ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

(2021 – 2022)

1η Σειρά Αναλυτικών Ασκήσεων

Ονοματεπώνυμο:

- Χρήστος Τσούφης

Αριθμός Μητρώου:

- 031 17 176

Στοιχεία Επικοινωνίας:

- el17176@mail.ntua.gr
- chris99ts@gmail.com

Άσκηση 1.1 (Maximum Likelihood estimation)

Έστω ότι η μεταβλητή x ακολουθεί μια κατανομή τύπου Erlang: $p(x|\theta) = \theta^2 x e^{(-\theta x)} u(x)$

όπου u(x) είναι η βηματική συνάρτηση $u(x) = \begin{cases} 1, if\ x > 0 \\ 0, if\ x < 0 \end{cases}$

Υποθέστε ότι n δείγματα $D = \{x_1, x_2, \dots, x_n\}$ *επιλέγονται ανεξάρτητα μεταξύ τους σύμφωνα με την κατανομή p(x|θ). Να βρείτε την εκτίμηση μέγιστης πιθανοφάνειας για την παράμετρο θ.*

Λύση:

Η πιθανότητα εκφράζεται ως εξής:

$$p(x|\theta) = p(\{x_1, x_2, \dots, x_N\}|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

Επομένως, η εκτίμηση της πιθανοφάνειας των δειγμάτων από την Κατανομή τύπου Erlang δίνεται από την σχέση:

$$p(D|\theta) = \prod_{i=1}^{N} (\theta^2 x_i e^{(-\theta x_i)} u(x_i))$$

Η σχέση αυτή είναι ισοδύναμη με την ακόλουθη συνάρτηση στην οποία έχει εφαρμοστεί λογάριθμος.

Ισχύει ότι, $\widehat{\theta_{ML}} = \operatorname{argmax} \ln p(D|\theta)$

$$L(\theta) = \ln \prod_{i=1}^{N} P(x_i|\theta) \Rightarrow L(\theta) = \sum_{i=1}^{N} \left(\ln \theta^2 + \ln x_i + \ln e^{-\theta x_i} + \ln u(x_i)\right) \Rightarrow$$

$$L(\theta) = \sum_{i=1}^{N} (2 \ln \theta + \ln x_i - \theta x_i) \Rightarrow L(\theta) = 2\, N \ln \theta + \sum_{i=1}^{N} \ln x_i - \theta \sum_{i=1}^{N} x_i$$

Συνεπώς, για την εκτίμηση της μέγιστης πιθανοφάνειας για την παράμετρο θ αρκεί κανείς να υπολογίσει πότε η παράγωγος ως προς θ ισούται με το μηδέν. Έτσι,

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \Rightarrow \frac{\partial}{\partial \theta}\left(2\, N \ln \theta + \sum_{i=1}^{N} \ln x_i - \theta \sum_{i=1}^{N} x_i\right) = 0 \Rightarrow$$

$$\frac{\partial}{\partial \theta}(2\, N \ln \theta) + \frac{\partial}{\partial \theta}\left(\sum_{i=1}^{N} \ln x_i\right) - \frac{\partial}{\partial \theta}\left(\theta \sum_{i=1}^{N} x_i\right) = 0 \Rightarrow$$

$$\frac{2N}{\theta} - \sum_{i=1}^{N} x_i = 0 \Rightarrow \theta = \frac{2N}{\sum_{i=1}^{N} x_i}$$

<u>Άσκηση 1.2 (Minimax Criterion)</u>

Αφού μελετήσετε την ενότητα 2.3.1 από το κεφάλαιο 2 του συγγράμματος [2] να λύσετε την ακόλουθη άσκηση. Έστω το minimax κριτήριο για τη zero-one συνάρτηση κόστους, δηλαδή $\lambda_{11} = \lambda_{22} = 0$ *και* $\lambda_{12} = \lambda_{21} = 1$.

1. Να δείξετε πως για τις περιοχές απόφασης ισχύει: $\int_{R_2} p(x|\omega_1)dx = \int_{R_1} p(x|\omega_2)dx$

2. Είναι η λύση πάντα μοναδική; Εάν όχι, δώστε ένα αντιπαράδειγμα.

<u>Λύση</u>:

1. Σε αυτή την άσκηση εξετάζεται το minimax κριτήριο για τη zero-one συνάρτηση. Με το minimax κριτήριο επιτυγχάνεται η σχεδίαση ενός ταξινομητή όπου στην χειρότερη περίπτωση για κάθε δείγμα, το μέγιστο ρίσκο είναι ελάχιστο από κάθε άλλο.

Το συνολικό ρίσκο εκφράζεται σύμφωνα με τη σχέση

$$R(a_i|x) = \sum_{j=1}^{c} \lambda(a_i|\omega_j)P(\omega_j|x)$$

Έστω ένας ταξινομητής δύο αποφάσεων. Τότε, είτε το ρίσκο R_1 αντιστοιχεί στην περιοχή απόφασης όπου ο ταξινομητής επιλέγει ω_1 είτε το ρίσκο R_2 αντιστοιχεί στην περιοχή απόφασης όπου ο ταξινομητής επιλέγει ω_2.

Συνεπώς, το ρίσκο αυτού του ταξινομητή θα εκφράζεται από τη σχέση:

$$R = \int_{R_1} [\lambda_{11}P(\omega_1)p(x|\omega_1) + \lambda_{12}P(\omega_2)p(x|\omega_2)]dx +$$

$$\int_{R_2} [\lambda_{21}P(\omega_1)p(x|\omega_1) + \lambda_{22}P(\omega_2)p(x|\omega_2)]dx$$

Και επειδή ισχύει ότι $P(\omega_2) = 1 - P(\omega_1)$, ισχύει ότι:

$$R\big(P(\omega_1)\big) = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx +$$

$$P(\omega_1)[(\lambda_{11} - \lambda_{22}) - (\lambda_{21} - \lambda_{11}) \int_{R_2} p(x|\omega_1)dx - (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx]$$

Από την παραπάνω σχέση, προκύπτει το minimax κριτήριο μηδενίζοντας την ακόλουθη σχέση.

$$I_{mm} = 0 \Rightarrow (\lambda_{11} - \lambda_{22}) - (\lambda_{21} - \lambda_{11}) \int_{R_2} p(x|\omega_1)dx - (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx] = 0$$

Οπότε, το minimax κριτήριο γίνεται:

$$R_{mm,1} = R_{mm,2} \Rightarrow \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx = \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{R_2} p(x|\omega_1)dx$$

Επιπροσθέτως, η zero-one συνάρτηση κόστους εκφράζεται από την σχέση:

$$\lambda\left(\alpha_i \middle| \omega_j\right) = \begin{cases} 0, \mathrm{i} = \mathrm{j} \\ 1, \mathrm{i} \neq \mathrm{j} \end{cases}$$

Ισοδύναμα, $\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = \lambda_{21} = 1$

Οπότε, η συνάρτηση κόστους μετασχηματίζεται ως εξής:

$$\mathrm{R}\big(\mathrm{P}(\omega_1)\big) = \int_{\mathrm{R}_1} \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx} + \int_{\mathrm{R}_2} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} =$$

$$\int_{\mathrm{R}_1} \big(1 - \mathrm{P}(\omega_1)\big)\mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx} + \int_{\mathrm{R}_2} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} =$$

$$\int_{\mathrm{R}_1} \mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx} + \mathrm{P}(\omega_1)\left[\int_{\mathrm{R}_2} \mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} - \int_{\mathrm{R}_1} \mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx}\right]$$

Σημειώνεται ότι μεταξύ των R_1, R_2, η πιθανότητα σφάλματος αλλάζει γραμμικά σε σχέση με την a-priori πιθανότητα $\mathrm{P}(\omega_1)$ και μεγιστοποιείται όταν πάρει την τιμή 1.

Επομένως, σύμφωνα με τις παραπάνω σχέσεις για τις περιοχές απόφασης θα ισχύει ότι:

$$\lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathrm{R}_1} \mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx} = \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathrm{R}_2} \mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} \Rightarrow$$

$$0 + (1 - 0) \int_{\mathrm{R}_1} \mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx} = 0 + (1 - 2) \int_{\mathrm{R}_2} \mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} \Rightarrow$$

$$\int_{\mathrm{R}_1} \mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx} = \int_{\mathrm{R}_2} \mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx}$$

2. Ένα αντιπαράδειγμα είναι το εξής:

Έστω οι ακόλουθες κατανομές

$$\mathrm{p}(\mathrm{x}|\omega_1) = \begin{cases} \frac{1}{2}, 0 \leq \mathrm{x} \leq 2 \\ 0, \text{otherwise} \end{cases} \;\&\; \mathrm{p}(\mathrm{x}|\omega_2) = \begin{cases} \frac{1}{2}, 1 \leq \mathrm{x} \leq 3 \\ 0, \text{otherwise} \end{cases}$$

Παρατηρείται τότε ότι

$$\int_{\mathrm{R}_1} \mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx} = \int_{\mathrm{R}_2} \mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx}$$

Οπότε, η λύση δεν είναι μοναδική.

Άσκηση 1.3 (Bayes Error)

Έστω δύο ανεξάρτητες κατανομές $p(x|\omega_i)$ με priors $P(\omega_i)$. Υποθέτουμε ότι οι κατανομές αυτές ορίζονται για ένα χώρο διάστασης d = 2.

1. Έστω ότι οι κατανομές σας είναι κανονικές (normal). Προβάλλετε τις κατανομές στη μια διάσταση και δείξτε πως το σφάλμα στη μία διάσταση είναι ίσο ή μεγαλύτερο από το σφάλμα Bayes στις δύο διαστάσεις.

2. Δείξτε το ίδιο αποτέλεσμα για οποιαδήποτε κατανομή στις δύο διαστάσεις.

Λύση:

1. Έστω μια line σε d = 2 – διαστάσεις που προβάλλεται σε μια line στην μια διάσταση.

Το conditional error σε d διαστάσεις είναι:

$$\mathrm{E_d} = \int_{\mathrm{line}} \min[\mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1), \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}|\omega_2)]\,\mathrm{dx} = \int_{\mathrm{G_1}} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} + \int_{\mathrm{G_2}} \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx}$$

Όπου $\mathrm{G_1}$, $\mathrm{G_2}$ είναι τα disjoint segments όπου οι posterior πιθανότητες κάθε κατανομής είναι ελάχιστες.

