Οι ακόλουθες λύσεις είναι συνοπτικές. Οι κανονικές λύσεις που παραδίδονται θα πρέπει να είναι πιο λεπτομερείς.

Άσκηση 1.1: (Discriminant functions for Bayesian classifier)
Κάνοντας απλή αντικατάσταση των δεδομένων στην εξίσωση $g(\mathbf{x}) = g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$:

$$(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{\mu}_j^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln\frac{P(\omega_i)}{P(\omega_j)} = 0$$

απ' όπου με κατάλληλες πράξεις προκύπτει η ζητούμενη μορφή του υπερεπιπέδου.

Άσκηση 1.2: (MLE-MAP)
α) Για τον εκτιμητή MLE της μέσης τιμής έχουμε ότι:

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\mathbb{R}} p(\theta|x_1, x_2, \ldots, x_N) = \cdots = \arg\max_{\theta\in\mathbb{R}} \prod_{k=1}^{N} p(x_k|\theta) = \cdots =$$
$$= \arg\min_{\theta\in\mathbb{R}} \sum_{k=1}^{N}(x_k - \theta)^2 = \cdots = \frac{1}{N}\sum_{k=1}^{N} x_k$$

β) Αντίστοιχα, για τον εκτιμητή MAP:

$$\hat{\theta}_{MAP} = \arg\max_{\theta\in\mathbb{R}} p(\theta|x_1, x_2, \ldots, x_N) = \cdots = \arg\max_{\theta\in\mathbb{R}} p(\theta)\prod_{k=1}^{N} p(x_k|\theta) = \cdots =$$
$$= \arg\min_{\theta\in\mathbb{R}} \left\{\frac{(\theta - v)^2}{2\beta^2} + \sum_{k=1}^{N}\frac{(x_k - \theta)^2}{2\sigma^2}\right\} = \cdots = \frac{\frac{1}{N}\sum_{k=1}^{N} x_k + \frac{\sigma^2 v}{N\beta^2}}{1 + \frac{\sigma^2}{N\beta^2}}$$

γ)

$$\lim_{N\to\infty} \hat{\theta}_{MAP} = \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} x_k = \lim_{N\to\infty} \hat{\theta}_{MLE}$$

Άσκηση 1.3: (Entropies of probability distributions)
α)

$$H_{\text{Gaussian}}[x] = \mathbb{E}\{-\ln p(x)\} = -\int_{-\infty}^{+\infty} p(x)\ln p(x)\,\mathrm{d}x = \cdots = \frac{1}{2}\ln(2\pi e\sigma^2)$$

$$H_{\text{Uniform}}[x] = \mathbb{E}\{-\ln p(x)\} = -\int_{-\infty}^{+\infty} p(x)\ln p(x)\,\mathrm{d}x = \cdots = \frac{1}{2}\ln\left\{(b-a)^2\right\} = \cdots = \frac{1}{2}\ln(12\sigma^2)$$

$$H_{\text{Triangular}}[x] = \mathbb{E}\{-\ln p(x)\} = -\int_{-\infty}^{+\infty} p(x)\ln p(x)\,\mathrm{d}x = \cdots = \frac{1}{2}\ln\left\{e\left(\frac{b-a}{2}\right)^2\right\} \overset{(*)}{=} \cdots = \frac{1}{2}\ln(6e\sigma^2)$$

(*) Εάν θεωρηθεί ότι η τριγωνική κατανομή είναι συμμετρική (ισοσκελής) στο διάστημα $[a, b]$.
β)

$$H_{\text{MVGaussian}}[x] = \mathbb{E}\{-\ln p(x)\} = \mathbb{E}\left\{\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} = \cdots =$$
$$= \frac{1}{2}\ln\left\{(2\pi)^d e^d|\boldsymbol{\Sigma}|\right\} = \frac{1}{2}\ln\left(|2\pi e\boldsymbol{\Sigma}|\right)$$

Άσκηση 1.4: (Bayesian parameter estimation)

