2, \frac{x_1^2 + x_2^2 - 5}{3}\right) = \frac{x_1^2 + x_2^2 - 5}{3}$$

και επομένως ο ταξινομητής μας κατηγοριοποιεί τα σημεία αναλόγα με το εαν πέφτουν εντός ή εκτός του κύκλου με εξίσωση $x_1^2 + x_2^2 = 5$. Έτσι για το σημείο $(\sqrt{2}, \sqrt{2})$ έχουμε ότι $g(\sqrt{2}, \sqrt{2}) = 4 < 5$, ενώ για το σημείο $(-2, 3)$ ότι $g(-2, 3) = 13 > 5$. Άρα ταξινομούνται στη δεύτερη και στην πρώτη κατηγορία αντίστοιχα.

Τέλος, το περιθώριο ρ του ταξινομητή ισούται με τη ρίζα του αντιστρόφου της ℓ_1 νόρμας του διανύσματος a, δηλαδή $\rho^2 = \frac{1}{\|a\|_1} = 1$ ([MRT18, p. 69]).

Άσκηση 4

β) Θα χτίσουμε το νευρωνικό μας δίκτυο φτιάχνοντας κατάλληλες χαρακτηριστικές συναρτήσεις. Τα επόμενα δύο Λήμματα μας δίνουν τα δομικά υλικά για την κατασκευή μας.

Λήμμα 1. *Έστω X διανυσματικός χώρος, $a \in \mathbb{R}$ και $f \in X^{\#}$ γραμμικό συναρτησιακό. Θέτουμε $H_{a,f} = \{x \in X : f(x) \geq a\}$ να είναι ο ημίχωρος που ορίζεται από το συναρτησιακό f και το a. Τότε υπάρχει νευρωνικό δίκτυο $N : X \to \mathbb{R}$ με $N(x) = I_{H_{a,f}}(x)$, όπου με I_A συμβολίζουμε τη δείκτρια συνάρτηση του συνόλου A.*

Απόδειξη: Έστω f, a όπως στην εκφώνηση. Θεωρούμε την step function $\sigma : \mathbb{R} \to \mathbb{R}$ με $\sigma = I_{[0,\infty)}$ και ορίζουμε $N(x) = \sigma(f(x) - a)$. Τότε $N(x) = 1$ αν και μόνο αν $f(x) \geq a$, και $N(x) = 0$ διαφορετικά, άρα το νευρωνικό δίκτυο N ικανοποιεί τη ζητούμενη ιδιότητα. □

Επομένως κάθε ημίχωρος μπορεί να αναπαρασταθεί πλήρως με κάποιο νευρωνικό δίκτυο. Το ίδιο ισχύει και για τις τομές/ενώσεις τους:

Λήμμα 2. *Αν X διανυσματικός χώρος και $A_1, \dots, A_n \subseteq X$ σύνολα που μπορούν να αναπαρασταθούν επακριβώς με ένα νευρωνικό δίκτυο, τότε το ίδιο ισχύει και για τα σύνολα $\cup_{i=1}^n A_i$ και $\cap_{i=1}^n A_i$.*

Απόδειξη: Έστω $N_1, \dots, N_n$ τα αντίστοιχα νευρωνικά δίκτυα που αναπαριστούν επακριβώς τα σύνολα $A_1, \dots, A_n$, δηλαδή $N_i(x) = I_{A_i}(x)$ για κάθε $i = 1, \dots, n$ και $x \in X$. Ορίζουμε το δίκτυο $N_{A_1 \cap \dots \cap A_n}(x) = \sigma(N_1(x) + \dots + N_n(x) - n)$. Τότε

$$N_{A_1 \cap \dots \cap A_n}(x) = \begin{cases} 1, & N_1(x) + \dots + N_n(x) \geq n, \\ 0, & N_1(x) + \dots + N_n(x) < n, \end{cases} = \begin{cases} 1, & N_1(x) = \dots = N_n(x) = 1, \\ 0, & \text{διαφορετικά}, \end{cases}$$
$$= \begin{cases} 1, & x \in A_1 \cap \dots \cap A_n, \\ 0, & \text{διαφορετικά}, \end{cases} = I_{A_1 \cap \dots \cap A_n}.$$

