ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

(2021 – 2022)

2η Σειρά Αναλυτικών Ασκήσεων

Ονοματεπώνυμο:

- Χρήστος Τσούφης

Αριθμός Μητρώου:

- 031 17 176

Στοιχεία Επικοινωνίας:

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

Άσκηση 2.1 (SVM)

Μας δίνονται N = 8 διανύσματα χαρακτηριστικών $\widetilde{x_n} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ *που προέρχονται από δύο κλάσεις* ω_1 *και* ω_2*:*

$$\omega_1: \left\{ \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \begin{pmatrix} -3 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ 3 \end{pmatrix} \right\}, \qquad z_1 = z_2 = z_3 = z_4 = -1$$

$$\omega_2: \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}, \qquad z_5 = z_6 = z_7 = z_8 = 1$$

όπου οι συντελεστές $z_n = \pm 1$ *υποδεικνύουν την κατηγορία καθενός από τα δείγματα. Στην περίπτωση ενός ταξινομητή SVM, στόχος είναι η εύρεση του διανύσματος βάρους w με το ελάχιστο μήκος, το οποίο να υπόκειται στους περιορισμούς* $z_n w^T y_n \geq 1$ *(n = 1, …, N). Τα διανύσματα w και* y_n *είναι επαυξημένα κατά* w_0 *(*$w = [w_0 \ \tilde{w}]^T$*) και* $y_{n,0} = 1$ *(*$y_n = [1 \ \widetilde{y_n}]^T$*), αντίστοιχα.*

(α) *Αρχικά, ελέγξτε εάν οι δύο κλάσεις είναι γραμμικώς διαχωρίσιμες, μέσω του σχεδιασμού των παραπάνω σημείων σε ένα γράφημα. Στη συνέχεια, στην περίπτωση όπου δεν είναι γραμμικώς διαχωρίσιμες, μετατρέψτε τα διανύσματα* $\widetilde{x_n}$ *σε έναν χώρο υψηλότερων διαστάσεων,* $y_n = \varphi(\widetilde{x_n})$*, χρησιμοποιώντας την εξής μορφή φ-functions 2ης τάξης:* $\varphi(x_1, x_2) = \left[1 \ x_1 \ x_2 \ \frac{x_1^2 + x_2^2 - 5}{3}\right]^T$*.*

(β) *Να προσδιοριστούν οι συντελεστές* a_n *(n = 1, …, N) του προβλήματος ελαχιστοποίησης (Υποκεφάλαιο 5.11.1 [2], [5, SVM]). Η λύση που βρήκατε είναι δεκτή και αν ναι, γιατί;*

(γ) *Να υπολογιστεί το ζητούμενο διάνυσμα βαρών w. Επαληθεύστε ότι ισχύουν όλες οι προϋποθέσεις* $z_n w^T y_n \geq 1$ *(n = 1, …, N).*

(δ) *Υπολογίστε το περιθώριο β του ταξινομητή.*

(ε) *Βρείτε τη συνάρτηση διαχωρισμού* $g(x_1, x_2) = w^T \varphi(x_1, x_2)$ *στον αρχικό χώρο* $x_1 - x_2$ *και σχεδιάστε την καμπύλη* $g(x_1, x_2) = 0$ *μαζί με τα 8 αρχικά σημεία σε κάποιο εργαλείο σχεδιασμού σε Η/Υ.*

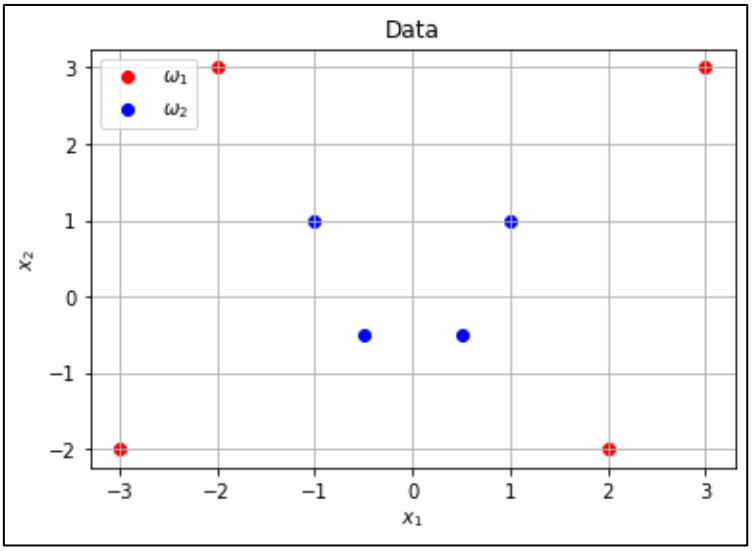
(στ) *Ποια είναι τα support vectors;*

(ζ) *Σε ποιες κατηγορίες ταξινομούνται τα σημεία* $\begin{pmatrix} \sqrt{2} \\ \sqrt{2} \end{pmatrix}$ *και* $\begin{pmatrix} -2 \\ 3 \end{pmatrix}$*;*

Σημείωση: Επεξηγήστε αναλυτικά τη διαδικασία που ακολουθήσατε για να φτάσετε στις λύσεις σας (π.χ. τις παραδοχές, τα σύμβολα και τους τύπους που τυχόν χρησιμοποιήσατε, τη σωστή αντικατάσταση μεταβλητών και δεικτών σε αυτούς, αριθμητικά αποτελέσματα ενδιάμεσων ή βοηθητικών μεταβλητών, κ.ά.). Στην περίπτωση επαναληπτικών αλγορίθμων, τα παραπάνω ισχύουν σε κάθε επανάληψη της εκτέλεσής τους.

Λύση:

(α) Για τον έλεγχο εάν οι δύο κλάσεις είναι γραμμικώς διαχωρίσιμες, σχεδιάστηκε το ακόλουθο διάγραμμα.



Εύκολα μπορεί να παρατηρήσει κανείς ότι τα σημεία δεν είναι γραμμικώς διαχωρίσιμα.

Συγκεκριμένα, υπάρχουν τα εξής σημεία στην ευθεία y = x, τα (3, 3), (1, 1), (-0.5, -0.5) με τα ακριανά να ανήκουν στην κλάση w_1 και τα εσωτερικά στην κλάση w_1. Είναι λοιπόν, προφανές ότι τα σημεία αυτά δεν μπορούν να χωριστούν με ευθεία γραμμή.

Εν συνεχεία, μετατρέπονται τα $\widetilde{x_n}$ σε έναν χώρο υψηλότερων διαστάσεων, $y_n = \varphi(\widetilde{x_n})$, χρησιμοποιώντας την εξής μορφή φ-functions 2ης τάξης:

$$\varphi(x_1, x_2) = \left[1 \;\; x_1 \;\; x_2 \;\; \frac{x_1^2 + x_2^2 - 5}{3}\right]^T$$

Οπότε, τα νέα δεδομένα θα είναι:

w_1:	(3, 3)	→	(1, 3, 3, 4.3)
	(2, 2)	→	(1, 2, -2, 1)
	(-3, -2)	→	(1, -3, -2, 2.7)
	(-2, 3)	→	(1, -2, 3, 2.7)
w_2:	(1, 1)	→	(1, 1, 1, -1)
	(0.5, -0.5)	→	(1, 0.5, -0.5, -1.5)
	(-0.5, -0.5)	→	(1, -0.5, -0.5, -1.5)
	(-1, 1)	→	(1, -1, 1, -1)

(β) Ο προσδιορισμός των συντελεστών a_n του προβλήματος ελαχιστοποίησης γίνεται ως εξής:

Είναι γνωστό ότι το width του margin ισούται με $\frac{1}{\|\vec{w}\|}$, οπότε για να μεγιστοποιηθεί το width αρκεί να ελαχιστοποιηθεί το $\|\vec{w}\|$ ή για ευκολία στις πράξεις, το $\frac{1}{2}\|\vec{w}\|^2$.

