1 Maximum Likelihood estimation

Εφόσον τα δείγματα x_i με $i = 1, \dots, n$ επιλέγονται από την κατανομή

$$p\,(x|\theta) = \theta^2 x e^{-\theta x} u(x), \tag{1.1}$$

με $u(x)$ τη βηματική συνάρτηση Heaviside, θα ισχύει πως $x > 0$, $\forall x \in \mathcal{D}$. Η πιθανοφάνεια, $\mathcal{L}\,(\mathcal{D})$, θα δίνεται ως

$$\mathcal{L}\,(\mathcal{D}) = \prod_{k=1}^{n} p\,(x_k|\theta) = \theta^{2n} \prod_{k=1}^{n} \left(x_k e^{-x_k \theta}\right), \tag{1.2}$$

με $x_k > 0$, $k = 1, \dots, n$. Σύμφωνα με την αρχή μέγιστης πιθανοφάνειας, η καλύτερη εκτίμηση για την παράμετρο θ είναι η $\hat{\theta}$ για την οποία ισχύει

$$\hat{\theta} = \mathrm{argmax}_{\theta} \mathcal{L}\,(\mathcal{D}). \tag{1.3}$$

Ο φυσικός λογάριθμος είναι γνησίως αύξουσα συνάρτηση του ορίσματός της, επομένως η μεγιστοποίηση της πιθανοφάνειας ισοδυναμεί με τη μεγιστοποίηση του φυσικού λογαρίθμου της, για τον οποίο ισχύει

$$\begin{aligned} \ln \mathcal{L} &= \ln\left(\theta^{2n}\right) + \ln\left[\prod_{k=1}^{n} \left(x_k e^{-x_k \theta}\right)\right] = 2n \ln\theta + \sum_{k=1}^{n} \ln\left(x_k e^{-x_k \theta}\right) = \\ &= 2n \ln\theta + \sum_{k=1}^{n} \ln x_k - \theta \sum_{k=1}^{n} x_k. \end{aligned} \tag{1.4}$$

Οι απαιτήσεις

$$\left.\frac{\partial\,(\ln \mathcal{L})}{\partial \theta}\right|_{\hat{\theta}} = 0 \quad \text{και} \quad \left.\frac{\partial^2\,(\ln \mathcal{L})}{\partial \theta^2}\right|_{\hat{\theta}} < 0 \tag{1.5}$$

δίνουν

$$\frac{2n}{\hat{\theta}} - \sum_{k=1}^{n} x_k = 0 \Leftrightarrow \hat{\theta} = \frac{2n}{\sum_{k=1}^{n} x_k} \quad \text{και} \quad -\frac{1}{\hat{\theta}^2} < 0, \tag{1.6}$$

αντίστοιχα. Η δεύτερη απαίτηση ικανοποιείται ταυτοτικά, άρα γίνεται πράγματι λόγος για μεγιστοποίηση, ενώ παρατηρώντας πως στην πρώτη σχέση εμφανίζεται η μέση τιμή των δειγμάτων

$$\mathbb{E}\{x\} = \frac{1}{n} \sum_{k=1}^{n} x_k, \tag{1.7}$$

η βέλτιστη εκτίμηση της Σχέσης (1.6) για τη θ γράφεται ως

$$\hat{\theta} = \frac{2}{\mathbb{E}\{x\}}. \tag{1.8}$$

2 Minimax Criterion

Έστω ένα πρόβλημα ταξινόμησης που περιλαμβάνει δύο κατηγορίες ω_1, ω_2. Κάθε δείγμα προς ταξινόμηση, x, μπορεί να ανήκει είτε στην ω_1, η οποία έχει κάποια a priori πιθανότητα $p(\omega_1)$, είτε στην ω_2, η οποία έχει κάποια a priori πιθανότητα $p(\omega_2)$. Συμβολίζοντας την απόφαση ταξινόμησης με α_i, τότε η ταξινόμηση ενός δείγματος x που ανήκει στην κατηγορία ω_j γίνεται ορθά όταν $i = j$ και λανθασμένα όταν $i \neq j$. Μπορεί, έτσι, κανείς να ορίσει μια συνάρτηση ρίσκου $R(a_i|x)$ ως

$$R(\alpha_i|x) = \lambda_{i1} p(\omega_1|x) + \lambda_{i2} p(\omega_2|x), \tag{2.1}$$

όπου τα λ_{ij} με $i, j \in \{1, 2\}$ αντιστοιχούν στο κόστος κάθε απόφασης α_i, δεδομένου πως το υπό μελέτη δείγμα ανήκει στην κατηγορία ω_j. Το συνολικό ρίσκο ταξινόμησης δίνεται τότε από τη σχέση

$$\begin{aligned} R = &\int_{\mathcal{R}_1} d\mathbf{x} \ [\lambda_{11} p(\omega_1)\, p(\mathbf{x}|\omega_1) + \lambda_{12} p(\omega_2)\, p(\mathbf{x}|\omega_2)] + \\ &+ \int_{\mathcal{R}_2} d\mathbf{x} \ [\lambda_{21} p(\omega_1)\, p(\mathbf{x}|\omega_1) + \lambda_{22} p(\omega_2)\, p(\mathbf{x}|\omega_2)], \end{aligned} \tag{2.2}$$

όπου $\mathcal{R}_1$ και $\mathcal{R}_2$ είναι οι περιοχές του χώρου που αντιστοιχούν στις αποφάσεις α_1 και α_2, αντίστοιχα.

2.1 Στην προκείμενη περίπτωση, τα λ_{ij} δίνονται ως

$$\lambda_{ij} = 1 - \delta_{ij}, \tag{2.3}$$

όπου δ_{ij} το δ-Kronecker. Έτσι, το συνολικό ρίσκο αφορά αποκλειστικά τις λανθασμένες αποφάσεις και παίρνει τη μορφή

$$R = p(\omega_2) \int_{\mathcal{R}_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2) + p(\omega_1) \int_{\mathcal{R}_2} d\mathbf{x}\, p(\mathbf{x}|\omega_1), \tag{2.4}$$

η οποία δεν είναι παρά το αναμενόμενο σφάλμα ταξινόμησης κατά Bayes. Λαμβάνοντας υπ' όψιν πως $p(\omega_2) = 1 - p(\omega_1)$, η σχέση αυτή γράφεται ισοδύναμα ως

$$R = \int_{\mathcal{R}_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2) + p(\omega_1) \left[\int_{\mathcal{R}_2} d\mathbf{x}\, p(\mathbf{x}|\omega_1) - \int_{\mathcal{R}_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2) \right]. \tag{2.5}$$

Για δεδομένες περιοχές απόφασης $\mathcal{R}_1$, $\mathcal{R}_2$, το ρίσκο εξαρτάται γραμμικά από την a priori πιθανότητα $p(\omega_1)$. Το κριτήριο minimax βασίζεται στην εξάλειψη της εξάρτησης του ρίσκου από την $p(\omega_1)$, μέσω μηδενισμού του σχετικού συντελεστή του στη Σχέση (2.5) (ισοδύναμη με το μηδενισμό της παραγώγου του R ως προς $p(\omega_1)$, δηλαδή ελαχιστοποίηση ως προς την a priori πιθανότητα). Προκύπτει, τότε, η λεγόμενη λύση minimax για τις περιοχές απόφασης $\mathcal{R}_1$, $\mathcal{R}_2$, η οποία στην προκείμενη περίπτωση αντιστοιχεί στην

$$\int_{\mathcal{R}_2} d\mathbf{x}\, p(\mathbf{x}|\omega_1) - \int_{\mathcal{R}_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2) = 0 \Leftrightarrow \int_{\mathcal{R}_2} d\mathbf{x}\, p(\mathbf{x}|\omega_1) = \int_{\mathcal{R}_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2). \tag{2.6}$$

2.2 Η λύση που προκύπτει στη Σχέση (2.6) για τις περιοχές απόφασης $\mathcal{R}_1, \mathcal{R}_2$ δεν είναι μοναδική. Αυτό γίνεται εύκολα αντιληπτό, εάν κανείς θεωρήσει $x \in \mathbb{R}$ και τις κατανομές

$$p\left(x|\omega_1\right) = \begin{cases} 1/\eta, & \text{εάν } 0 \leq x \leq \eta \\ 0, & \text{αλλιώς} \end{cases} \tag{2.7}$$

και

$$p\left(x|\omega_2\right) = \begin{cases} 1/\eta, & \text{εάν } -\eta/2 \leq x \leq \eta/2 \\ 0, & \text{αλλιώς} \end{cases}, \tag{2.8}$$

όπου $\eta > 0$ και οι a priori πιθανότητες θεωρούνται ίσες, δηλαδή $p\left(\omega_1\right) = p\left(\omega_1\right) = 0.5$. Η «φυσική» επιλογή για τις περιοχές απόφασης είναι η $\mathcal{R}_1 = (-\eta/2, \eta/4)$ και ως άμεση συνέπεια $\mathcal{R}_2 = (\eta/4, \eta)$ (βλ. Εικόνα 2.1 (a)). Στην περίπτωση αυτή, η Σχέση (2.6) πράγματι ισχύει, αφού

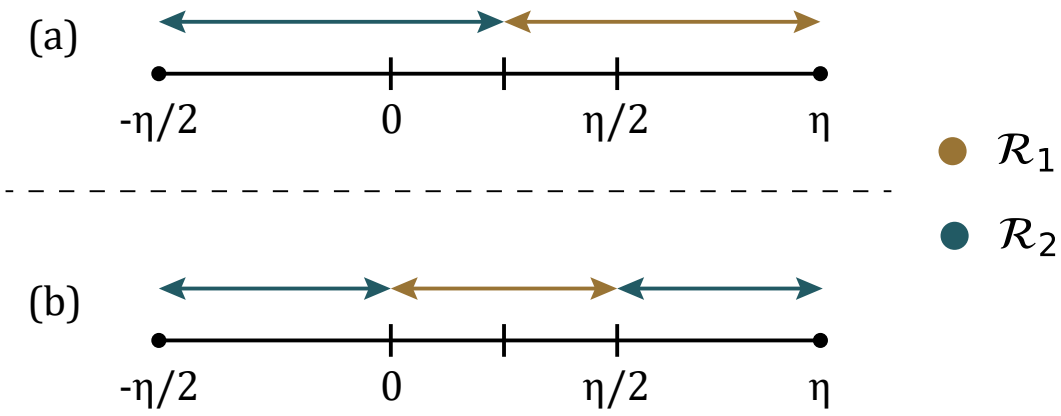
$$\int_{\mathcal{R}_2} dx\, p\left(x|\omega_1\right) = \frac{1}{\eta}\int_{\eta/4}^{\eta} dx = \frac{3}{4} \quad \text{και} \quad \int_{\mathcal{R}_1} dx\, p\left(x|\omega_2\right) = \frac{1}{\eta}\int_{-\eta/2}^{\eta/4} dx = \frac{3}{4}. \tag{2.9}$$

Παρ' όλα αυτά, η Σχέση (2.6) ισχύει και στη «λιγότερο φυσική» επιλογή $\mathcal{R}_1 = (0, \eta/2)$ και επομένως $\mathcal{R}_2 = (-\eta/2, 0) \cup (\eta/2, \eta)$ (βλ. Εικόνα 2.1 (b)), χωρίς φυσικά αυτό να σημαίνει πως ελαχιστοποιείται έτσι και το συνολικό ρίσκο. Πράγματι,

$$\int_{\mathcal{R}_2} dx\, p\left(x|\omega_1\right) = \underbrace{\int_{-\eta/2}^{0} dx\, p\left(x|\omega_1\right)}_{=0} + \int_{\eta/2}^{\eta} dx\, p\left(x|\omega_1\right) = \frac{1}{\eta}\int_{\eta/2}^{\eta} dx = \frac{1}{2} \tag{2.10}$$

και

$$\int_{\mathcal{R}_1} dx\, p\left(x|\omega_2\right) = \frac{1}{\eta}\int_{0}^{\eta/2} dx = \frac{1}{2}. \tag{2.11}$$



Εικόνα 2.1: Απεικόνιση των περιοχών απόφασης για κάθε κατηγορία, για τις περιπτώσεις: (a) $\mathcal{R}_1 = (-\eta/2, \eta/4), \mathcal{R}_2 = (\eta/4, \eta)$ και (b) $\mathcal{R}_1 = (0, \eta/2), \mathcal{R}_2 = (-\eta/2, 0) \cup (\eta/2, \eta)$.

3 Bayes Error

Εφόσον αποδειχθεί το ζητούμενο για τη γενική περίπτωση κατανομών $p(\mathbf{x}|\omega_1)$, $p(\mathbf{x}|\omega_2)$, η απόδειξη θα αφορά και την ειδική περίπτωση όπου

$$p(\mathbf{x}|\omega_1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \quad \text{και} \quad p(\mathbf{x}|\omega_2) \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2). \tag{3.1}$$

Ως εκ τούτου, θεωρούνται δύο ανεξάρτητες κατανομές $p(\mathbf{x}|\omega_1)$, $p(\mathbf{x}|\omega_2)$, όπου $\mathbf{x}$ τυχαία μεταβλητή με $\mathbf{x} \in \mathbb{R}^2$ και με a priori πιθανότητες $p(\omega_1)$ και $p(\omega_2)$, αντίστοιχα. Εξ ορισμού, το σφάλμα κατά Bayes $P(e)$ για την ταξινόμηση κάποιου νέου δείγματος, $\mathbf{x}$, ισούται με

$$P(e) = p(\omega_1) \int_{\mathcal{R}_2} d\mathbf{x}\, p(\mathbf{x}|\omega_1) + p(\omega_2) \int_{\mathcal{R}_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2), \tag{3.2}$$

όπου $\mathcal{R}_i$, $i = 1, 2$ είναι οι σχετικές περιοχές απόφασης στον $\mathbb{R}^2$. Έστω τώρα ένας προβολικός τελεστής, $\hat{T}$, ο οποίος προβάλλει διανύσματα $\mathbf{x}$ του $\mathbb{R}^2$ στον $\mathbb{R}$ ως $\hat{T}\mathbf{x} = y$, με $y \in \mathbb{R}$. Στην περίπτωση αυτή, για την κατανομή π της τυχαίας μεταβλητής y ισχύει

$$\int_{\mathcal{S}} dy\, \pi(y) = \int_{\hat{T}^{-1}(\mathcal{S})} d\mathbf{x}\, p(\mathbf{x}), \tag{3.3}$$

όπου $\mathcal{S} \subseteq \mathbb{R}$ και $T^{-1}(\mathcal{S}) \subseteq \mathbb{R}^2$ είναι η αντίστροφη εικόνα του $\mathcal{S}$[1]. Αντί για την ταξινόμηση η οποία δίνει το σφάλμα της (3.2), αν κανείς θεωρήσει την εναλλακτική ταξινόμηση της εκφώνησης, κατά την οποία κάθε δείγμα $\mathbf{x}$ προβάλλεται μέσω του $\hat{T}$ σε κάποιο y και στη συνέχεια το y ταξινομείται κατά Bayes βάσει των κατανομών $\pi(y|\omega_1)$ και $\pi(y|\omega_2)$, τότε το αντίστοιχο σφάλμα ταξινόμησης, $\tilde{P}(e)$, αντιστοιχεί σε

$$\tilde{P}(e) = p(\omega_1) \int_{\rho_2} dy\, \pi(y|\omega_1) + p(\omega_2) \int_{\rho_1} dy\, \pi(y|\omega_2), \tag{3.4}$$

όπου ρ_1 και ρ_2 είναι οι περιοχές απόφασης για την ταξινόμηση στον $\mathbb{R}$. Βάσει της Σχέσης (3.3), η παραπάνω μπορεί να γραφεί ως

$$\begin{aligned} \tilde{P}(e) &= p(\omega_1) \int_{\hat{T}^{-1}(\rho_2)} d\mathbf{x}\, p(\mathbf{x}|\omega_1) + p(\omega_2) \int_{\hat{T}^{-1}(\rho_1)} d\mathbf{x}\, p(\mathbf{x}|\omega_2) \\ &= p(\omega_1) \int_{\tilde{\mathcal{R}}_2} d\mathbf{x}\, p(\mathbf{x}|\omega_1) + p(\omega_2) \int_{\tilde{\mathcal{R}}_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2), \end{aligned} \tag{3.5}$$

όπου

$$\tilde{\mathcal{R}}_i = \hat{T}^{-1}(\rho_i). \tag{3.6}$$

Εφόσον το σφάλμα της (3.2) είναι το σφάλμα κατά Bayes, αντιστοιχεί στο ελάχιστο δυνατό σφάλμα της ταξινόμησης των $\mathbf{x}$. Επομένως, οι περιοχές απόφασης $\tilde{\mathcal{R}}_i$ του εναλλακτικού τρόπου ταξινόμησης, όπου τα $\mathbf{x}$ πρώτα προβάλλονται στη 1 διάσταση, θα μπορούν να γραφούν ως

$$\tilde{\mathcal{R}}_i = \mathcal{R}_i + \Gamma_i, \tag{3.7}$$

[1] Σημειώνεται εδώ πως ο συμβολισμός αυτός δεν αφορά την πράξη του τελεστή $\hat{T}^{-1}$, αφού δεν έχει γίνει καμία υπόθεση για την αντιστρεψιμότητά του - ειδικά δεδομένου πως αποτελεί προβολικό τελεστή. Η $T^{-1}(\mathcal{S})$ δεν είναι παρά το σύνολο $\{\mathbf{x} : \hat{T}\mathbf{x} \in \mathcal{S}\}$.