Εντελώς όμοια, για το δίκτυο $N_{A_1 \cup \dots \cup A_n}(x) = \sigma(N_1(x) + \dots + N_n(x) - 1)$ έχουμε

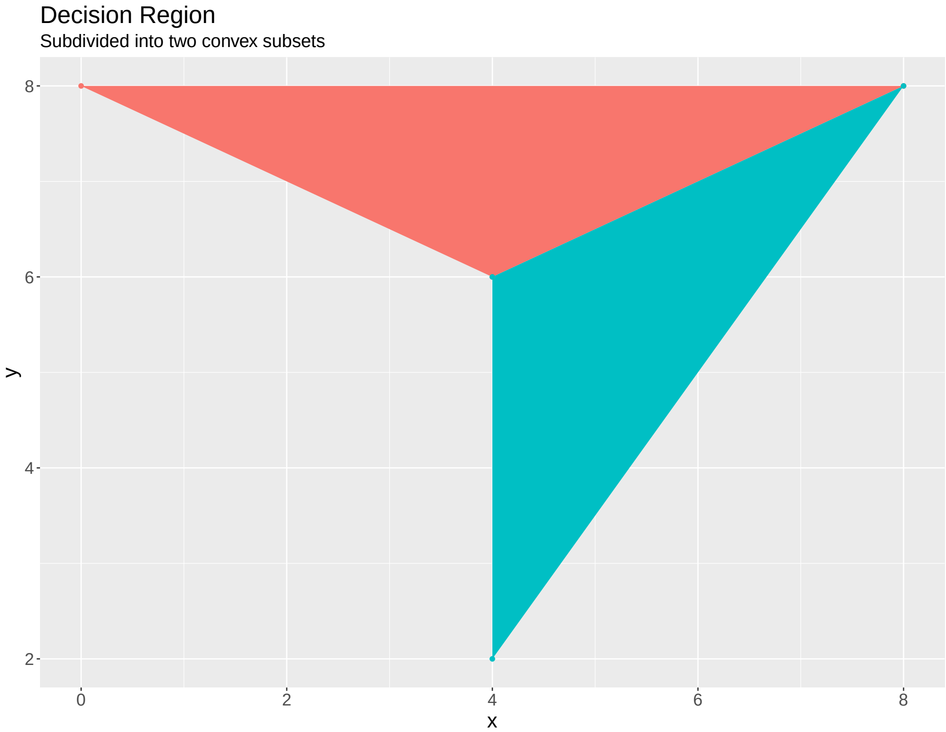
$$N_{A_1 \cup \dots \cup A_n}(x) = \begin{cases} 1, & N_1(x) + \dots + N_n(x) \geq 1, \\ 0, & N_1(x) + \dots + N_n(x) < 1, \end{cases} = \begin{cases} 1, & N_i(x) = 1 \text{ για κάποιο } i, \\ 0, & N_i(x) = 0 \text{ για κάθε } i, \end{cases} = I_{A_1 \cup \dots \cup A_n},$$

και επομένως αναπαριστά την ένωσή τους. □

Έστω $\mathcal{W}$ σύνολο υποσυνόλων του X. Ορίζουμε $\mathcal{W}^{\cup} = \{\cup_{i=1}^n A_i : n \in \mathbb{N}, A_i \in \mathcal{W}\}$ και $\mathcal{W}^{\cap} = \{\cap_{i=1}^n A_i : n \in \mathbb{N}, A_i \in \mathcal{W}\}$. Εφαρμόζοντας επαγωγικά τα παραπάνω δύο Λήμματα, έχουμε ότι ολόκληρη η ιεραρχία των ακολούθων συνόλων αναπαρίσταται επακριβώς από κάποιο νευρωνικό δίκτυο:

$$\begin{aligned} \mathcal{W}_0 &= \{H_{a,f} : a \in \mathbb{R}, f \in X^{\#}\} \\ \mathcal{W}_1 &= \mathcal{W}_0 \cup \mathcal{W}_0^{\cup} \cup \mathcal{W}_0^{\cap} \\ &\vdots \\ \mathcal{W}_{n+1} &= \mathcal{W}_n \cup \mathcal{W}_n^{\cup} \cup \mathcal{W}_n^{\cap} \\ &\vdots \end{aligned}$$

Έτσι αν θέσουμε $\mathcal{W} = \cup_{n=1}^{\infty} \mathcal{W}_n$, κάθε στοιχείο του είναι επακριβώς αναπαραστάσιμο. Η ιεραρχία αυτή φαίνεται να μη δίνει καινούργια στοιχεία από το $\mathcal{W}_3$ και έπειτα.