α) Επειδή $\mathbf{a}^\top \mathbf{A}\mathbf{a} \in \mathbb{R}$ και $\mathrm{tr}[\mathbf{XY}] = \mathrm{tr}[\mathbf{YX}]$:

$$\mathbf{a}^\top \mathbf{A}\mathbf{a} = \mathrm{tr}\left[\mathbf{a}^\top \mathbf{A}\mathbf{a}\right] = \mathrm{tr}\left[\mathbf{A}\mathbf{a}\mathbf{a}^\top\right]$$

β) Χρησιμοποιώντας την ανεξαρτησία των n δειγμάτων και το (α) ερώτημα αποδεικνύεται το ζητούμενο.

γ) Χρησιμοποιώντας τις σχέσεις $\mathrm{tr}[\mathbf{A}] = \lambda_1 + \lambda_2 + \cdots + \lambda_d$ και $|\mathbf{A}| = \lambda_1\lambda_2 \ldots \lambda_d$ αποδεικνύεται το ζητούμενο.

δ) Αρχικά, εξισώνοντας με μηδέν τις παραγώγους $\frac{\partial p(\mathbf{x}_1,\ldots,\mathbf{x}_n|\Sigma)}{\partial \lambda_i}$ προκύπτει ότι $\lambda_1 = \cdots = \lambda_d = 1$. Στη συνέχεια, πρέπει να αποδειχθεί ότι ο πίνακας $\mathbf{A}$ διαγωνοποιείται, δηλαδή $\mathbf{A} = \mathbf{PDP}^{-1}$. Αυτό δεν είναι προφανές (hint: να χρησιμοποιηθεί το γεγονός ότι ο πίνακας $\hat{\mathbf{\Sigma}}$ είναι θετικά ορισμένος και συμμετρικός). Στη συνέχεια, επειδή έχουμε ήδη αποδείξει ότι $\mathbf{D} = \mathbf{I}$:

$$\mathbf{\Sigma}_{MLE}^{-1}\hat{\mathbf{\Sigma}} = \mathbf{A} = \mathbf{PDP}^{-1} = \cdots = \mathbf{I} \qquad \Longrightarrow \qquad \mathbf{\Sigma}_{MLE} = \hat{\mathbf{\Sigma}}$$

Άσκηση 1.5: (Probabilistic discriminative models: Logistic Regression)

Αντικαθιστούμε $y_n = \sigma(\mathbf{w}^\top \phi_n)$. Με απλή παραγώγιση της $E(\mathbf{w})$ ως προς $\mathbf{w}$ προκύπτει το ζητούμενο.

Άσκηση 1.6: (Perceptrons)

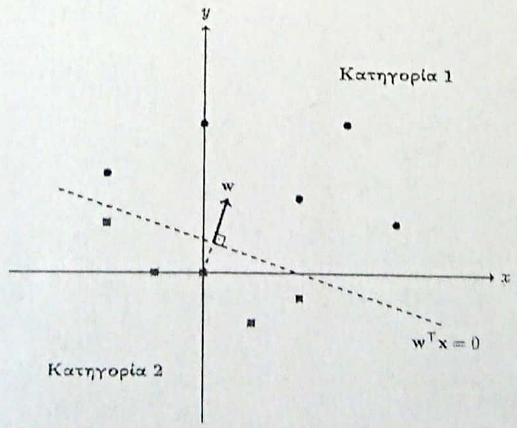
Προκειμένου να εκτελέσουμε τον reward & punishment αλγόριθμο, επιτρέπουμε στο διάνυσμα παραμέτρων $\mathbf{w} = [b \quad w_1 \quad w_2]^\top$ να έχει μία bias τιμή b και, αντίστοιχα, επεκτείνουμε κατά μία συνιστώσα τα διανύσματα των χαρακτηριστικών $\mathbf{x}_t = [1 \quad x_{t,1} \quad x_{t,2}]^\top$. Επομένως:

Για $t = 0$: $\quad \mathbf{w}(0) = [0 \quad 0 \quad 0]^\top \quad \mathbf{x}_0 = [1 \quad 2 \quad 1.5]^\top \quad \mathbf{w}(0)^\top \mathbf{x}_0 = 0 \quad \mathbf{w}_1 = [1 \quad 2 \quad 1.5]^\top$

Για $t = 1$: $\quad \mathbf{w}(1) = [1 \quad 2 \quad 1.5]^\top \quad \mathbf{x}_1 = [1 \quad 4 \quad 1]^\top \quad \mathbf{w}(1)^\top \mathbf{x}_1 = 10.5 \quad \mathbf{w}_2 = [1 \quad 2 \quad 1.5]^\top$

Συνεχίζοντας την επαναληπτική διαδικασία, προκύπτει ότι το τελικό διάνυσμα παραμέτρων είναι το $\mathbf{w} = [-2 \quad 1 \quad 3]^\top$, επομένως η ευθεία που διαχωρίζει τα δεδομένα μας είναι η:

$$-2 + x + 3y = 0$$



Σχήμα 1: Η τελική ευθεία που διαχωρίζει τα δεδομένα.

Οι ακόλουθες λύσεις είναι συνοπτικές. Οι κανονικές λύσεις που παραδίδονται θα πρέπει να είναι πιο λεπτομερείς.

Άσκηση 2.1: (Expectation-Maximization)

Από τα δεδομένα $\mathcal{D}$, διαχωρίζουμε τις μεταβλητές στα σύνολα $\mathcal{D}_{good} = \{x_1^{(1)}, x_2^{(1)}, x_1^{(2)}, x_2^{(2)}, x_1^{(3)}\}$ και $\mathcal{D}_{bad} = \{x_2^{(3)}\}$.

(α) E step: Υπολογισμός της εκτιμώμενης τιμής:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^0) = \mathbb{E}_{\mathcal{D}_{bad}}\{\ln p(\mathcal{D}; \boldsymbol{\theta} | \mathcal{D}_{good}; \boldsymbol{\theta}^0)\} = \int_{-\infty}^{+\infty} p(\mathcal{D}_{bad}|\mathcal{D}_{good}; \boldsymbol{\theta}^0) \ln p(\mathcal{D}; \boldsymbol{\theta})\, \mathrm{d}x_2^{(3)} =$$

$$= \int_{-\infty}^{+\infty} p(x_2^{(3)}; \theta_2^0) \ln\{p(x_1^{(1)}; \theta_1) p(x_2^{(1)}; \theta_2) p(x_1^{(2)}; \theta_1) p(x_2^{(2)}; \theta_2) p(x_1^{(3)}; \theta_1) p(x_2^{(3)}; \theta_2)\}\, \mathrm{d}x_2^{(3)} =$$

$$= \cdots = 3\ln\theta_1 - 3\ln\theta_2 - 7\theta_1$$

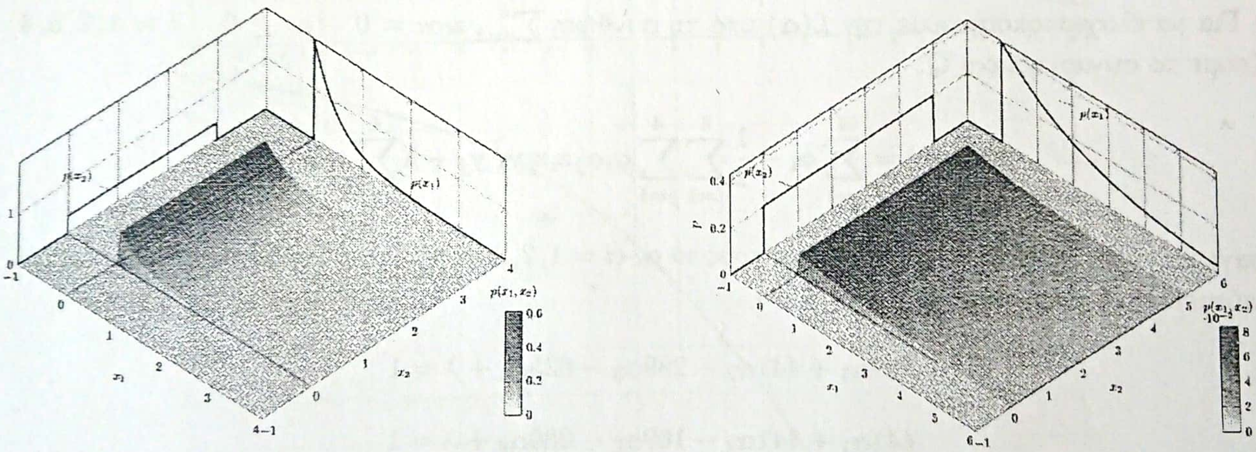
(β) M step: Μεγιστοποίηση της προηγούμενης έκφρασης ως προς θ_1 και θ_2 δίνει ότι:

$$\theta_1 = \frac{3}{7}$$

$$\theta_2 = \max_i x_2^{(i)} = 5$$

καθώς η Q δεν παρουσιάζει ακρότατο ως προς θ_2. Άρα $\boldsymbol{\theta} = (\frac{3}{7}, 5)^\top$.

(γ)



Σχήμα 1: Η κατανομή $p(x_1, x_2)$ πριν και μετά την εκτίμηση των παραμέτρων.

Άσκηση 2.2: (Basis vectors - Principal Component Analysis)

(α) Το διάνυσμα $\mathbf{x}$ γράφεται ως $\mathbf{x} = \sum_{i=0}^{N-1} y_i \mathbf{e}_i$, και επομένως $\hat{\mathbf{x}} = \sum_{i=0}^{m-1} y_i \mathbf{e}_i$, όπου $y_i = \mathbf{e}_i^\top \mathbf{x}$. Για το μέσο τετραγωνικό σφάλμα θα έχουμε ότι:

$$J = \mathbb{E}\left\{||\mathbf{x} - \hat{\mathbf{x}}||^2\right\} = \mathbb{E}\left\{(\mathbf{x} - \hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}})\right\} = \cdots = \sum_{i=0}^{N-1} \mathbb{E}\{y_i^2\} - \sum_{i=0}^{m-1} \mathbb{E}\{y_i^2\} = \sum_{i=m}^{N-1} \mathbb{E}\{y_i^2\} =$$

$$= \cdots = \sum_{i=m}^{N-1} \mathbf{e}_i^\top \mathbb{E}\{\mathbf{x}\mathbf{x}^\top\} \mathbf{e}_i = \sum_{i=m}^{N-1} \mathbf{e}_i^\top R_x \mathbf{e}_i$$

Προκειμένου να ελαχιστοποιηθεί το σφάλμα υπό τις συνθήκες $\mathbf{e}_i^\top \mathbf{e}_i = 1$, ορίζουμε τη συνάρτηση Lagrange:

$$L = \sum_{i=m}^{N-1} \mathbf{e}_i^\top R_x \mathbf{e}_i - \sum_{i=0}^{N-1} \lambda_i (\mathbf{e}_i^\top \mathbf{e}_i - 1)$$

Παραγωγίζοντας ως προς τα $\mathbf{e}_i$ προκύπτει ότι:

$$R_x \mathbf{e}_i = \lambda_i \mathbf{e}_i \qquad \forall i = m, \ldots, N-1$$

Επομένως τα $\mathbf{e}_i$ είναι τα ιδιοδιανύσματα του πίνακα R_x.

(β)

$$J = \sum_{i=m}^{N-1} \mathbf{e}_i^\top R_x \mathbf{e}_i = \sum_{i=m}^{N-1} \mathbf{e}_i^\top \lambda_i \mathbf{e}_i = \sum_{i=m}^{N-1} \lambda_i$$

Η παραπάνω ποσότητα περιέχει τις $N-m$ μικρότερες ιδιοτιμές. Άρα η βάση θα περιέχει τις m μεγαλύτερες ιδιοτιμές.