Ταυτόχρονα, για τα δεδομένα είναι επιθυμητό να ισχύει ότι:

$$z_i \cdot \tilde{w} \cdot \tilde{y_i} \geq 1 \Leftrightarrow z_i(w \cdot y_i + w_0) \geq 1$$

Έτσι, εφαρμόζεται η μέθοδος Πολλαπλασιαστών Lagrange:

$$L(w, a) = \frac{1}{2}\|\vec{w}\|^2 - \sum_i a_i(z_i\vec{w}\vec{y_i} + w_0 - 1)$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \vec{w} - \sum_i a_i z_i \vec{y_i} = 0 \Rightarrow \vec{w} = \sum_i a_i z_i \vec{y_i} \quad (1)$$

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow \sum_i a_i z_i = 0$$

Οπότε, αντικαθιστώντας τις παραπάνω σχέσεις στο L, προκύπτει:

$$L(w, a) = \frac{1}{2}\left(\sum_i a_i z_i \vec{y_i}\right)\left(\sum_j a_j z_j \vec{y_j}\right) - \sum_i a_i z_i \vec{y_i}\left(\sum_j a_j z_j \vec{y_j}\right) - \left(\sum_j a_j z_j \vec{y_j}\right) w_0 + \sum_i a_i \Rightarrow$$

$$\Rightarrow L(w, a) = \sum_i a_i - \frac{1}{2}\left(\sum_i \left(\sum_j a_i a_j z_i z_j \vec{y_i}\vec{y_j}\right)\right)$$

$$\text{subject to: } \sum_i a_i z_i = 0$$

Lagrangian Multiplier λ:

$$L(w, a) = \sum_i a_i - \frac{1}{2}\left(\sum_i \left(\sum_j a_i a_j z_i z_j \vec{y_i}\vec{y_j}\right)\right) - \lambda \sum_i a_i z_i$$

Έτσι, εξισώνοντας τις μερικές παραγώγους των a_i, $\frac{\partial L}{\partial a_i} = 0$, προκύπτει το παρακάτω γραμμικό σύστημα της μορφής: $A \cdot x = b$

Παρατηρείται ότι ο A δεν είναι αντιστρέψιμος, όμως το σύστημα έχει λύση καθώς το b ανήκει στον χώρος στηλών του A. Μάλιστα, έχει άπειρες λύσεις καθώς αν x_0 μια λύση του $A \cdot x = b$ τότε για κάθε λύση $\tilde{x}$ του ομογενούς συστήματος $A \cdot x = 0$, η $\tilde{x} + x_0$ είναι λύση του $A \cdot x = b$.

Τελικά, μια λύση του συστήματος θα είναι:

```
[-0.  0.14678   -0.  0.10013   0.16178   0.0107   0.0012   0.07323]
```

(γ) Ο υπολογισμός του διανύσματος βαρών w γίνεται ως εξής από την (1):

$$\vec{w} = \sum_i a_i z_i \vec{y_i}$$

Οπότε,

```
[0.   0.22223    -0.66665333]
```

Για το w_0, παίρνοντας έναν εκ των περιορισμών $z_i(\vec{w} \cdot \vec{y_i} + w_0) \geq 1$ για κάποιο y_i που είναι support vector, η σχέση θα ισχύει ως ισότητα:

$$z_1(\vec{w} \cdot \vec{y_1} + w_0) = 1 \Rightarrow w_0 = 0.11$$

Το επαυξημένο διάνυσμα βαρών είναι το:

```
[0.111     0.    0.22223    -0. 66665333]
```

(δ) Το περιθώριο β του ταξινομητή θα είναι $\beta = \frac{1}{|w|}$:

```
1.4056180082863914
```

(ε) Η συνάρτηση διαχωρισμού $g(x_1, x_2) = w^T \varphi(x_1, x_2)$ στον αρχικό χώρο $x_1 - x_2$ θα είναι:

$$g(x_1, x_2) = w^T \varphi(x_1, x_2) = [w_0\ w_1\ w_2\ w_3]^T \left[1\ x_1\ x_2\ \frac{x_1^2 + x_2^2 - 5}{3}\right] =$$

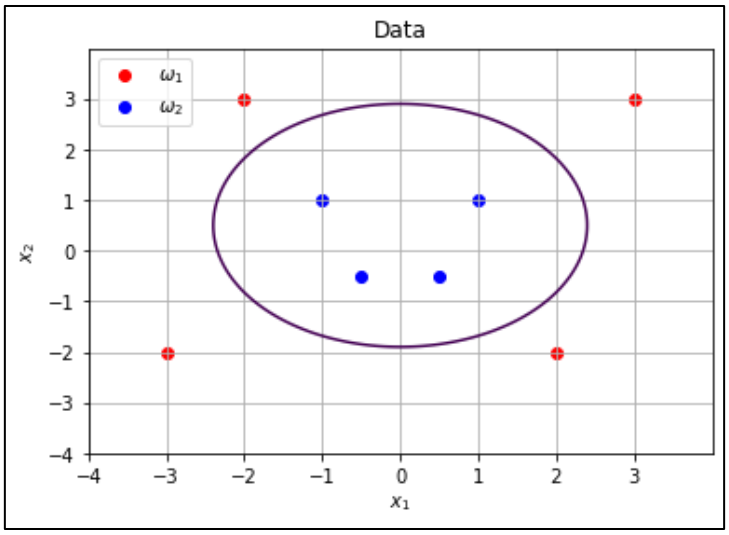
$$= w_0 + w_1 x_1 + w_2 x_2 + w_3 \frac{x_1^2 + x_2^2 - 5}{3} =$$

$$= \frac{w_3}{3} x_1^2 + \frac{w_3}{3} x_2^2 + w_1 x_1 + w_2 x_2 + w_0 - \frac{5 w_3}{3} =$$

$$= -0.222217777777778 x_1^2 - 0.222217777777778 x_2^2 + 0.22223 x_2 + 1.22208888888889$$

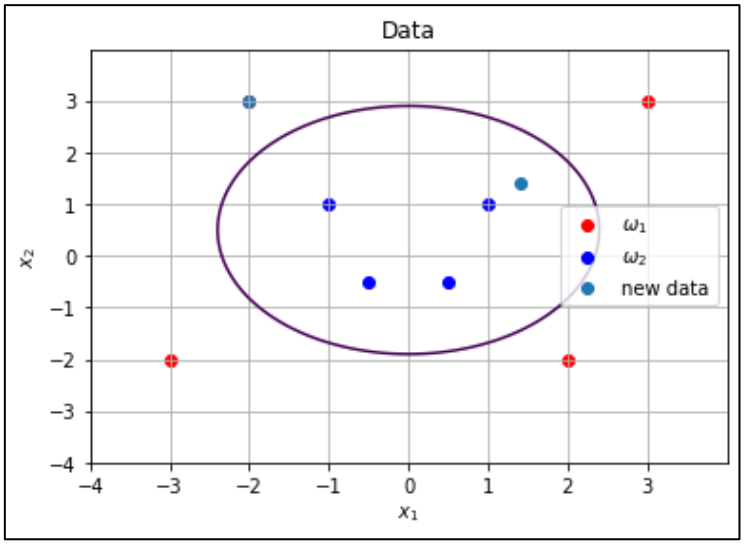
$$= -0.2223 x_1^2 - 0.2223 x_2^2 + 0.2224 x_2 + 1.222$$

Η καμπύλη $g(x_1, x_2) = 0$ μαζί με τα 8 αρχικά σημεία θα είναι:



(στ) Αλγεβρικά, τα support vectors θα είναι όλα τα σημεία που ο πολλαπλασιαστής Lagrange τους είναι θετικός. Εδώ και όλα εκτός από x_1, x_3 σημεία είναι support vectors.

(ζ) Τα έξτρα σημεία $\begin{pmatrix}\sqrt{2}\\ \sqrt{2}\end{pmatrix}$ και $\begin{pmatrix}-2\\ 3\end{pmatrix}$ θα ταξινομηθούν ως εξής:



Άσκηση 2.2 (HMM)

Έστω ένα κρυφό μοντέλο Markov λ = (A, B, π) με τρεις καταστάσεις 1, 2, 3 και δύο τύπους παρατηρήσεων H και T. Δίνεται ο παρακάτω πίνακας μεταβάσεων A όπου $A_{ij} = p(q_t = j, q_{t-1} = i)$

$$\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

Δίνονται επίσης οι ακόλουθες πιθανότητες των παρατηρήσεων B

P(O|q)			
O / q	*1*	*2*	*3*
H	*0.5*	*0.75*	*0.25*
T	*0.5*	*0.25*	*0.75*

Οι a-priori πιθανότητες π είναι $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$.

Έστω ότι η παρατηρούμενη ακολουθία είναι O = (H, T, H). Υπολογίστε:

1. *Την πιθανότητα P(O | λ) χρησιμοποιώντας τον forward αλγόριθμο.*

2. *Την πιθανότητα P(O | λ) χρησιμοποιώντας τον backward αλγόριθμο.*

3. *Υπολογίστε την πιο πιθανή σειρά καταστάσεων δεδομένης της ακολουθίας O χρησιμοποιώντας τον αλγόριθμο Viterbi.*

4. *Θέλουμε να χρησιμοποιήσουμε την ακόλουθη σειρά παρατηρήσεων*

HHHHHHTTHHHHHHHH

για να εκπαιδεύσουμε το κρυφό μοντέλο Markov. Αντί για τον πολύπλοκο αλγόριθμο forward–backward, θα ακολουθήσουμε τον εξής ψευδο–ΕΜ αλγόριθμο που αναφέρεται συχνά ως εκπαίδευση Viterbi, (Viterbi-training).