Σχήμα 3: Το σύνολό μας γράφεται ως η ένωση δύο κυρτών πολυτόπων.

Κάθε κυρτό πολύτοπο, δηλαδή κάθε σύνολο που μπορεί να γραφεί ως η κυρτή θήκη πεπερασμένου το πλήθος σημείων, γράφεται ως η τομή των ημιχώρων που το περιέχουν [AB06, p. 302], δηλαδή ανήκει στο $\mathcal{W}_1$ της προηγούμενης ιεραρχίας. Το σχήμα της άσκησης γράφεται ως ένωση δύο κυρτών πολυτόπων (βλ. Σχήμα 3) επομένως μπορεί να αναπαρασταθεί επακριβώς σύμφωνα με τα προηγούμενα αφού ανήκει στο $\mathcal{W}_2$.

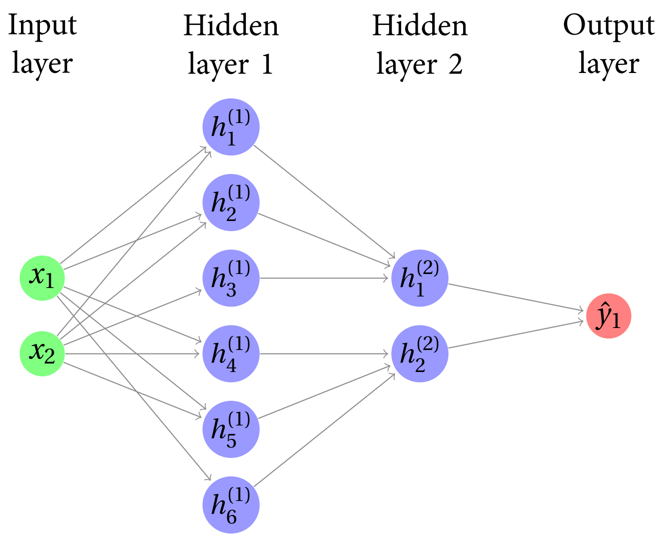
Για το κόκκινο κομμάτι του Σχήματος, ορίζουμε τις συναρτήσεις

$$
\begin{aligned}
h_1^1(x) &= \sigma\left(x_2 + \frac{1}{2}x_1 - 8\right),\\
h_2^1(x) &= \sigma\left(x_2 - \frac{1}{2}x_1 - 4\right),\\
h_3^1(x) &= \sigma(8 - x_2)
\end{aligned}
$$

και εν συνεχεία την $h_1^2(x) = \sigma(h_1^1(x) + h_2^1(x) + h_3^1(x) - 3)$, δηλαδή την τομή των 3 ημίχωρων. Αντίστοιχα για το μπλε κομμάτι έχουμε τους ημίχωρους

$$
\begin{aligned}
h_4^1(x) &= \sigma\left(\frac{1}{2}x_1 - x_2 + 4\right),\\
h_5^1(x) &= \sigma\left(x_2 - \frac{3}{2}x_1 + 4\right),\\
h_6^1(x) &= \sigma(x_1 - 4)
\end{aligned}
$$

και την τομή τους $h_2^2(x) = \sigma(h_4^1(x) + h_5^1(x) + h_6^1(x) - 3)$. Το τελικό μας σχήμα δίνεται από την ένωση των δύο αυτών συνόλων, δηλαδή την $\hat{y}_1(x) = \sigma(h_1^2(x) + h_2^2(x) - 1)$. Σχηματικά, το νευρωνικό μας δίκτυο απεικονίζεται στο παρακάτω διάγραμμα.