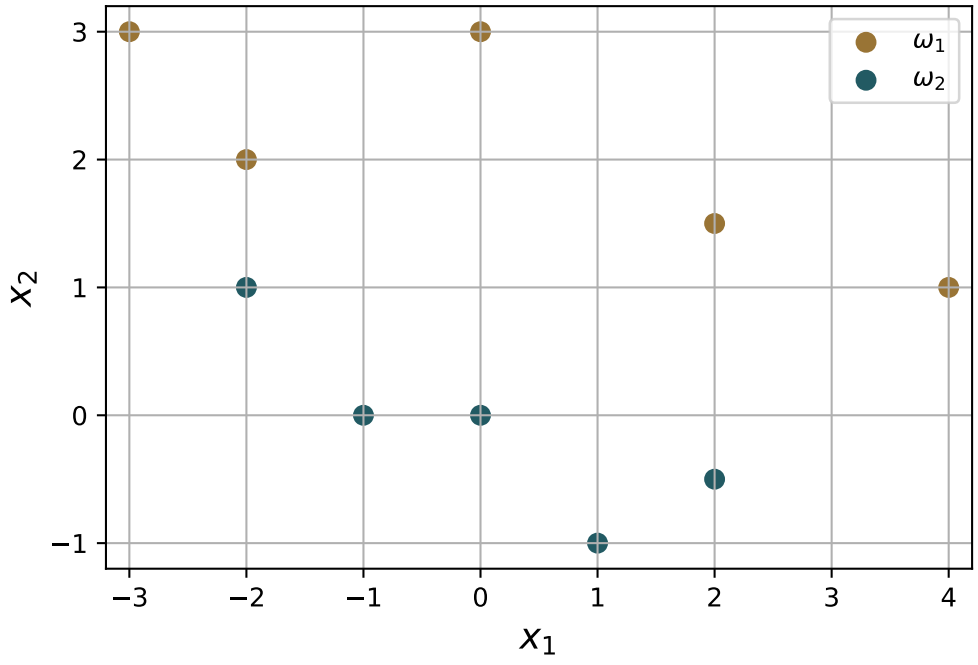
όπου στην περίπτωση που οι δύο ταξινομητές είναι ισοδύναμοι (βέλτιστη περίπτωση) θα ισχύει $\Gamma_i = \emptyset, i = 1, 2$, διαφορετικά η Σχέση (3.5) θα ισοδυναμεί με

$$\begin{aligned}\tilde{P}(e) &= p(\omega_1) \int_{\mathcal{R}_2+\Gamma_2} d\mathbf{x}\, p(\mathbf{x}|\omega_1) + p(\omega_2) \int_{\mathcal{R}_1+\Gamma_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2) \\ &= P(e) + p(\omega_1) \int_{\Gamma_2} d\mathbf{x}\, p(\mathbf{x}|\omega_1) + p(\omega_2) \int_{\Gamma_1} d\mathbf{x}\, p(\mathbf{x}|\omega_2) \\ &\geq P(e), \end{aligned} \tag{3.8}$$

αποδεικνύοντας έτσι το ζητούμενο, ότι δηλαδή το σφάλμα ταξινόμησης στη μία διάσταση κατόπιν προβολής είναι μεγαλύτερο ή ίσο από το σφάλμα κατά Bayes στις δύο διαστάσεις. Το αποτέλεσμα αυτό είναι απολύτως αναμενόμενο και σε διαισθητικό επίπεδο. Συγκεκριμένα, η πράξη προβολής $\hat{T} : \mathbb{R}^\mu \to \mathbb{R}^\nu$ οδηγεί στην καλύτερη περίπτωση σε διατήρηση και σε όλες τις υπόλοιπες σε απώλεια πληροφορίας όταν $\mu > \nu$. Έτσι, αφού η ταξινόμηση κατόπιν προβολής γίνεται με πληροφορία το πολύ ίση με όση υπήρχε πριν την προβολή, είναι λογικό το σφάλμα ταξινόμησης να είναι τουλάχιστον ίσο (διαφορετικά μεγαλύτερο) με το αντίστοιχο σφάλμα χωρίς την προβολή.

4 Perceptrons

Τα σημεία που αντιστοιχούν στα δοσμένα διανύσματα χαρακτηριστικών της μορφής $[x_1, x_2]^\mathsf{T}$ απεικονίζονται στο γράφημα της Εικόνας 4.1.



Εικόνα 4.1: Απεικόνιση των σημείων που αντιστοιχούν στα διανύσματα χαρακτηριστικών. Τα σημεία που αντιστοιχούν στην κλάση ω_1 απεικονίζονται με χρυσό χρώμα, ενώ τα σημεία που αντιστοιχούν στην κλάση ω_2 απεικονίζονται με μπλε χρώμα. Είναι εμφανές πως οι δύο κλάσεις είναι γραμμικά διαχωρίσιμες.

Οπτικά, είναι φανερό πως τα σημεία προέρχονται από γραμμικά διαχωρίσιμες κλάσεις. Ο διαχωρισμός τους μπορεί να πραγματοποιηθεί μέσω ενός αλγόριθμου τύπου Perceptron, ο οποίος είναι βέβαιο πως θα συγκλίνει σε πεπερασμένο αριθμό βημάτων, δεδομένου πως οι κλάσεις είναι πράγματι γραμμικά διαχωρίσιμες. Σκοπός του αλγορίθμου αυτού είναι η εύρεση μιας ευθείας διαχωρισμού της μορφής

$$g(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + w_0, \tag{4.1}$$

όπου $\mathbf{x} = [x_1, x_2]^\top$ και $\mathbf{w}$ είναι το λεγόμενο διάνυσμα βαρών, δηλαδή το διάνυσμα οι συνιστώσες του οποίου αποτελούν τους συντελεστές των x_i, $i = 1, 2$. Για ένα διάνυσμα $\mathbf{x}^{(1)}$ που προέρχεται από την κλάση ω_1 θα πρέπει να ισχύει $g\left(\mathbf{x}^{(1)}\right) \geq 0$, ενώ για ένα διάνυσμα $\mathbf{x}^{(2)}$ που προέρχεται από την κλάση ω_2 θα πρέπει να ισχύει $g\left(\mathbf{x}^{(2)}\right) \leq 0$. Εάν η σταθερά w_0 ενσωματωθεί στο διάνυσμα βαρών και κάθε διάνυσμα χαρακτηριστικών επαυξηθεί ως $[x_1, x_2, 1]^\top$, τότε οι συνθήκες αυτές γράφονται ισοδύναμα

$$\mathbf{x} \in \omega_1 \Rightarrow \mathbf{w} \cdot \mathbf{x} > 0 \quad \text{και} \quad \mathbf{x} \in \omega_2 \Rightarrow \mathbf{w} \cdot \mathbf{x} < 0. \tag{4.2}$$

Έτσι, ο διαχωρισμός των σημειών μέσω μιας ευθείας διαχωρισμού ανάγεται στον υπολογισμό ενός κατάλληλου διανύσματος βαρών, $\mathbf{w}$, με την εξίσωση $\mathbf{w} \cdot \mathbf{x} = 0$ να αντιστοιχεί στην ευθεία διαχωρισμού. Η αρχή λειτουργίας του αλγορίθμου Perceptron είναι η ακόλουθη: κατά το t-οστό βήμα επανάληψης, το διάνυσμα βαρών (επαυξημένο κατάλληλα ώστε να περιλαμβάνει και τη σταθερά w_0 ως τρίτη συνιστώσα) συμβολίζεται με $\mathbf{w}(t)$. Σε κάθε βήμα ταξινομείται ένα μόνο δείγμα, το οποίο αντιστοιχεί στο διάνυσμα χαρακτηριστικών $\mathbf{x}_t$, επαυξημένο κατάλληλα ώστε να περιλαμβάνει το 1 ως τρίτη συνιστώσα, και η ταξινόμηση γίνεται βάσει του υπολογισμού του εσωτερικού γινομένου $\mathbf{w}(t) \cdot \mathbf{x}_t$. Στην περίπτωση επιτυχούς ταξινόμησης (βλ. Σχέση (4.2)) το διάνυσμα βαρών δε μεταβάλλεται και ο αλγόριθμος προχωρά στην ταξινόμηση του επόμενου δείγματος με $\mathbf{w}(t+1) = \mathbf{w}(t)$. Στην περίπτωση λανθασμένης ταξινόμησης, το διάνυσμα βαρών μεταβάλλεται σύμφωνα με τον ακόλουθο κανόνα, ανάλογα με το λάθος:

$$\mathbf{w}(t+1) = \begin{cases} \mathbf{w}(t) + \rho \mathbf{x}_t, & \text{εάν } \mathbf{x}_t \in \omega_1, \ \text{αλλά } \mathbf{w}(t) \cdot \mathbf{x}_t \leq 0 \\ \mathbf{w}(t) - \rho \mathbf{x}_t, & \text{εάν } \mathbf{x}_t \in \omega_2, \ \text{αλλά } \mathbf{w}(t) \cdot \mathbf{x}_t \geq 0 \end{cases} \tag{4.3}$$

όπου ρ είναι μια σταθερά που στα πλαίσια της άσκησης θεωρείται ίση με τη μονάδα ($\rho = 1$). Τα διανύσματα χαρακτηριστικών ταξινομούνται κυκλικά και ένας πλήρης κύκλος (εποχή) αντιστοιχεί στην ταξινόμηση κάθε δείγματος της συλλογής. Η σύγκλιση του αλγορίθμου επέρχεται όταν κάθε δείγμα ταξινομείται σωστά, δηλαδή όταν το διάνυσμα βαρών δε μεταβάλλεται για 10 διαδοχικές ταξινομήσεις (όπου 10 είναι το πλήθος των δεδομένων).

Αρχικοποιώντας το διάνυσμα βαρών ως $\mathbf{w} = [0, 0, 0]^\top$, οι πράξεις για κάθε κύκλο (εποχή) του αλγορίθμου παρουσιάζονται στις επόμενες σελίδες, μαζί με την απεικόνιση της αντίστοιχης ευθείας $\mathbf{w} \cdot \mathbf{x} = 0$, προκειμένου να φανεί η πρόοδος του αλγορίθμου ανά εποχή. Ο αλγόριθμος συγκλίνει πρακτικά στο τέλος της 4ης εποχής, αλλά τυπικά έπειτα από 5 εποχές συνολικά, δίνοντας διάνυσμα βαρών

$$\mathbf{w} = [1, 4.5, -4]^\top, \tag{4.4}$$

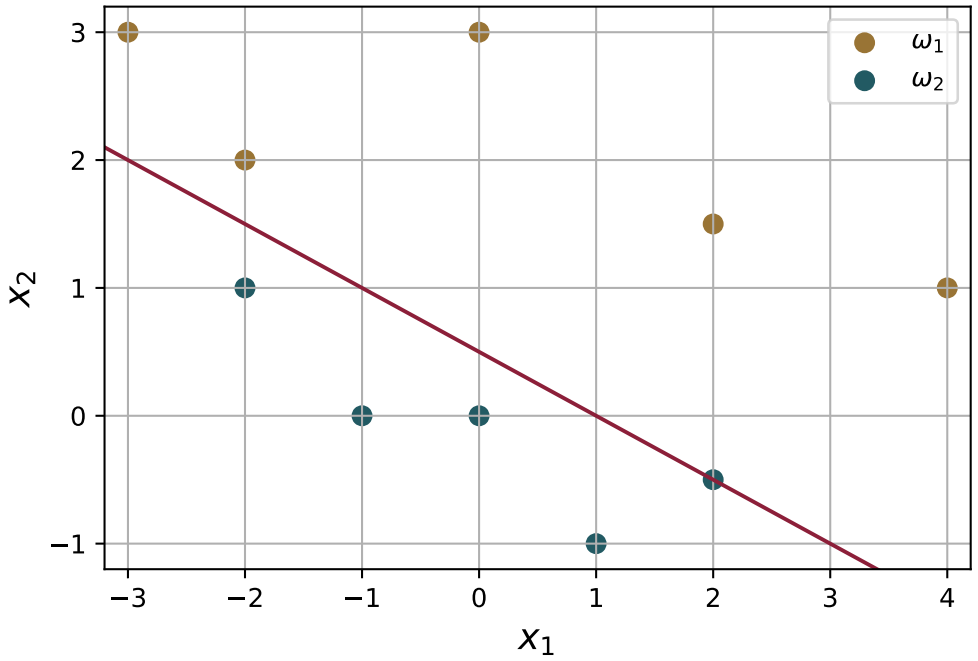
το οποίο φυσικά αντιστοιχεί στην ευθεία διαχωρισμού με εξίσωση

$$x_1 + 4.5x_2 - 4 = 0, \tag{4.5}$$

η οποία απεικονίζεται και στο γράφημα της Εικόνας 4.5.

Εποχή 1

$$\mathbf{w}(1) \cdot \mathbf{x}_1 = [0,0,0] \cdot [2,1.5,1] = 0.0, \text{ άρα } \mathbf{w}(2) = \mathbf{w}(1) + \mathbf{x}_1 = [2,1.5,1]^\mathsf{T}$$
$$\mathbf{w}(2) \cdot \mathbf{x}_2 = [2,1.5,1] \cdot [2,-0.5,1] = 4.25, \text{ άρα } \mathbf{w}(3) = \mathbf{w}(2) - \mathbf{x}_2 = [0,2.0,0]^\mathsf{T}$$
$$\mathbf{w}(3) \cdot \mathbf{x}_3 = [0,2.0,0] \cdot [-3,3,1] = 6.0, \text{ άρα } \mathbf{w}(4) = \mathbf{w}(3)$$
$$\mathbf{w}(4) \cdot \mathbf{x}_4 = [0,2.0,0] \cdot [1,-1,1] = -2.0, \text{ άρα } \mathbf{w}(5) = \mathbf{w}(4)$$
$$\mathbf{w}(5) \cdot \mathbf{x}_5 = [0,2.0,0] \cdot [0,3,1] = 6.0, \text{ άρα } \mathbf{w}(6) = \mathbf{w}(5)$$
$$\mathbf{w}(6) \cdot \mathbf{x}_6 = [0,2.0,0] \cdot [-1,0,1] = 0.0, \text{ άρα } \mathbf{w}(7) = \mathbf{w}(6) - \mathbf{x}_6 = [1,2.0,-1]^\mathsf{T}$$
$$\mathbf{w}(7) \cdot \mathbf{x}_7 = [1,2.0,-1] \cdot [4,1,1] = 5.0, \text{ άρα } \mathbf{w}(8) = \mathbf{w}(7)$$
$$\mathbf{w}(8) \cdot \mathbf{x}_8 = [1,2.0,-1] \cdot [0,0,1] = -1.0, \text{ άρα } \mathbf{w}(9) = \mathbf{w}(8)$$
$$\mathbf{w}(9) \cdot \mathbf{x}_9 = [1,2.0,-1] \cdot [-2,2,1] = 1.0, \text{ άρα } \mathbf{w}(10) = \mathbf{w}(9)$$
$$\mathbf{w}(10) \cdot \mathbf{x}_{10} = [1,2.0,-1] \cdot [-2,1,1] = -1.0, \text{ άρα } \mathbf{w}(1) = \mathbf{w}(10)$$



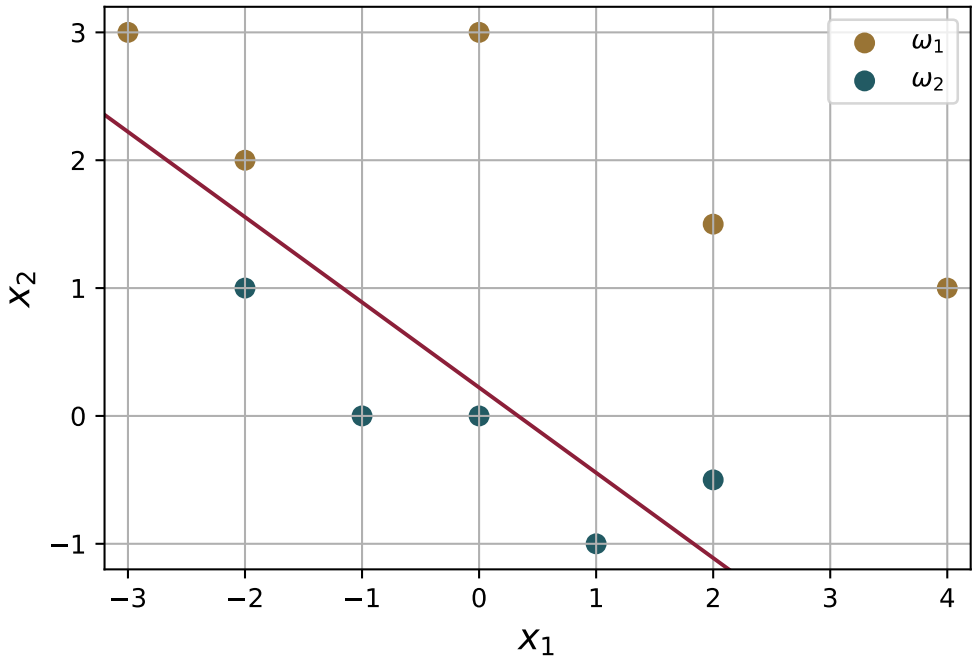
Εικόνα 4.2: Απεικόνιση των σημείων και της προτεινόμενης ευθείας διαχωρισμού στο τέλος της πρώτης εποχής.

Στο τέλος της πρώτης εποχής, η εξίσωση της ευθείας είναι η $x_1 + 2x_2 + 1 = 0$ και το γράφημα απεικονίζεται στην Εικόνα 4.2. Είναι εμφανές πως η ευθεία διαχωρισμού διέρχεται από ένα εκ των σημείων, άρα το σημείο αυτό δε μπορεί να ταξινομηθεί σωστά σε κάποια από τις δύο κατηγορίες.