- *Expectation step: Αποκωδικοποίηση χρησιμοποιώντας τον αλγόριθμο Viterbi.*
- *Maximization step: Μεγιστοποίηση συνολικής πιθανότητας καταστάσεων και παρατηρήσεων ML.*

Δηλαδή ο αλγόριθμος εκπαίδευσης Viterbi, βρίσκει την πιο πιθανή σειρά καταστάσεων δεδομένης της σειράς παρατηρήσεων και στη συνέχεια μεγιστοποιεί την συνολική πιθανότητα της σειράς καταστάσεων $q_0 q_1 q_2$ *που μόλις υπολογίσαμε και παρατηρήσεων* $O_0 O_1 O_2$ *που δίδονται. Το δεύτερο βήμα του αλγορίθμου είναι η συνήθης εκπαίδευση με μεγιστοποίηση πιθανότητας Maximum-Likelihood-training που εφαρμόζεται σε κάθε Μπεϋσιανό δίκτυο με όλες τις παραμέτρους (q, o) παρατηρήσιμες. Εκτελέστε δύο επαναλήψεις του αλγορίθμου Viterbi training χρησιμοποιώντας τη δοθείσα ακολουθία. (Hint: Μπορείτε να λύσετε την άσκηση γραφικά σε ένα Τρελλις χωρίς πολλές πράξεις).*

5. *Ποιες είναι οι κύριες διαφορές μεταξύ του forward–backward και Viterbi–training και ποιος αλγόριθμος αναμένεται να έχει καλύτερα αποτελέσματα.*

Λύση:

1. Αναζητείται η πιθανότητα της ακολουθίας O = (H, T, H) χρησιμοποιώντας τον forward αλγόριθμο.

Έστω HMM – 1 οπότε ισχύουν οι σχέσεις:

$$q_{t-1}||q_{t+1}|q_t$$

$$o_t||o_{t-1}|q_t, q_{t-1}$$

$$o_t||q_{t-1}|q_t$$

Επομένως, από τον κανόνα της αλυσίδας και το θεώρημα του Bayes, θα ισχύει ότι:

$$P(O) = \sum_Q P(O, Q) = \sum_Q P(o_0 o_1 o_2 q_0 q_1 q_2) = \sum_Q P(o_0 o_1 o_2 | q_0 q_1 q_2) P(q_0 q_1 q_2) \Rightarrow$$

$$\Rightarrow P(O) = \sum_{q_0 q_1 q_2} \prod_{t=0}^{2} P(o_t|q_t) P(q_t|q_{t-1}) \Rightarrow$$

$$\Rightarrow P(O) = \sum_{q_2} b_{q_2}(o_2) \left[\sum_{q_1} a_{q_1 q_2} b_{q_1}(o_1) \sum_{q_0} \left(a_{q_0 q_1}(o_0) \pi_{q_0} \right) \right]$$

Επιπλέον, ισχύουν οι σχέσεις:

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

$$P(O|\lambda) = \sum_{i=1}^{N} a_T(i)$$

Αυτή η έκφραση αποτελεί τον forward αλγόριθμο.

Στην συνέχεια, υπολογίζονται αναλυτικά όλα τα βήματα.

$$a_1(1) = \pi_1 b_1(o_1) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

$$a_1(2) = \pi_2 b_2(o_2) = \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{4}$$

$$a_1(3) = \pi_3 b_3(o_3) = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$$

$$a_2(1) = (a_1(1)a_{11} + a_1(2)a_{21} + a_1(3)a_{31}) b_1(o_2) = \frac{1}{9} \cdot \frac{3}{2} \cdot \frac{1}{2}$$

$$a_2(2) = (a_1(1)a_{12} + a_1(2)a_{22} + a_1(3)a_{32}) b_2(o_2) = \frac{1}{9} \cdot \frac{3}{2} \cdot \frac{1}{4}$$

$$a_2(3) = (a_1(1)a_{13} + a_1(2)a_{23} + a_1(3)a_{33}) b_3(o_2) = \frac{1}{9} \cdot \frac{3}{2} \cdot \frac{3}{4}$$

$$a_3(1) = (a_3(1)a_{11} + a_3(2)a_{21} + a_3(3)a_{31})b_1(o_3) = \frac{1}{9} \cdot \frac{9}{4} \cdot \frac{1}{2}$$

$$a_3(2) = (a_3(1)a_{12} + a_3(2)a_{22} + a_3(3)a_{32})b_2(o_3) = \frac{1}{9} \cdot \frac{9}{4} \cdot \frac{3}{4}$$

$$a_3(3) = (a_3(1)a_{13} + a_3(2)a_{23} + a_3(3)a_{33})b_3(o_3) = \frac{1}{9} \cdot \frac{9}{4} \cdot \frac{1}{4}$$

$$P(O|\lambda) = \sum_{i=1}^{N} a_T(i) = a_3(1) + a_3(2) + a_3(3) = \frac{1}{8}$$

2. Αναζητείται η πιθανότητα της ακολουθίας O = (H, T, H) χρησιμοποιώντας τον backward αλγόριθμο.

Εδώ, ισχύουν οι εξής σχέσεις:

$$\beta_t(j) = \sum_{i=1}^{N} a_{ji} bi(o_{t+1})\beta_{t+1}(i)$$

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1)\beta_1(i)$$

Στην συνέχεια, υπολογίζονται αναλυτικά όλα τα βήματα.

$$\beta_3(1) = 1$$

$$\beta_3(2) = 1$$

$$\beta_3(3) = 1$$

$$\beta_2(1) = a_{11}b_1(o_3)\beta_3(1) + a_{12}b_2(o_3)\beta_3(2) + a_{13}b_3(o_3)\beta_3(3) = \frac{1}{3} \cdot \frac{3}{2}$$

$$\beta_2(2) = a_{21}b_1(o_3)\beta_3(1) + a_{22}b_2(o_3)\beta_3(2) + a_{23}b_3(o_3)\beta_3(3) = \frac{1}{3} \cdot \frac{3}{2}$$

$$\beta_2(3) = a_{31}b_1(o_3)\beta_3(1) + a_{32}b_2(o_3)\beta_3(2) + a_{33}b_3(o_3)\beta_3(3) = \frac{1}{3} \cdot \frac{3}{2}$$

$$\beta_1(1) = a_{11}b_1(o_2)\beta_2(1) + a_{12}b_2(o_2)\beta_2(2) + a_{13}b_3(o_2)\beta_2(3) = \frac{1}{3} \cdot \frac{1}{2}$$

$$\beta_1(2) = a_{21}b_1(o_2)\beta_2(1) + a_{22}b_2(o_2)\beta_2(2) + a_{23}b_3(o_2)\beta_2(3) = \frac{1}{3} \cdot \frac{1}{2}$$

$$\beta_1(3) = a_{31}b_1(o_2)\beta_2(1) + a_{32}b_2(o_2)\beta_2(2) + a_{33}b_3(o_2)\beta_2(3) = \frac{1}{3} \cdot \frac{1}{2}$$

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1)\beta_1(i) = \pi_1 b_1(o_1)\beta_1(1) + \pi_2 b_2(o_1)\beta_2(1) + \pi_3 b_3(o_1)\beta_3(1) = \frac{1}{8}$$

3. Σε αυτό το βήμα υπολογίζεται η πιο πιθανή σειρά καταστάσεων δεδομένης της ακολουθίας O χρησιμοποιώντας τον αλγόριθμο Viterbi. Ο αλγόριθμος Viterbi εφαρμόζεται για την εύρεση της πιο πιθανής αλληλουχίας καταστάσεων μέσω δεδομένης ακολουθίας παρατηρήσεων.

Περιγραφή αλγορίθμου Viterbi: Έστω ένα μοντέλο λ με πίνακες λ = (A, B, π). Οι πιθανότητες για κάθε πιθανή μετάβαση από κάθε κατάσταση υπολογίζονται με την ακόλουθη σχέση για όλες τις καταστάσεις I και τις χρ. στιγμές t: $\delta_t(i) = \max_{q_0 \dots q_1} P(q_0 \dots q_1, q_{t=i}, o_1 \dots o_t|\lambda)$. Επομένως, αρχικά πολλαπλασιάζεται η πιθανότητα εκκίνησης (π) από κάθε κατάσταση με την πιθανότητα μετάβασης σε κάθε επόμενη κατάσταση (A) και το γινόμενο αντιστοιχίζεται στο $\delta_t(i)$. Η διαδικασία επαναλαμβάνεται μέχρι το τέλος του χρόνου T αντιστοιχίζοντας στο $\delta_t(i)$ το μέγιστο από όλα τα γινόμενα της πιθανότητας μετάβασης από την παρούσα κατάσταση επί το προηγούμενο $\delta_{t-1}(i)$ επί την πιθανότητα εκπομπής της παρατήρησης στην θέση t(B), για μετάβαση σε κάθε κατάσταση. Έτσι, για την εύρεση του πιο πιθανού μονοπατιού αποθηκεύεται η πιθανότερη κάθε φορά μετάβαση.