Στην μια διάσταση, θα ισχύει:

$$\mathrm{E_{d-1}} = \int_{\mathrm{line}} \min[\mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1), \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}|\omega_2)]\,\mathrm{dx}$$

Στο άπειρο ισχύει ότι:

$$\int_{\mathrm{line}} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} < \int_{\mathrm{line}} \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}|\omega_2)\mathrm{dx}$$

Κι έτσι,

$$\mathrm{E_{d-1}} = \int_{\mathrm{line}} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx}$$

Εάν $\mathrm{G_1} + \mathrm{G_2}$ ορίζουν ολόκληρη την line, τότε

$$\mathrm{E_{d-1}} = \int_{\mathrm{G_1+G_2}} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} = \int_{\mathrm{G_1}} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} + \int_{\mathrm{G_2}} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} =$$

$$\int_{\mathrm{G_1}} \mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1)\mathrm{dx} + \int_{\mathrm{line}} [\mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}|\omega_2) + |\mathrm{f}(\mathrm{x})|]\mathrm{dx} = \mathrm{E_d} + \int_{\mathrm{G_2}} |\mathrm{f}(\mathrm{x})|\mathrm{dx} \geq \mathrm{E_d}$$

Επιπλέον, χρησιμοποιήθηκε το γεγονός ότι $\mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1) = [\mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}|\omega_2) + |\mathrm{f}(\mathrm{x})|]$ στο $\mathrm{G_2}$ για κάποια άγνωστη συνάρτηση f.

Με άλλα λόγια, για τις πιθανότητες θα ισχύει ότι:

$$p(x|\omega_1) = \frac{1}{2\pi\sigma_{1x}\sigma_{1y}} \exp\left\{-\frac{(x-\mu_{1x})^2}{2\sigma_{1x}^2} - \frac{\left(y-\mu_{1y}\right)^2}{2\sigma_{1y}^2}\right\}$$

$$p(x|\omega_2) = \frac{1}{2\pi\sigma_{2x}\sigma_{2y}} \exp\left\{-\frac{(x-\mu_{2x})^2}{2\sigma_{2x}^2} - \frac{\left(y-\mu_{2y}\right)^2}{2\sigma_{2y}^2}\right\}$$

Οπότε, το σφάλμα θα είναι

$$P_{2d} = \int_{-\infty}^{+\infty} \min[P(\omega_1)p(x|\omega_1), P(\omega_2)p(x|\omega_2)]\,dx = \int_{G_2} P(\omega_1)p(x|\omega_1)dx + \int_{G_1} P(\omega_2)p(x|\omega_2)dx$$

Επομένως, αντικαθιστώντας με την εξίσωση της ευθείας (line) y = ax + b για την 1 διάσταση, προκύπτει:

$$P_{1d} = \int_{line} \min[P(\omega_1)p(x|\omega_1), P(\omega_2)p(x|\omega_2)]\,dx$$

Εάν,

$$\int_{line} P(\omega_1)p(x|\omega_1)dx > \int_{line} P(\omega_2)p(x|\omega_2)dx$$

Τότε,

$$P_{1d}(error) = \int_{line} P(\omega_1)p(x|\omega_1)dx$$

Ενώ, εάν $G_1 + G_2$ ορίζουν ολόκληρη την line, τότε

$$P_{1d} = \int_{G_1+G_2} P(\omega_1)p(x|\omega_1)dx = \int_{G_1} P(\omega_1)p(x|\omega_1)dx + \int_{G_2} P(\omega_1)p(x|\omega_1)dx > P_{2d}$$

2. Σε πραγματικά προβλήματα Αναγνώρισης Προτύπων σπάνια υπάρχουν οι πραγματικές κατανομές, αντιθέτως υπάρχουν απλώς εκτιμήσεις. Η παραπάνω δήλωση δεν ισχύει εάν οι κατανομές που χρησιμοποιούνται είναι απλώς εκτιμήσεις. Όμως, αφού αποδείχθηκε για οποιαδήποτε κατανομή, τότε θα ισχύει και για Bayes distribution.

Άσκηση 1.4 (Perceptrons)

Μας δίνονται 10 διανύσματα χαρακτηριστικών που προέρχονται από δύο κλάσεις ω_1 και ω_2:

$$\omega_1: [2, 1.5]^T, [4, 1]^T, [-3, 3]^T, [-2, 2]^T, [0, 3]^T$$

$$\omega_2: [-1, 0]^T, [0, 0]^T, [2, -0.5]^T, [1, -1]^T, [-2, 1]^T$$

Αρχικά, ελέγξτε εάν οι δύο κλάσεις είναι γραμμικά διαχωρίσιμες, μέσω του σχεδιασμού των παραπάνω σημείων σε ένα γράφημα. Στη συνέχεια, χρησιμοποιήστε τον ακόλουθο αλγόριθμο εκπαίδευσης ενός perceptron με ρ = 1 και $w(0) = [0, 0]^T$, για να σχεδιάσετε μία ευθεία γραμμή που να διαχωρίζει τις δύο κλάσεις. Εάν το συγκεκριμένο διάνυσμα βαρών w δεν επαρκεί (να σχολιαστεί με μία πρόταση το γιατί), τότε να χρησιμοποιηθεί ως αρχικό διάνυσμα βαρών το $w(0) = [0, 0, 0]^T$, κάνοντας την κατάλληλη επαύξηση ταυτόχρονα και στα διανύσματα χαρακτηριστικών. Τέλος, να δοθεί με μορφή εξίσωσης η διαχωριστική καμπύλη που αντιστοιχεί στο υπολογισθέν διάνυσμα βαρών.

***Αλγόριθμος Perceptron**: Έστω w(t) η εκτίμηση του διανύσματος βάρους και x_t το αντίστοιχο διάνυσμα χαρακτηριστικών στο t-οστό βήμα επανάληψης. Ο αλγόριθμος έχει ως εξής:*

$$w(t+1) = w(t) + \rho x_t \text{ αν } x_t \in \omega_1 \text{ και } w(t)^T x_t \leq 0$$

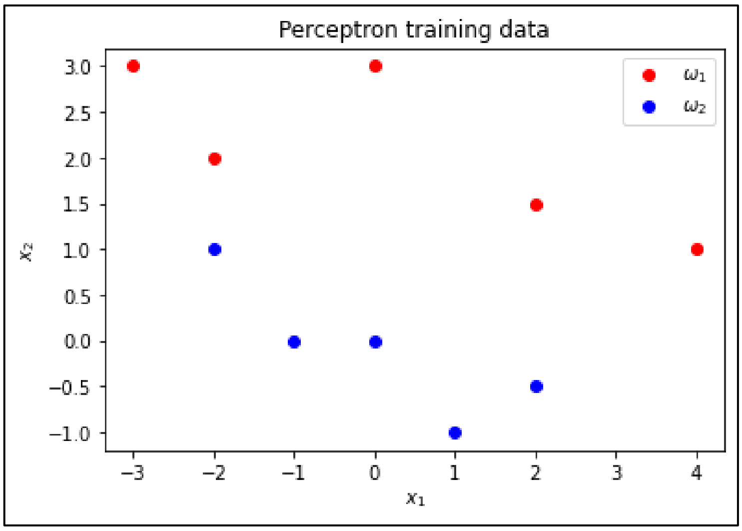
$$w(t+1) = w(t) - \rho x_t \text{ αν } x_t \in \omega_2 \text{ και } w(t)^T x_t \geq 0$$

$$w(t+1) = w(t) \text{ αλλιώς}$$

Ο ανωτέρω αλγόριθμος έχει την μορφή αλγορίθμων τύπου reward and punishment. Δηλαδή, αν το τωρινό δείγμα εκπαίδευσης ταξινομηθεί σωστά, τότε δεν γίνεται τίποτα (reward = no action). Αλλιώς, αν το δείγμα δεν ταξινομηθεί σωστά, η τιμή του διανύσματος βάρους μεταβάλλεται προσθέτοντας (αφαιρώντας) μία τιμή ανάλογη του x_t (punishment = correction cost).

Λύση:

Αρχικά, από τον σχεδιασμό των σημείων προκύπτει το ακόλουθο γράφημα.



Παρατηρείται ότι το data set είναι linearly separable.

Ύστερα, πραγματοποιείται η εκπαίδευση του perceptron με ρ = 1 και $w(0) = [0, 0]^T$ για την εύρεση μιας ευθείας που να διαχωρίζει τις δύο κλάσεις. Στο perceptron δίνονται εναλλάξ samples και από τις δύο κλάσεις για γρηγορότερη σύγκλιση. Έτσι, προκύπτει:

x_i	**class**	$\mathbf{w(t)^T x_t}$	**classification**	$\mathbf{w(t+1) = w(t) \pm \rho x_t}$
Epoch 1				
(2.0, 1.5)	ω1	0.0	False	(2.0, 1.5)
(-1, 0)	ω2	-2.0	Right	(2.0, 1.5)
(4, 1)	ω1	9.5	Right	(2.0, 1.5)
(0, 0)	ω2	0.0	False	(2.0, 1.5)
(-3, 3)	ω1	-1.5	False	(-1.0, 4.5)
(2.0, -0.5)	ω2	-4.25	Right	(-1.0, 4.5)
(-2, 2)	ω1	11.0	Right	(-1.0, 4.5)
(1, -1)	ω2	-5.5	Right	(-1.0, 4.5)
(0, 3)	ω1	13.5	Right	(-1.0, 4.5)
(-2, 1)	ω2	6.5	False	(1.0, 3.5)
Epoch 2				
(2.0, 1.5)	ω1	7.25	Right	(1.0, 3.5)
(-1, 0)	ω2	-1.0	Right	(1.0, 3.5)
(4, 1)	ω1	7.5	Right	(1.0, 3.5)
(0, 0)	ω2	0.0	False	(1.0, 3.5)
(-3, 3)	ω1	7.5	Right	(1.0, 3.5)
(2.0, -0.5)	ω2	0.25	False	(-1.0, 4.0)
(-2, 2)	ω1	10.0	Right	(-1.0, 4.0)
(1, -1)	ω2	-5.0	Right	(-1.0, 4.0)
(0, 3)	ω1	12.0	Right	(-1.0, 4.0)
(-2, 1)	ω2	6.0	False	(1.0, 3.0)
Epoch 3				
(2.0, 1.5)	ω1	6.5	Right	(1.0, 3.0)
(-1, 0)	ω2	-1.0	Right	(1.0, 3.0)
(4, 1)	ω1	7.0	Right	(1.0, 3.0)
(0, 0)	ω2	0.0	False	(1.0, 3.0)
(-3, 3)	ω1	6.0	Right	(1.0, 3.0)
(2.0, -0.5)	ω2	0.5	False	(-1.0, 3.5)
(-2, 2)	ω1	9.0	Right	(-1.0, 3.5)
(1, -1)	ω2	-4.5	Right	(-1.0, 3.5)
(0, 3)	ω1	10.5	Right	(-1.0, 3.5)
(-2, 1)	ω2	5.5	False	(1.0, 2.5)

Παρατηρείται ότι από την Epoch 2, τα σημεία [2, -0.5] και [-2, 1] δεν ταξινομούνται σωστά επειδή η εξίσωση της ευθείας διαχωρισμού διέρχεται από την αρχή των αξόνων.