(γ)

$$\sum_{i=0}^{m-1} \mathrm{Var}\{y_i\} = \cdots = \sum_{i=0}^{m-1} \lambda_i$$

Εφ'όσον έχουν επιλεχθεί οι m μεγαλύτερες ιδιοτιμές για τη βάση, θα μεγιστοποιείται και το παραπάνω άθροισμα.

Άσκηση 2.3: (Support Vector Machines - SVM)

(β) Για να ελαχιστοποιήσουμε την $L(\boldsymbol{\alpha})$ υπό τη συνθήκη $\sum_{i=1}^{4} z_i \alpha_i = 0 \quad (\alpha_i \geq 0, \quad i = 1,2,3,4)$ ορίζουμε το συναρτησιακό Q:

$$Q(\boldsymbol{\alpha}, \lambda) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_i \alpha_j z_i z_j \mathbf{y}_i^\top \mathbf{y}_j + \lambda \sum_{i=1}^{4} z_i \alpha_i$$

Παραγωγίζοντας την παραπάνω σχέση ως προς τα α_i $(i = 1,2,3,4)$ και λ και εξισώνοντας με το μηδέν, προκύπτει το εξής σύστημα:

$$729\alpha_1 + 441\alpha_2 - 289\alpha_3 - 625\alpha_4 + \lambda = 1$$

$$441\alpha_1 + 441\alpha_2 - 169\alpha_3 - 289\alpha_4 + \lambda = 1$$

$$-289\alpha_1 - 169\alpha_2 + 121\alpha_3 + 225\alpha_4 - \lambda = 1$$

$$-625\alpha_1 - 289\alpha_2 + 225\alpha_3 + 729\alpha_4 - \lambda = 1$$

$$-\alpha_1 + -\alpha_2 + \alpha_3 + \alpha_4 = 0$$

με τελική τη λύση την

$$\begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \lambda \end{bmatrix}^\top = \begin{bmatrix} 0.0118 & 0.0039 & 0.0087 & 0.0070 & -2.4475 \end{bmatrix}^\top$$

η οποία ικανοποιεί όλους τους περιορισμούς που τέθηκαν.

(γ) Το ζητούμενο διάνυσμα βαρών $\mathbf{a}_r$ δίνεται από τη σχέση:

$$\mathbf{a}_r = \sum_{i=1}^{4} \alpha_i z_i \mathbf{y}_{i,r} = \begin{bmatrix} -\frac{17}{7544}\sqrt{2} & \frac{135}{7544}\sqrt{2} & -\frac{751}{7544}\sqrt{2} & -\frac{89}{7544} & -\frac{789}{7544} \end{bmatrix}^\top$$

Για την εύρεση του a_0 χρησιμοποιείται το γεγονός ότι ένα οποιοδήποτε διάνυσμα $\mathbf{y}_i$ αποτελεί support vector (αφού όλα τα $\alpha_i > 0$):

$$z_1[a_0 \quad \mathbf{a}_r]^\top \mathbf{y}_1 - 1 = 0$$

απ' όπου προκύπτει ότι $a_0 = \frac{2308}{943}$. Άρα, τελικά:

$$\mathbf{a} = \begin{bmatrix} \frac{2308}{943} & -\frac{17}{7544}\sqrt{2} & \frac{135}{7544}\sqrt{2} & -\frac{751}{7544}\sqrt{2} & -\frac{89}{7544} & -\frac{789}{7544} \end{bmatrix}^\top$$