Άσκηση 5

Αν και δε ζητείται στην άσκηση, εξηγούμε γιατί η περιθώρια συνάρτηση πιθανότητας είναι κανονική. Για την Z έχουμε ότι $p(z) \propto e^{-\frac{1}{2}z'z}$, ενώ για τη X δεδομένης της Z ισχύει ότι

$$p(x\,|\,z) \propto e^{-\frac{1}{2\sigma^2}(x-Wz-\mu)'(x-Wz-\mu)} = e^{-\frac{1}{2\sigma^2}(x-\mu)'(x-\mu)}e^{-\frac{1}{2\sigma^2}(z'W'Wz-2z'W'(x-\mu))}.$$

Συνδυάζοντάς τες,

$$\begin{aligned} p(x) &= \int p(x\,|\,z)p(z)dz \propto \int e^{-\frac{1}{2\sigma^2}(x-\mu)'(x-\mu)}e^{-\frac{1}{2\sigma^2}(z'W'Wz+\sigma^2 z'z-2z'W'(x-\mu))}dz \\ &= e^{-\frac{1}{2\sigma^2}(x-\mu)'(x-\mu)}\int e^{-\frac{1}{2\sigma^2}(z'(W'W+\sigma^2 I)z-2z'W'(x-\mu))}dz. \end{aligned} \tag{3}$$

Γενικά, αν μια τ.μ. W ακολουθεί την πολυδιάστατη κανονική κατανομή $N(\rho, \frac{1}{\sigma^2}\Sigma)$, τότε

$$p(w) \propto e^{-\frac{1}{2\sigma^2}(z'\Sigma^{-1}z-2z'\Sigma^{-1}\rho+\rho'\Sigma^{-1}\rho)}. \tag{4}$$

Συγκρίνοντας την ολοκληρωταία ποσότητα της (3) με την (4) παρατηρούμε ότι αντιστοιχεί σε κανονική κατανομή με $\Sigma^{-1} = W'W + \sigma^2 I$ και μέση τιμή $\rho = \Sigma W(x-\mu)$. Πολλαπλασιάζουμε και διαιρούμε επίσης με $e^{\frac{1}{2\sigma^2}(\rho'\Sigma^{-1}\rho)}$, το οποίο βγαίνει εκτός του ολοκληρώματος ως σταθερά, ώστε να εμφανιστεί το πλήρες ανάπτυγμα του εκθετικού μέσα στο ολοκλήρωμα. Πλέον το ολοκλήρωμα ισούται με μια σταθερά, την οποία μπορούμε να αγνοήσουμε. Έχουμε λοιπόν ότι

$$p(x) \propto e^{-\frac{1}{2\sigma^2}((x-\mu)'(x-\mu)-\rho'\Sigma^{-1}\rho)} = e^{-\frac{1}{2\sigma^2}((x-\mu)'B(x-\mu))},$$

όπου $B = I - W'(W'W+\sigma^2 I)^{-1}W$. Δηλαδή πολυδιάστατη κανονική με μέση τιμή μ και πίνακα συνδιασπορών $\frac{1}{\sigma^2}B^{-1}$. Για τον υπολογισμό του αντιστρόφου του $W'W+\sigma^2 I$ θα χρησιμοποιήσουμε την ταυτότητα των Sherman- Morrison-Woodbury[†] η οποία δίνει ότι

$$(W'W+\sigma^2 I)^{-1} = \frac{1}{\sigma^2}I - \frac{1}{\sigma^2}IW'\left(I+WW'\frac{1}{\sigma^2}\right)^{-1}W\frac{1}{\sigma^2}I = \frac{1}{\sigma^2}I - \frac{1}{\sigma^2}W'(WW'+\sigma^2 I)W = \frac{1}{\sigma^2}B.$$

[†] Ταυτότητα Sherman-Morrison-Woodbury [GVL, p. 65]: $(A+UV')^{-1} = A^{-1} - A^{-1}U(I+V'A^{-1}U)^{-1}V'A^{-1}$.

Άρα η X ακολουθεί πολυδιάστατη κανονική με μέση τιμή μ και πίνακα συνδιασπορών $(W'W+\sigma^2 I)$.