Εποχή 2

$$\mathbf{w}(1) \cdot \mathbf{x}_1 = [1,2.0,-1] \cdot [2,1.5,1] = 4.0, \text{ άρα } \mathbf{w}(2) = \mathbf{w}(1)$$
$$\mathbf{w}(2) \cdot \mathbf{x}_2 = [1,2.0,-1] \cdot [2,-0.5,1] = 0.0, \text{ άρα } \mathbf{w}(3) = \mathbf{w}(2) - \mathbf{x}_2 = [-1,2.5,-2]^\mathsf{T}$$
$$\mathbf{w}(3) \cdot \mathbf{x}_3 = [-1,2.5,-2] \cdot [-3,3,1] = 8.5, \text{ άρα } \mathbf{w}(4) = \mathbf{w}(3)$$

$$\mathbf{w}(4) \cdot \mathbf{x}_4 = [-1, 2.5, -2] \cdot [1, -1, 1] = -5.5, \text{ άρα } \mathbf{w}(5) = \mathbf{w}(4)$$
$$\mathbf{w}(5) \cdot \mathbf{x}_5 = [-1, 2.5, -2] \cdot [0, 3, 1] = 5.5, \text{ άρα } \mathbf{w}(6) = \mathbf{w}(5)$$
$$\mathbf{w}(6) \cdot \mathbf{x}_6 = [-1, 2.5, -2] \cdot [-1, 0, 1] = -1.0, \text{ άρα } \mathbf{w}(7) = \mathbf{w}(6)$$
$$\mathbf{w}(7) \cdot \mathbf{x}_7 = [-1, 2.5, -2] \cdot [4, 1, 1] = -3.5, \text{ άρα } \mathbf{w}(8) = \mathbf{w}(7) + \mathbf{x}_7 = [3, 3.5, -1]^\mathsf{T}$$
$$\mathbf{w}(8) \cdot \mathbf{x}_8 = [3, 3.5, -1] \cdot [0, 0, 1] = -1.0, \text{ άρα } \mathbf{w}(9) = \mathbf{w}(8)$$
$$\mathbf{w}(9) \cdot \mathbf{x}_9 = [3, 3.5, -1] \cdot [-2, 2, 1] = 0.0, \text{ άρα } \mathbf{w}(10) = \mathbf{w}(9) + \mathbf{x}_9 = [1, 5.5, 0]^\mathsf{T}$$
$$\mathbf{w}(10) \cdot \mathbf{x}_{10} = [1, 5.5, 0] \cdot [-2, 1, 1] = 3.5, \text{ άρα } \mathbf{w}(1) = \mathbf{w}(10) - \mathbf{x}_{10} = [3, 4.5, -1]^\mathsf{T}$$



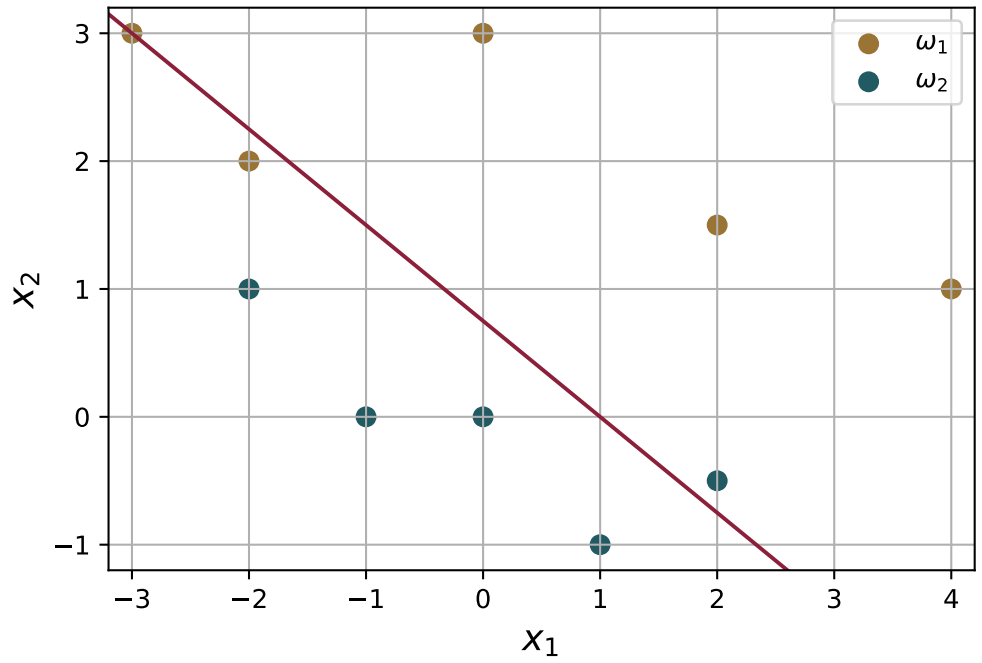
Εικόνα 4.3: Απεικόνιση των σημείων και της προτεινόμενης ευθείας διαχωρισμού στο τέλος της δεύτερης εποχής.

Στο τέλος της δεύτερης εποχής, η εξίσωση της ευθείας είναι η $3x_1 + 4.5x_2 - 1 = 0$ και το γράφημα απεικονίζεται στην Εικόνα 4.3. Τώρα, η ευθεία διαχωρισμού χωρίζει το επίπεδο με τέτοιο τρόπο, ώστε ένα εκ των σημείων να ταξινομείται λανθασμένα.

Εποχή 3

$$\mathbf{w}(1) \cdot \mathbf{x}_1 = [3, 4.5, -1] \cdot [2, 1.5, 1] = 11.75, \text{ άρα } \mathbf{w}(2) = \mathbf{w}(1)$$
$$\mathbf{w}(2) \cdot \mathbf{x}_2 = [3, 4.5, -1] \cdot [2, -0.5, 1] = 2.75, \text{ άρα } \mathbf{w}(3) = \mathbf{w}(2) - \mathbf{x}_2 = [1, 5.0, -2]^\mathsf{T}$$
$$\mathbf{w}(3) \cdot \mathbf{x}_3 = [1, 5.0, -2] \cdot [-3, 3, 1] = 10.0, \text{ άρα } \mathbf{w}(4) = \mathbf{w}(3)$$
$$\mathbf{w}(4) \cdot \mathbf{x}_4 = [1, 5.0, -2] \cdot [1, -1, 1] = -6.0, \text{ άρα } \mathbf{w}(5) = \mathbf{w}(4)$$
$$\mathbf{w}(5) \cdot \mathbf{x}_5 = [1, 5.0, -2] \cdot [0, 3, 1] = 13.0, \text{ άρα } \mathbf{w}(6) = \mathbf{w}(5)$$
$$\mathbf{w}(6) \cdot \mathbf{x}_6 = [1, 5.0, -2] \cdot [-1, 0, 1] = -3.0, \text{ άρα } \mathbf{w}(7) = \mathbf{w}(6)$$

$$\mathbf{w}(7) \cdot \mathbf{x}_7 = [1, 5.0, -2] \cdot [4, 1, 1] = 7.0, \text{ άρα } \mathbf{w}(8) = \mathbf{w}(7)$$
$$\mathbf{w}(8) \cdot \mathbf{x}_8 = [1, 5.0, -2] \cdot [0, 0, 1] = -2.0, \text{ άρα } \mathbf{w}(9) = \mathbf{w}(8)$$
$$\mathbf{w}(9) \cdot \mathbf{x}_9 = [1, 5.0, -2] \cdot [-2, 2, 1] = 6.0, \text{ άρα } \mathbf{w}(10) = \mathbf{w}(9)$$
$$\mathbf{w}(10) \cdot \mathbf{x}_{10} = [1, 5.0, -2] \cdot [-2, 1, 1] = 1.0, \text{ άρα } \mathbf{w}(1) = \mathbf{w}(10) - \mathbf{x}_{10} = [3, 4.0, -3]^\mathsf{T}$$

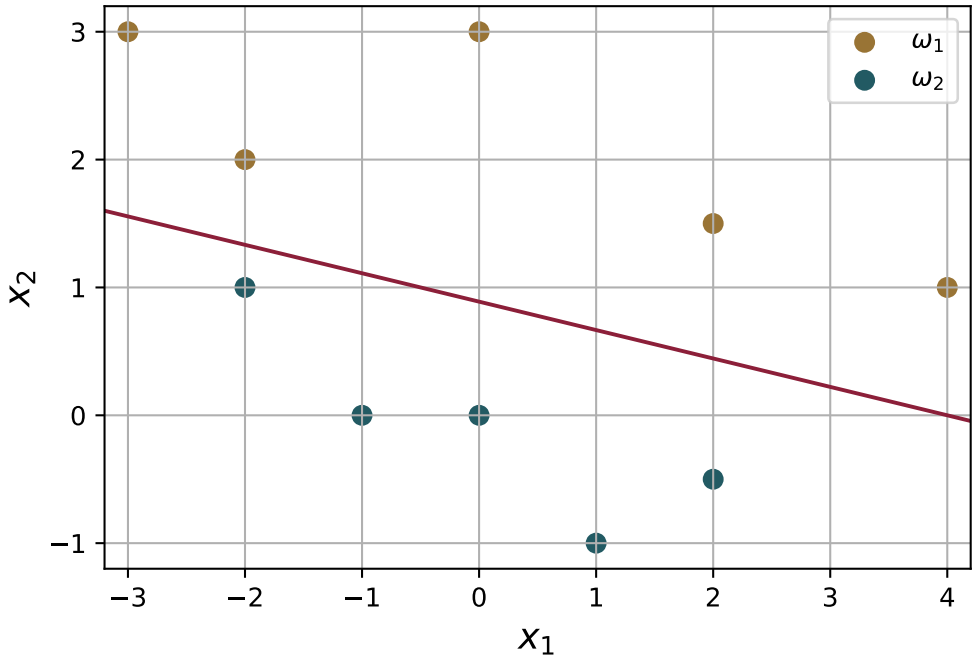


Εικόνα 4.4: Απεικόνιση των σημείων και της προτεινόμενης ευθείας διαχωρισμού στο τέλος της τρίτης εποχής.

Στο τέλος της τρίτης εποχής, η εξίσωση της ευθείας είναι η $3x_1 + 4x_2 - 3 = 0$ και το γράφημα απεικονίζεται στην Εικόνα 4.4. Η ευθεία διαχωρισμού είναι τέτοια, ώστε η ταξινόμηση δύο σημείων να γίνεται λανθασμένα και ένα τρίτο σημείο να μην ταξινομείται, καθώς αποτελεί σημείο της ευθείας.

Εποχή 4

$$\mathbf{w}(1) \cdot \mathbf{x}_1 = [3, 4.0, -3] \cdot [2, 1.5, 1] = 9.0, \text{ άρα } \mathbf{w}(2) = \mathbf{w}(1)$$
$$\mathbf{w}(2) \cdot \mathbf{x}_2 = [3, 4.0, -3] \cdot [2, -0.5, 1] = 1.0, \text{ άρα } \mathbf{w}(3) = \mathbf{w}(2) - \mathbf{x}_2 = [1, 4.5, -4]^\mathsf{T}$$
$$\mathbf{w}(3) \cdot \mathbf{x}_3 = [1, 4.5, -4] \cdot [-3, 3, 1] = 6.5, \text{ άρα } \mathbf{w}(4) = \mathbf{w}(3)$$
$$\mathbf{w}(4) \cdot \mathbf{x}_4 = [1, 4.5, -4] \cdot [1, -1, 1] = -7.5, \text{ άρα } \mathbf{w}(5) = \mathbf{w}(4)$$
$$\mathbf{w}(5) \cdot \mathbf{x}_5 = [1, 4.5, -4] \cdot [0, 3, 1] = 9.5, \text{ άρα } \mathbf{w}(6) = \mathbf{w}(5)$$
$$\mathbf{w}(6) \cdot \mathbf{x}_6 = [1, 4.5, -4] \cdot [-1, 0, 1] = -5.0, \text{ άρα } \mathbf{w}(7) = \mathbf{w}(6)$$
$$\mathbf{w}(7) \cdot \mathbf{x}_7 = [1, 4.5, -4] \cdot [4, 1, 1] = 4.5, \text{ άρα } \mathbf{w}(8) = \mathbf{w}(7)$$
$$\mathbf{w}(8) \cdot \mathbf{x}_8 = [1, 4.5, -4] \cdot [0, 0, 1] = -4.0, \text{ άρα } \mathbf{w}(9) = \mathbf{w}(8)$$
$$\mathbf{w}(9) \cdot \mathbf{x}_9 = [1, 4.5, -4] \cdot [-2, 2, 1] = 3.0, \text{ άρα } \mathbf{w}(10) = \mathbf{w}(9)$$
$$\mathbf{w}(10) \cdot \mathbf{x}_{10} = [1, 4.5, -4] \cdot [-2, 1, 1] = -1.5, \text{ άρα } \mathbf{w}(1) = \mathbf{w}(10)$$



Εικόνα 4.5: Απεικόνιση των σημείων και της προτεινόμενης ευθείας διαχωρισμού στο τέλος της τέταρτης εποχής.

Στο τέλος της τέταρτης εποχής, η εξίσωση της ευθείας είναι η $x_1 + 4.5x_2 - 4 = 0$ και το γράφημα απεικονίζεται στην Εικόνα 4.5. Φαίνεται πως η ευθεία διαχωρισμού είναι κατάλληλη ώστε όλα τα δείγματα να ταξινομούνται σωστά στις κατηγορίες τους. Παρ' όλα αυτά, ο αλγόριθμος θεωρείται πως έχει συγκλίνει όταν το διάνυσμα βαρών δε μεταβάλεται σε έναν πλήρη κύκλο. Η τελευταία μεταβολή του διανύσματος βαρών πραγματοποιήθηκε στο βήμα $t =$ της 4ης εποχής, το οποίο αντιστοιχούσε στην ταξινόμηση του σημείου $\mathbf{x}_2 = [2, -0.5, 1]^\mathsf{T}$. Τυπικά, λοιπόν, θα πρέπει να ξεκινήσει και μια 5η εποχή, κατά την οποία θα επιχειρηθεί η ταξινόμηση των $\mathbf{x}_1$ και $\mathbf{x}_2$. Εάν αυτή γίνει επιτυχώς, τότε μπορεί κανείς να πει με βεβαιότητα πως επήλθε σύγκλιση.

Εποχή 5

$$\mathbf{w}(1) \cdot \mathbf{x}_1 = [1, 4.5, -4] \cdot [2, 1.5, 1] = 4.75, \quad \text{άρα } \mathbf{w}(2) = \mathbf{w}(1)$$
$$\mathbf{w}(2) \cdot \mathbf{x}_2 = [1, 4.5, -4] \cdot [2, -0.5, 1] = -4.25, \quad \text{άρα } \mathbf{w}(3) = \mathbf{w}(2)$$

Πράγματι, τα δύο δείγματα ταξινομούνται σωστά, επομένως ο αλγόριθμος συγκλίνει στο διάνυσμα βαρών της Σχέσης (4.4), το οποίο αντιστοιχεί στην ευθεία διαχωρισμού της Σχέσης (4.5).

5 Kullback-Leibler Divergence

Αρχικά, θα πρέπει να σημειωθεί πως εφόσον η $p_1(\mathbf{x})$ αντιστοιχεί σε Γκαουσιανή και η $p_2(\mathbf{x})$ είναι η συνάρτηση που πρέπει να προσεγγιστεί, η μετρική Kullback-Leibler που πρέπει να ελαχιστοποιηθεί είναι η

$$D_{\mathrm{KL}}\left(p_2(\mathbf{x}), p_1(\mathbf{x})\right) = \int d\mathbf{x}\, p_2(\mathbf{x}) \ln \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} \tag{5.1}$$

και όχι η $D_{\mathrm{KL}}\left(p_1(\mathbf{x}), p_2(\mathbf{x})\right)$ που αναγράφεται στην εκφώνηση. Ένας απλός τρόπος να φανεί αυτό είναι το ακόλουθο αντιπαράδειγμα. Έστω $x \in \mathbb{R}$. Τότε $p_1(x) \sim \mathcal{N}\left(\mu, \sigma^2\right)$ και η $p_2(x)$ είναι η κατανομή που πρέπει να προσεγγιστεί βάσει της Γκαουσιανής. Μπορεί κανείς να υποθέσει πως η $p_2(x)$ είναι μια Λαπλασιανή με μέση τιμή 0 και διακύμανση $2b^2$. Παίρνοντας για την Γκαουσιανή μέση τιμή $\mu = 0$, εάν η εκφώνηση της άσκησης είναι σωστή, τότε η $D_{\mathrm{KL}}\left(p_1(\mathbf{x}), p_2(\mathbf{x})\right)$ θα πρέπει να ελαχιστοποιείται όταν $\sigma^2 = 2b^2$. Αντικαθιστώντας στην Σχέση $D_{\mathrm{KL}}\left(p_1(\mathbf{x}), p_2(\mathbf{x})\right)$ της εκφώνησης τις

$$p_1(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad \text{και} \quad p_2(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right), \tag{5.2}$$

προκύπτει

$$D_{\mathrm{KL}}\left(p_1(\mathbf{x}), p_2(\mathbf{x})\right) = -\frac{1}{2} + \sqrt{\frac{2}{\pi}}\frac{\sigma}{b} + \ln b - \ln\left(\sqrt{2\pi}\sigma\right). \tag{5.3}$$

Ελαχιστοποιώντας αυτήν την έκφραση είτε ως προς b είτε ως προς σ, προκύπτει η εξίσωση

$$\sigma^2 = \frac{\pi}{2} b^2 \tag{5.4}$$

και δεδομένου πως $\mathrm{Var}\left[\mathcal{N}\left(\mu, \sigma^2\right)\right] = \sigma^2$ και $\mathrm{Var}[\mathrm{Laplace}\,(\mu, b)] = 2b^2$ η Σχέση (5.4) οδηγεί σε αποτέλεσμα για το οποίο ισχύει

$$\mathrm{Var}\left[\mathcal{N}\left(\mu, \sigma^2\right)\right] \neq \mathrm{Var}\left[\mathrm{Laplace}\,(\mu, b)\right], \tag{5.5}$$

σε αντίθεση με αυτό που ζητείται προς απόδειξη από την εκφώνηση. Εάν, αντίθετα, οι Σχέσεις (5.2) αντικατασταθούν στην $D_{\mathrm{KL}}\left(p_2(\mathbf{x}), p_1(\mathbf{x})\right)$ της Σχέσης (5.1), τότε προκύπτει

$$D_{\mathrm{KL}}\left(p_2(\mathbf{x}), p_1(\mathbf{x})\right) = -1 + \frac{b^2}{\sigma^2} - \ln(2b) + \ln\left(\sqrt{2\pi}\sigma\right). \tag{5.6}$$

Η ελαχιστοποίηση αυτής της έκφρασης είτε ως προς b είτε ως προς σ οδηγεί στην εξίσωση

$$\sigma^2 = 2b^2, \tag{5.7}$$

η οποία ισοδυναμεί με την προς απόδειξη σχέση

$$\mathrm{Var}\left[\mathcal{N}\left(\mu, \sigma^2\right)\right] = \mathrm{Var}\left[\mathrm{Laplace}\,(\mu, b)\right]. \tag{5.8}$$

Δεδομένων αυτών, έστω τώρα $\mathbf{x} \in \mathbb{R}^d$ και $p_2(\mathbf{x})$ μια τυχαία κατανομή, η οποία θα προσεγγιστεί με την $p_1(\mathbf{x})$, για την οποία ισχύει

$$p_1(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\mathbf{\Sigma})}} \cdot \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right), \tag{5.9}$$

μέσω ελαχιστοποίησης της $D_{\mathrm{KL}}(p_2(\mathbf{x}), p_1(\mathbf{x}))$. Ισχύει

$$\begin{aligned} D_{\mathrm{KL}}(p_2(\mathbf{x}), p_1(\mathbf{x})) &= \underbrace{\int d\mathbf{x}\, p_2(\mathbf{x}) \ln p_2(\mathbf{x})}_{C} - \int d\mathbf{x}\, p_2(\mathbf{x}) \ln p_1(\mathbf{x}) \\ &= C + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_2}\left\{d\ln(2\pi) + \ln[\det(\mathbf{\Sigma})] + (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \\ &= C + \frac{d}{2}\ln(2\pi) + \frac{1}{2}A(\boldsymbol{\mu}, \mathbf{\Sigma}), \end{aligned} \tag{5.10}$$

όπου

$$A(\boldsymbol{\mu}, \mathbf{\Sigma}) = \mathbb{E}_{\mathbf{x}\sim p_2}\left\{\ln[\det(\mathbf{\Sigma})] + (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{5.11}$$