Έπειτα, για την εκπαίδευση του perceptron, το πρώτο πράγμα που πρέπει να σημειωθεί είναι ότι τα διάνυσμα βάρους $w(0) = [0, 0]^T$ δεν αρκεί καθώς δεν περιέχει έναν bias όρο. Χωρίς τον bias όρο, η γραμμή που σχηματίζεται από τον perceptron είναι υποχρεωμένη τα περάσει από το origin και έτσι η εκπαίδευση είτε δεν είναι αρκετά flexible είτε είναι αδύνατη. Για αυτόν τον λόγο, απαιτείται επαύξηση του διανύσματος βαρών οπότε επεκτείνονται όλα τα διανύσματα σύμφωνα με την ακόλουθη φόρμα της discriminant function.

$$g(x) = w^T x + b$$

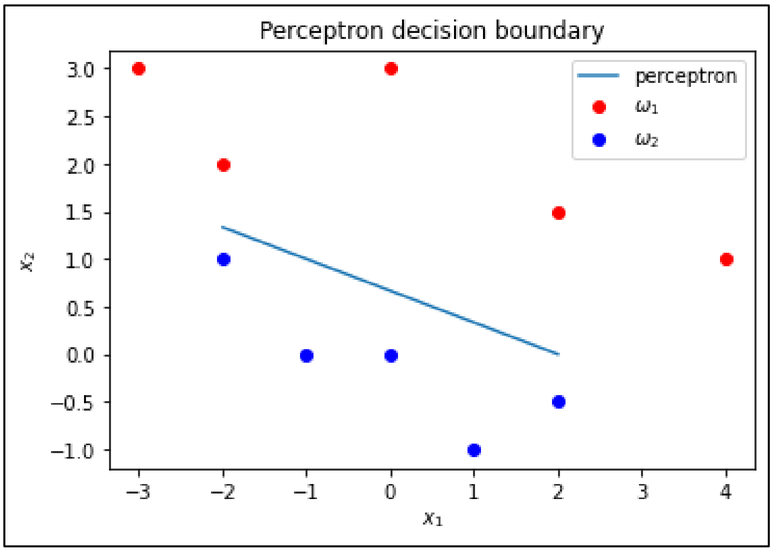
$$g(x) = w_e^T x_e$$

$$w_e = [b\ w]^T$$

$$x_e = [1\ x]^T$$

Έτσι, κάνοντας merge όλα τα δεδομένα σε ένα πίνακα, για καλύτερο χειρισμό, ορίζεται η perceptron training function. Η κύρια λειτουργία της είναι να κάνει feed στον perceptron συνεχόμενα με το data set και να σταματά την εκπαίδευση όταν κάνει classify όλα τα samples σωστά.

Εκτελώντας τον κώδικα προκύπτει η ακόλουθη εικόνα:



Μπορεί να παρατηρήσει κανείς ότι η εξίσωση που περιγράφει το perceptron decision boundary θα είναι:

$$b + w_1 x_1 + w_2 x_2 = 0 \Rightarrow -3 + 2x_1 + 4.5x_2 = 0$$

Παρατηρείται μάλιστα ότι εκτελώντας παραπάνω από 1 φορές το συγκεκριμένο κομμάτι κώδικα παρατηρείται ότι τα w_1, w_1, b αλλάζουν (λ.χ. [-2 2 4.5], [-4 3 6.5] κλπ.).

Παρακάτω φαίνονται αναλυτικά τα βήματα του αλγορίθμου:

x_i	**class**	$\mathbf{w(t)^T x_t}$	**classification**	$\mathbf{w(t+1) = w(t) \pm \rho x_t}$
Epoch 1				
(2.0, 1.5, 1.0)	ω1	0.0	False	(2.0, 1.5, 1.0)
(-1, 0, 1)	ω2	-1.0	Right	(2.0, 1.5, 1.0)
(4, 1, 1)	ω1	10.5	Right	(2.0, 1.5, 1.0)
(0, 0, 1)	ω2	1.0	False	(2.0, 1.5, 0.0)
(-3, 3, 1)	ω1	-1.5	False	(-1.0, 4.5, 1.0)
(2.0, -0.5, 1.0)	ω2	-3.25	Right	(-1.0, 4.5, 1.0)
(-2, 2, 1)	ω1	12.0	Right	(-1.0, 4.5, 1.0)
(1, -1, 1)	ω2	-4.5	Right	(-1.0, 4.5, 1.0)
(0, 3, 1)	ω1	14.5	Right	(-1.0, 4.5, 1.0)
(-2, 1, 1)	ω2	7.5	False	(1.0, 3.5, 0.0)
Epoch 2				
(2.0, 1.5, 1.0)	ω1	7.25	Right	(1.0, 3.5, 0.0)
(-1, 0, 1)	ω2	-1.0	Right	(1.0, 3.5, 0.0)
(4, 1, 1)	ω1	7.5	Right	(1.0, 3.5, 0.0)
(0, 0, 1)	ω2	0.0	False	(1.0, 3.5, -1.0)
(-3, 3, 1)	ω1	6.5	Right	(1.0, 3.5, -1.0)
(2.0, -0.5, 1.0)	ω2	-0.75	Right	(1.0, 3.5, -1.0)
(-2, 2, 1)	ω1	4.0	Right	(1.0, 3.5, -1.0)
(1, -1, 1)	ω2	-3.5	Right	(1.0, 3.5, -1.0)
(0, 3, 1)	ω1	9.5	Right	(1.0, 3.5, -1.0)
(-2, 1, 1)	ω2	0.5	False	(3.0, 2.5, -2.0)
Epoch 3				
(2.0, 1.5, 1.0)	ω1	7.75	Right	(3.0, 2.5, -2.0)
(-1, 0, 1)	ω2	-5.0	Right	(3.0, 2.5, -2.0)
(4, 1, 1)	ω1	12.5	Right	(3.0, 2.5, -2.0)
(0, 0, 1)	ω2	-2.0	Right	(3.0, 2.5, -2.0)
(-3, 3, 1)	ω1	-3.5	False	(0.0, 5.5, -1.0)
(2.0, -0.5, 1.0)	ω2	-3.75	Right	(0.0, 5.5, -1.0)
(-2, 2, 1)	ω1	10.0	Right	(0.0, 5.5, -1.0)
(1, -1, 1)	ω2	-6.5	Right	(0.0, 5.5, -1.0)
(0, 3, 1)	ω1	15.5	Right	(0.0, 5.5, -1.0)
(-2, 1, 1)	ω2	4.5	False	(2.0, 4.5, -2.0)
Epoch 4				
(2.0, 1.5, 1.0)	ω1	8.75	Right	(2.0, 4.5, -2.0)
(-1, 0, 1)	ω2	-4.0	Right	(2.0, 4.5, -2.0)
(4, 1, 1)	ω1	10.5	Right	(2.0, 4.5, -2.0)
(0, 0, 1)	ω2	-2.0	Right	(2.0, 4.5, -2.0)
(-3, 3, 1)	ω1	5.5	Right	(2.0, 4.5, -2.0)
(2.0, -0.5, 1.0)	ω2	-0.25	Right	(2.0, 4.5, -2.0)
(-2, 2, 1)	ω1	3.0	Right	(2.0, 4.5, -2.0)
(1, -1, 1)	ω2	-4.5	Right	(2.0, 4.5, -2.0)
(0, 3, 1)	ω1	11.5	Right	(2.0, 4.5, -2.0)
(-2, 1, 1)	ω2	-1.5	Right	(2.0, 4.5, -2.0)

<u>Άσκηση 1.5 (Kullback-Leibler Divergence)</u>

Προκειμένου να συγκρίνουμε πολυδιάστατες κατανομές πιθανότητας που ορίζονται στον ίδιο χώρο έχουνε οριστεί αρκετές "μετρικές". Μια από αυτές είναι η Kullback-Leibler που ορίζεται ως εξής:

$$D_{KL}\big(p_1(x), p_2(x)\big) = \int p_1(x) ln \frac{p_1(x)}{p_2(x)} dx$$

Έστω τώρα ότι θέλουμε να προσεγγίσουμε μια οποιαδήποτε κατανομή $p_1(x)$ με μια πολυδιάστατη Γκαουσιανή $p_2(x) \sim \mathcal{N}(\mu, \Sigma)$. Να δείξετε ότι οι τιμές που οδηγούν στη μικρότερη τιμής για την Kullback-Leibler divergence είναι οι "προφανείς" διαισθητικά τιμές:

$$\mu = \mathbb{E}_{x \sim p_2} p_2[x] \ \&\Sigma = \mathbb{E}_{x \sim p_2} p_2[(x-\mu)(x-\mu)^T]$$

<u>Λύση</u>:

Έστω ότι $\mathrm{p_2(x) \equiv p(x|\omega_2)} \sim \mathcal{N}(\mu, \Sigma)$ αλλά $\mathrm{p_1(x) \equiv p(x|\omega_1)}$ είναι αυθαίρετη.