Εδώ, ισχύουν οι εξής σχέσεις:

$$\delta_{t+1}(j) = \max_i\left(\delta_t(i)a_{ij}\right)b_j(o_{t+1})$$

$$P(O|\lambda) = \max_i(\delta_T(i))$$

Στην συνέχεια, υπολογίζονται αναλυτικά όλα τα βήματα.

$$\delta_1(1) = \pi_1 b_1(o_1) = \frac{1}{3}\cdot\frac{1}{2} = \frac{1}{6}$$

$$\delta_1(2) = \pi_2 b_2(o_1) = \frac{1}{3}\cdot\frac{3}{4} = \frac{1}{4}$$

$$\delta_1(3) = \pi_3 b_3(o_1) = \frac{1}{3}\cdot\frac{1}{4} = \frac{1}{12}$$

$$\delta_2(1) = \max(\delta_1(1)a_{11}, \delta_1(2)a_{21}, \delta_1(3)a_{31})b_1(o_2) = \frac{1}{9}\cdot\frac{3}{4}\cdot\frac{1}{2}$$

$$\delta_2(2) = \max(\delta_1(1)a_{12}, \delta_1(2)a_{22}, \delta_1(3)a_{32})b_2(o_2) = \frac{1}{9}\cdot\frac{3}{4}\cdot\frac{1}{4}$$

$$\delta_2(3) = \max(\delta_1(1)a_{13}, \delta_1(2)a_{23}, \delta_1(3)a_{33})b_3(o_2) = \frac{1}{9}\cdot\frac{3}{4}\cdot\frac{3}{4}$$

$$\delta_3(1) = \max(\delta_2(1)a_{11}, \delta_2(2)a_{21}, \delta_2(3)a_{31})b_1(o_3) = \frac{1}{9}\cdot\frac{9}{16}\cdot\frac{1}{2}$$

$$\delta_3(2) = \max(\delta_2(1)a_{12}, \delta_2(2)a_{22}, \delta_2(3)a_{32})b_2(o_3) = \frac{1}{9}\cdot\frac{9}{16}\cdot\frac{3}{4}$$

$$\delta_3(3) = \max(\delta_2(1)a_{13}, \delta_2(2)a_{23}, \delta_2(3)a_{33})b_3(o_3) = \frac{1}{9}\cdot\frac{9}{16}\cdot\frac{1}{4}$$

$$P(O|\lambda) = \max_i(\delta_T(i)) = \max_i\left(\delta_3(1), \delta_3(2), \delta_3(3)\right) = \frac{1}{64}$$

$$\widehat{Q} = \arg\max_Q P(Q|O,\lambda) = \arg\max_Q(\delta_3(1), \delta_3(2), \delta_3(3)) = (q_1 = 2, q_2 = 3, q_3 = 2)$$

Τα αποτελέσματα αυτά επιβεβαιώνονται και από τον κώδικα:

```
Best path: ['state 2', 'state 3', 'state 2']

Probability value:  0.015625
```

4. Σε αυτή την περίπτωση οι πιθανότητες μετάβασης και εκπομπής δεν είναι γνωστές και για αυτό θα πρέπει να εκτιμηθούν μέσω μιας γνωστής ακολουθίας παρατηρήσεων.

Περιγραφή Viterbi training: Οι αρχικές τιμές για τις πιθανότητες μετάβασης και εκπομπής έχουν δοθεί στην εκφώνηση. Οι τιμές που επιλέγονται είναι πολύ σημαντικές για το αποτέλεσμα που θα προκύψει. Σημειώνεται ότι οι πιθανότητες εκκίνησης θεωρήθηκαν ίσες από κάθε κατάσταση. Ύστερα, πραγματοποιείται Viterbi decoding ώστε να βρεθεί το καλύτερο μονοπάτι και η αντίστοιχη πιθανότητα. Στην συνέχεια, με loop βρίσκεται το καλύτερο μονοπάτι εσωτερικών καταστάσεων και γίνεται προσαρμογή των τιμών των παραμέτρων του μοντέλου ως εξής. Πρώτα, $a_{ij} = \frac{n_{i\to j}}{n_i}$, όπου $n_{i\to j}$ ο αριθμός μεταβάσεων από την κατάσταση i στην κατάσταση j με το καλύτερο μονοπάτι και n_i ο αριθμός των φορών που η κατάσταση εκκίνησης είναι η i και μετά $b_{kj} = \frac{n_{j\to k}}{n_j}$, όπου $n_{j\to k}$ ο αριθμός εκπομπής της παρατήρησης k όταν το σύστημα βρίσκεται στην κατάσταση j και n_j ο αριθμός των φορών που το σύστημα βρέθηκε στην κατάσταση j. Έπειτα, εφαρμόζεται ξανά Viterbi decoding και υπολογίζεται το νέο μονοπάτι και η πιθανότητα. Μάλιστα, εάν η πιθανότητα έχει μικρή διαφορά με την προηγούμενη, τότε ο αλγόριθμος συγκλίνει και το loop τερματίζεται ενώ αν αυτό δεν συμβεί τότε τερματίζει μετά από ένα προκαθορισμένο όριο από loops.

Τα αποτελέσματα που προκύπτουν από τον κώδικα είναι:

```
Best path probability 0.0000000002

Best path probability 0.0024531840

Best path probability 0.0024531840
```

5. Οι κύριες διαφορές μεταξύ forward – backward και Viterbi – training είναι οι εξής:

- Ο αλγόριθμος Baum-Welch ή forward-backward είναι ουσιαστικά μια εφαρμογή του EM σε HMM και συγκλίνει κάθε φορά σε κάποιο τοπικό μέγιστο. Ωστόσο, απαιτεί διπλή διάσχιση πάνω στα δεδομένα σε κάθε βήμα και για αυτό το λόγο έχει μεγάλη πολυπλοκότητα.
- Ο αλγόριθμος Viterbi training ή Segmental k-means training είναι πρακτικά μια εφαρμογή ψευδό-EM. Ουσιαστικά, προσεγγίζει τις παραμέτρους με MLE και για αυτό το λόγο είναι ταχύτερος αλλά λιγότερος ακριβής.

Ο αλγόριθμος που αναμένεται να έχει καλύτερα αποτελέσματα σε γενικές περιπτώσεις είναι ο Viterbi training διότι εκμεταλλεύεται την υπολογιστική ισχύ και θεωρεί ότι μπορεί να παρατηρήσει την ακολουθία κρυφών καταστάσεων στο HMM. Όμως, εάν το αρχικό μοντέλο δεν είναι καλό, τότε αναγκαστικά θα πρέπει να χρησιμοποιηθεί ο Baum-Welch.

Άσκηση 2.3 (CART)

Στην ηλεκτρονική σελίδα μιας ταξιδιωτικής υπηρεσίας έχετε εισάγει ένα αυτόματο σύστημα διαλόγου για την εξυπηρέτηση των χρηστών κατά την κράτηση εισιτηρίων. Για την αξιολόγηση του συστήματος λαμβάνετε υπόψιν τις παρακάτω μετρικές από τις διαδράσεις των χρηστών με το σύστημα:

- *SATISFACTION: Αν έμεινε ο χρήστης ικανοποιημένος από τη διάδραση ([Y]ES/[N]O)*
- *WORD_ACCURACY: Το ποσοστό των λέξεων που αναγνωρίστηκαν επιτυχώς από το σύστημα αναγνώρισης φωνής ([0 − 100])*
- *TASK_COMPLETION: Αν ολοκληρώθηκε επιτυχώς η κράτηση ([Y]ES/[N]O)*
- *TASK_DURATION: Ο χρόνος που διήρκεσε η διάδραση σε λεπτά*

Κατά την αξιολόγηση ενός αυτόματου συστήματος διαλόγου λαμβάνετε τις παρακάτω απαντήσεις

SATISFACTION	*WORD_ACCURACY*	*TASK_COMPLETION*	*TASK_DURATION*
Y	*100*	*Y*	*3*
Y	*100*	*Y*	*2*
Y	*90*	*N*	*4*
N	*95*	*N*	*2*
N	*80*	*Y*	*5*
Y	*85*	*Y*	*5*
Y	*80*	*Y*	*1*
N	*85*	*N*	*3*
Y	*95*	*N*	*4*

Με βάση αυτόν τον πίνακα αξιολόγησης θέλετε να αποφασίσετε σε ποια κατεύθυνση θα επενδύσετε για τη βελτίωση του συστήματος ώστε να μεγιστοποιήσετε την ικανοποίηση των χρηστών. Συγκεκριμένα θέλετε να χρησιμοποιήσετε τον αλγόριθμο CART για να αποφασίσετε ποια από τις παραμέτρους WORD_ACCURACY, TASK_COMPLETION και TASK_DURATION επηρεάζει περισσότερο αν ένας χρήστης έχει μείνει ικανοποιημένος ή όχι από τη διάδραση.