Η ελαχιστοποίηση της $D_{\mathrm{KL}}(p_2(\mathbf{x}), p_1(\mathbf{x}))$ ως προς $\boldsymbol{\mu}$ και $\mathbf{\Sigma}$ ισοδυναμεί με την ελαχιστοποίηση της $A(\boldsymbol{\mu}, \mathbf{\Sigma})$ ως προς $\boldsymbol{\mu}$ και $\mathbf{\Sigma}$, αντίστοιχα. Δεδομένου πως ο πίνακας $\mathbf{\Sigma}$ αντιστοιχεί σε πίνακα συνδιακύμανσης, είναι συμμετρικός. Έτσι, προκύπτουν

$$\frac{\partial}{\partial \boldsymbol{\mu}}\left[(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] = -\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \left(\mathbf{\Sigma}^{-1}\right)^\top(\mathbf{x}-\boldsymbol{\mu}) = -2\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{5.12}$$

και

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{\Sigma}}\left\{\ln[\det(\mathbf{\Sigma})] + (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} = \frac{1}{\det(\mathbf{\Sigma})}\frac{\partial \det(\mathbf{\Sigma})}{\partial \mathbf{\Sigma}} + \frac{\partial}{\partial \mathbf{\Sigma}}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right] = \\ &\frac{1}{\det(\mathbf{\Sigma})}\mathrm{adj}(\mathbf{\Sigma}) - \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} = \frac{1}{\det(\mathbf{\Sigma})}\cdot\det(\mathbf{\Sigma})\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} = \\ &\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} = \mathbf{\Sigma}^{-1}\left[\mathbb{1} - (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}\right]. \end{aligned} \tag{5.13}$$

Χρησιμοποιώντας τα αποτελέσματα αυτά, η ελαχιστοποίηση της $D_{\mathrm{KL}}(p_2(\mathbf{x}), p_1(\mathbf{x}))$ δίνει

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\mu}}\left[D_{\mathrm{KL}}(p_2(\mathbf{x}), p_1(\mathbf{x}))\right] = \mathbb{0} \Leftrightarrow \frac{\partial A}{\partial \boldsymbol{\mu}} = \mathbb{0} \Leftrightarrow \mathbb{E}_{\mathbf{x}\sim p_2}\left\{-2\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} = \mathbb{0} \Leftrightarrow \\ &\mathbf{\Sigma}^{-1}\mathbb{E}_{\mathbf{x}\sim p_2}\{(\mathbf{x}-\boldsymbol{\mu})\} = \mathbb{0} \Leftrightarrow \mathbf{\Sigma}\cdot\mathbf{\Sigma}^{-1}\mathbb{E}_{\mathbf{x}\sim p_2}\{(\mathbf{x}-\boldsymbol{\mu})\} = \mathbf{\Sigma}\cdot\mathbb{0} \Leftrightarrow \mathbb{E}_{\mathbf{x}\sim p_2}\{(\mathbf{x}-\boldsymbol{\mu})\} = \mathbb{0} \Leftrightarrow \\ &\mathbb{E}_{\mathbf{x}\sim p_2}\{\mathbf{x}\} = \boldsymbol{\mu} \end{aligned} \tag{5.14}$$

και

$$\frac{\partial}{\partial \boldsymbol{\Sigma}}\left[D_{\mathrm{KL}}\left(p_2(\mathbf{x}), p_1(\mathbf{x})\right)\right] = \mathbb{0} \Leftrightarrow \frac{\partial A}{\partial \boldsymbol{\Sigma}} = \mathbb{0} \Leftrightarrow \mathbb{E}_{\mathbf{x}\sim p_2}\left\{\boldsymbol{\Sigma}^{-1}\left[\mathbb{1} - (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right]\right\} = \mathbb{0} \Leftrightarrow$$

$$\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\mathbf{x}\sim p_2}\left\{\mathbb{1} - (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right\} = \mathbb{0} \Leftrightarrow \boldsymbol{\Sigma}\cdot\boldsymbol{\Sigma}^{-1}\mathbb{E}_{\mathbf{x}\sim p_2}\left\{\mathbb{1} - (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right\} = \boldsymbol{\Sigma}\cdot\mathbb{0} \Leftrightarrow$$

$$\mathbb{E}_{\mathbf{x}\sim p_2}\left\{\mathbb{1} - (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right\} = \mathbb{0} \Leftrightarrow \mathbb{E}_{\mathbf{x}\sim p_2}\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}\right\} = \mathbb{1} \Leftrightarrow$$

$$\mathbb{E}_{\mathbf{x}\sim p_2}\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\right\}\boldsymbol{\Sigma}^{-1}\cdot\boldsymbol{\Sigma} = \mathbb{1}\cdot\boldsymbol{\Sigma} \Leftrightarrow \boldsymbol{\Sigma} = \mathbb{E}_{\mathbf{x}\sim p_2}\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\top}\right\}, \tag{5.15}$$

αποδεικνύοντας έτσι τα ζητούμενα.

6 Linear Regression and the LMS Algorithm

6.1 Το υπό μελέτη μοντέλο είναι το απλό γραμμικό μοντέλο, για $x_i, y_i \in \mathbb{R}$ με $i = 1, \dots, 10$. Σε ό,τι αφορά τα δεδομένα (x_i, y_i), αυτά θα υπακούν τη σχέση

$$y = w_0 + w_1 x + \varepsilon_i, \tag{6.1}$$

όπου τα ε_i αποτελούν τα τυχαία σφάλματα, για τα οποία ισχύει βάσει υπόθεσης γκαουσιανού θορύβου πως

$$\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right). \tag{6.2}$$

Δεδομένου αυτού, οι παρατηρήσεις y_i θα ακολουθούν επίσης μια κανονική κατανομή και συγκεκριμένα

$$y_i \sim \mathcal{N}\left(w_0 + w_1 x_i, \sigma^2\right). \tag{6.3}$$

Δεδομένης της ανεξαρτησίας μεταξύ των δεδομένων, η πιθανοφάνεια $\mathcal{L}$ θα δίνεται ως το γινόμενο

$$\begin{aligned} \mathcal{L}\left(y_1, \dots, y_{10} | w_0, w_1, x_1, \dots, x_{10}\right) &= \prod_{i=1}^{10} \mathcal{N}\left(w_0 + w_1 x_i, \sigma^2\right) \\ &= \frac{1}{(2\pi)^5 \sigma^{10}} \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{10} \left(y_i - w_0 - w_1 x_i\right)^2\right]. \end{aligned} \tag{6.4}$$

Εφόσον ο φυσικός λογάριθμος είναι γνησίως αύξουσα συνάρτηση, η μεγιστοποίηση της πιθανοφάνειας ισοδυναμεί με τη μεγιστοποίηση του λογαρίθμου της, για τον οποίο ισχύει

$$\ln \mathcal{L} = -5 \ln (2\pi) - 10 \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{10} \left(y_i - w_0 - w_1 x_i\right)^2. \tag{6.5}$$

Οι παράμετροι ως προς τις οποίες πρέπει να μεγιστοποιηθεί ο λογάριθμος της πιθανοφάνειας είναι τα w_0 και w_1. Θέτοντας

$$S\left(w_0, w_1\right) = \sum_{i=1}^{10} \left(y_i - w_0 - w_1 x_i\right)^2 \tag{6.6}$$

και δεδομένου πως

$$\frac{\partial}{\partial w_0} \left(-5 \ln (2\pi) - 10 \ln \sigma\right) = \frac{\partial}{\partial w_1} \left(-5 \ln (2\pi) - 10 \ln \sigma\right) = 0 \tag{6.7}$$

ταυτοτικά, συμπεραίνει κανείς πως η μεγιστοποίηση της πιθανοφάνειας ισοδυναμεί με την ελαχιστοποίηση της $S\left(w_0, w_1\right)$, δεδομένου του αρνητικού προσήμου στον τρίτο όρο της Σχέσης (6.5). Η παράσταση αυτή δεν είναι παρά το άθροισμα των τετραγώνων των αποκλίσεων μεταξύ των παρατηρήσεων y_i και των εκτιμήσεων του απλού γραμμικού μοντέλου $w_0 + w_1 x_i$, επομένως, πράγματι, η μεγιστοποίηση της πιθανοφάνειας ισοδυναμεί με τη μέθοδο ελαχίστων τετραγώνων.

6.2 Η μέθοδος ελαχίστων τετραγώνων αποτελεί την ελαχιστοποίηση της $S(w_0, w_1)$. Συμβολίζοντας με $(\hat{w}_0, \hat{w}_1)$ τις τιμές των w_0 και w_1 που ελαχιστοποιούν την ποσότητα αυτή, ισχύουν

$$\left.\frac{\partial S(w_0, w_1)}{\partial w_0}\right|_{(\hat{w}_0, \hat{w}_1)} = 0 \Leftrightarrow -2\sum_{i=1}^{10}(y_i - \hat{w}_0 - \hat{w}_1 x_i) = 0 \Leftrightarrow$$

$$\Leftrightarrow 10\hat{w}_0 + \hat{w}_1 \sum_{i=1}^{10} x_i = \sum_{i=1}^{10} y_i. \tag{6.8}$$

και

$$\left.\frac{\partial S(w_0, w_1)}{\partial w_1}\right|_{(\hat{w}_0, \hat{w}_1)} = 0 \Leftrightarrow -2\sum_{i=1}^{10} x_i (y_i - \hat{w}_0 - \hat{w}_1 x_i) = 0 \Leftrightarrow$$

$$\Leftrightarrow \hat{w}_0 \sum_{i=1}^{10} x_i + \hat{w}_1 \sum_{i=1}^{10} {x_i}^2 = \sum_{i=1}^{10} y_i x_i. \tag{6.9}$$

Με βάση τα δοσμένα δεδομένα, ισχύουν:

$$\sum_{i=1}^{10} x_i = 0.38 + 0.44 + 0.48 + 0.54 + 0.58 + 0.64 + 0.71 + 0.76 + 0.82 + 0.96 = 6.31$$

$$\sum_{i=1}^{10} y_i = 2.05 + 2.23 + 2.13 + 2.33 + 2.67 + 2.68 + 2.81 + 2.97 + 3.12 + 3.20 = 26.19$$

$$\sum_{i=1}^{10} {x_i}^2 = 0.38^2 + 0.44^2 + 0.48^2 + 0.54^2 + 0.58^2 + 0.64^2 + 0.71^2 + 0.76^2 + 0.82^2 + 0.96^2$$

$$= 4.2817$$

$$\sum_{i=1}^{10} x_i y_i = 0.38 \cdot 2.05 + 0.44 \cdot 2.23 + 0.48 \cdot 2.13 + 0.54 \cdot 2.33 + 0.58 \cdot 2.67 + 0.64 \cdot 2.68 +$$

$$+ 0.71 \cdot 2.81 + 0.76 \cdot 2.97 + 0.82 \cdot 3.12 + 0.96 \cdot 3.20 = 17.1873 \tag{6.10}$$

Οι Σχέσεις (6.8) και (6.9) συνιστούν, επομένως, το ακόλουθο σύστημα:

$$\begin{aligned} 10\hat{w}_0 + 6.31\hat{w}_1 &= 26.19 \\ 6.31\hat{w}_0 + 4.2817\hat{w}_1 &= 17.1873 \end{aligned} \tag{6.11}$$

Οι σχετικές ορίζουσες του συστήματος υπολογίζονται ως

$$\begin{aligned} D &= 10 \cdot 4.2817 - 6.31^2 = 3.0009 \\ D_x &= 26.19 \cdot 4.2817 - 6.31 \cdot 17.1873 = 3.6859 \\ D_y &= 10 \cdot 17.1873 - 6.31 \cdot 26.19 = 6.6141, \end{aligned} \tag{6.12}$$

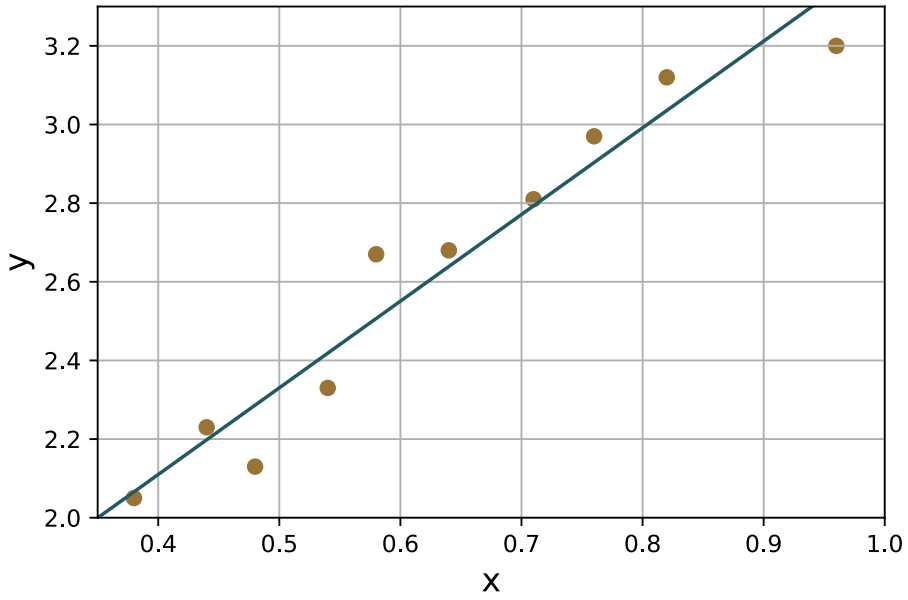
συνεπώς η λύση του συστήματος αντιστοιχεί στις τιμές

$$\hat{w}_0 = \frac{D_x}{D} \simeq 1.228 \quad \text{και} \quad \hat{w}_1 = \frac{D_y}{D} \simeq 2.204. \tag{6.13}$$

Δεδομένου πως ο πίνακας των δεύτερων παραγώγων της S είναι θετικά ορισμένος στο σημείο $(\hat{w}_0, \hat{w}_1)$, είναι βέβαιο πως η διαδικασία στασιμοποίησης της S ισοδυναμεί με ελαχιστοποίηση (και όχι μεγιστοποίηση). Συνεπώς, το βέλτιστο μοντέλο βάσει μεθόδου ελαχίστων τετραγώνων είναι το

$$\hat{y} = 1.228 + 2.204x. \tag{6.14}$$

Στην Εικόνα 6.1 απεικονίζονται τα σημεία που αντιστοιχούν στα δεδομένα (x_i, y_i), καθώς και η ευθεία της Σχέσης (6.14).



Εικόνα 6.1: Απεικόνιση των σημείων που αντιστοιχούν στα δεδομένα μαζί με την ευθεία της Σχέσης (6.14) που προκύπτει με τη μέθοδο ελαχίστων τετραγώνων.

6.3 Το επόμενο βήμα αποτελεί ο υπολογισμός των συντελεστών w_0 και w_1 χρησιμοποιώντας τον αλγόριθμο Least Mean Squares (LMS). Θεωρώντας νέα διανύσματα, $\mathbf{x}_i = [1, x_i]^\mathsf{T}$, μπορεί κανείς να ορίσει ένα αντίστοιχο διάνυσμα βαρών $\mathbf{w} = [w_0, w_1]^\mathsf{T}$, συνιστώσες του οποίου αποτελούν οι προς προσδιορισμό συντελεστές, ώστε η Σχέση (6.6) να πάρει τη μορφή

$$S(\mathbf{w}) = \sum_{i=1}^{10} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2. \tag{6.15}$$

Η αρχή λειτουργίας του εν λόγω αλγορίθμου αποτελεί στην ουσία περίπτωση gradient descent και είναι η ακόλουθη: ξεκινώντας από μια αυθαίρετη τιμή του $\mathbf{w}$, η οποία καλείται $\mathbf{w}(1)$, το διάνυσμα βαρών μεταβάλλεται σε κάθε βήμα, t, του αλγορίθμου, σύμφωνα με τον κανόνα

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \rho(t)(y_t - \mathbf{w}(t) \cdot \mathbf{x}_t)\mathbf{x}_t, \tag{6.16}$$

όπου ο συμβολισμός $\mathbf{x}_t$ και y_t υποδεικνύει πως σε κάθε βήμα τα δεδομένα λαμβάνονται ανά ένα και κυκλικά και μια πλήρης εποχή αντιστοιχεί σε 10 βήματα, όπου και τα 10 δεδομένα έχουν περάσει από τον αλγόριθμο. Επιπλέον η $\rho\,(t)$ είναι μια συνάρτηση κόστους, η οποία μπορεί να μεταβάλλεται ανά βήμα. Αξίζει να σημειωθεί πως

$$\nabla S\,(\mathbf{w}) \sim \sum_{t=1}^{10} (y_i - \mathbf{w} \cdot \mathbf{x}_i)\,\mathbf{x}_i, \tag{6.17}$$

το οποίο είναι και ο λόγος για τον οποίο ο αλγόριθμος LMS εντάσσεται στην ευρύτερη κατηγορία των gradient descent αλγορίθμων και έχει τη μορφή της Σχέσης (6.16). Η βασική του εννοιολογική διαφορά με τη μέθοδο ελαχίστων τετραγώνων είναι πως έχει στοχαστική φύση και επεξεργάζεται ένα δείγμα τη φορά αντί για το σύνολο των δειγμάτων, αφού σε κάθε βήμα δε λαμβάνεται υπ' όψιν ολόκληρο το άθροισμα της Σχέσης (6.17), παρά μόνο ένας όρος του.

Αρχικοποιώντας το διάνυσμα $\mathbf{w}$ στην τιμή $\mathbf{w}\,(1) = [0, 0]^\mathsf{T}$ και ακολουθώντας την πρόταση των [1] προκειμένου να επιτευχθεί σύγκλιση, επιλέγεται

$$\rho(t) = 1/t. \tag{6.18}$$

Τα αποτελέσματα για την πρώτη εποχή λειτουργίας του αλγορίθμου (ένας πλήρης κύκλος στα δεδομένα) παρουσιάζονται παρακάτω.