Η Kullback-Leibler Divergence από $\mathrm{p_1(x)}$ σε $\mathrm{p_2(x)}$ είναι:

$$\mathrm{D_{KL}\big(p_1(x), p_2(x)\big) = \int p_1(x) \ln p_1(x)\, dx + \int p_1(x)[d \ln(2\pi) + \ln|\Sigma| + (x-\mu)^T \Sigma^{-1} (x-\mu)] dx}$$

Παραπάνω χρησιμοποιήθηκε το γεγονός ότι η $\mathrm{p_2}$ είναι μια Gaussian, τέτοια ώστε:

$$\mathrm{p_2(x) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right]}$$

Οπότε, στόχος είναι η αναζήτηση μ και Σ τέτοιων ώστε να ελαχιστοποιείται η "απόσταση". Αυτό γίνεται μηδενίζοντας την παράγωγο ώστε να προκύψει το εξής:

$$\mathrm{\frac{\partial}{\partial \mu} D_{KL}\big(p_1(x), p_2(x)\big) = -\int \Sigma^{-1}(x-\mu) p_1(x) dx = 0 \Rightarrow \Sigma^{-1} \int p_1(x)(x-\mu) dx = 0}$$

Υποθέτοντας ότι Σ είναι non-singular, η εξίσωση γίνεται:

$$\mathrm{\int p_1(x)(x-\mu) dx = E_1[x-\mu] = 0 \Rightarrow E_1[x] = \mu}$$

Αυτό σημαίνει ότι η mean της δεύτερης κατανομής πρέπει να είναι η ίδια με την Gaussian.

Τώρα θα υπολογιστεί η covariance της δεύτερης κατανομής. Αυτό επιτυγχάνεται μηδενίζοντας την παράγωγο της Kullback-Leibler Divergence. Για διευκόλυνση στους συμβολισμούς και χωρίς βλάβη της γενικότητας, προκύπτει ότι B = Σ Οπότε, δεδομένου ότι $\mathrm{\frac{\partial |B|}{\partial B} = |B| B^{-1}}$ και ότι $\mathrm{B = \Sigma^{-1}}$ είναι συμμετρικός, προκύπτει:

$$\mathrm{\frac{\partial}{\partial B} D_{KL}\big(p_1(x), p_2(x)\big) = 0 \Rightarrow -\int p_1(x)[-B^{-1} + (x-\mu)(x-\mu)^T] dx = 0}$$

Έτσι, $\mathrm{E_1[\Sigma - (x-\mu)(x-\mu)^T] = 0 \Rightarrow E_1[(x-\mu)(x-\mu)^T] = \Sigma}$

Συνεπώς, δείχθηκε ότι η covariance της δεύτερης κατανομής πρέπει να κάνει match με την Gaussian.

Άσκηση 1.6 (Linear Regression and the LMS Algorithm)

Θεωρήστε το ακόλουθο μοντέλο γραμμικής παλινδρόμησης: $y = w_0 + w_1 x$

καθώς και το παραπάνω σύνολο δεδομένων στη μορφή (x, y):

$$\left\{\begin{pmatrix}0.38\\2.05\end{pmatrix},\begin{pmatrix}0.44\\2.23\end{pmatrix},\begin{pmatrix}0.48\\2.13\end{pmatrix},\begin{pmatrix}0.54\\2.33\end{pmatrix},\begin{pmatrix}0.58\\2.67\end{pmatrix},\begin{pmatrix}0.64\\2.68\end{pmatrix},\begin{pmatrix}0.71\\2.81\end{pmatrix},\begin{pmatrix}0.76\\2.97\end{pmatrix},\begin{pmatrix}0.82\\3.12\end{pmatrix},\begin{pmatrix}0.96\\3.20\end{pmatrix}\right\}$$

για το οποίο γνωρίζουμε ότι έχει προκύψει έπειτα από την επίδραση γκαουσιανού θορύβου.

1. Να αποδείξετε ότι η λύση μέγιστης πιθανοφάνειας είναι ισοδύναμη με τη λύση που προκύπτει από τη μέθοδο των ελαχίστων τετραγώνων.

2. Να υπολογίσετε τους άγνωστους συντελεστές χρησιμοποιώντας τη μέθοδο των ελαχίστων τετραγώνων.

3. Χρησιμοποιώντας ένα από τα παραπάνω δεδομένα τη φορά, και εφαρμόζοντας τον αλγόριθμο Least Mean Squares (LMS) (Υποκεφάλαιο 5.8 του βιβλίου [2]), να υπολογίσετε εκ νέου τους άγνωστους συντελεστές. Να επαναληφθεί η διαδικασία για τουλάχιστον μία εποχή χειροκίνητα.

4. Γράψτε πρόγραμμα που να εφαρμόζει τον LMS στο συγκεκριμένο πρόβλημα, ώστε να βρείτε μια καλύτερη προσέγγιση.

5. Να σχεδιάσετε σε ένα κοινό γράφημα τις ευθείες που βρήκατε σαν λύσεις.

Λύση:

1. 1ος Τρόπος:

Έστω ότι παρατηρούνται δεδομένα $y_i = w_1 x_i + w_0 + z_i$, όπου $z_i \sim \mathcal{N}(0, \sigma^2)$.

Η κατανομή που ακολουθεί ο θόρυβος, χωρίς βλάβη της γενικότητας, έχει μηδενική μέση (αλλιώς θα μπορούσε να αλλάξει κατάλληλα το w_0.

Η πιθανότητα να έρθει δείγμα είναι ίση με:

$$P = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{z_i^2}{2\sigma^2}\right\} \Rightarrow P = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - w_1 x_i - w_0)^2}{2\sigma^2}\right\} \Rightarrow$$

$$\log P = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w_1 x_i - w_0)^2$$

Από την παραπάνω σχέση παρατηρείται ότι μόνο ο όρος με το sum εξαρτάται από τα w_0, w_1. Οπότε, για δεδομένη διασπορά σ^2, τα w_0, w_1 που μεγιστοποιούν την πιθανοφάνεια του δείγματος είναι εκείνα που ελαχιστοποιούν τον παραπάνω όρο, δηλαδή:

$$\sum_{i=1}^{n} (y_i - w_1 x_i - w_0)^2 = \sum_{i=1}^{n} (y_i - (w_1 x_i - w_0))^2$$

Δηλαδή, οι M.L.E. εκτιμητές των w_0, w_1 είναι η λύση της μεθόδου ελαχίστων τετραγώνων.

2^{ος} Τρόπος:

Εναλλακτικά, για $y = w_1 x + w_0$ (1) και δεδομένου ότι τα δείγματα έχουν προκύψει από την επίδραση γκαουσιανού θορύβου θα ισχύει ότι $E(y) = x_i w_i$. Δηλαδή,

$$\sum_{i=1}^{N} f\left(\frac{y_i}{x_i}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - w_1 x_i)^2}{2\sigma^2}\right\} \Rightarrow$$

$$\ln\left(f\left(\frac{y_i}{x_i}\right)\right) = \frac{10}{\sqrt{2\pi}\sigma}\ln(2\pi) - \frac{10}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{10}(y_i - w_1 x_i)^2 \qquad (2)$$

Συνεπώς, για την εύρεση της μέγιστης πιθανοφάνειας, η παράγωγος θα είναι:

$$\frac{\partial}{\partial w_1}\left(f\left(\frac{y_i}{x_i}\right)\right) = 0 \Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{10}(y_i - w_1 x_i)x_i^T = 0 \Rightarrow$$

$$\sum_{i=1}^{10} x_i^T y_i - \sum_{i=1}^{10} x_i^T x_i w_1 = 0 \Rightarrow w_1 = \sum_{i=1}^{10}\left(x_i^T x_i\right)^{-1} \cdot \sum_{i=1}^{10} x_i^T y_i \Rightarrow$$

$$w_1 = (x^T x)^{-1}(x^T y) \qquad (3)$$

Έπειτα, η τιμή του w_1 που υπολογίστηκε, αντικαθίσταται την (1) και τελικά υπολογίζεται το w_0.

Με την μέθοδο ελαχίστων τετραγώνων: $S = \sum_{i=1}^{10}(y_i - w_1 x_i)^2 \qquad (4)$

Επομένως, η παράγωγος θα είναι: $\frac{\partial}{\partial w_1}(S) = 0 \Rightarrow 2\sum_{i=1}^{10}(y_i - w_1 x_i)x_i^T = 0 \Rightarrow$

$$\sum_{i=1}^{10} y_i x_i^T - \sum_{i=1}^{10} x_i^T x_i w_1 = 0 \Rightarrow w_1 = \sum_{i=1}^{10}\left(x_i^T x_i\right)^{-1} \cdot \sum_{i=1}^{10} y_i x_i^T \Rightarrow$$

$$w_1 = (x^T x)^{-1}(y x^T) \qquad (5)$$

Οπότε, παρατηρείται ότι οι σχέσεις (3) και (5) είναι ίδιες οπότε η μέθοδος ελαχίστων τετραγώνων δίνει το ίδιο αποτέλεσμα με τη λύση της μέγιστης πιθανοφάνειας.

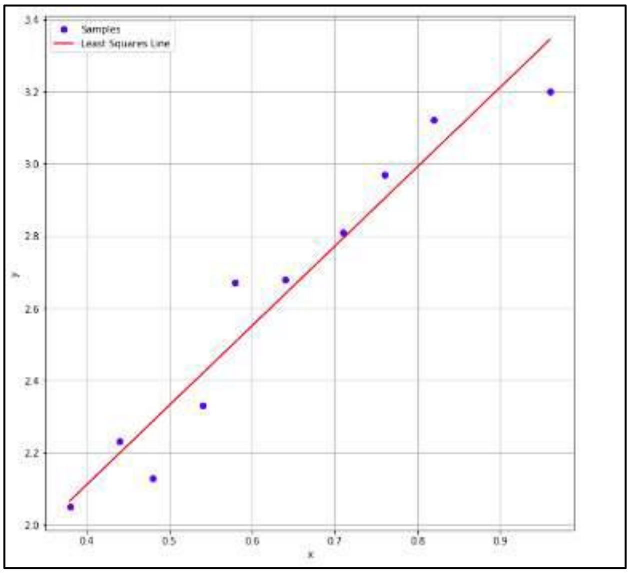
2. Από την μέθοδο ελαχίστων τετραγώνων είναι γνωστό πως:

$$w_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$w_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

Όπου, $\sum_{i=1}^{n} x_i = 6.31$, $\sum_{i=1}^{n} y_i = 26.19$, $\sum_{i=1}^{n} x_i^2 = 4.282$, $\sum_{i=1}^{n} x_i y_i = 17.187$

Οπότε $w_0 = 1.228$ και $w_1 = 2.204$.