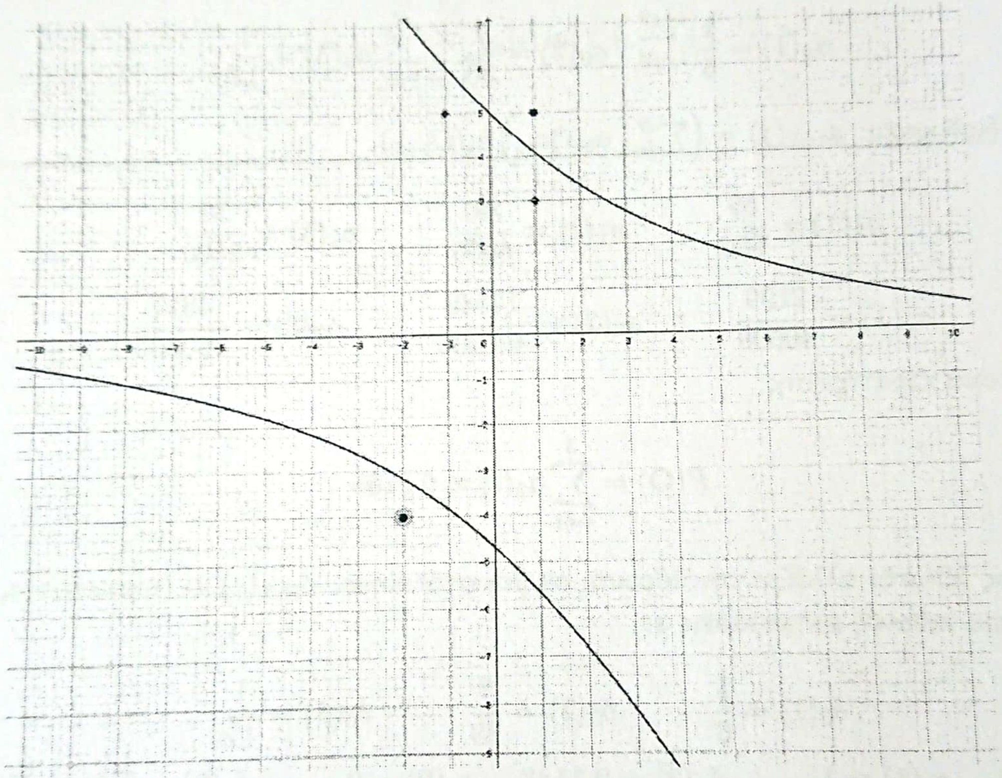
Τώρα ικανοποιούνται όλες οι προϋποθέσεις $z_i \mathbf{a}^\top \mathbf{y}_i \geq 1$ $(i = 1, \dots, 4)$.

δ)

$$\beta = \frac{1}{||\mathbf{a}_r||} = 5.6301$$

ε)

$$g(x_1, x_2) = \mathbf{a}^\top \phi(x_1, x_2) = \mathbf{a}^\top \begin{bmatrix} 1 & \sqrt{2}x_1 & \sqrt{2}x_2 & \sqrt{2}x_1x_2 & x_1^2 & x_2^2 \end{bmatrix}^\top$$



Σχήμα 2: Η καμπύλη διαχωρισμού $g(x_1, x_2) = 0$.

στ) Όλα τα δεδομένα $\mathbf{y}_i$ αποτελούν support vectors, διότι $\alpha_i > 0$ $(i = 1, \dots, 4)$.

Άσκηση 2.4: (Conditional independence)

α) Υπολογίζουμε με τη σειρά τις εξής πιθανότητες:

$$p(G = 1) = \cdots = 0.791$$

$$p(D = 0) = \cdots = 0.3254$$

$$p(G = 0|F = 0) = \cdots = 0.705$$

$$p(D = 0|F = 0) = \cdots = 0.623$$

Άρα τελικά:

$$p(F=0|D=0) = \frac{p(D=0|F=0)p(F=0)}{p(D=0)} = 0.383$$

β)

$$p(G=0|B=0) = \cdots = 0.76$$

$$p(D=0, B=0) = \cdots = 0.0328$$

Άρα τελικά:

$$p(F=0|D=0, B=0) = \frac{p(F=0)p(B=0)\sum_G p(D=0|G)p(G|F=0, B=0)}{p(D=0, B=0)} = 0.2073$$

Η δεύτερη πιθανότητα προκύπτει μικρότερη λόγω του **explaining away**.