Εποχή 1

$$\mathbf{w}\,(2) = [0, 0]^\mathsf{T} + \frac{1}{1}\,(2.05 - [0, 0] \cdot [1, 0.38])\,[1, 0.38]^\mathsf{T} = [2.05, 0.779]^\mathsf{T}$$

$$\mathbf{w}\,(3) = [2.05, 0.779]^\mathsf{T} + \frac{1}{2}\,(2.23 - [2.05, 0.779] \cdot [1, 0.44])\,[1, 0.38]^\mathsf{T} = [1.969, 0.743]^\mathsf{T}$$

$$\mathbf{w}\,(4) = [1.969, 0.743]^\mathsf{T} + \frac{1}{3}\,(2.13 - [1.969, 0.743] \cdot [1, 0.48])\,[1, 0.38]^\mathsf{T} = [1.904, 0.712]^\mathsf{T}$$

$$\mathbf{w}\,(5) = [1.904, 0.712]^\mathsf{T} + \frac{1}{4}\,(2.33 - [1.904, 0.712] \cdot [1, 0.54])\,[1, 0.38]^\mathsf{T} = [1.914, 0.718]^\mathsf{T}$$

$$\mathbf{w}\,(6) = [1.914, 0.718]^\mathsf{T} + \frac{1}{5}\,(2.67 - [1.914, 0.718] \cdot [1, 0.58])\,[1, 0.38]^\mathsf{T} = [1.982, 0.757]^\mathsf{T}$$

$$\mathbf{w}\,(7) = [1.982, 0.757]^\mathsf{T} + \frac{1}{6}\,(2.68 - [1.982, 0.757] \cdot [1, 0.64])\,[1, 0.38]^\mathsf{T} = [2.018, 0.78]^\mathsf{T}$$

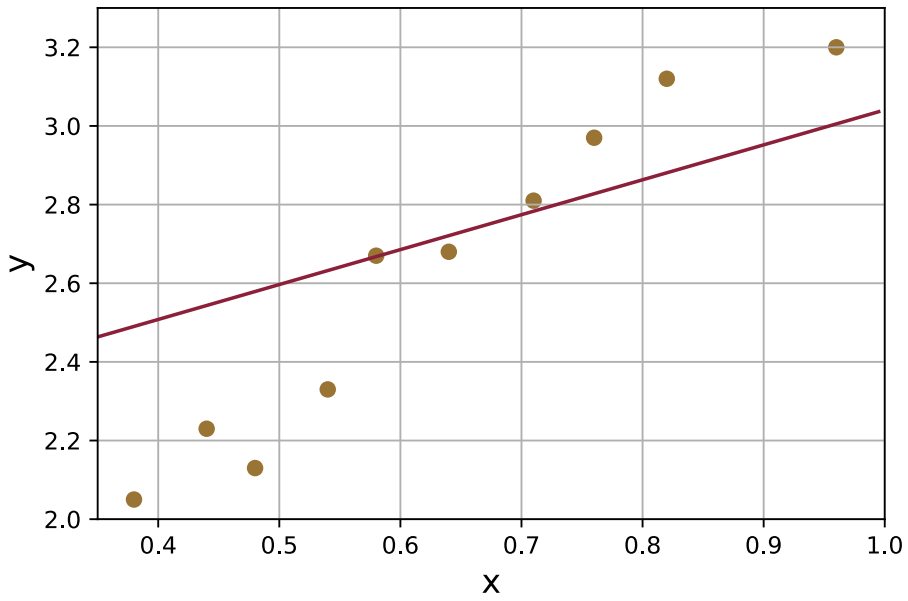
$$\mathbf{w}\,(8) = [2.018, 0.78]^\mathsf{T} + \frac{1}{7}\,(2.81 - [2.018, 0.78] \cdot [1, 0.71])\,[1, 0.38]^\mathsf{T} = [2.052, 0.804]^\mathsf{T}$$

$$\mathbf{w}\,(9) = [2.052, 0.804]^\mathsf{T} + \frac{1}{8}\,(2.97 - [2.052, 0.804] \cdot [1, 0.76])\,[1, 0.38]^\mathsf{T} = [2.09, 0.833]^\mathsf{T}$$

$$\mathbf{w}\,(10) = [2.09, 0.833]^\mathsf{T} + \frac{1}{9}\,(2.12 - [2.09, 0.833] \cdot [1, 0.82])\,[1, 0.38]^\mathsf{T} = [2.129, 0.865]^\mathsf{T}$$

$$\mathbf{w}\,(1) = [2.129, 0.865]^\mathsf{T} + \frac{1}{10}\,(3.2 - [2.129, 0.865] \cdot [1, 0.96])\,[1, 0.38]^\mathsf{T} = [2.153, 0.888]^\mathsf{T}$$

Δεδομένου του διανύσματος βάρους που προκύπτει στο τέλος της πρώτης εποχής, $\mathbf{w} = [2.153, 0.888]^\mathsf{T}$, η βέλτιστη ευθεία σύμφωνα με τον αλγόριθμο LMS απεικονίζεται στο γράφημα της Εικόνας 6.2.



Εικόνα 6.2: Απεικόνιση των σημείων που αντιστοιχούν στα δεδομένα μαζί με την ευθεία που προκύπτει στο τέλος της πρώτης εποχής του αλγορίθμου LMS.

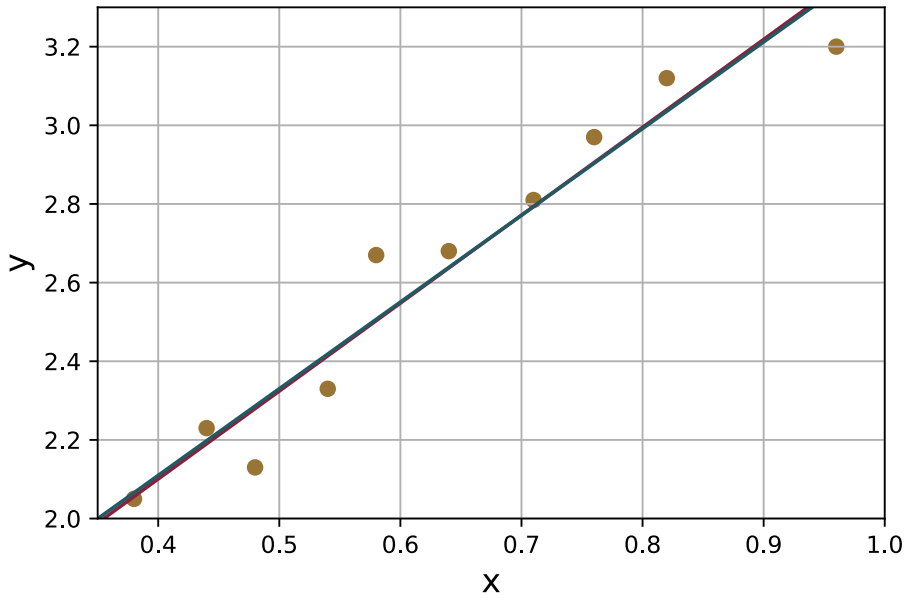
6.4 Γίνεται εμφανές πως ένας μόνο κύκλος του αλγορίθμου δεν επαρκεί προκειμένου να λάβει κανείς ένα ικανοποιητικό αποτέλεσμα. Για το σκοπό αυτό, ο αλγόριθμος αυτός αναπτύχθηκε σε γλώσσα Python και εφαρμόστηκε είτε για 500 εποχές, είτε μέχρι να επέλθει σύγκλιση. Στην περίπτωση αυτή, η σύγκλιση ισοδυναμεί με τη διόρθωση που πραγματοποιείται στο τέλος κάθε επανάληψης να είναι μικρότερη από ένα κατώφλι, το οποίο επιλέχθηκε ίσο με $5 \cdot 10^{-4}$, δεδομένου πως τα αποτελέσματα παρουσιάζονται, ως αυτό το σημείο, με ακρίβεια τριών δεκαδικών. Ο κώδικας φαίνεται παρακάτω.

```
1  import numpy as np
2
3  data = [(0.38,2.05),(0.44,2.23),(0.48,2.13),(0.54,2.33),(0.58,2.67),\
4          (0.64,2.68),(0.71,2.81),(0.76,2.97),(0.82,3.12),(0.96,3.20)]
5
6  Xs = []
7  Ys = []
8
9  for i in range(10):
10     Xs.append(data[i][0])
11     Ys.append(data[i][1])
12
13 cycles = 500
14 epsilon = 0.0005
15 correction = 1.0
16 cycle = 0
17 weight = [0,0]
18
19 while cycle < cycles and abs(correction) > epsilon:
20     for i in range(10):
21         if abs(correction) < epsilon:
22             break
23
```

```
24          rho = 1/(i + 1)
25          x_i, y_i = Xs[i], Ys[i]
26          update = rho*(y_i - weight[0] - weight[1]*x_i)
27
28          weight[0] += update
29          weight[1] += update*x_i
30          correction = update*np.sqrt(1 + x_i**2)
31
32      cycle += 1
33
34  print(weight)
```

6.5 Με το πέρας λειτουργίας του αλγορίθμου, το διάνυσμα βαρών που προκύπτει είναι το $\mathbf{w} = [1.206, 2.236]^{\mathsf{T}}$. Στο γράφημα της Εικόνας 6.3 απεικονίζεται με κόκκινο χρώμα η ευθεία που προέκυψε μέσω της μεθόδου ελαχίστων τετραγώνων και με μπλε χρώμα η ευθεία που προέκυψε μέσω της μεθόδου LMS, δηλαδή η

$$\tilde{y} = 1.206 + 2.236x. \tag{6.19}$$



Εικόνα 6.3: Απεικόνιση των σημείων που αντιστοιχούν στα δεδομένα μαζί με την ευθεία της Σχέσης (6.19), η οποία προέκυψε κατόπιν σύγκλισης του αλγορίθμου LMS (κόκκινο χρώμα), καθώς και την ευθεία της Σχέσης (6.14), η οποία προέκυψε μέσω μεθόδου ελαχίστων τετραγώνων (μπλε χρώμα).

Είναι ολοφάνερο πως όταν ο αλγόριθμος LMS τρέχει για περισσότερες εποχές, το αποτέλεσμα που δίνει είναι σημαντικά καλύτερο από το αντίστοιχο στο τέλος της πρώτης μόνο εποχής. Μάλιστα, στην προκείμενη περίπτωση, οπτικά η μία ευθεία μοιάζει να πέφτει ακριβώς επάνω στην άλλη, αν και από τις εξισώσεις τους φαίνεται πως δεν είναι ταυτόσημες.

7 Bayes meets k-NN

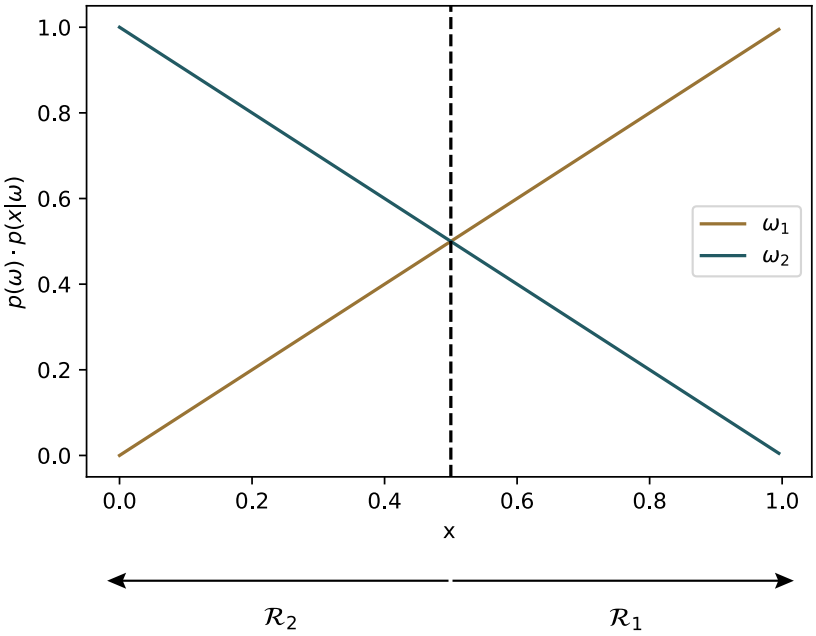
7.1 Δεδομένων των a-priori πιθανοτήτων για τις δύο κατηγορίες και τις συναρτήσεις κατανομών τους, το σημείο απόφασης κατά Bayes, x_0^*, θα προκύψει ως η λύση της εξίσωσης

$$p(\omega_1)p(x|\omega_1) = p(\omega_2)p(x|\omega_2). \tag{7.1}$$

Συγκεκριμένα,

$$\frac{1}{2} \cdot 2x_0^* = \frac{1}{2} \cdot \left(2 - 2x_0^*\right) \Leftrightarrow x_0^* = \frac{1}{2}, \tag{7.2}$$

το οποίο σημαίνει πως τα x με $0 \leq x < 0.5$ ταξινομούνται στην κατηγορία ω_2, ενώ τα x με $0.5 < x \leq 1$ ταξινομούνται στην κατηγορία ω_1, όπως εξάλλου φαίνεται και από το γράφημα της Εικόνας 7.1.

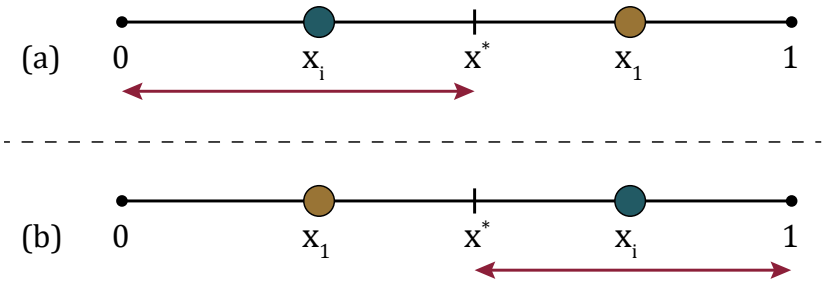


Εικόνα 7.1: Διάγραμμα $p(\omega_i)p(x|\omega_i)$ vs x για $i = 1, 2$ και απεικόνιση των περιοχών ταξινόμησης $\mathcal{R}_1, \mathcal{R}_2$ βάσει του σημείου απόφασης x_0^*.

Σε ό,τι αφορά το σφάλμα ταξινόμησης Bayes, αυτό μπορεί να υπολογιστεί ως το σχετικό εμβαδόν της περιοχής επικάλυψης των δύο καμπυλών της Εικόνας 7.1, η οποία δεν είναι παρά ένα τρίγωνο με βάση 1 και ύψος 0.5. Ακολουθώντας την αναλυτική προσέγγιση, γράφει κανείς:

$$\begin{aligned} p\left(\text{error}|x\right) &= p\left(\omega_1\right) \int_{\mathcal{R}_2} dx\, p\left(x|\omega_1\right) + p\left(\omega_2\right) \int_{\mathcal{R}_1} dx\, p\left(x|\omega_2\right) = \\ &= \frac{1}{2} \int_0^{0.5} dx\, 2x + \frac{1}{2} \int_{0.5}^{1} dx\, \left(2 - 2x\right) = \frac{x^2}{2}\Big|_0^{0.5} + \frac{2x - x^2}{2}\Big|_{0.5}^{1} = \\ &= 0.25. \end{aligned} \tag{7.3}$$

7.2 Η νέα ταξινόμηση γίνεται μέσω επιλογής ίσου πλήθους σημείων εκπαίδευσης από κάθε κατανομή, έστω n, κατασκευάζοντας έτσι έναν ταξινομητή τύπου 1-NN ο οποίος έχει εκπαιδευτεί σε $2n$ συνολικά δεδομένα. Στην περίπτωση $n = 1$, όπου ένα σημείο εκπαίδευσης επιλέγεται για κάθε κατηγορία, οι πιθανές καταστάσεις απεικονίζονται στα σχήματα της Εικόνας 7.2.



Εικόνα 7.2: Εκπαίδευση του ταξινομητή 1-NN χρησιμοποιώντας ένα σημείο από κάθε κατηγορία. Το σφάλμα του ταξινομητή ισοδυναμεί με το να βρεθεί το σημείο προς ταξινόμηση σε κάποια από τις κόκκινες περιοχές.

Η ακολουθούμενη σύμβαση είναι για δείγματα που προήλθαν από την κατανομή για την κατηγορία ω_1 να χρησιμοποιούνται αριθμοί ως δείκτες, ενώ για δείγματα που προήλθαν από την κατανομή για την κατηγορία ω_2 να χρησιμοποιούνται λατινικοί χαρακτήρες. Έτσι, το x_1 της Εικόνας 7.2 αντιστοιχεί σε δείγμα της ω_1, ενώ το x_i αντιστοιχεί σε δείγμα της ω_2. Υποθέτοντας πως το σημείο προς ταξινόμηση, x, ανήκει στην κατηγορία ω_1, το αναμενόμενο σφάλμα του ταξινομητή θα ισοδυναμεί με την πιθανότητα το σημείο να ταξινομηθεί (λανθασμένα) στην κατηγορία ω_2. Αυτό θα συμβεί εάν η απόστασή του από το x_i είναι μικρότερη από την απόστασή του από το x_1, εφόσον δηλαδή ισχύει

$$\|x_i - x\|_2 < \|x_1 - x\|_2. \tag{7.4}$$

Οι περιοχές στις οποίες εάν βρεθεί το x θα ταξινομηθεί λανθασμένα απεικονίζονται στην Εικόνα 7.2, με το σημείο απόφασης, x^*, να είναι απλώς η θέση στο μέσο των x_1 και x_i μετρώντας από το 0, δηλαδή

$$x^* = \frac{x_i + x_1}{2}. \tag{7.5}$$

Στην περίπτωση αυτή, το σφάλμα $P_1(e)$ θα ισούται με

$$P_1(e) = I_a + I_b, \tag{7.6}$$

όπου

$$I_a = \int_0^1 dx_1\, p\,(x_1|\omega_1) \int_0^{x_1} dx_i\, p\,(x_i|\omega_2) \int_0^{\frac{x_i+x_1}{2}} dx\, p(x|\omega_1), \tag{7.7}$$

$$I_b = \int_0^1 dx_i\, p\,(x_i|\omega_2) \int_0^{x_i} dx_1\, p\,(x_1|\omega_1) \int_{\frac{x_i+x_1}{2}}^1 dx\, p(x|\omega_1), \tag{7.8}$$

με τους δείκτες να αντιστοιχούν στις περιπτώσεις όπως απεικονίζονται στην Εικόνα 7.2. Σε ό,τι αφορά τον υπολογισμό των ολοκληρωμάτων, ισχύει

$$\begin{aligned}
I_a &= 4\int_0^1 dx_1\, x_1 \int_0^{x_1} dx_i\ (1-x_i) \int_0^{\frac{x_i+x_1}{2}} dx\, 2x = 4\int_0^1 dx_1\, x_1 \int_0^{x_1} dx_i\ (1-x_i)\left[x^2\right]_0^{\frac{x_i+x_1}{2}} \\
&= \int_0^1 dx_1\, x_1 \int_0^{x_1} dx_i\ (1-x_i)\left({x_i}^2 + {x_1}^2 + 2x_1x_i\right) \\
&= \int_0^1 dx_1\, x_1 \int_0^{x_1} dx_i\ \left[{x_1}^2 + x_i\left(2x_1 - {x_1}^2\right) + {x_i}^2(1-2x_1) - {x_i}^3\right] \\
&= \int_0^1 dx_1\, x_1 \left[{x_1}^3 + \frac{{x_1}^2}{2}\left(2x_1 - {x_1}^2\right) + \frac{{x_1}^3}{3}(1-2x_1) - \frac{{x_1}^4}{4}\right] \\
&= \int_0^1 dx_1\ \left[{x_1}^4 + {x_1}^4 - \frac{{x_1}^5}{2} + \frac{{x_1}^4}{3} - \frac{2{x_1}^5}{3} - \frac{{x_1}^5}{4}\right] = \int_0^1 dx_1\ \left[\frac{7{x_1}^4}{3} - \frac{17{x_1}^5}{12}\right] \\
&= \frac{83}{360} \simeq 0.2306
\end{aligned} \tag{7.9}$$