3. Έστω η εξίσωση Y·a = b και αναζητείται το διάνυσμα βάρους α.

Όπου,

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.38 & 0.44 & 0.48 & 0.54 & 0.58 & 0..64 & 0.71 & 0.76 & 0.82 & 0.96 \end{bmatrix}^T \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} =$$

$$= [2.05\ 2.23\ 2.13\ 2.33\ 2.67\ 2.68\ 2.81\ 2.97\ 3.12\ 3.20]^T$$

Επίσης, η συνάρτηση κόστους δίνεται από τη σχέση:

$$J_s(a) = \|Ya - b\|^{\ 2} = \sum_{i=1}^{n} (a^t y_i - b_i)^2 = \sum_{i=1}^{n} (\omega_0 + \omega_1 x_i - y_i)^2$$

Αυτή ελαχιστοποιείται με gradient decent.

Ο αλγόριθμος LMS είναι λέει ότι $a(0)$ είναι αυθαίρετο και επίσης ότι:

$$a(k+1) = a(k) + \eta(k)(b_k - a(k)^T y^k) y^k$$

Οπότε, για $\eta(k) = \frac{1}{k+1}$, $\alpha(0) = \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ προκύπτει ότι $y^k = \begin{pmatrix} 1 \\ x_k \end{pmatrix}$ και το βήμα εκπαίδευσης:

$$a(k+1) = a(k) + \frac{1}{k+1}\left(b_k - a(k)^T \begin{bmatrix} 1 \\ x_k \end{bmatrix}\right) \begin{bmatrix} 1 \\ x_k \end{bmatrix}$$

Ακολούθως υπολογίζονται τα α ως εξής:

$$a(1) = \begin{bmatrix} 1.335 \\ 1.127 \end{bmatrix}, a(2) = \begin{bmatrix} 1.468 \\ 1.186 \end{bmatrix}, a(3) = \begin{bmatrix} 1.491 \\ 1.197 \end{bmatrix}, a(4) = \begin{bmatrix} 1.530 \\ 1.218 \end{bmatrix}, a(5) = \begin{bmatrix} 1.602 \\ 1.260 \end{bmatrix},$$

$$a(6) = \begin{bmatrix} 1.641 \\ 1.285 \end{bmatrix}, a(7) = \begin{bmatrix} 1.673 \\ 1.307 \end{bmatrix}, a(8) = \begin{bmatrix} 1.707 \\ 1.333 \end{bmatrix}, a(9) = \begin{bmatrix} 1.739 \\ 1.359 \end{bmatrix}, a(10) = \begin{bmatrix} 1.753 \\ 1.373 \end{bmatrix}$$

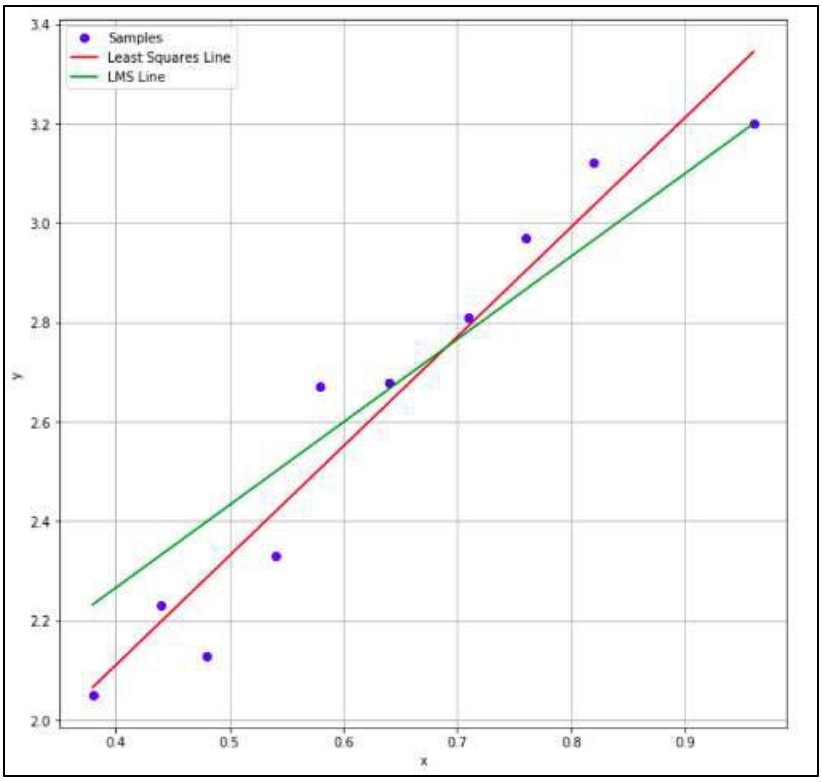
Και ενδεικτικά, από τον αναλυτικό υπολογισμό για κάθε Epoch, προκύπτει:

k	**x(k)**	**b(k)**	**a(k)**	**θ(k)**
Epoch: 1				
1	0.38	2.05	(1.335, 1.127)	(+0.0009, +0.0009)
2	0.44	2.23	(1.468, 1.186)	(+0.0009, +0.0009)
3	0.48	2.13	(1.491, 1.197)	(+0.0009, +0.0009)
4	0.54	2.33	(1.530, 1.218)	(+0.0009, +0.0009)
5	0.58	2.67	(1.602, 1.260)	(+0.0009, +0.0009)
6	0.64	2.68	(1.641, 1.285)	(+0.0009, +0.0009)
7	0.71	2.81	(1.673, 1.307)	(+0.0009, +0.0009)
8	0.76	2.97	(1.707, 1.333)	(+0.0009, +0.0009)
9	0.82	3.12	(1.739, 1.359)	(+0.0009, +0.0009)
10	0.96	3.20	(1.753, 1.373)	(+0.0009, +0.0009)
…				

Αναλυτικότερα φαίνονται στο συνοδευόμενο αρχείο.

4. Η πλήρης εκτέλεση του LMS φαίνεται στο συνοδευόμενο αρχείο.

5. Το γράφημα θα είναι το εξής:



Άσκηση 1.7 (Bayes meets k-NN)

Έστω δύο ταξινομητές οι οποίοι βασίζονται σε δείγματα που προέρχονται από δύο κατανομές

$$p(x|\omega_1) = \begin{cases} 2x, for\ 0 \le x \le 1 \\ 0, otherwise \end{cases}$$

$$p(x|\omega_2) = \begin{cases} 2-2x, for\ 0 \le x \le 1 \\ 0, otherwise \end{cases}$$

Επιπλέον οι a-priori πιθανότητες για κάθε κατηγορία είναι $p[\omega_1] = p[\omega_2] = \frac{1}{2}$.

1. Ποιος είναι ο κανόνας απόφασης (decision rule) κατά Bayes και ποιο το σφάλμα ταξινόμησης Bayes;

2. Έστω ότι δειγματοληπτούμε ένα τυχαίο σημείο από την κατηγορία ω_1 *και ένα από την* ω_2*. Με βάση τα σημεία αυτά κατασκευάζουμε έναν ταξινομητή κοντινότερου γείτονα (nearest-neighbor-classifier). Υποθέτουμε τώρα πως δειγματοληπτούμε ένα επιπλέον test σημείο για μια από τις δύο κατηγορίες, π.χ.* ω_1*. Βρείτε το αναμενόμενο σφάλμα (expected error)* $P_1(e)$ *(ο δείκτης υποδηλώνει το πλήθος σημείων) μέσω ολοκλήρωσης.*

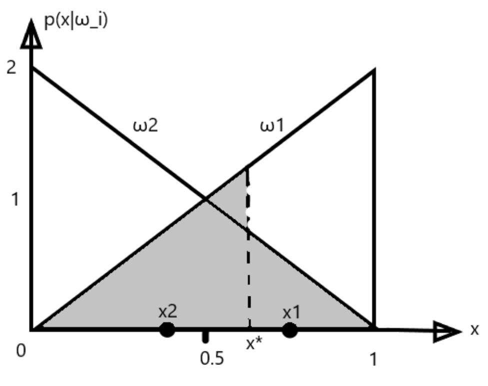
3. Επαναλάβετε τη διαδικασία για δύο σημεία εκπαίδευσης (training) από κάθε κατηγορία και ένα μόνο σημείο αποτίμησης (test point) και βρείτε το $P_2(e)$.

4. Γενικεύστε για n το πλήθος σημεία, ώστε να βρείτε το $P_n(e)$.

5. Μελετήστε τι συμβαίνει στο όριο $\lim_{n\to\infty} P_n(e)$ *σε σχέση με το Bayes error.*

Λύση:

Έστω ότι $p[\omega_1] = p[\omega_2] = \frac{1}{2}$ και ότι οι κατανομές που δίνονται είναι οι ακόλουθες.



1. Παρατηρώντας το παραπάνω διάγραμμα, από την συμμετρία του προβλήματος προκύπτει ότι το Bayes decision boundary είναι x* = 0.5.