1. *Ποιες είναι οι δύο κατηγορίες ω_1 και ω_2 που θέλω να ταξινομήσω με αυτό το δέντρο απόφασης. Ποιες είναι οι παράμετροι και οι ερωτήσεις του δέντρου απόφασης.*

2. *Κατασκευάστε το δυαδικό δέντρο απόφασης για τις κατηγορίες SATISFACTION = Y και SATISFACTION = N με τις παραμέτρους (ερωτήσεις) WORD_ACCURACY, TASK_COMPLETION, TASK_DURATION ώστε να ελαχιστοποιήσετε την εντροπία σε κάθε σημείο απόφασης (entropy impurity).*

3. *Ποια είναι η πρώτη (πιο σημαντική), δεύτερη, τρίτη ερώτηση του δέντρου. Πως επηρεάζει η επιτυχία της συναλλαγής, η ποιότητα του αναγνωριστή φωνής και η διάρκεια της συνομιλίας, την ικανοποίηση του συνομιλητή από το σύστημα διαλόγου.*

Λύση:

1. Οι κατηγορίες ω_1 και ω_2 που είναι επιθυμητό να ταξινομηθούν με αυτό το δέντρο απόφασης είναι οι εξής:

- κατηγορία ω_1: SATISFACTION = Y
- κατηγορία ω_2: SATISFACTION = N

Οι παράμετροι θα είναι οι παρατηρήσεις:

- WORD_ACCURACY
- TASK_COMPLETION
- TASK_DURATION

Τέλος, οι αντίστοιχες ερωτήσεις σε αυτό το δέντρο απόφασης θα είναι:

- WORD_ACCURACY > τιμή, όπου τιμή ∈ [0, 100]
- TASK_COMPLETION = ([Y]ES/[N]O)
- TASK_DURATION > τιμή, όπου τιμή ∈ [0, 5]

2. Ο πίνακας για κάθε κατηγορία θα είναι ο εξής:

SATISFACTION	WORD_ACCURACY	TASK_COMPLETION	TASK_DURATION
Y	100	Y	3
Y	100	Y	2
Y	90	N	4
Y	85	Y	5
Y	80	Y	1
Y	95	N	4

SATISFACTION	WORD_ACCURACY	TASK_COMPLETION	TASK_DURATION
N	95	N	2
N	80	Y	5
N	85	N	3

Είναι γνωστό ότι αλγόριθμος CART ελέγχει τη μικρότερη εντροπία σε κάθε κόμβο όμως δεν εγγυάται και την συνολική μικρότερη εντροπία.

Για κάθε κόμβο του δέντρου, ο δεξιός απόγονος εκφράζει το SATISFACTION = Y ενώ ο αριστερός εκφράζει το SATISFACTION = N.

Για την εντροπία στο σύνολο των κατηγοριών ισχύει:

$$\mathrm{i(N)} = -[\mathrm{P}(\omega_1)\log \mathrm{P}(\omega_1) + \mathrm{P}(\omega_2)\log \mathrm{P}(\omega_2)]$$

Μετά από δοκιμές, επιλέχθηκαν οι εξής ερωτήσεις:

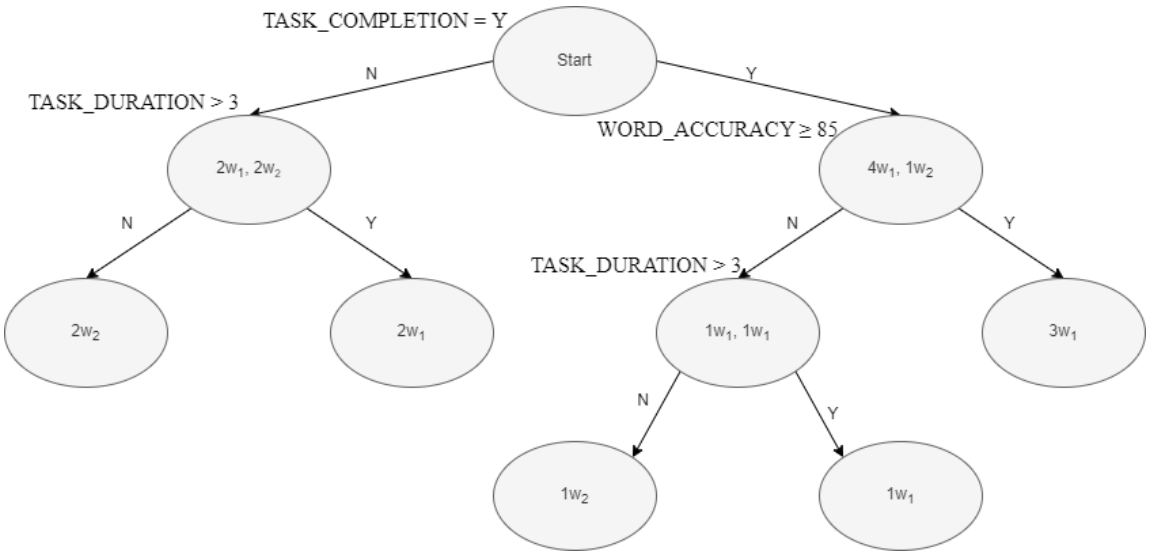
- Ερώτηση κόμβου (Start): TASK_COMPLETION = Y
  Οπότε, στον δεξιό κόμβο (1) η εντροπία θα είναι:
  $$\mathrm{i(N)} = -\left[\frac{4}{5}\log\frac{4}{5} + \frac{1}{5}\log 1.5\right] = 0.72$$
  Και στον αριστερό κόμβο (2) η εντροπία θα είναι:
  $$\mathrm{i(N)} = -\left[\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right] = 1$$

- Ερώτηση κόμβου (1): WORD_ACCURACY ≥ 85
  Οπότε, στον δεξιό κόμβο θα υπάρχουν μόνο δεδομένα της ομάδας 1 και η εντροπία θα είναι:
  $$\mathrm{i(N)} = 0$$
  Και, στον αριστερό κόμβο (3) η εντροπία θα είναι:
  $$\mathrm{i(N)} = 1$$

- Ερώτηση κόμβου (2): TASK_DURATION > 3
  Οπότε, αναλόγως την απάντηση οι κατηγορίες διαχωρίζονται τελείως και η εντροπία θα είναι:
  $$\mathrm{i(N)} = 0$$

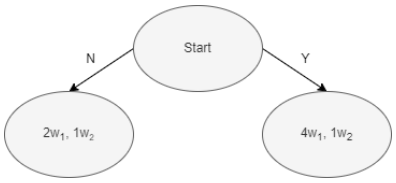
- Ερώτηση κόμβου (3): TASK_DURATION > 3
  Οπότε, αναλόγως την απάντηση οι κατηγορίες διαχωρίζονται τελείως και η εντροπία θα είναι:
  $$\mathrm{i(N)} = 0$$

Οπότε, το δυαδικό δέντρο απόφασης θα είναι το εξής:



3. Το κατά πόσο επηρεάζει η κάθε παράμετρος εξαρτάται από την εντροπία των παιδιών του κόμβου μετά την εφαρμογή της. Αναλυτικότερα,

- Ερώτηση: TASK_COMPLETION



Εδώ, στον δεξιό κόμβο η εντροπία θα είναι:

$$\mathrm{i(N)} = 0.72$$

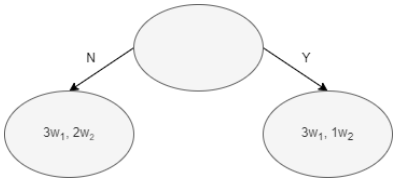
Και στον αριστερό κόμβο η εντροπία θα είναι:

$$\mathrm{i(N)} = 1$$

Επομένως, η συνολική εντροπία θα είναι:

$$\mathrm{i(N)} = \frac{5}{9} \cdot 0.72 + \frac{4}{9} \cdot 1 = 0.84$$

- Ερώτηση: TASK_DURATION



Εδώ, στον δεξιό κόμβο η εντροπία θα είναι:

$$\mathrm{i(N)} = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.81$$

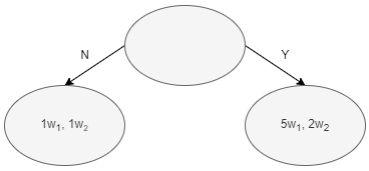
Και στον αριστερό κόμβο η εντροπία θα είναι:

$$\mathrm{i(N)} = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.97$$

Επομένως, η συνολική εντροπία θα είναι:

$$\mathrm{i(N)} = \frac{4}{9} \cdot 0.81 + \frac{5}{9} \cdot 0.97 = 0.90$$

- Ερώτηση: WORD_ACCURACY



Εδώ, στον δεξιό κόμβο η εντροπία θα είναι:

$$\mathrm{i(N)} = -\frac{5}{7}\log\frac{5}{7} - \frac{2}{7}\log\frac{2}{7} = 0.86$$

Και στον αριστερό κόμβο η εντροπία θα είναι:

$$\mathrm{i(N)} = \frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2} = 1$$