και

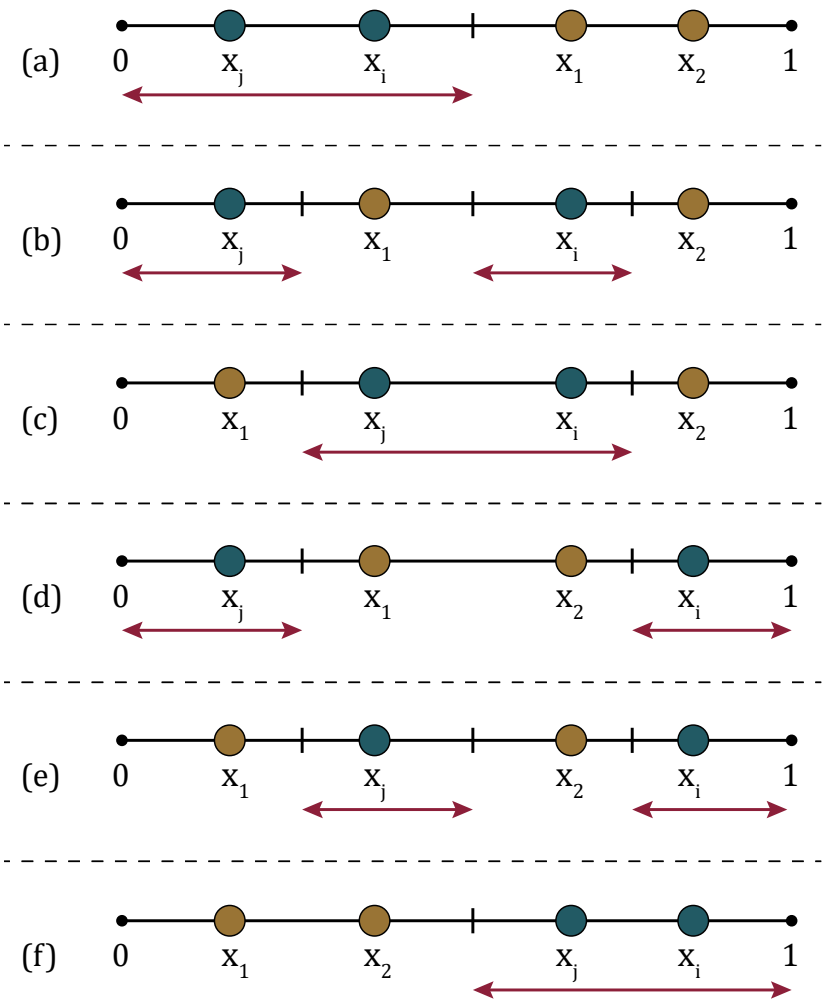
$$\begin{aligned}
I_b &= 4\int_0^1 dx_i\ (1-x_i) \int_0^{x_i} dx_1\, x_1 \int_{\frac{x_i+x_1}{2}}^1 dx\, 2x = 4\int_0^1 dx_i\ (1-x_i) \int_0^{x_i} dx_1\, x_1 \left[x^2\right]_{\frac{x_i+x_1}{2}}^1 \\
&= \int_0^1 dx_i\ (1-x_i) \int_0^{x_i} dx_1\, x_1 \left(4 - {x_i}^2 - {x_1}^2 - 2x_1x_i\right) \\
&= \int_0^1 dx_i\ (1-x_i) \int_0^{x_i} dx_1\ \left[x_1\left(4 - {x_i}^2\right) - 2x_i{x_1}^2 - {x_1}^3\right] \\
&= \int_0^1 dx_i\ (1-x_i) \left[2{x_i}^2 - \frac{{x_i}^4}{2} - \frac{2{x_i}^4}{3} - \frac{{x_i}^4}{4}\right] \\
&= \int_0^1 dx_i\ \left[2{x_i}^2 - 2{x_i}^3 - \frac{17{x_i}^4}{12} + \frac{17{x_i}^5}{12}\right] = \left[\frac{2{x_i}^3}{3} - \frac{1{x_i}^4}{2} - \frac{17{x_i}^5}{60} + \frac{17{x_i}^6}{72}\right]_0^1 \\
&= \frac{43}{360} \simeq 0.1194.
\end{aligned} \tag{7.10}$$

Έτσι, προκύπτει

$$P_1(e) = I_a + I_b = \frac{126}{360} = 0.35. \tag{7.11}$$

7.3 Κατ' αντίστοιχο τρόπο εργάζεται κανείς για την περίπτωση $n = 2$. Σε αυτήν, οι πιθανοί συνδυασμοί που πρέπει να ληφθούν υπ' όψιν απεικονίζονται στην Εικόνα 7.3, όπου ξανά σε κάθε περίπτωση τα κόκκινα διαστήματα αντιστοιχούν στις τιμές του x για τις οποίες η ταξινόμηση δε θα γίνει επιτυχώς. Οι περιοχές απόφασης ορίζονται ξανά ως τα μέσα μεταξύ δύο δειγμάτων εκπαίδευσης διαφορετικού τύπου, μόνο που σε ορισμένες περιπτώσεις (π.χ. περίπτωση (b)) είναι περισσότερες από μια, επομένως υπάρχουν δύο σχετικά ολοκληρώματα ως προς τη μεταβλητή x. Το αντίστοιχο αναμενόμενο σφάλμα ταξινόμησης είναι εδώ ίσο με

$$P_2(e) = q\left(I_a + I_b + I_c + I_d + I_e + I_f\right), \tag{7.12}$$



Εικόνα 7.3: Εκπαίδευση του ταξινομητή 1-NN χρησιμοποιώντας δύο σημεία από κάθε κατηγορία (χρυσό για την ω_1 και μπλε για την ω_2). Το σφάλμα του ταξινομητή ισοδυναμεί με το να βρεθεί το σημείο προς ταξινόμηση σε κάποια από τις κόκκινες περιοχές.

όπου ο παράγοντας q είναι μια σταθερά που αντιστοιχεί σε ένα βάρος. Συγκεκριμένα, σε αντίθεση με την περίπτωση $n = 1$, η δεικτοδότηση που πραγματοποιείται εδώ έχει σημασία: έχει επιλεγεί σε κάθε περίπτωση $x_2 > x_1$ σε ό,τι αφορά τα δείγματα της κατηγορίας ω_1 και αντίστοιχα $x_i > x_j$ σε ό,τι αφορά τα δείγματα της κατηγορίας ω_2. Φυσικά, με την εναλλαγή των x_1 και x_2, το αποτέλεσμα θα ήταν το ίδιο, επομένως η συνεισφορά του αντίστοιχου ολοκληρώματος στο συνολικό σφάλμα πρέπει να ληφθεί υπ' όψιν πολλαπλασιάζοντας με έναν παράγοντα 2. Το ίδιο ακριβώς ισχύει και για την εναλλαγή των x_i και x_j, συνεπώς για τον παράγοντα q ισχύει $q = 4$. Αντίστοιχα, για την περίπτωση $n = 3$, ο παράγοντας αυτός θα ήταν $q = 36$, ενώ για τη γενική περίπτωση ισχύει

$$q = (n!)^2 . \tag{7.13}$$

Τα αντίστοιχα ολοκληρώματα υπολογίζονται διαδοχικά όπως και στην περίπτωση $n = 2$ και ορίζονται ως

$$I_a = \int_0^1 dx_2\, p\,(x_2|\omega_1) \int_0^{x_2} dx_1\, p\,(x_1|\omega_1) \int_0^{x_1} dx_i\, p\,(x_i|\omega_2) \int_0^{x_i} dx_j\, p\,(x_j|\omega_2) \cdot$$
$$\cdot \int_0^{\frac{x_i+x_1}{2}} dx\, p\,(x|\omega_1) \simeq 0.0401, \tag{7.14}$$

$$I_b = \int_0^1 dx_2\, p\,(x_2|\omega_1) \int_0^{x_2} dx_i\, p\,(x_i|\omega_2) \int_0^{x_i} dx_1\, p\,(x_1|\omega_1) \int_0^{x_1} dx_j\, p\,(x_j|\omega_2) \cdot$$
$$\cdot \left(\int_0^{\frac{x_1+x_j}{2}} dx\, p\,(x|\omega_1) + \int_{\frac{x_1+x_i}{2}}^{\frac{x_i+x_2}{2}} dx\, p\,(x|\omega_1) \right) \simeq 0.0061 + 0.0122 = 0.0183, \tag{7.15}$$

$$I_c = \int_0^1 dx_2\, p\,(x_2|\omega_1) \int_0^{x_2} dx_i\, p\,(x_i|\omega_2) \int_0^{x_i} dx_j\, p\,(x_j|\omega_2) \int_0^{x_j} dx_1\, p\,(x_1|\omega_1) \cdot$$
$$\cdot \int_{\frac{x_i+x_1}{2}}^{\frac{x_i+x_2}{2}} dx\, p\,(x|\omega_1) \simeq 0.0068, \tag{7.16}$$

$$I_d = \int_0^1 dx_i\, p\,(x_i|\omega_2) \int_0^{x_i} dx_2\, p\,(x_2|\omega_1) \int_0^{x_2} dx_1\, p\,(x_1|\omega_1) \int_0^{x_1} dx_j\, p\,(x_j|\omega_2) \cdot$$
$$\cdot \left(\int_0^{\frac{x_1+x_j}{2}} dx\, p\,(x|\omega_1) + \int_{\frac{x_2+x_i}{2}}^{1} dx\, p\,(x|\omega_1) \right) \simeq 0.0024 + 0.0101 = 0.0125, \tag{7.17}$$

$$I_e = \int_0^1 dx_i\, p\,(x_i|\omega_2) \int_0^{x_i} dx_2\, p\,(x_2|\omega_1) \int_0^{x_2} dx_j\, p\,(x_j|\omega_2) \int_0^{x_j} dx_1\, p\,(x_1|\omega_1) \cdot$$
$$\cdot \left(\int_{\frac{x_1+x_j}{2}}^{\frac{x_2+x_j}{2}} dx\, p\,(x|\omega_1) + \int_{\frac{x_2+x_i}{2}}^{1} dx\, p\,(x|\omega_1) \right) \simeq 0.0009 + 0.0037 = 0.0046, \tag{7.18}$$

$$I_f = \int_0^1 dx_i\, p\,(x_i|\omega_2) \int_0^{x_i} dx_j\, p\,(x_j|\omega_2) \int_0^{x_j} dx_2\, p\,(x_2|\omega_1) \int_0^{x_2} dx_1\, p\,(x_1|\omega_1) \cdot$$
$$\cdot \int_{\frac{x_2+x_j}{2}}^{1} dx\, p\,(x|\omega_1) \simeq 0.0026. \tag{7.19}$$

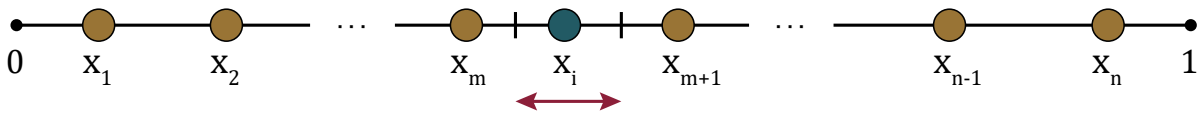
Συνολικά, προκύπτει

$$P_2(e) = 4\,(0.0401 + 0.0183 + 0.0068 + 0.0125 + 0.0046 + 0.0026) = 0.3396. \tag{7.20}$$

7.4 Η γενίκευση των παραπάνω στην περίπτωση τυχαίου n δεν είναι τετριμμένη και ενδέχεται να μη μπορεί να πραγματοποιηθεί με την άνω προσέγγιση. Στα πλαίσια μιας εργασίας ενός συνόλου ασκήσεων με δεδομένη προθεσμία η διερεύνηση του ζητήματος δε μπορούσε να ολοκληρωθεί χρονικά, παρ' όλα αυτά παρακάτω παρουσιάζεται μια πιθανή κατεύθυνση. Στην περίπτωση των n δειγμάτων από κάθε κατηγορία, ενδέχεται το πρόβλημα (n δείγματα ω_1, n δείγματα ω_2) να μπορεί να απλοποιηθεί ως (n δείγματα ω_1, 1 δείγμα ω_2) σε ό,τι αφορά τα δεδομένα εκπαίδευσης. Στην περίπτωση αυτή, μπορεί να οριστεί το «τυχαίο διάγραμμα» ως

$$I(m) = 2^{n+1} \int_0^1 dx_n \int_0^{x_n} dx_{n-1} \dots \int_0^{x_{m+2}} dx_{m+1} \int_0^{x_{m+1}} dx_i \int_0^{x_i} dx_m \dots \int_0^{x_2} dx_1 \cdot$$
$$\cdot (1 - x_i) \left(\prod_{k=1}^{n} x_k \right) \int_{\frac{x_m + x_i}{2}}^{\frac{x_{m+1} + x_i}{2}} dx\, 2x, \tag{7.21}$$

το οποίο απεικονίζεται στην Εικόνα 7.4. Η συνεισφορά του στο συνολικό αναμενόμενο σφάλμα θα έρχεται με βάρος $q = [(n+1)!]^2$ επί κάποιο επιπλέον απαραίτητο βάρος $w(m)$ ώστε να μπορεί να γίνει η αναγωγή 1 δείγμα κατηγορίας $\omega_2 \to n$ δείγματα κατηγορίας ω_2. Στην περίπτωση αυτή, το διάγραμμα μπορεί να διαβαστεί ως «το ενδεχόμενο το δείγμα της κατηγορίας ω_1 να βρίσκεται εκατέρωθεν των δειγμάτων x_m και x_{m+1}, όπου $1 \leq m \leq n$».



Εικόνα 7.4: Διάγραμμα που αντιστοιχεί στην περίπτωση το υπό μελέτη δείγμα της κατηγορίας ω_2, έστω x_i, να βρίσκεται εκατέρωθεν των x_m και x_{m+1}.

Τότε, η περιοχή απόφασης θα οριοθετείται από τις τιμές 0.5 $(x_m + x_i)$ και 0.5 $(x_i + x_{m+1})$. Το συνολικό αναμενόμενο σφάλμα θα δίνεται ως το άθροισμα σε όλα αυτά τα διαγράμματα, επί τα αντίστοιχα βάρη, κατ' αναλογία με τα διαγράμματα Feynman στη Φυσική Στοιχειωδών Σωματιδίων:

$$P_n(e) = q \sum_{m=1}^{n} w(m) \cdot I(m). \tag{7.22}$$

7.5 Προκειμένου να υπάρχει αποτέλεσμα στην περίπτωση $n \to \infty$, υπάρχουν δύο πιθανές διαδρομές. Η μία είναι η προσομοίωση Monte Carlo του συγκεκριμένου προβλήματος, ο κώδικας για την οποία δίνεται στην επόμενη σελίδα σε γλώσσα Python. Η λογική είναι η ακόλουθη: για καθεμιά από τις δύο κατηγορίες επιλέγονται n δείγματα από κάθε κατανομή. Κάθε επιλεγμένο δείγμα αποτελεί μια τυχαία μεταβλητή που ανήκει στο διάστημα [0,1], επομένως η προσομοίωση γίνεται βάσει της αντιστροφής της αθροιστικής συνάρτησης πιθανότητας. Συγκεκριμένα, για την κατανομή $p\,(x|\omega_1)$ ισχύει

$$F_1\,(x) = \int_0^x d\xi\; 2\xi = x^2, \tag{7.23}$$

ενώ για την $p\,(x|\omega_2)$ ισχύει

$$F_2\,(x) = \int_0^x d\xi\; (2 - 2\xi) = 2x - x^2. \tag{7.24}$$

Θέτοντας $y = F_i\,(x)$, με $i = 1, 2$, και απαιτώντας $y \in [0, 1]$, προκύπτει πως τα δείγματα από τις κατηγορίες ω_1 και ω_2 αντιστοιχούν στις

$$y_{\omega_1}\,(x) = \sqrt{x} \quad \text{και} \quad y_{\omega_2}\,(x) = 1 - \sqrt{1 - x}, \tag{7.25}$$

αντίστοιχα.

```
import random
import numpy as np

def omega_1(r):
    return np.sqrt(r)

def omega_2(r):
    return 1 - np.sqrt(1-r)

# Number of runs
N_tests = 10000000
correct = 0
N = 1000

for j in range(N_tests):

    samples = []
    labels = []

    for i in range(N):
        sample1 = omega_1(random.random())
        samples.append(sample1)
        labels.append(1)
        sample2 = omega_2(random.random())
        samples.append(sample2)
        labels.append(2)

    new_x = omega_1(random.random())
    npsamples = np.array(samples)

    if labels[np.argmin(abs(npsamples-new_x))] == 1:
        correct += 1

print("The average error = "+str(1-(correct/N_tests)))
```

Κάθε δείγμα κάθε κατανομής δεικτοδοτείται βάσει της κατανομής από την οποία προήλθε. Στη συνέχεια, ένα νέο δείγμα, x, επιλέγεται από την κατανομή ω_1. Δεδομένου του πιο κοντινού γείτονα του δείγματος αυτού, πραγματοποιείται η ταξινόμησή του βάσει του 1-NN και εάν η ταξινόμηση είναι επιτυχής, ένας δείκτης αυξάνεται κατά 1 μονάδα. Η διαδικασία αυτή επαναλαμβάνεται 10^7 φορές, προκειμένου το αποτέλεσμα να είναι στατιστικά έμπιστο, ως ίσο με (1 - επιτυχείς ταξινομήσεις/σύνολο επαναλήψεων). Πράγματι, για $n = 1$ η προσομοίωση δίνει $P_1(e) = 0.350074$, ενώ για $n = 2$ δίνει $P_2(e) = 0.339592$, τα οποία είναι πρακτικά ίσα με τις τιμές που υπολογίστηκαν αναλυτικά. Για την εύρεση του ορίου $n \to \infty$ επιλέγεται ο αριθμός $n = 1000$, οπότε το αποτέλεσμα είναι

$$\lim_{n\to\infty} [P_n(e)] = \frac{1}{3}. \tag{7.26}$$

Γενικά, ο αριθμός 1000 σε καμία περίπτωση δε θεωρείται τόσο μεγάλος ώστε να προσεγγίζει

το $+\infty$, όμως (όπως θα εξηγηθεί και παρακάτω), η $P_n(e)$ φθίνει ως $\sim 1/n^2$, επομένως το αποτέλεσμα που προκύπτει έχει ακρίβεια υψηλότερη από αυτήν με την οποία έχουν δοθεί όλα τα αποτελέσματα ως τώρα.