Επιπλέον, το σφάλμα είναι η περιοχή που φαίνεται σκιασμένη με σκούρο γκρι χρώμα, δια ολόκληρη την περιοχή πιθανότητα. Έτσι,

$$\mathrm{P}^* = \int_0^1 \min[\mathrm{P}(\omega_1)\mathrm{p}(\mathrm{x}|\omega_1), \mathrm{P}(\omega_2)\mathrm{p}(\mathrm{x}|\omega_2)]\,\mathrm{dx} =$$

$$\mathrm{P}(\omega_1)\int_0^{0.5} 2\mathrm{xdx} + \mathrm{P}(\omega_2)\int_{0.5}^1 (2-2\mathrm{x})\mathrm{dx} = 0.5\frac{1}{4} + 0.5\frac{1}{4} = \frac{1}{4}$$

2. Έστω ότι επιλέγεται τυχαία ένα σημείο από την class ω_1 σύμφωνα με την $\mathrm{p}(\mathrm{x}|\omega_1)$ και ένα άλλο σημείο από την class ω_2 σύμφωνα με την $\mathrm{p}(\mathrm{x}|\omega_2)$. Τότε, το error θα είναι:

$$\mathrm{P}_1(\mathrm{e}) = \mathrm{P}(\omega_1)\int_0^1 \mathrm{p}(\mathrm{x}|\omega_1)\mathrm{p}(\mathrm{y} \in \omega_2)\mathrm{dx} + \mathrm{P}(\omega_2)\int_0^1 \mathrm{p}(\mathrm{x}|\omega_2)\mathrm{p}(\mathrm{y} \in \omega_1)\mathrm{dx} =$$

$$= \int_0^1 \mathrm{p}(\mathrm{x}|\omega_2)\mathrm{p}(\mathrm{y} \in \omega_1)\mathrm{dx} = \int_0^1 \mathrm{p}(\mathrm{x}|\omega_2)\int_0^1 \mathrm{p}(\omega_1|\mathrm{y}_1)\mathrm{dy}_1\mathrm{dx} = \frac{1}{2}$$

3. Από το ερώτημα (4.) προκύπτει ότι για n = 2:

$$\mathrm{P}_2(\mathrm{e}) = \frac{1}{3} + \frac{1}{(2+1)(2+3)} + \frac{1}{2(2+2)(2+3)} = \frac{51}{120} = 0.425$$

4. Η γενίκευση προκύπτει παρατηρώντας από την συμμετρία ότι το test point ανήκει στην class ω_2. Σε αυτή την περίπτωση, η πιθανότητα του error είναι η πιθανότητα όπου το κοντινότερο point του δείγματος του test point είναι στο ω_1. Έτσι, η πιθανότητα του error θα είναι:

$$\mathrm{P}_\mathrm{n}(\mathrm{e}) = \int_0^1 \mathrm{P}(\mathrm{x}|\omega_2)\mathrm{P}[\text{κοντινότερο } \mathrm{y}_\mathrm{i} \text{ στο x που είναι στην } \omega_1]\mathrm{dx} =$$

$$\int_0^1 \mathrm{P}(\,\mathrm{x}|\omega_2)\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P}\big[\mathrm{y}_\mathrm{i} \in \omega_1 \text{ και } \mathrm{y}_\mathrm{i} \text{ πιο κοντά στο x από το } \mathrm{y}_\mathrm{j}, \forall \mathrm{i} \neq \mathrm{j}\big]\mathrm{dx} =$$

$$\int_0^1 \mathrm{P}(\mathrm{x}|\omega_2)\mathrm{nP}[\mathrm{y}_1 \in \omega_1 \text{ και } |\mathrm{y}_1 - \mathrm{x}| > |\mathrm{y}_\mathrm{i} - \mathrm{x}|, \forall \mathrm{i} > 1]\mathrm{dx} =$$

$$\int_0^1 \mathrm{P}(\mathrm{x}|\omega_2)\mathrm{n}\int_0^1 \mathrm{P}(\omega_1|\mathrm{y}_1)\mathrm{P}[|\mathrm{y}_\mathrm{i} - \mathrm{x}| > |\mathrm{y}_1 - \mathrm{x}|, \forall \mathrm{i} > 1]\mathrm{dy}_1\mathrm{dx} =$$

$$\int_0^1 \mathrm{P}(\mathrm{x}|\omega_2)\mathrm{n}\int_0^1 \mathrm{P}(\omega_1|\mathrm{y}_1)\mathrm{P}[|\mathrm{y}_2 - \mathrm{x}| > |\mathrm{y}_1 - \mathrm{x}|]^{\mathrm{n}-1}\mathrm{dy}\,\mathrm{dx}$$

Για τον υπολογισμό της πιθανότητα $P[|y_2 - x| > |y_1 - x|]$, χωρίζεται το ολοκλήρωμα σε 6 cases, οι οποίες αναλύονται στην συνέχεια.

Case 1: $x \in [0, 0.5]$, $0 < y_1 < x$: $P[|y_2 - x| > |y_1 - x|] = 1 + 2y_1 - 2x$

Case 2: $x \in [0, 0.5]$, $0 < y_1 < 2x$: $P[|y_2 - x| > |y_1 - x|] = 1 - 2y_1 + 2x$

Case 3: $x \in [0, 0.5]$, $2x < y_1 < 1$: $P[|y_2 - x| > |y_1 - x|] = 1 - y_1$

Case 4: $x \in [0.5, 1]$, $0 < y_1 < 2x\text{-}1$: $P[|y_2 - x| > |y_1 - x|] = y_1$

Case 5: $x \in [0.5, 1]$, $2x\text{-}1 < y_1 < x$: $P[|y_2 - x| > |y_1 - x|] = 1 + 2y_1 - 2x$

Case 6: $x \in [0.5, 1]$, $x < y_1 < 1$: $P[|y_2 - x| > |y_1 - x|] = 1 - 2y_1 + 2x$

Έπειτα, αντικαθίστανται αυτές οι τιμές στο παραπάνω ολοκλήρωμα και έτσι σπάει σε 6 επιμέρους cases ως εξής:

$$\begin{aligned}
P_n(e) = \int_0^{0.5} P(x|\omega_2) n \Bigg[& \int_0^x P(\omega_1|y_1)(1 + 2y_1 - 2x)^{n-1}\, dy_1 + \int_x^{2x} P(\omega_1|y_1)(1 + 2x - 2y_1)^{n-1} dy_1 \\
& + \int_{2x}^1 P(\omega_1|y_1)(1 - y_1)^{n-1} dy_1 \Bigg] dx \\
& + \int_{0.5}^1 P(x|\omega_2) n \Bigg[\int_0^{2x-1} P(\omega_1|y_1) y_1^{n-1} dy_1 + \int_{2x-1}^x P(\omega_1|y_1)(1 + 2y_1 - 2x)^{n-1} dy_1 \\
& + \int_x^1 P(\omega_1|y_1)(1 + 2x - 2y_1)^{n-1} dy_1 \Bigg] dx
\end{aligned}$$

Η πυκνότητα και η posterior πιθανότητα δίνονται ίσες με $p(x|\omega_2) = 2(1 - x)$ και $P(\omega_2|y) = y$ για $x \in [0, 1]$ και $y \in [0, 1]$. Οπότε, αντικαθιστώντας αυτές τις τιμές στο παραπάνω ολοκλήρωμα, προκύπτει:

$$\begin{aligned}
P_n(e) = \int_0^{0.5} 2n(1 - x) \Bigg[& \int_0^x y_1(1 + 2y_1 - 2x)^{n-1}\, dy_1 + \int_x^{2x} y_1(1 + 2x - 2y_1)^{n-1} dy_1 \\
& + \int_{2x}^1 y_1(1 - y_1)^{n-1} dy_1 \Bigg] dx
\end{aligned}$$

$$+ \int_{0.5}^1 2n(1 - x) \left[\int_0^{2x-1} y_1^n dy_1 + \int_{2x-1}^x y_1(1 + 2y_1 - 2x)^{n-1} dy_1 + \int_x^1 y_1(1 + 2x - 2y_1)^{n-1} dy_1 \right] dx$$

Στην συνέχεια, πρέπει να υπολογιστούν τα εξής ολοκληρώματα με διαφορετικά όρια.

Το πρώτο είναι: $\int_a^b y_1(1+2y_1-2x)^{n-1}dy_1$

Έστω η συνάρτηση $h(y_1) = 1+2y_1-2x \Rightarrow y_1 = \frac{h+2x-1}{2}$ & $dy_1 = \frac{dh}{2}$. Έτσι, προκύπτει ότι:

$$\int_a^b y_1(1+2y_1-2x)^{n-1}dy_1 = \frac{1}{4}\int_{h(a)}^{h(b)} (h+2x-1)h^{n-1}dh = \frac{1}{4}[\frac{2x-1}{n}h^n + \frac{1}{n+1}u^{n+1}]_{h(a)}^{h(b)}$$

Το δεύτερο είναι: $\int_a^b y_1(1+2x-2y_1)^{n-1}dy_1$

Έστω η συνάρτηση $h(y_1) = 1+2x-2y_1 \Rightarrow y_1 = \frac{1+2x-h}{2}$ & $dy_1 = -\frac{dh}{2}$. Έτσι, προκύπτει ότι:

$$\int_a^b y_1(1+2x-2y_1)^{n-1}dy_1 = -\frac{1}{4}\int_{h(a)}^{h(b)} (1+2x-h)h^{n-1}dh = -\frac{1}{4}[\frac{2x+1}{n}h^n - \frac{1}{n+1}u^{n+1}]_{h(a)}^{h(b)}$$

Έτσι, χρησιμοποιώντας τα παραπάνω ολοκληρώματα, μπορεί κανείς να υπολογίσει τρία από τα 6 επιμέρους ολοκληρώματα του $P_n(e)$:

$$\int_0^x y_1(1+2y_1-2x)^{n-1}dy_1 = \frac{1}{4}[\frac{2x-1}{n}h^n + \frac{1}{n+1}h^{n+1}]_{h(x)=1}^{h(o)=1-2x} =$$

$$\frac{1}{4}\left(\frac{2x+1}{n} + \frac{1}{n+1}\right) + \frac{1}{4}(1-2x)^{n+1}(\frac{1}{n} - \frac{1}{n+1})$$

$$\int_x^{2x} y_1(1+2x-2y_1)^{n-1}dy_1 = -\frac{1}{4}[\frac{2x+1}{n}h^n - \frac{1}{n+1}h^{n+1}]_{h(x)=1}^{h(2x)=1-2x} =$$

$$\frac{1}{4}\left(\frac{2x+1}{n} - \frac{1}{n+1}\right) - \frac{1}{2n}(1-2x)^n + \frac{1}{4}(1-2x)^{n+1}(\frac{1}{n} + \frac{1}{n+1})$$

$$\int_{2x}^1 y_1(1-y_1)^{n-1}dy_1 = \int_0^{1-2x} (1-h)h^{n-1}dh = [\frac{1}{n}h^n - \frac{1}{n+1}h^{n+1}]_0^{1-2x} =$$

$$\frac{1}{n}(1-2x)^n - \frac{1}{n+1}(1-2x)^{n+1}$$

Έπειτα, προσθέτοντας αυτά τα τρία ολοκληρώματα και μετά από τον υπολογισμό του αθροίσματος προκύπτει:

$$\frac{x}{n} + \frac{1}{2n}(1-2x)^n + \left(\frac{1}{2n} - \frac{1}{n+1}\right)(1-2x)^{n+1}$$