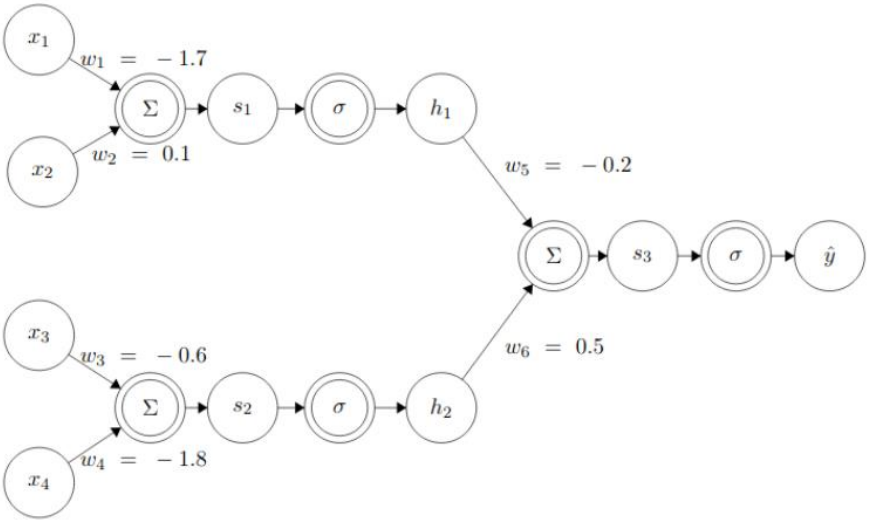
Επομένως, η συνολική εντροπία θα είναι:

$$\mathrm{i(N)} = \frac{7}{9} \cdot 0.86 + \frac{2}{9} \cdot 1 = 0.89$$

Συνεπώς, με αύξουσα σειρά θα είναι: TASK_DURATION > 3, WORD_ACCURACY ≥ 85 και TASK_COMPLETION = Y, αφού μεγαλύτερη εντροπία συνεπάγεται μικρότερη επιρροή.

Άσκηση 2.4 (MLP Backpropagation)

Υποθέστε ότι έχουμε το ακόλουθο γράφο (computation graph) που περιγράφει ένα νευρωνικό δίκτυο. Οι κόμβοι που βρίσκονται σε μονό κύκλο υποδηλώνουν μεταβλητές (για παράδειγμα η x_1 είναι μια μεταβλητή εισόδου, η h_1 είναι μια ενδιάμεση μεταβλητή και $\hat{y}$ είναι μια μεταβλητή εξόδου). Οι κόμβοι που βρίσκονται μέσα σε διπλό κύκλο υποδηλώνουν συναρτήσεις (για παράδειγμα το Σ υπολογίζει το άθροισμα των εισόδων του και η σ αναπαριστά τη συνάρτηση logistic σ(x) = $\frac{1}{1+e^{-x}}$. Οι ακμές που έχουν βάρη w_i υποδηλώνουν πολλαπλασιασμό της μεταβλητής εισόδου με το w_i.



Θεωρήστε ότι η συνάρτηση για το L2 Loss δίνεται από τη σχέση $L(y, \hat{y}) = \|y - \hat{y}\|_2^2$. Επίσης, υποθέστε ότι μας δίνονται τα δεδομένα ενός δείγματος $(x_1, x_2, x_3, x_4) = (-0.5, 1.4, 0.9, -3)$ με τιμή για το πραγματικό label ίση με 0.5. Χρησιμοποιήστε τον αλγόριθμο backpropagation για να υπολογίσετε τη μερική παράγωγο $\frac{\partial L}{\partial w_1}$.

Σημείωση: το gradient για τη συνάρτηση L2 loss είναι ίσο με $2\|y - \hat{y}\|$.

Λύση:

Πρώτα, εφαρμόζεται η forward διαδικασία για τον υπολογισμό της πρόβλεψης $\hat{y}$ αλλά και των ενδιάμεσων μεταβλητών $(s_1, s_2, h_1, h_2, h_3)$. Έτσι,

$$s_1 = w_1 x_1 + w_2 x_2 = (-1.7) \cdot (-0.5) + (0.1) \cdot (1.4) = 0.99$$

$$h_1 = \sigma(s_1) = \frac{1}{1 + e^{-s_1}} = 0.729$$

$$s_2 = w_3 x_3 + w_4 x_4 = (-0.6) \cdot (0.9) + (-1.8) \cdot (-3) = 4.86$$

$$h_2 = \sigma(s_2) = \frac{1}{1 + e^{-s_2}} = 0.992$$

$$s_3 = w_5 h_1 + w_6 h_2 = (-0.2) \cdot (0.729) + (0.5) \cdot (0.992) = 0.350$$

$$\hat{y} = \sigma(s_3) = \frac{1}{1 + e^{-s_2}} = 0.586$$

Έπειτα, εφαρμόζεται η backward διαδικασία και υπολογίζεται το $\frac{\partial L}{\partial w_1}$ εφαρμόζοντας τον κανόνα της αλυσίδας:

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial t_3} \frac{\partial h_1}{\partial t_1} \frac{\partial t_1}{\partial w_1}$$

Όπου,

$$\frac{\partial L}{\partial \hat{y}} = 2\|y - \hat{y}\| = 2|0.5 - 0.586| = 2 \cdot 0.086 = 0.172$$

$$\frac{\partial \hat{y}}{\partial t_3} = \sigma(t_3)\big(1 - \sigma(t_3)\big) = \hat{y}(1 - \hat{y}) = 0.586(1 - 0.586) = 0.243$$

$$\frac{\partial t_3}{\partial h_1} = w_5 = -0.2$$

$$\frac{\partial h_1}{\partial t_1} = \sigma(t_1)\big(1 - \sigma(t_1)\big) = h_1(1 - h_1) = 0.729(1 - 0.729) = 0.198$$

$$\frac{\partial t_1}{\partial w_1} = x_1 = -0.5$$

Επομένως,

$$\frac{\partial L}{\partial w_1} = (0.172) \cdot (0.243) \cdot (-0.2) \cdot (0.198) \cdot (-0.5) = 0.000827$$

Άσκηση 2.5 (KLT – PCA)

Υποθέτουμε τυχαία διανύσματα $x \in \mathbb{C}^d$ *(που μπορεί να παριστάνουν χαρακτηριστικά σε πρόβλημα αναγνώρισης προτύπων, γενικά με μιγαδικές τιμές). Ο αντίστοιχος χώρος Hilbert έχει εσωτερικό γινόμενο* $\langle x, y \rangle = \mathcal{E}\{y^H x\}$

Η ενέργεια του κάθε τυχαίου διανύσματος ισούται με $\langle x, x \rangle = \mathcal{E}\{x^H x\} = \mathcal{E}\left\{||x||^2\right\}$

Θέλουμε να βρούμε ένα unitary γραμμικό μετασχηματισμό (πίνακα) A ώστε τα μετασχηματισμένα διανύσματα $y = A^H x,\ A^{-1} = A^H$

να έχουν δύο ιδιότητες:

1. Να έχουν ορθογώνιες συνιστώσες

2. αν κρατήσουμε μόνο τις πρώτες p < d συνιστώσες να έχουμε Mean Squared Error (MSE)

Η λύση και βέλτιστη επιλογή είναι ο Karhunen Loeve μετασχηματισμός KLT, γνωστός και ως Principal Component Analysis (PCA). Συγκεκριμένα, επιλέγομε ως στήλες του A τα ορθοκανονικά διανύσματα $\{e_1, \dots, e_d\}$ *που είναι τα ιδιοδιανύσματα του* $R_x = \mathcal{E}\{xx^H\}$*. Από αυτά, τα πρώτα p ιδιοδιανύσματα* $\{e_1, \dots, e_p\}$ *αντιστοιχούν στις p μεγαλύτερες ιδιοτιμές* $\lambda_1 \geq \dots \geq \lambda_p$*. Η τάξης-p βέλτιστη προσέγγιση* $\hat{x}$ *και το αντίστοιχο ελάχιστο MSE J είναι:*

$$\hat{x} = \sum_{k=1}^{p} y_k e_k,\ \ y_k = e_k^H x \quad (1) \quad \& \quad J = \mathcal{E}\{||x - \hat{x}||^2\} \quad (2)$$

Με τις ανωτέρω επιλογές, τα μετασχηματισμένα χαρακτηριστικά $\{y_i\}$ *είναι ορθογώνια.*

ΠΡΟΣ ΑΠΟΔΕΙΞΗ*: Inductive or Batch matrix solution of PCA*

1. *Υποθέτοντας ότι έχουμε λύσει το πρόβλημα PCA για την περίπτωση του p = 1, αποδείξτε τη γενική περίπτωση όπου το p είναι οποιοσδήποτε αριθμός 1 < p < d χρησιμοποιώντας επαγωγή.*