Σε ό,τι αφορά τη δεύτερη διαδρομή προς αυτό το αποτέλεσμα, αυτή βασίζεται στην ανάλυση μιας δημοσίευσης των T. M. Cover και P. E. Hart [2], στην οποία παρουσιάζεται το πρόβλημα αυτό, αλλά με μια πολύ διαφορετική αρχική υπόθεση. Στη δημοσίευση αυτή, γίνεται η υπόθεση πως η δειγματοληψία γίνεται τυχαία από κάθε κατανομή (με πιθανότητα 50%, μιας και οι a priori τους είναι ίσες και οι κατανομές εμφανίζουν τη συμμετρικότητα που απεικονίζεται στην Εικόνα 7.1), ούτως ώστε ο ταξινομητής να εκπαιδευτεί με n συνολικά δεδομένα εκπαίδευσης, τα οποία δεν είναι απαραίτητο να έχουν αναλογία 1:1 ως προς τις κατηγορίες. Με άλλα λόγια, αν και η πιθανότητα είναι εξαιρετικά μικρή, είναι πιθανό ο ταξινομητής να εκπαιδευτεί με n δεδομένα της μιας κατηγορίας και 0 δεδομένα της άλλης. Στην περίπτωση όπου τα δεδομένα εκπαίδευσης είναι λίγα, τα αποτελέσματα για το αναμενόμενο σφάλμα αναμένεται να διαφέρουν σημαντικά από τα αντίστοιχα της περίπτωσης αυτής της άσκησης. Όμως για μεγάλες τιμές του n αναμένεται οι δύο περιπτώσεις να είναι ισοδύναμες, μιας και ο ταξινομιτής θα εκπαιδευτεί με πρακτικά ισάριθμα δείγματα από κάθε κατηγορία. Σκοπός της ακόλουθης ανάλυσης είναι, λοιπόν, ο αναλυτικός υπολογισμός του $P_n(e)$ για την περίπτωση της τυχαίας δειγματοληψίας, και στη συνέχεια ο αναλυτικός υπολογισμός του ορίου $n \to \infty$, το αποτέλεσμα του οποίου οφείλει να ταυτίζεται με το αποτέλεσμα της παρούσας άσκησης.

Ο 1-NN ταξινομητής εκπαιδεύεται στην περίπτωση αυτή με $2n$ δείγματα $x_1, x_2, \dots, x_{2n}$, καθένα εκ των οποίων έχει ίση πιθανότητα να προέρχεται από την ω_1 ή την ω_2. Θεωρώντας και πάλι πως το προς ταξινόμηση δείγμα, x, ανήκει στην ω_1, το αναμενόμενο σφάλμα θα δίνεται από τη σχέση

$$P_n(e) = \sum_{i=1}^{2n} p\left(x \in \omega_1 \text{ και } x_i \in \omega_2 \text{ με } \|x_i - x\|_2 < \|x_j - x\|_2,\ \forall j \neq i\right), \tag{7.27}$$

όπου η άθροιση γίνεται επάνω σε όλα τα δείγματα που μπορεί να ικανοποιούν την προϋπόθεση αυτή, δηλαδή $2n$. Μάλιστα, λόγω της φύσης της δειγματοληψίας και της συμμετρικότητας των κατανομών, καθένας όρος του αθροίσματος θα είναι ίσος με τους υπόλοιπους, οπότε η Σχέση (7.27) μπορεί να απλοποιηθεί ως

$$P_n(e) = 2n \cdot p\left(x \in \omega_1 \text{ και } x_i \in \omega_2 \text{ με } \|x_i - x\|_2 < \|x_j - x\|_2,\ \forall j \neq i\right), \tag{7.28}$$

όπου έχει επιλεγεί ένα συγκεκριμένο x_i το οποίο απαιτείται να ανήκει στην κατηγορία ω_2, ώστε να είναι αυτό που θα καθορίσει τη λανθασμένη ταξινόμηση του x. Προκειμένου, φυσικά, να συμβεί αυτό, θα πρέπει το x_i να βρίσκεται πιο κοντά στο x από οποιοδήποτε εκ των $2n-1$ υπόλοιπων x_j $(i \neq j)$. Δεδομένων των παρατηρήσεων αυτών, η Σχέση (7.28) γράφεται ισοδύναμα

$$P_n(e) = 2n \cdot \int_0^1 dx\, p\left(x|\omega_1\right) \int_0^1 dx_i\, p\left(\omega_2|x_i\right) \prod_{j \neq i} p\left(\|x_i - x\|_2 < \|x_j - x\|_2\right) \tag{7.29}$$

όπου το γινόμενο εμφανίζεται δεδομένου πως το ενδεχόμενο το x_i να βρίσκεται πιο κοντά στο x απ' ότι το x_a είναι ανεξάρτητο από το ενδεχόμενο το x_i να βρίσκεται πιο κοντά στο x απ' ότι το x_b, εφόσον $a \neq b$. Μάλιστα, οι πιθανότητες αυτές είναι ισοδύναμες, επομένως το αναμενόμενο σφάλμα απλοποιείται περαιτέρω ως

$$P_n(e) = 2n \cdot \int_0^1 dx\, p\,(x|\omega_1) \int_0^1 dx_i\, p\,(\omega_2|x_i) \left[p\left(\|x_i - x\|_2 < \|x_j - x\|_2\right)\right]^{2n-1}. \qquad (7.30)$$

Στη φάση αυτή, το πρόβλημα των $2n$ δειγμάτων εκπαίδευσης έχει αναχθεί σε ένα πρόβλημα 2 μόνο δειγμάτων εκπαίδευσης, επομένως αρκεί απλώς ο υπολογισμός της πιθανότητας το x_i να βρίσκεται πιο κοντά στο x απ' ότι το x_j. Πρέπει, επομένως, να ληφθούν υπ' όψιν όλα τα πιθανά ενδεχόμενα για τη σχετική θέση των x, x_i, x_j επάνω στο ευθύγραμμο τμήμα που εκτείνεται από το 0 έως το 1 [2]. Τα ενδεχόμενα αυτά ανήκουν σε 2 μεγάλες οικογένειες, με την πρώτη να αντιστοιχεί στην περίπτωση $0 \leq x < 0.5$ και τη δεύτερη στην περίπτωση $0.5 < x \leq 1$. Σε καθεμιά εξ αυτών και για δεδομένη τιμή του x, υπάρχουν συνολικά τρεις οικογένειες περιπτώσεων για τη σχετική θέση του x_i, δεδομένων των οποίων καθορίζονται μονοσήμαντα τα διαστήματα στα οποία μπορεί να ανήκει το x_j, προκειμένου να ισχύει $\|x_i - x\|_2 < \|x_j - x\|_2$. Συγκεκριμένα:

- Εάν $0 < x < 0.5$ και $x_i < x$, τότε $\|x_i - x\|_2 = x - x_i$, οπότε:

$$\begin{gathered}\|x_i - x\|_2 < \|x_j - x\|_2 \;\Leftrightarrow\; \|x_j - x\|_2 > x - x_i \;\Leftrightarrow \\ \Leftrightarrow\; x_j - x > x - x_i \quad \text{ή} \quad x_j - x < x_i - x \;\Leftrightarrow\; x_j > 2x - x_i \quad \text{ή} \quad x_j < x_i.\end{gathered} \qquad (7.31)$$

- Εάν $0 < x < 0.5$ και $x_i > x$, τότε $\|x_i - x\|_2 = x_i - x$, οπότε:

$$\begin{gathered}\|x_i - x\|_2 < \|x_j - x\|_2 \;\Leftrightarrow\; \|x_j - x\|_2 > x_i - x \;\Leftrightarrow \\ \Leftrightarrow\; x_j - x > x_i - x \quad \text{ή} \quad x_j - x < x - x_i \;\Leftrightarrow\; x_j > x_i \quad \text{ή} \quad x_j < 2x - x_i,\end{gathered} \qquad (7.32)$$

  όπου η δεύτερη λύση, $x_j < 2x - x_i$, ορίζεται μόνο για $x_i < 2x$, καθώς πρέπει πάντα να ισχύει $x_j \geq 0$.

- Εάν $0.5 < x < 1$ και $x_i < x$, τότε $\|x_i - x\|_2 = x - x_i$, οπότε:

$$\begin{gathered}\|x_i - x\|_2 < \|x_j - x\|_2 \;\Leftrightarrow\; \|x_j - x\|_2 > x - x_i \;\Leftrightarrow \\ \Leftrightarrow\; x_j - x > x - x_i \quad \text{ή} \quad x_j - x < x_i - x \;\Leftrightarrow\; x_j > 2x - x_i \quad \text{ή} \quad x_j < x_i,\end{gathered} \qquad (7.33)$$

  όπου η πρώτη λύση, $x_j > 2x - x_i$, ορίζεται μόνο για $x_i < 2x - 1$, καθώς πρέπει πάντα να ισχύει $x_j \leq 1$.

- Εάν $0 < x < 0.5$ και $x_i > x$, τότε $\|x_i - x\|_2 = x_i - x$, οπότε:

$$\begin{gathered}\|x_i - x\|_2 < \|x_j - x\|_2 \;\Leftrightarrow\; \|x_j - x\|_2 > x_i - x \;\Leftrightarrow \\ \Leftrightarrow\; x_j - x > x_i - x \quad \text{ή} \quad x_j - x < x - x_i \;\Leftrightarrow\; x_j > x_i \quad \text{ή} \quad x_j < 2x - x_i.\end{gathered} \qquad (7.34)$$

Οι 6 συνολικά οικογένειες περιπτώσεων απεικονίζονται στον ακόλουθο πίνακα (Πίνακας 7.1). Σε καθένα από τα διαστήματα αυτά, η πιθανότητα $p\left(\|x_i - x\|_2 < \|x_j - x\|_2\right)$ υπολογίζεται απλώς ως ο λόγος του μήκους των διαστημάτων στα οποία μπορεί να ανήκει το x_j προκειμένου να ισχύει $\|x_i - x\|_2 < \|x_j - x\|_2$, προς το συνολικό μήκος του ευθύγραμμου τμήματος στο οποίο κείνται τα x, x_i, x_j, το οποίο ισούται με τη μονάδα.

[2] Ο λόγος για τον οποίο η προσέγγιση που ακολουθήθηκε στα πρώτα ερωτήματα της άσκησης δε μπορεί να εφαρμοστεί εδώ, είναι πως αν τα άκρα ολοκλήρωσης ως προς x και x_i δεν είναι πάντοτε ίδια, τότε η πιθανότητα $p\left(\|x_i - x\|_2 < \|x_j - x\|_2\right)$ δε μπορεί να γραφεί σε μια κλειστή, σταθερή μορφή.

x	x_i	x_j
$(0, 0.5)$	$(0, x)$	$(0, x_i) \cup (2x - x_i, 1)$
$(0, 0.5)$	$(x, 2x)$	$(0, 2x - x_i) \cup (x_i, 1)$
$(0, 0.5)$	$(2x, 1)$	$(x_i, 1)$
$(0.5, 1)$	$(0, 2x - 1)$	$(0, x_i)$
$(0.5, 1)$	$(2x - 1, x)$	$(0, x_i) \cup (2x - x_i, 1)$
$(0.5, 1)$	$(x, 1)$	$(0, 2x - x_i) \cup (x_i, 1)$

Πίνακας 7.1: Επιτρεπτές τιμές του x_j για κάθε πιθανή σχετική θέση των x και x_i.

Δεδομένων αυτών, το αναμενόμενο σφάλμα έχει τη μορφή

$$P_n(e) = 2n\,(I_1 + I_2 + I_3 + I_4 + I_5 + I_6)\,, \tag{7.35}$$

όπου τα I_1, I_2, I_5, I_6 υπολογίζονται αλλάζοντας τη μεταβλητή x_i σε $\xi = 1 \pm 2x_i \mp 2x$ και στη συνέχεια τη x σε $\phi = 1 \pm 2x$. Για παράδειγμα,

$$\begin{aligned}
I_1 &= \int_0^{0.5} dx\, 2x \int_0^x dx_i\ (1 - x_i)\,(1 - 2x + 2x_i)^{2n-1} = -\frac{1}{2}\int_0^{0.5} dx\, x \int_{1-2x}^1 d\xi\ (\xi + 2x - 3)\,\xi^{2n-1} \\
&= -\frac{1}{2}\int_0^{0.5} dx\, x \left[\frac{\xi^{2n+1}}{2n+1} + (2x-3)\,\frac{\xi^{2n}}{2n}\right]_{1-2x}^1 \\
&= -\frac{1}{2}\int_0^{0.5} dx\, x \left(\frac{1}{2n+1} + \frac{2x-3}{2n} - \frac{(1-2x)^{2n+1}}{2n+1} + \frac{(1-2x)^{2n+1}}{2n} + \frac{2\,(1-2x)^{2n}}{2n}\right) \\
&= -\frac{1}{16}\left(\frac{1}{2n+1} - \frac{3}{2n}\right) - \frac{1}{48n} - \frac{1}{2}\left(\frac{1}{2n} - \frac{1}{2n+1}\right)\int_0^{0.5} dx\, x\,(1-2x)^{2n+1} - \\
&- \frac{1}{2n}\int_0^{0.5} dx\, x\,(1-2x)^{2n} \\
&= -\frac{1}{16}\left(\frac{1}{2n+1} - \frac{3}{2n}\right) - \frac{1}{48n} - \frac{1}{8}\left(\frac{1}{2n} - \frac{1}{2n+1}\right)\int_0^{1} d\phi\ (1-\phi)\,\phi^{2n+1} - \\
&- \frac{1}{8n}\int_0^{1} d\phi\ (1-\phi)\,\phi^{2n} \\
&= \frac{1}{16}\left(\frac{3}{2n} - \frac{1}{2n+1}\right) - \frac{1}{48n} - \frac{1}{8}\left(\frac{1}{2n} - \frac{1}{2n+1}\right)\left(\frac{1}{2n+2} - \frac{1}{2n+3}\right) - \\
&- \frac{1}{8n}\left(\frac{1}{2n+1} - \frac{1}{2n+2}\right).
\end{aligned} \tag{7.36}$$

Κατ' αντίστοιχο τρόπο προκύπτουν:

$$I_2 = \frac{1}{16}\left(\frac{1}{2n+1}+\frac{1}{2n}\right)-\frac{1}{48n}-\frac{1}{8}\left(\frac{1}{2n}+\frac{1}{2n+1}\right)\left(\frac{1}{2n+2}-\frac{1}{2n+3}\right), \tag{7.37}$$

$$I_5 = \frac{3}{16}\left(\frac{3}{2n}-\frac{1}{2n+1}\right)-\frac{7}{48n}+\frac{1}{8}\left(\frac{1}{2n}+\frac{1}{2n+1}\right)\left(\frac{1}{2n+2}+\frac{1}{2n+3}\right)- \\ -\frac{1}{8n}\left(\frac{1}{2n+1}+\frac{1}{2n+2}\right) \tag{7.38}$$

$$I_6 = \frac{3}{16}\left(\frac{1}{2n}+\frac{1}{2n+1}\right)-\frac{7}{48n}+\frac{1}{8}\left(\frac{1}{2n}-\frac{1}{2n+1}\right)\left(\frac{1}{2n+2}+\frac{1}{2n+3}\right). \tag{7.39}$$

Ο υπολογισμός των I_3 και I_4 πραγματοποιείται ακόμη πιο άμεσα, δίνοντας

$$I_3 = \frac{1}{2\,(2n+1)}\left(\frac{1}{2n+2}-\frac{1}{2n+3}\right), \tag{7.40}$$

$$I_4 = \frac{1}{2}\left(\frac{1}{2n\,(2n+1)}+\frac{1}{2n\,(2n+2)}-\frac{1}{(2n+1)\,(2n+2)}-\frac{1}{(2n+1)\,(2n+3)}\right). \tag{7.41}$$

Έτσι, η Σχέση (7.35) παίρνει την τελική μορφή

$$\begin{aligned} P_n(e) &= 2n\left(\frac{1}{6n}+\frac{1}{4n\,(2n+2)}+\frac{1-2n}{4n\,(2n+3)\,(2n+1)}\right) \\ &= \frac{1}{3}+\frac{1}{4\,(n+1)}+\frac{1-2n}{2\,(2n+3)\,(2n+1)} \\ &= \frac{1}{3}+\frac{1}{(2n+1)\,(2n+3)}+\frac{1}{4\,(n+1)\,(2n+3)}. \end{aligned} \tag{7.42}$$

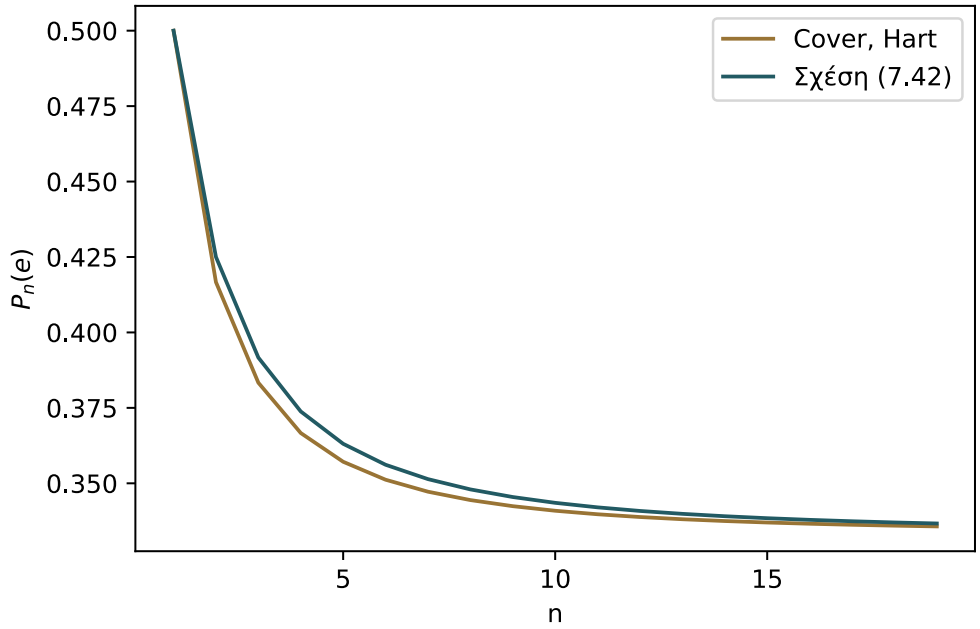
Όπως αναφέρθηκε και παραπάνω, το μη σταθερό μέρος της $P_n(e)$ φθίνει ως $\sim 1/n^2$, αιτιολογώντας έτσι το να θέσει κανείς την τιμή $n = 1000$ προκειμένου να πάρει μια ιδέα σχετικά με το τι συμβαίνει στο όριο $n \to \infty$. Δεδομένης της αναλυτικής Σχέσης (7.42), το όριο αυτό μπορεί να υπολογιστεί ακριβώς, επαληθεύοντας έτσι τη Σχέση (7.26) και την υπόθεση πως στο όριο αυτό τα δύο προβλήματα είναι ισοδύναμα.