Για κάθε $D\times M$ πίνακα W, ο πίνακας WW' είναι προφανώς συμμετρικός και θετικά ημιορισμένος: Αν $x\in\mathbb{R}^D$, τότε $x'WW'x = \|W'x\|^2 \geq 0$. Με τη σειρά του, ο πίνακας $C = WW'+\sigma^2 I$ είναι συμμετρικός και θετικά ορισμένος. Πράγματι, $C' = (WW')' + \sigma^2 I' = WW' + \sigma^2 I = C$, άρα συμμετρικός. Αν λ ιδιοτιμή του C με αντίστοιχο ιδιοδιάνυσμα το x_0, τότε $WW'x_0 + \sigma^2 x_0 = \lambda x_0$, δηλαδή $WW'x_0 = (\lambda-\sigma^2)x_0$ και το $\lambda-\sigma^2$ αποτελεί ιδιοτιμή του WW'. Αφού ο τελευταίος είναι θετικά ημιορισμένος, θα πρέπει $\lambda \geq \sigma^2 > 0$. Καθώς ο C έχει όλες του τις ιδιοτιμές αυστηρά θετικές, είναι θετικά ορισμένος.

Έστω $X \sim N(\mu, C)$ πολυδιάστατη κανονική τ.μ. Η συνάρτηση κατανομής της $p(x)$ είναι ανάλογη της $p(x) \propto e^{-\frac{1}{2}(x-\mu)'C^{-1}(x-\mu)}$. Έστω τώρα δείγμα $(x_i)_{i=1}^N$ από αυτήν την κατανομή. Τότε η πιθανοφάνεια του δείγματος είναι ανάλογη του $p(\tilde{x}) \propto e^{-\sum_{i=1}^N \frac{1}{2}(x_i-\mu)'C^{-1}(x_i-\mu)}$. Άρα

$$l(\tilde{x},\mu) = \ln p(\tilde{x},\mu) \propto \sum_{i=1}^N \left(-x_i'C^{-1}x_i + 2\mu'C^{-1}x_i - \mu'C^{-1}\mu\right) \Rightarrow$$

$$\nabla_\mu l(\tilde{x},\mu) \propto 2C^{-1}\sum_{i=1}^N x_i - 2NC^{-1}\mu.$$

Μηδενίζοντας την τελευταία σχέση προκύπτει ότι το $\hat{\mu} = \frac{\sum_{i=1}^N x_i}{N}$ αποτελεί στάσιμο σημείο. Καθώς $\nabla_\mu l(\tilde{x},\mu) \propto -2NC^{-1}$ αρνητικά ορισμένος,[‡] το στάσιμο σημείο που βρέθηκε αποτελεί μοναδικό μέγιστο.

Άσκηση 6

α) Για κάθε πρότυπο k, ορίζουμε ως A_k τους δείκτες για τους οποίους η αντίστοιχη παρατήρηση ανήκει στο εν λόγω πρότυπο, $A_k = \{i : x_i \text{ belongs to } k \text{ pattern}\}$. Συμβολίζουμε επίσης με $\#A_k$ το πλήθος αυτών των δεικτών. Η προς μεγιστοποίηση συνάρτηση $\mathcal{E}(H,\mu)$ ισούται με

$$\begin{aligned}
\mathcal{E}(H,\mu) &= \sum_{k=1}^K\sum_{n=1}^N y_{kn}\sum_{i=1}^D\left[\mu_k(i)\ln\frac{\mu_k(i)}{x_n(i)} - \mu_k(i) + x_n(i)\right]\\
&= \sum_{k=1}^K\sum_{n\in A_k}\sum_{i=1}^D\left[\mu_k(i)\ln\frac{\mu_k(i)}{x_n(i)} - \mu_k(i) + x_n(i)\right]\\
&= \sum_{k=1}^K\sum_{i=1}^D\sum_{n\in A_k}\left(\mu_k(i)\ln\mu_k(i) - \mu_k(i)\ln x_n(i) - \mu_k(i) + x_n(i)\right)\\
&= \sum_{k=1}^K\sum_{i=1}^D\left(\#A_k\mu_k(i)\ln\mu_k(i) - \mu_k(i)\sum_{n\in A_k}\ln x_n(i) - \#A_k\mu_k(i) + \sum_{n\in A_k}x_n(i)\right)\\
&= \sum_{k=1}^K\sum_{i=1}^D\left(\mu_k(i)\ln\frac{\mu_k(i)^{\#A_k}}{\prod_{n\in A_k}x_n(i)} - \#A_k\mu_k(i) + \sum_{n\in A_k}x_n(i)\right). \qquad (5)
\end{aligned}$$