2. *Δείξτε ότι η ελάχιστη τιμή για το σφάλμα J της PCA ως προς τα* e_i*, και δεδομένου του περιορισμού ορθοκανονικότητας, αποκτάται όταν τα* e_i *είναι ιδιοδιανύσματα του πίνακα συσχέτισης (συνδιασποράς). Για να πετύχουμε αυτό πρέπει να εισάγουμε τον πίνακα πολλαπλασιαστών Lagrange H, ένα για κάθε περιορισμό, έτσι ώστε η τροποποιημένη μετρική παραμόρφωσης να δίνεται από τον τύπο*

$$\tilde{J} = trace\{U^H R U\} + trace\{H(I - U^H U)\}$$

όπου το U είναι ένας πίνακας $d \times (d - p)$*, του οποίου οι στήλες είναι τα διανύσματα* e_i*. Τώρα, ελαχιστοποιώντας το* $\tilde{J}$ *ως προς το U, δείξτε ότι η λύση ικανοποιεί την εξίσωση RU = UH. Προφανώς, μια πιθανή λύση αποκτάται όταν οι στήλες του U είναι ιδιοδιανύσματα του πίνακα R, στην οποία περίπτωση ο H είναι ένας διαγώνιος πίνακας που περιέχει τις αντίστοιχες ιδιοτιμές. Για να αποκτήσετε μια γενική λύση, δείξτε ότι ο H μπορεί να θεωρηθεί ένας συμμετρικός πίνακας, και χρησιμοποιώντας την τεχνική επέκτασης ιδιοδιανυσμάτων δείξτε ότι η γενική λύση της RU = UH οδηγεί στην ίδια τιμή του* $\tilde{J}$ *με την ειδική λύση όπου οι στήλες του U είναι τα ιδιοδιανύσματα του R. Δείχνοντας ότι όλες οι λύσεις είναι ισοδύναμες μπορούμε να επιλέξουμε τη βολική λύση των ιδιοδιανυσμάτων.*

Λύση:

1. Η απόδειξη θα γίνει με επαγωγή.

- Έστω ότι το πρόβλημα PCA έχει λυθεί για p = 1.
- Έστω ότι υπάρχει λύση p για το PCA. Τότε, θα ισχύει ότι:

$$\vec{y_i} \text{ ορθογώνια} \rightarrow \widehat{\vec{x_i}} = \vec{m} + \sum_{j=1}^{p} a_j \vec{e_j}, i = 1, \dots, N$$

Και επίσης ότι το σφάλμα προσέγγισης τάξης p δίνεται από τον τύπο:

$$J_{min} = \sum_{i=p+1}^{d} \lambda_i \text{, όπου } \lambda_1 > \lambda_2 > \cdots > \lambda_d \text{ οι ιδιοτιμές του } R_x$$

Με $\vec{x_0} = \vec{m}$ από ανάλυση μηδενικής τάξης, προκύπτει:

$$J_0^{min}(\vec{x_0}) = J_0^{min}(\vec{x_0}) = 0$$

Επίσης,

$$a_j = \langle \vec{e_j}, x_j - m - \sum_{i=1}^{j} a_i \vec{e_i} \rangle, \text{με } \vec{e_i} = u\left[s_{i-1}^{(\lambda max)}\right]$$

- Τότε, για p + 1 θα ισχύει ότι:

$$\widehat{\vec{x}_{p+1}^{(i)}} = \vec{m} + \sum_{j=1}^{p+1} a_j \vec{e_j}$$

Οπότε,

$$\min J_{p+1}(\vec{a_1}, \vec{a_2}, \dots, \vec{a_{p+1}}, \vec{e_1}, \dots, \vec{e_{p+1}}) = 0 \Leftrightarrow \sum_{i=1}^{N} \left\| \vec{x_i} - \widehat{\vec{x}_{p+1}^{(i)}} \right\|^2 = 0 \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=1}^{N} \left\| \vec{x_i} - \vec{m} - \sum_{j=1}^{p+1} a_j^{(i)} \vec{e_j} \right\|^2 = 0 \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=1}^{N} \left\| \vec{x_i} - \vec{m} - \sum_{j=1}^{p} a_j^{(i)} \vec{e_j} \right\|^2 + \sum_{i=1}^{N} a_{p+1}^{(i)} \left\| \vec{e_j} \right\|^2 - 2 \sum_{i=1}^{N} a_{p+1}^{(i)} \overrightarrow{e_{p+1}} \left[\vec{x_i} - \vec{m} - \sum_{j=1}^{p} a_j^{(i)} \vec{e_j} \right] = 0 \Leftrightarrow$$

$$\Leftrightarrow \frac{\partial J_{p+1}}{\partial \overrightarrow{a_{p+1}}} = 0 \Leftrightarrow \sum_{i=1}^{N} 2 a_{p+1}^{(i)} \left\| \overrightarrow{e_{p+1}} \right\|^2 - 2 \sum_{i=1}^{N} \overrightarrow{e_{p+1}}^T \left[\vec{x_i} - \vec{m} - \sum_{j=1}^{p} a_j^{(i)} \vec{e_j} \right] = 0 \Leftrightarrow$$

$$\Leftrightarrow a_{p+1}^{(i)} = \overrightarrow{e_{p+1}}^T \left[\vec{x_i} - \vec{m} - \sum_{j=1}^{p} a_j^{(i)} \vec{e_j} \right] = \langle \widehat{\vec{x}_p^{(i)}}, \overrightarrow{e_{p+1}} \rangle$$

Σημειώνεται ότι για το $\overrightarrow{e_{p+1}}$ θεωρήθηκε ότι:

$$S_p = \sum_{j=1}^{p} \left[\hat{\vec{x}}_p'^{(i)}, \hat{\vec{x}}_p'^{T(i)} \right]$$

Οπότε, εφαρμόζοντας $a_{p+1}^{(i)}$:

$$J_{p+1} = \overrightarrow{e_{p+1}}^{T} S_p \overrightarrow{e_{p+1}} + \sum_{i=1}^{N} {a_{p+1}^{(i)}}^2 \left\| \overrightarrow{e_{p+1}} \right\|^2 - 2 \sum_{i=1}^{N} \left(\overrightarrow{e_{p+1}}^{T} \widehat{\vec{x}_p^{(i)}} \right)^T \overrightarrow{e_{p+1}} \widehat{\vec{x}_p^{(i)}}$$

Και για να ελαχιστοποιηθεί:

$$\frac{\partial J_{p+1}}{\partial \overrightarrow{a_{p+1}}} = 0 \Leftrightarrow 2S_p \overrightarrow{e_{p+1}} - 2\lambda_{p+1} \overrightarrow{e_{p+1}} = 0 \Leftrightarrow S_p \overrightarrow{e_{p+1}} = \lambda_{p+1} \overrightarrow{e_{p+1}}$$

Συνεπώς, με την Επαγωγή αποδείχθηκε ότι ο KLT έχει:

$$\overrightarrow{y_i} \text{ ασυσχέτιστα}$$

$$\text{προσεγγίσεις τάξης p: } \widehat{\vec{x}_i} = \vec{m} + \sum_{j=1}^{p} a_j \vec{e_j} \text{ με } J_{min} = \sum_{i}^{d} \lambda_i$$

Μάλιστα, οι $\vec{e_j}$ λύσεις αντιστοιχούν στα ιδιοδιανύσματα του Salter – variance των $\hat{\vec{x}}$. Επιπλέον, το σφάλμα J είναι αναμενόμενο να έχει ως ελάχιστη τιμή το άθροισμα των μικρότερων ιδιοτιμών λ_i που δεν αντιστοιχούν στα $\vec{e_j}$ αφού "κατασκευάζεται" ανά τάξη προσέγγισης.

2. Για την απόδειξη αρχικά αναλύονται οι πίνακες U και R και υπολογίζεται το ίχνος τους.

Δίνεται η σχέση:

$$\tilde{J} = \text{trace}\{U^H R U\} + \text{trace}\{H(I - U^H U)\}$$

Στην συνέχεια, υπολογίζεται η επιθυμητή σχέση ελαχιστοποιώντας το $\tilde{J}$.

Πρώτα, ελαχιστοποιείται το $\tilde{J}$ ως προς U:

$$\frac{\partial \tilde{J}}{\partial U} = 0 \Leftrightarrow \frac{\partial}{\partial U}[\text{trace}\{U^H R U\}] + \frac{\partial}{\partial U}[\text{trace}\{H(I - U^H U)\}] = 0 \Leftrightarrow$$

$$\xLeftrightarrow{\text{Matrix Cookbook}} RU - UH = 0 \Leftrightarrow RU = UH \qquad (1)$$

Μια λύση είναι:

$$U = \left[e_1(R) | \dots | e_{d-p}(R) \right]$$

$$H = dg\{\lambda_i(R)\}, i = 1, \dots, d - p$$

Εάν H συμμετρικός, τότε τα ιδιοδιανύσματα του είναι ορθοκανονικά, οπότε:

$$S = \left[e_{1\,H} | \dots | e_{d-p\,H} \right]$$

Και θα είναι και unitary, οπότε:

$$(1) \Rightarrow H = U^H R U = U^{-1} R U \qquad (2)$$

Όπου, για την πρώτη λύση ο $U^{H}RU$ θα είναι διαγώνιος.