Ύστερα πραγματοποιείται ο υπολογισμός των άλλων 3 επιμέρους ολοκληρωμάτων ως εξής:

$$\int_0^{2x-1} y_1^n dy_1 = \frac{1}{n+1}(2x-1)^{n+1}$$

$$\int_{2x-1}^{x} y_1(1+2y_1-2x)^{n-1}dy_1 = \frac{1}{4}[\frac{2x-1}{n}h^n + \frac{1}{n+1}h^{n+1}]_{h(2x-1)=2x-1}^{h(x)=1} =$$
$$\frac{1}{4}\left(\frac{2x-1}{n}+\frac{1}{n+1}\right) - \frac{1}{4}(2x-1)^{n+1}(\frac{1}{n}+\frac{1}{n+1})$$

$$\int_{x}^{1} y_1(1+2x-2y_1)^{n-1}dy_1 = -\frac{1}{4}[\frac{2x+1}{n}h^n - \frac{1}{n+1}h^{n+1}]_{h(x)=1}^{h(1)=2x-1} =$$
$$\frac{1}{4}\left(\frac{2x+1}{n}-\frac{1}{n+1}\right) - \frac{1}{2n}(2x-1)^n - \frac{1}{4}(2x-1)^{n+1}(\frac{1}{n}-\frac{1}{n+1})$$

Μετά, προσθέτοντας αυτά τα τρία ολοκληρώματα και μετά από τον υπολογισμό του αθροίσματος προκύπτει:

$$\frac{x}{n} - \frac{1}{2n}(2x-1)^n - \left(\frac{1}{2n}-\frac{1}{n+1}\right)(2x-1)^{n+1}$$

Έτσι, το sum που ακολουθεί θα είναι ίσο με:

$$\int_0^{0.5} 2n(1-x)\left[\int_0^{x} y_1(1+2y_1-2x)^{n-1}\,dy_1 + \int_x^{2x} y_1(1+2x-2y_1)^{n-1}dy_1 + \int_{2x}^{1} y_1(1-y_1)^{n-1}dy_1\right]dx =$$

$$\int_0^{0.5} 2n(1-x)\left[\frac{x}{n}+\frac{1}{2n}(1-2x)^n + \left(\frac{1}{2n}-\frac{1}{n+1}\right)(1-2x)^{n+1}\right]dx$$

Και θέτοντας $h = 1-2x \Rightarrow x = \frac{1-h}{2}$ & $dx = -\frac{dh}{2}$ & $1-x = \frac{1+h}{2}$ προκύπτει:

$$\int_0^1 n\left(\frac{1+h}{2}\right)[\frac{1-h}{2n}+\frac{1}{2n}h^n + \left(\frac{1}{2n}-\frac{1}{n+1}\right)h^{n+1}]dh$$

Παρομοίως, το ακόλουθο θα είναι ίσο με:

$$\int_{0.5}^{1} 2n(1-x)\left[\int_0^{2x-1} y_1^n dy_1 + \int_{2x-1}^{x} y_1(1+2y_1-2x)^{n-1}dy_1 + \int_x^1 y_1(1+2x-2y_1)^{n-1}dy_1\right]dx$$
$$= \int_{0.5}^{1} 2n(1-x)\left[\frac{x}{n}-\frac{1}{2n}(2x-1)^n - \left(\frac{1}{2n}-\frac{1}{n+1}\right)(2x-1)^{n+1}\right]dx$$

Και θέτοντας $h = 2x - 1 \Rightarrow x = \frac{1+h}{2}$ & $dx = \frac{dh}{2}$ & $1 - x = \frac{1-h}{2}$ προκύπτει:

$$\int_0^1 n(\frac{1-h}{2})\left[\frac{1+h}{2n} - \frac{1}{2n}h^n - \left(\frac{1}{2n} - \frac{1}{n+1}\right)h^{n+1}\right]dh$$

Τελικά, θα ισχύει ότι:

$$P_n(e) = \int_0^1 n\left(\frac{1+h}{2}\right)\left[\frac{1-h}{2n} + \frac{1}{2n}h^n + \left(\frac{1}{2n} - \frac{1}{n+1}\right)h^{n+1}\right]dh$$
$$+ \int_0^1 n\left(\frac{1-h}{2}\right)\left[\frac{1+h}{2n} - \frac{1}{2n}h^n - \left(\frac{1}{2n} - \frac{1}{n+1}\right)h^{n+1}\right]dh =$$

$$n\int_0^1 \frac{1}{2n}(1-h^2) + \frac{1}{2n}h^{n+1} + \left(\frac{1}{2n} - \frac{1}{n+1}\right)h^{n+2}dh =$$

$$n[\frac{1}{2n}\left(h - \frac{h^3}{3}\right)\frac{1}{2n(n+2)}h^{n+2} + \frac{1}{n+3}\left(\frac{1}{2n} - \frac{1}{n+1}\right)h^{n+3}]_0^1 =$$

$$n\left[\frac{1}{2n}\frac{2}{3} + \frac{1}{2n(n+2)} + \frac{1}{n+3}\left(\frac{1-n}{2n(n+1)}\right)\right] = \frac{1}{3} + \frac{(n+1)(n+3) - (n-1)(n+2)}{2(n+1)(n+2)(n+3)} =$$

$$\frac{1}{3} + \frac{3n+5}{2(n+1)(n+2)(n+3)} = \frac{1}{3} + \frac{1}{(n+1)(n+3)} + \frac{1}{2(n+2)(n+3)}$$

Παρατηρείται ότι αυτό συγκλίνει στο $\frac{1}{n^2}$.

Μπορεί να επαληθεύσει κανείς αυτό το αποτέλεσμα και για την περίπτωση του $n = 1$. Είναι αναμενόμενο το $P_1(e) = \frac{1}{2}$ σε αυτή την περίπτωση δεδομένου ότι η true label στο test point είτε θα κάνει match είτε mismatch με το μοναδικό sample point με ίση πιθανότητα, και ο παραπάνω τύπος το επιβεβαιώνει, καθώς:

$$P_1(e) = \frac{1}{3} + \frac{1}{(1+1)(1+3)} + \frac{1}{2(1+2)(1+3)} = \frac{1}{3} + \frac{1}{8} + \frac{1}{24} = \frac{1}{2}$$

5. Το όριο για άπειρα δεδομένα θα είναι απλώς: $\lim_{n\to\infty} P_n(e) = \frac{1}{3}$

Παρατηρείται ότι είναι μεγαλύτερο από το Bayes error, όπως είναι αναμενόμενο. Μάλιστα, αυτή η λύση δείχνει κατά τα όρια, όπως προκύπτουν από την θεωρία. Δηλαδή:

$$P^* \leq P \leq P^*(2 - 2P^*) \Rightarrow \frac{1}{4} \leq \frac{1}{3} \leq \frac{3}{8}$$

Άσκηση 1.8 (EM)

Θεωρήστε τα δεδομένα $D = \left\{\binom{1}{2}, \binom{4}{5}, \binom{1}{*}\right\}$*, όπου το * υποδηλώνει μια άγνωστη τιμή χαρακτηρισμού. Υποθέτουμε πως τα σημεία αυτά έχουν προέρθει από μια διδιάστατη διαχωρίσιμη κατανομή* $p(x_1, x_2) = p(x_1)p(x_2)$*, με*

$$p(x_1) \sim \begin{cases} \frac{1}{\theta_1} e^{(-\theta_1 x_1)}, if\ x_1 \geq 0 \\ 0, otherwise \end{cases}$$

$$p(x_2) \sim U(0, \theta_2) = \begin{cases} \frac{1}{\theta_2}, if\ 0 \leq x_2 \leq \theta_2 \\ 0, otherwise \end{cases}$$

1. Ξεκινώντας από μια αρχική εκτίμηση $\theta^0 = \binom{2}{3}$ *υπολογίστε αναλυτικά το* $Q(\theta, \theta^0)$ *που αποτελεί το βήμα Ε του αλγορίθμου ΕΜ.*

2. Βρείτε το θ που μεγιστοποιεί το $Q(\theta, \theta^0)$ *που αποτελεί το βήμα Μ του αλγορίθμου ΕΜ.*

3. Σχεδιάστε ένα γράφημα που να δείχνει την κατανομή $p(x_1, x_2)$ *πριν και μετά την εκτίμηση των παραμέτρων.*

Λύση:

Το data set είναι $D = \left\{\binom{1}{2}, \binom{4}{5}, \binom{1}{*}\right\}$, όπου το * υποδηλώνει μια άγνωστη τιμή για το στοιχείο x_2 του τρίτου data point.