Η συνάρτηση $f(x) = x\ln\frac{x^m}{b} - mx$ για $m\in\mathbb{N}$, έχει παράγωγο ίση με $f'(x) = m\ln x - \ln b$ με ρίζα στο $x_0 = b^{1/m}$, και δεύτερη παράγωγο ίση με $f''(x) = \frac{m}{x} > 0$ για κάθε $x > 0$. Άρα ελαχιστοποιείται

[‡] Οι ιδιοτιμές λ_i του C είναι αυστηρά θετικές. Οι ιδιοτιμές του C^{-1} είναι οι $\frac{1}{\lambda_i}$, οι οποίες επίσης είναι αυστηρά θετικές, άρα ο C^{-1} θετικά ορισμένος και ο $-C^{-1}$ αρνητικά ορισμένος.

στο εν λόγω σημείο. Εφαρμόζοντας το αποτέλεσμα αυτό για $b = \prod_{n\in A_k} x_n(i)$ και $m = \#A_k$, προκύπτει ότι κάθε παράσταση εντός της παρένθεσης στη σχέση (5) ελαχιστοποιείται για $\mu_k(i) = \left(\prod_{n\in A_k} x_n(i)\right)^{1/\#A_k}$ δηλαδή στο γεωμετρικό μέσο των παρατηρήσεων που έχουν προέλθει από το αντίστοιχο πρότυπο.

β) Επιλύοντας την εξίσωση $\sum_n y_{kn} \frac{\partial D(\mu_k, x_n)}{\partial \mu_k} = 0$ βρίσκουμε

$$\sum_n y_{kn} \frac{\partial D(\mu_k, x_n)}{\partial \mu_k} = \sum_n y_{nk} \left(\ln \mu_k - \ln x_n\right) = \#A_k \ln \mu_k - \sum_{n\in A_k} \ln x_n = 0,$$

δηλαδή ότι $\mu_k = \left(\prod_{n\in A_k} x_n\right)^{1/\#A_k}$. Άρα ο αλγόριθμος GHAS είναι ο εξής:

Generalized Hard Algorithmic Scheme	
Βήμα 1	Αρχικοποιούμε με κάποιο διάνυσμα $\mu = (\mu_k)_{k=1}^K$.
Βήμα 2	Για κάθε $i = 1, \dots, N$ και κάθε $k = 1, \dots, K$, θέτουμε $y_{kn} = 1$ αν το πρότυπο που βρίσκεται πλησιέστερα (σύμφωνα με την $\mathcal{D}$) στην παρατήρηση x_n είναι το μ_k και μηδέν διαφορετικά.
Βήμα 3	Υπολογίζουμε τα καινούργια μ_k σύμφωνα με τον τύπο $\mu_k = \left(\prod_{n\in A_k} x_n\right)^{1/\#A_k}$ όπου A_k δηλώνει το σύνολο δεικτών για τους οποίους οι αντίστοιχες παρατηρήσεις κατηγοριοποιήθηκαν στο k πρότυπο στο περασμένο βήμα.
Βήμα 4	Επιστρέφουμε στο Βήμα 2.

Άσκηση 7

α) Έστω X, Y ανεξάρτητες τ.μ. με μηδενικές μέσες τιμές. Τότε για το άθροισμά τους ισχύει ότι

$$\begin{aligned}
\mathbb{E}[(X+Y)^4] &= \mathbb{E}[X^4] + \mathbb{E}[Y^4] + 6\mathbb{E}[X^2Y^2] + \mathbb{E}[4X^3Y] + \mathbb{E}[4XY^3] \\
&= \mathbb{E}[X^4] + \mathbb{E}[Y^4] + 6\mathbb{E}[X^2]\mathbb{E}[Y^2] + 4\mathbb{E}[X^3]\mathbb{E}[Y] + 4\mathbb{E}[X]\mathbb{E}[Y^3] \\
&= \mathbb{E}[X^4] + \mathbb{E}[Y^4] + 6\mathbb{E}[X^2]\mathbb{E}[Y^2],
\end{aligned}$$