Άσκηση 2.5: (Hidden Markov Models)

α) Forward αλγόριθμος - αρχικοποίηση: $a_0(i) = \pi_i b_i(O_0)$

$$a_0(1) = \frac{1}{6} \qquad a_0(2) = \frac{7}{30} \qquad a_0(3) = \frac{1}{12}$$

Επαναληπτική διαδικασία: $a_{t+1}(j) = \left(\sum_{i=1}^N a_t(i)a_{ij}\right) b_j(O_{t+1})$

$$a_1(1) = \frac{59}{600} \qquad a_1(2) = \frac{791}{6000} \qquad a_1(3) = \frac{59}{2400}$$

$$a_2(1) = \frac{2199}{40000} \qquad a_2(2) = \frac{12549}{400000} \qquad a_2(3) = \frac{4827}{160000}$$

Τελικό αποτέλεσμα παρατήρησης:

$$P(\mathbf{O}) = \sum_{i=1}^{3} a_2(i) \approx 0.1165$$

β) Ο αλγόριθμος Viterbi αλλάζει την άθροιση σε μεγιστοποίηση: $\delta_{t+1}(j) = (\max_i \delta_t(i)a_{ij})\, b_j(O_{t+1})$ και διατηρεί τις πιο πιθανές καταστάσεις s:

$$\delta_0(1) = \frac{1}{6} \qquad \delta_0(2) = \frac{7}{30} \qquad \delta_0(3) = \frac{1}{12}$$

$$\delta_1(1) = 0.0667, \quad s_1(1) = 1 \qquad \delta_1(2) = 0.1143, \quad s_1(2) = 2 \qquad \delta_1(3) = 0.0146, \quad s_1(3) = 3$$

$$\delta_2(1) = 0.0267, \quad s_2(1) = 1 \qquad \delta_2(2) = 0.0240, \quad s_2(2) = 2 \qquad \delta_2(3) = 0.0086, \quad s_2(3) = 2$$

Συνεπώς, η πιο πιθανή ακολουθία καταστάσεων είναι η $1 \to 1 \to 1$, με πιθανότητα 0.0267.

γ) Από την αποκωδικοποίηση Viterbi προκύπτει ότι η πιο πιθανή ακολουθία καταστάσεων είναι η $1 \to 1 \to 1 \to \cdots \to 1$ (E-step). Επομένως, οι μόνες παράμετροι του μοντέλου που μπορούν να υπολογιστούν είναι οι (M-step):

$$a_{11} = 1 \qquad a_{12} = 0 \qquad a_{13} = 0$$

$$p(O=H|q=1) = \frac{8}{15} \qquad p(O=T|q=1) = \frac{7}{15}$$

Επαναλαμβάνοντας τη διαδικασία E-M, καταλήγουμε και πάλι στην ίδια ακολουθία καταστάσεων, άρα ο αλγόριθμος εκπαίδευσης τερματίζει.

δ) Ο forward-backward αλγόριθμος είναι γενικά πιο αργός λόγω των δύο περασμάτων που εκτελεί, αλλά ταυτόχρονα και πιο ακριβής στα αποτελέσματα σε σχέση με τον pseudo-EM αλγόριθμο που εφαρμόστηκε.

Άσκηση 2.6: (Cross-entropy)

α) Έχουμε:

$$J = -\sum_{i=1}^{N}\sum_{k=1}^{k_L} y_k(i)\ln \hat{y}_k(i) - \sum_{i=1}^{N}\sum_{k=1}^{k_L} y_k(i)\ln y_k(i)$$

Για δυαδικές επιθυμητές τιμές εξόδου ο πρώτος όρος είναι 0 ενώ ο δεύτερος όρος έχει ελάχιστη τιμή μηδέν.

β) Έχουμε:

$$\hat{y}_k(i) = y_k(i) + e_k(i)$$

όπου $e_k(i)$ το σφάλμα. Με κατάλληλες πράξεις στη συνάρτηση εντροπίας βρίσκουμε:

$$J = -\sum_{i=1}^{N}\sum_{k=1}^{k_L} y_k(i)\ln(1 + \frac{e_k(i)}{y_k(i)})$$