1. Για το E step του αλγορίθμου EM θα ισχύει:

$$Q(\theta; \theta^0) = E_{x_{32}}\left[\ln p(D; \theta) | \theta^0, D_g\right] = E_{x_{32}}\left[\ln p(x_g, x_b; \theta) | \theta^0, D_g\right] =$$

$$\int_{-\infty}^{+\infty} \left[\sum_{i=1}^{3} \ln p(x_i|\theta) + \ln p(x_3|\theta)\right] p(x_{32}|x_{31} = 2; \theta^0) dx_{32} =$$

$$\sum_{i=1}^{2} \ln p(x_i|\theta) + \int_{-\infty}^{+\infty} \ln p\left(\begin{bmatrix} 1 \\ x_{32} \end{bmatrix} \middle| \theta\right) \cdot \frac{p\left(\binom{1}{x_{32}} \middle| \theta^0\right)}{\int_{-\infty}^{+\infty} p\left(\binom{1}{x'_{32}} \middle| \theta^0\right) d'_{x_{32}}} dx_{32} =$$

$$\sum_{i=1}^{2} \ln p(x_i|\theta) + \int_{-\infty}^{+\infty} \ln p\left(\begin{bmatrix} 1 \\ x_{32} \end{bmatrix} \middle| \theta\right) p_{x_{32}}(x_{32}|\theta^0) dx_{32} =$$

$$\sum_{i=1}^{2} \ln p(x_i|\theta) + \int_{0}^{\min(\theta_2, 3)} \ln\left(\frac{1}{\theta_1 \theta_2} e^{-1/\theta_1}\right) \cdot \frac{1}{3} dx_{32}$$

Οπότε, υπάρχουν τα εξής cases:

1ο case: $\min(\theta_2, 3) = \theta_2$:

$$Q(\theta;\theta^0) = \sum_{i=1}^{2} \ln p(x_i|\theta) + \frac{\theta_2}{3} \ln\left(\frac{1}{\theta_1\theta_2} e^{-\frac{1}{\theta_1}}\right)$$

2ο case: $\min(\theta_2, 3) = 3$:

$$Q(\theta;\theta^0) = \sum_{i=1}^{2} \ln p(x_i|\theta) + \ln\left(\frac{1}{\theta_1\theta_2} e^{-\frac{1}{\theta_1}}\right)$$

Όπου, χρησιμοποιήθηκε το γεγονός ότι $p\left(\begin{pmatrix} 1 \\ x_{32} \end{pmatrix} \middle| \theta^0\right) \neq 0$ για $0 \leq x_{32} \leq \theta_2$ και $p_{x_{32}}(x_{32}|\theta^0) \neq 0$ για $0 \leq x_{32} \leq \theta_2 = 3$.

Επίσης, ισχύει ότι

$$\sum_{i=1}^{2} \ln p(x_i|\theta) = \ln \frac{1}{\theta_1\theta_2} e^{-1/\theta_1} + \ln \frac{1}{\theta_1\theta_2} e^{-4/\theta_1} = 2 \ln \frac{1}{\theta_1\theta_2} e^{-5/\theta_1}$$

Οπότε, τελικά θα ισχύει ότι:

1ο case: $2 \leq \theta_2 \leq 3$:

$$Q(\theta;\theta^0) = 2 \ln\left(\frac{1}{\theta_1\theta_2} e^{-\frac{5}{\theta_1}}\right) + \frac{\theta_2}{3} \ln\left(\frac{1}{\theta_1\theta_2} e^{-\frac{1}{\theta_1}}\right)$$

2ο case: $\theta_2 > 3$:

$$Q(\theta;\theta^0) = 3 \ln\left(\frac{1}{\theta_1\theta_2} e^{-\frac{6}{\theta_1}}\right)$$

Παρατήρηση:

Το $Q(\theta;\theta^0)$ ίσως και να μπορεί να υπολογιστεί και όπως φαίνεται στην συνέχεια (?).

$$Q(\theta;\theta^0) = \ln p(x_1|\theta) + \ln p(x_2|\theta) + 2e^2 \int_{-\infty}^{+\infty} \ln p\left(\begin{pmatrix}1\\x_{32}\end{pmatrix}\middle|\theta\right)\cdot p\left(\begin{pmatrix}1\\x_{32}\end{pmatrix}\middle|\theta^0\right) dx_{32} =$$

$$\ln p(x_1|\theta) + \ln p(x_2|\theta) + 2e^2 \int_0^{\min(\theta_2,3)} \ln p\left(\begin{pmatrix}1\\x_{32}\end{pmatrix}\middle|\begin{matrix}\theta_1\\\theta_2\end{matrix}\right)\cdot p\left(\begin{pmatrix}1\\x_{32}\end{pmatrix}\middle|\theta^0 = \begin{pmatrix}2\\3\end{pmatrix}\right) dx_{32}$$

Οπότε, προκύπτουν τα εξής cases:

1º case: $\theta_2 \leq 2$:

$$Q(\theta;\theta^0) = 0$$

2º case: $2 < \theta_2 \leq 3$:

$$Q(\theta;\theta^0) = \ln p(x_1|\theta) + 2e^2 \int_0^{\theta_2} \ln p\left(\begin{pmatrix}1\\x_{32}\end{pmatrix}\middle|\begin{matrix}\theta_1\\\theta_2\end{matrix}\right)\cdot p\left(\begin{pmatrix}1\\x_{32}\end{pmatrix}\middle|\theta^0 = \begin{pmatrix}2\\3\end{pmatrix}\right) dx_{32} =$$

$$\ln p(x_1|\theta) + \frac{\theta_2}{3}\ln\left(\frac{1}{\theta_1\theta_2}e^{-\theta_1}\right)$$

3º case: $\theta_2 > 3$:

$$Q(\theta;\theta^0) = \ln p(x_1|\theta) + \ln p(x_2|\theta) + 2e^2 \int_0^{3} \ln p\left(\begin{pmatrix}1\\x_{32}\end{pmatrix}\middle|\begin{matrix}\theta_1\\\theta_2\end{matrix}\right)\cdot p\left(\begin{pmatrix}1\\x_{32}\end{pmatrix}\middle|\theta^0 = \begin{pmatrix}2\\3\end{pmatrix}\right) dx_{32} =$$

$$\ln p(x_1|\theta) + \ln p(x_2|\theta) + \ln\left(\frac{1}{\theta_1\theta_2}e^{-\theta_1}\right)$$

Το αποτέλεσμα αυτό είναι σε συνάρτηση με το θ_1, θ_2 για τα διάφορα cases.

Σημειώνεται ότι η συνθήκη κανονικοποίησης είναι $\int_{-\infty}^{+\infty} p(x_1)dx_1 = 1 \Rightarrow \int_{-\infty}^{+\infty} \frac{1}{\theta_1} e^{-\theta_1 x_1} dx_1 = 1$ και έχει λύση για $\theta_1 = 1$.

2. Σε αυτό το βήμα, αναζητείται το μέγιστο της Q. Έτσι,

1º case:

$$\frac{\partial Q}{\partial \theta_2} = -\frac{2}{\theta_2} - \frac{1}{3} - \frac{\theta_2}{3\theta_1} + \frac{\ln\frac{1}{\theta_1\theta_2}}{3}$$

2º case:

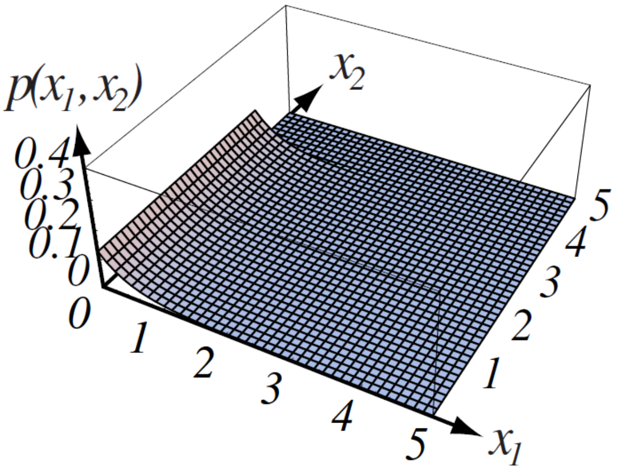
$$\frac{\partial Q}{\partial \theta_2} = -\frac{3}{\theta_2}$$

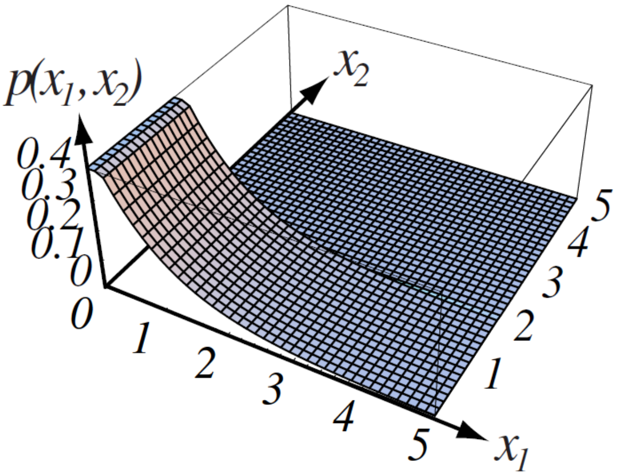
Παρατηρείται από το D ότι θα πρέπει $\theta_2 \geq 5$.

Επίσης, επειδή η Q είναι γνησίως φθίνουσα, αφού $\frac{\partial Q}{\partial \theta_2} < 0$, το μέγιστό της θα βρίσκεται για $\theta_2 = 5$. Οπότε, για να βρεθεί και το θ_1, θα πρέπει:

$$\frac{\partial Q}{\partial \theta_1} = -\frac{3}{\theta_1} + \frac{6}{\theta_1^2} = 0 \Rightarrow \theta_1 = 2$$

3. Το γράφημα που δείχνει την κατανομή $p(x_1, x_2)$ πριν και μετά την εκτίμηση των παραμέτρων είναι το ακόλουθο:





Resources:

1. [1] Γ. Καραγιάννης και Γ. Σταϊνχάουερ, Αναγνώριση Προτύπων και Μάθηση Μηχανών, ΕΜΠ, 2001.
2. [2] R. O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, Wiley, 2001.
3. [3] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
4. [4] S. Theodoridis and K. Koutroumbas, Pattern Recognition, 4th Edition Academic Pres, Elsevier, 2009. Ελληνική μετάφραση: απόδοση-επιμέλεια-πρόλογος ελληνικής έκδοσης Α.
5. Πικράκης, Κ. Κουτρουμπάς, Θ. Γιαννακόπουλος, Επιστημονικές Εκδόσεις Π.Χ. Πασχαλίδης-Broken Hill Publishers LTD, 2012.
6. https://math.stackexchange.com/questions/2470505/maximum-likelihood-for-an-erlang-distribution?fbclid=IwAR1g3WZhMW6OdaT3rYPC6_gOsYbNimOLhT4w6zlpgLf-6zbW9qRPUSu4lOA
7. https://courses.media.mit.edu/2010fall/mas622j/ProblemSets/ps4/ps4.pdf