λόγω ανεξαρτησίας και μηδενικών μέσων τιμών αντίστοιχα. Επιπλέον,

$$(\mathbb{E}[(X+Y)^2])^2 = \mathbb{E}[X^2]^2 + \mathbb{E}[Y^2]^2 + 2\mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Άρα για την κύρτωση του αθροίσματός τους θα έχουμε ότι

$$\begin{aligned}
\mathrm{kurt}(X+Y) &= \mathbb{E}[(X+Y)^4] - 3(\mathbb{E}[(X+Y)^2])^2 = \mathbb{E}[X^4] + \mathbb{E}[Y^4] - 3\mathbb{E}[X^2]^2 - 3\mathbb{E}[Y^2]^2 \\
&= \mathrm{kurt}(X) + \mathrm{kurt}(Y).
\end{aligned}$$

Να σημειωθεί ότι ενώ η κύρτωση είναι αθροιστική συνάρτηση για ανεξάρτητες μεταβλητές, δεν είναι γραμμική συνάρτηση ούτε καν για αυτές, αφού $\mathrm{kurt}(\lambda X) = \lambda^4 \mathrm{kurt}(X)$ για κάθε $\lambda \in \mathbb{R}$ και τ.μ. X.

β) Θεωρούμε $S_1, \dots, S_N$ ανεξάρτητες τ.μ. με μοναδιαίες διασπορές, καθώς και τη μείξη τους $X = \sum_{i=1}^N w_i S_i$. Για να ισούται η διασπορά της X με ένα, θα πρέπει

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^N w_i S_i\right) = \sum_{i=1}^N w_i^2 \mathrm{Var}(S_i) = 1.$$

Καθώς $\mathrm{Var}(S_i) = 1$ για κάθε i, η τελευταία συνθήκη απλοποιείται περαιτέρω στην $\sum_{i=1}^N w_i^2 = 1$.

Για το τελευταίο ερώτημα, η προηγούμενη συνθήκη κανονικοποίησης επιβάλλει ότι $w_1 = \ldots = w_N = \frac{1}{\sqrt{N}}$, επομένως το μείγμα μας είναι το $X_N = \frac{1}{\sqrt{N}} \sum_{i=1}^N S_i$. Εφαρμόζοντας την προηγούμενη παρατήρηση για το ότι $\mathrm{kurt}(\lambda X) = \lambda^4 \mathrm{kurt}(x)$ για κάθε $\lambda \in \mathbb{R}$, έχουμε ότι $\mathrm{kurt}(X_N) = \frac{1}{N^2} \mathrm{kurt}(S_1 + \ldots + S_N)$. Επιπλέον, οι $S_1, \ldots, S_N$ είναι ανεξάρτητες, άρα από το προηγούμενο ερώτημα

$$\mathrm{kurt}(X_N) = \frac{1}{N^2} \mathrm{kurt}(S_1 + \ldots + S_N) = \frac{1}{N^2} \sum_{i=1}^N \mathrm{kurt}(S_i)$$

και καθώς $\mathrm{kurt}(S_i) \in [-a, a]$ για κάθε $i = 1, \ldots, N$ και κάθε $N \in \mathbb{N}$, θα έχουμε ότι

$$-\frac{1}{N} a \leq \mathrm{kurt}(X_N) \leq \frac{1}{N} a,$$

για κάθε $N \in \mathbb{N}$. Παίρνοντας όρια στην τελευταία σχέση, προκύπτει ότι $\lim_{N \to \infty} \mathrm{kurt}(X_N) = 0$.

1 References

[AB06] C. Aliprantis, K. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, 3rd ed., 2006. doi: 10.1007/978-3-662-03004-2 Cited on p. 1, 5

[GVL] G. Golub, C. Van Loan, *Matrix Computations,* The Johns Hopkins University Press, 2013. isbn: 9781421407944 Cited on p. 6

[MRT18] M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of Machine Learning,* MIT Press, 2018. isbn: 9780262018258 Cited on p. 1, 3

[Sun96] R. Sundaram, *A First Course in Optimization Theory,* Cambridge University Press, 1996. isbn: 9780511804526 Cited on p. 2