Αξίζει στο σημείο αυτό να σημειωθεί πως το τελικό αυτό αποτέλεσμα διαφέρει από το αντίστοιχο των [2]. Καταρχάς, προκειμένου η σύγκριση να γίνει σε κοινό έδαφος, θα πρέπει στο αποτέλεσμά τους να γίνει η αντικατάσταση $n \to 2n$, καθώς εκείνο το αποτέλεσμα αφορά σύνολο n δειγμάτων, ενώ εδώ γίνεται αναφορά σε n δείγματα κάθε κατηγορίας, δηλαδή σύνολο $2n$ δειγμάτων. Έτσι, το ισοδύναμο δικό τους αποτέλεσμα είναι το

$$\tilde{P}_n(e) = \frac{1}{3}+\frac{1}{(2n+1)\,(2n+2)}. \tag{7.43}$$

Παρότι για υψηλές τιμές του n η απόκλιση είναι σημαντική, για μικρότερες τιμές η απόκλιση είναι μη αμελητέα, όπως φαίνεται και στην Εικόνα 7.5.

Προκειμένου να αξιολογηθεί το ποιο αποτέλεσμα είναι ορθότερο για να περιγράψει το πρόβλημα που έθεσαν οι [2] πραγματοποιήθηκε μια επιπλέον προσομοίωση Monte Carlo, ο κώδικας για την οποία φαίνεται παρακάτω.



Εικόνα 7.5: Σύγκριση της Σχέσης (7.42) με τη Σχέση (41) των Cover, Hart.

```
import random
import numpy as np

def omega_1(r):
    return np.sqrt(r)

def omega_2(r):
    return 1 - np.sqrt(1-r)

# Number of runs
N_tests = 10000000
correct = 0
N = 1

for j in range(N_tests):

    samples = []
    labels = []

    for i in range(2*N):
        # In this case, N generates TOTAL SAMPLES instead of pairs,
        # that's why the range is (2N).
        pointer = random.random()
        if pointer >= 0.5:
            sample = omega_1(random.random())
            labels.append(1)
        else:
            sample = omega_2(random.random())
            labels.append(2)
        samples.append(sample)

```

```
32      new_x = omega_1(random.random())
33      npsamples = np.array(samples)
34
35      if labels[np.argmin(abs(npsamples-new_x))] == 1:
36          correct += 1
37
38  print("The average error = "+str(1-(correct/N_tests)))
```

Η διαφορά με την προηγούμενη προσομοίωση είναι πως τα $2n$ δείγματα επιλέγονται κάθε φορά τυχαία από κάθε κατανομή, με πιθανότητα 50%, δεδομένης της συμμετρικότητάς τους, καθώς και του γεγονότος πως έχουν ίσες a priori πιθανότητες. Τα αποτελέσματα των προσομοιώσεων για μικρές τιμές του n και 10^7 επαναλήψεις φαίνονται στον ακόλουθο πίνακα (Πίνακας 7.2), μαζί με τις αντίστοιχες τιμές που προκύπτουν από τη Σχέση (7.42), καθώς και τη Σχέση (41) των [2] (δηλαδή τη Σχέση (7.43) του παρόντος κειμένου).

n	Monte Carlo	Σχέση (7.42)	Σχέση (41) Cover, Hart
1	0.4248	0.425	0.4167
2	0.3738	0.3738	0.3667
3	0.3561	0.3562	0.3512
4	0.3478	0.3479	0.3444
5	0.3433	0.3435	0.3409
6	0.3409	0.3408	0.3388
7	0.3392	0.3391	0.3375
8	0.3377	0.3379	0.3366
9	0.3372	0.3370	0.336
10	0.3363	0.3364	0.3355

Πίνακας 7.2: Υπολογισμός του $P_n(e)$ για διάφορες τιμές του n, χρησιμοποιώντας την προσομοίωση Monte Carlo, τη Σχέση (7.42), καθώς και τη Σχέση (41) των Cover, Hart. Τονίζεται πως το n έχει προσαρμοστεί ώστε να αντιστοιχεί σε σύνολο ζευγών δειγμάτων, και όχι σε σύνολο δειγμάτων (τα οποία ισούνται με $2n$).

Γίνεται εμφανές πως η προσομοίωση συμφωνεί σε πολύ μεγαλύτερο βαθμό με το αποτέλεσμα της Σχέσης (7.42), σε σχέση με αυτό των Cover, Hart, γεγονός το οποίο υποδεικνύει πως ενδεχομένως να παρέλειψαν κάτι στην ανάλυσή τους.

8 EM

Θα πρέπει αρχικά να τονιστεί πως η δοσμένη

$$p\left(x_1\right)=\frac{1}{\theta_1}\exp\left(-\theta_1 x_1\right),\ x_1\geq 0 \tag{8.1}$$

αντιστοιχεί σε συνάρτηση πυκνότητας πιθανότητας μόνο εάν $\theta_1=1$, διαφορετικά ισχύει

$$\int_0^{+\infty} dx\ p\left(x\right)\neq 1. \tag{8.2}$$

Προκειμένου η άσκηση να έχει το νόημα της εφαρμογής του αλγορίθμου ΕΜ για την εύρεση των βέλτιστων παραμέτρων θ_1, θ_2, θα πρέπει αυτές να είναι όντως παράμετροι. Έτσι, στα ακόλουθα θεωρείται πως στην άσκηση υπάρχει τυπογραφικό και στην πραγματικότητα η $p\left(x_1\right)$ αντιστοιχεί στη γνωστή εκθετική συνάρτηση κατανομής

$$p\left(x_1\right)=\theta_1\exp\left(-\theta_1 x_1\right),\ x_1\geq 0, \tag{8.3}$$

ή ισοδύναμα

$$p\left(x_1\right)=\frac{1}{\theta_1}\exp\left(-\frac{x_1}{\theta_1}\right),\ x_1\geq 0, \tag{8.4}$$

η οποία είναι κανονικοποιημένη στη μονάδα για κάθε τιμή της παραμέτρου $\theta_1>0$. Οφείλει κανείς να λάβει υπ' όψιν και τις δύο περιπτώσεις διορθώσεων καθώς, παρότι η μέθοδος θα πρέπει να συγκλίνει στις ίδιες κατανομές ανεξαρτήτως του πώς ορίζει κανείς τη θ_1, η συνάρτηση Q του πρώτου ερωτήματος θα έχει διαφορετική μορφή σε κάθε περίπτωση.

8.1 Τα δεδομένα χωρίζονται σε δύο σύνολα, αναλόγως με το τι πληροφορία λείπει από τα διανύσματα χαρακτηριστικών, ως εξής:

$$\mathcal{D}_g=\left\{\mathbf{x}^{(1)},\mathbf{x}^{(2)},x_1^{(3)}\right\}\quad \text{και}\quad \mathcal{D}_b=\left\{x_2^{(3)}\right\}. \tag{8.5}$$

Ορίζοντας το διάνυσμα $\boldsymbol{\theta}=\left[\theta_1,\theta_2\right]^\top$ και την αρχική εκτίμησή του $\boldsymbol{\theta}^0=\left[2,3\right]^\top$, για την εκτιμώμενη τιμή Q (βήμα Ε) ισχύει:

$$\begin{aligned}Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^0\right)&=\mathbb{E}_{\mathcal{D}_b}\left\{\ln p\left(\mathcal{D};\boldsymbol{\theta}\mid\mathcal{D}_g;\boldsymbol{\theta}^0\right)\right\}=\underbrace{\sum_{k=1}^{2}\ln p\left(\mathbf{x}^{(k)}|\boldsymbol{\theta}\right)}_{\mathrm{A}}+\mathbb{E}_{x_2^{(3)}}\left\{\ln p\left(\mathbf{x}^{(3)};\boldsymbol{\theta}\mid x_2^{(3)};\boldsymbol{\theta}^0\right)\right\}\\&=\mathrm{A}+\int_{-\infty}^{\infty}dx_2^{(3)}\ \ln p\left(\mathbf{x}^{(3)}\mid\boldsymbol{\theta}\right)\cdot p\left(x_2^{(3)}\mid\boldsymbol{\theta}^0;x_1^{(3)}=1\right).\end{aligned} \tag{8.6}$$

Δεδομένου πως η κατανομή $p\left(x_1,x_2\right)$ είναι διαχωρίσιμη, για την $p\left(\mathbf{x}^{(3)}\mid\boldsymbol{\theta}\right)$ ισχύει

$$\begin{aligned}p\left(\mathbf{x}^{(3)}\mid\boldsymbol{\theta}\right)&=\ln p\left(x_1^{(3)}|\theta_1\right)+\ln p\left(x_2^{(3)}|\theta_2\right)\\&=\ln p\left(x_1^{(3)}|\theta_1\right)-\ln\theta_2\cdot H\left(\theta_2-x_2^{(3)}\right)\ H\left(x_2^{(3)}\right),\end{aligned} \tag{8.7}$$

όπου $H\,(x)$ είναι η βηματική συνάρτηση Heaviside και στην ουσία το γινόμενο που εμφανίζεται στη Σχέση (8.7) φροντίζει η πιθανότητα αυτή να ορίζεται μόνο για τιμές $0 \leq x_2^{(3)} \leq \theta_2$. Σε ό,τι αφορά την $p\left(x_2^{(3)} \mid \boldsymbol{\theta}^0; x_1^{(3)} = 1\right)$, ισχύει

$$p\left(x_2^{(3)} \mid \boldsymbol{\theta}^0; x_1^{(3)} = 1\right) = \frac{1}{3} H\left(3 - x_2^{(3)}\right)\, H\left(x_2^{(3)}\right), \tag{8.8}$$

κατ' αντιστοιχία με το αποτέλεσμα της Σχέσης (8.7), δεδομένου πως $\theta_2^0 = 3$. Λαμβάνοντας τα πάντα υπ' όψιν, παίρνει κανείς

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\right) = \mathrm{A} + \frac{1}{3}\left[\ln p\left(x_1^{(3)}|\theta_1\right) - \ln \theta_2\right] \int_{-\infty}^{\infty} dx_2^{(3)}\, H\left(\theta_2 - x_2^{(3)}\right)\, H\left(3 - x_2^{(3)}\right)\, H\left(x_2^{(3)}\right), \tag{8.9}$$

καθώς $H^2\,(x) = H\,(x)$. Στο σημείο αυτό, τυπικά θα έπρεπε κανείς να πάρει περιπτώσεις σχετικά με τις επιτρεπτές τιμές της θ_2, ώστε να σπάσει το ολοκλήρωμα της (8.9) κατάλληλα. Εάν όμως κανείς παρατηρήσει πως στο σύνολο $\mathcal{D}_g$ υπάρχει η τιμή $x_2 = 5$, τότε εύλογα θα συμπεράνει πως $\theta_2 \geq 5$, δεδομένου πως εάν $\theta_2 < 5$, τότε η τιμή αυτή δε θα μπορούσε να παρατηρηθεί. Λαμβάνοντας αυτό υπ' όψιν, γίνεται εμφανές πως

$$H\left(\theta_2 - x_2^{(3)}\right)\, H\left(3 - x_2^{(3)}\right) = H\left(3 - x_2^{(3)}\right), \tag{8.10}$$

συνεπώς ισχύει

$$\int_{-\infty}^{\infty} dx_2^{(3)}\, H\left(\theta_2 - x_2^{(3)}\right)\, H\left(3 - x_2^{(3)}\right)\, H\left(x_2^{(3)}\right) = \int_0^3 dx_2^{(3)} = 3. \tag{8.11}$$

Έτσι, η Σχέση (8.9) λαμβάνει τη μορφή

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\right) &= \underbrace{\ln p\left(x_1^{(1)}|\theta_1\right) + \ln p\left(x_1^{(2)}|\theta_1\right) - \ln \theta_2 - \ln \theta_2}_{\mathrm{A}} + \ln p\left(x_1^{(3)}|\theta_1\right) - \ln \theta_2 \\ &= \ln p\left(x_1^{(1)}|\theta_1\right) + \ln p\left(x_1^{(2)}|\theta_1\right) + \ln p\left(x_1^{(3)}|\theta_1\right) - 3 \ln \theta_2. \end{aligned} \tag{8.12}$$

Στο σημείο αυτό καλείται κανείς να «διορθώσει» κατάλληλα την εκφώνηση. Χρησιμοποιώντας τη διόρθωση της Σχέσης (8.3) βρίσκει κανείς

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\right) &= \ln \theta_1 - \theta_1 + \ln \theta_1 - 4\theta_1 + \ln \theta_1 - \theta_1 - 3 \ln \theta_2 \\ &= 3 \ln \theta_1 - 6\theta_1 - 3 \ln \theta_2, \end{aligned} \tag{8.13}$$

ενώ για την περίπτωση της Σχέσης (8.4) προκύπτει

$$\begin{aligned} Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^0\right) &= -\ln \theta_1 - \frac{1}{\theta_1} - \ln \theta_1 - \frac{4}{\theta_1} - \ln \theta_1 - \frac{1}{\theta_1} - 3 \ln \theta_2 \\ &= -3 \ln \theta_1 - \frac{6}{\theta_1} - 3 \ln \theta_2, \end{aligned} \tag{8.14}$$

όπου τονίζεται πως οι υπολογισμοί αυτοί αφορούν $\theta_2 \geq 5$ και $\theta_1 > 0$.

8.2 Προχωρώντας στη μεγιστοποίηση (βήμα M) της Q, σε κάθε περίπτωση ισχύει πως

$$Q \sim -3 \ln \theta_2, \tag{8.15}$$

γεγονός που σημαίνει πως η Q είναι μια γνησίως φθίνουσα συνάρτηση ως προς τη μεταβλητή θ_2 και αφού η Q μπορεί να γραφεί ως ένα άθροισμα της μορφής $f(\theta_1) + g(\theta_2)$, όπου f, g κατάλληλες συναρτήσεις, η μεγιστοποίησή της ως προς τη θ_2 θα επιτυγχάνεται όταν αυτή λαμβάνει την ελάχιστη επιτρεπόμενη τιμή της, ανεξαρτήτως της τιμής της θ_1. Έτσι, συμπεραίνει κανείς πως $\theta_2 = 5$. Σε ό,τι αφορά τη μεταβλητή θ_1, η απαίτηση $\partial Q / \partial \theta_1 = 0$ δίνει στην περίπτωση της Σχέσης (8.13) την εξίσωση

$$\frac{3}{\theta_1} - 6 = 0, \tag{8.16}$$

ενώ στην περίπτωση της Σχέσης (8.14) την εξίσωση

$$-\frac{3}{\theta_1} + \frac{6}{\theta_1^2} = 0. \tag{8.17}$$

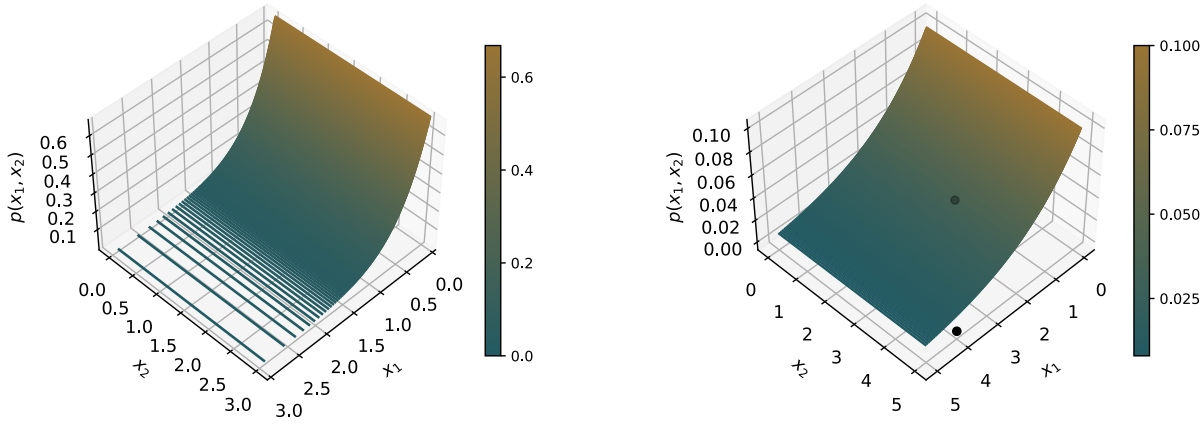
Όπως αναμένεται, η λύση της (8.16) είναι η $\theta_1 = 0.5$, ενώ η λύση της (8.17) είναι η $\theta_1 = 2$, ώστε η προκύπτουσα κατανομή $p(x_1)$ να είναι κοινή σε κάθε περίπτωση. Έτσι, ο αλγόριθμος EM δίνει αποτέλεσμα

$$p(x_1, x_2) = \begin{cases} 0.1\, e^{-0.5 \cdot x_1}, & \text{εάν } x_1 \geq 0 \text{ και } 0 \leq x_2 \leq 5 \\ 0, & \text{αλλιώς} \end{cases}. \tag{8.18}$$

8.3 Πριν την εκτίμηση των παραμέτρων, βάσει των αρχικών τιμών $\boldsymbol{\theta}^0$, η κατανομή $p(x_1, x_2)$ ισούται με

$$p(x_1, x_2) = \begin{cases} \frac{2}{3}\, e^{-2 \cdot x_1}, & \text{εάν } x_1 \geq 0 \text{ και } 0 \leq x_2 \leq 3 \\ 0, & \text{αλλιώς} \end{cases}. \tag{8.19}$$

Η γραφική παράσταση της $p(x_1, x_2)$ πριν και μετά την εκτίμηση φαίνεται στο αριστερό και το δεξί τμήμα της Εικόνας 8.1, αντίστοιχα, μαζί με τα $\mathbf{x}^{(1)}$ και $\mathbf{x}^{(2)}$.



Εικόνα 8.1: Απεικόνιση της $p(x_1, x_2)$ πριν (αριστερά) και μετά (δεξιά) την εκτίμηση των παραμέτρων θ_1, θ_2, μαζί με τα σημεία $\mathbf{x}^{(1)}$ και $\mathbf{x}^{(2)}$.

Αναφορες

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience; 2nd edition, 2000.

[2] T. M. Cover and P. E. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967. DOI: 10.1109/TIT.1967.1053964.