Ακόμη,

$$S^{H}S = I \;\; \& \;\; S^{H}HS = \Lambda \qquad (3)$$

Έστω ότι υπάρχει η γενική εξίσωση αλλά ο H δεν φτιάχνεται από τις $\lambda_{i(R)}$ ούτε ο U από τα $\overrightarrow{e_{i(R)}}$. Τότε,

$$H = U^{H}RU \Leftrightarrow HS = U^{H}RUS \Leftrightarrow S^{H}HS = S^{H}(U^{H}RU)S \Leftrightarrow$$

$$\overset{(3)}{\Leftrightarrow} S^{H}U^{H}RUS = \Lambda = dg\{\lambda_{i(H)}\} \Leftrightarrow (US)^{-1}RUS = dg\{\lambda_{i(H)}\}$$

<u>Εναλλακτικά</u>,

$$(1) \Rightarrow RU = UH \Rightarrow H = U^{H}RU$$

Όμως, εάν $H = U^{*}RU$

Τότε,

$$\sum_{i=1}^{d-p} \lambda_{i(H)}\, \overrightarrow{e_{i\,H}}\, \overrightarrow{e_{i\,H}}^{*} = \sum_{i=1}^{d-p} \lambda_{i(R)}\, \overrightarrow{e_{i\,R}}\, \overrightarrow{e_{i\,R}}^{*}$$

Και θεωρώντας, $H = S\Lambda S^{-1}$

Τότε,

$$\vec{V} = SS^{-1}\vec{V} = S\vec{C} = C_1\overrightarrow{e_{1\,H}} + \cdots + C_n\overrightarrow{e_{n\,H}}$$

$$\vec{C} = S^{-1}\vec{V}$$

Εάν ο H είναι συμμετρικός τότε τα ιδιοδιανύσματα του είναι ορθοκανονικά:

$$S = \left[e_{1\,H}\right| \ldots \left|e_{d-p\,H}\right] \;\; \& \;\; S^{H} = S$$

Και θα είναι και unitary, οπότε:

$$S^{*}(U^{*}RU)S = S^{*}HS = \Lambda$$

$$H = U^{*}RU \;\mu\varepsilon\; \lambda_i H = \lambda R$$

$$H = S\Lambda S^{H} \;\; \& \;\; \Lambda = S^{*}HS$$

Οπότε,

$$\vec{V} = UU^{T}\vec{V} = U\vec{C} = C_1\overrightarrow{e_{1\,H}} + \cdots + C_{d-p}\overrightarrow{e_{d-p\,H}}$$

$$H\vec{V} = S\Lambda\vec{C} \;\; \& \;\; \overrightarrow{V_i} = S\overrightarrow{C_i}$$

$$U = [S\overrightarrow{C_1}| \ldots |S\overrightarrow{C_{d-p}}] = S[\overrightarrow{C_1}| \ldots |\overrightarrow{C_{d-p}}]$$

Εάν ο H δεν είναι διαγώνιος αλλά ερμιτιανός, μπορεί να παραγοντοποιηθεί ως: $H = F^{H}\Lambda F$ όπου, F ο πίνακας ιδιοδιανυσμάτων του H.

Έπειτα, ελαχιστοποιώντας πάλι το ∫ προκύπτει η επιθυμητή σχέση.

Άσκηση 2.6 (Graphical Models)

Δίδεται ένα μοντέλο Markov με μνήμη 2 (MM–2) με δύο καταστάσεις 1 και 2. Γνωρίζουμε ότι για μία σειρά από καταστάσεις του μοντέλου MM–2 ισχύει η σχέση ανεξαρτησίας

$$p(q_t|q_{t-1}, q_{t-2}, q_{t-3}, \dots) = p(q_t|q_{t-1}, q_{t-2})$$

1. *Σχεδιάστε το Μπεϋζιανό δίκτυο που αντιστοιχεί στο μοντέλο Markov με μνήμη 2 (MM–2)*

2. *Παρατηρούμε την ακόλουθη σειρά από καταστάσεις*

$$(1, 1, 1, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1, 1, 2)$$

Υπολογίστε τις πιθανότητες μετάβασης του MM–2 χρησιμοποιώντας την αρχή της μεγιστοποίησης της πιθανότητας των παρατηρήσεων Maximum Likelihood. Δίνεται ότι:

$$p(q_0 = 1) = p(q_0 = 2) = 0.5$$

3. *Υπολογίστε την πιθανότητα να μείνουμε στην κατάσταση 1, 10 φορές δεδομένου ότι είμαστε ήδη στην κατάσταση 1, δηλαδή*

$$p(q_1 = 1, q_2 = 1, q_3 = 1, \dots, q_{10} = 1, q_{11} = 2|q_0 = 1)$$

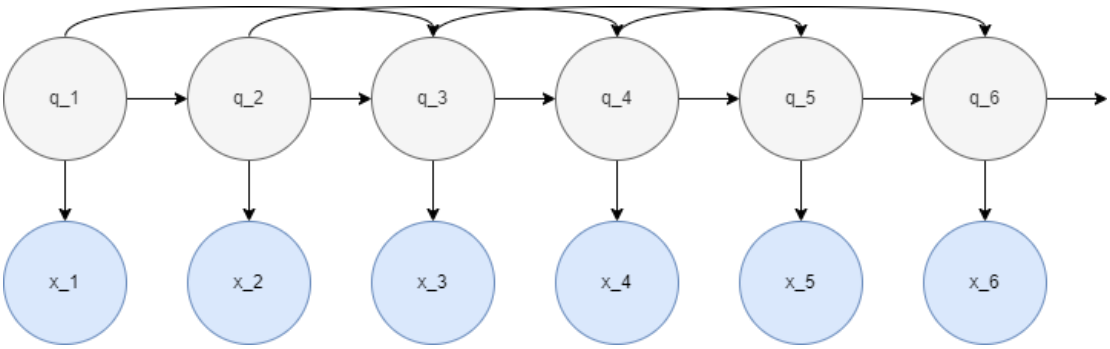
4. *θεωρείστε ότι τα δεδομένα που δίνονται στο ερώτημα 2 είναι για ένα μοντέλο Markov με μνήμη 1 (MM-1).*

(α΄) *Υπολογίστε τον πίνακα μετάβασης του νέου μοντέλου.*

(β΄) *Υπολογίστε πάλι την πιθανότητα του ερωτήματος 3 και συγκρίνετε τα αποτελέσματα.*

Λύση:

1. Το Μπεϋζιανό δίκτυο που αντιστοιχεί στο μοντέλο Markov με μνήμη 2 (MM–2) είναι το ακόλουθο:



Σημειώνεται ότι οι μεταβάσεις καταστάσεων που συμβολίζονται με q_i είναι ανεξάρτητες όλων των προηγούμενων καταστάσεων, δεδομένων όμως των δύο προηγούμενων καταστάσεων. Επίσης, το ίδιο θα ισχύει και για τις παρατηρήσεις x_i.

2. Η σειρά από καταστάσεις είναι D = (1, 1, 1, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 1, 1, 1, 1, 2).

Ο υπολογισμός της πιθανότητας μετάβασης μεταξύ δύο καταστάσεων (MM–2) γίνεται με την μέθοδο της μεγιστοποίησης της πιθανότητας – Maximum Likelihood (ML).

$$P(D|\theta) = p(q_n|q_{n-1}q_{n-2};\theta)\cdot p(q_{n-1}|q_{n-2}q_{n-3};\theta)\cdot \ldots \cdot p(q_1|q_0;\theta)\cdot p(q_0;\theta)$$

Έτσι, η λογαριθμική πιθανότητα με παραμέτρους θ για τον υπολογισμό των αγνώστων πιθανοτήτων μετάβασης m_{ij} θα είναι:

$$\log(P(D|\theta)) = \sum_{t=2}^{t=T} \log(p(q_t|q_{t-1}q_{t-2};\theta)) + \log\big(p(q_{t-1}|q_{t-2}q_{t-3};\theta)\big) + \cdots + \log(p(q_0;\theta))$$

Και το μέγιστο αυτής της πιθανότητας υπολογίζεται από την παραγώγιση της σχέσης ως προς m_{ij}.

Για την διευκόλυνση στις πράξεις, τέθηκε

$$\log(p(q_t|q_{t-1}q_{t-2})) = d(q_t = r, q_{t-1} = i, q_{t-2} = j)\log m_{rij}$$

Όπου, με d συμβολίζεται η δείκτρια συνάρτηση και επιπλέον ισχύει ότι:

$$\sum_r m_{rij} = 1, \quad 0 \le m_{rij} \le 1$$

Οπότε, η σχέση γίνεται:

$$\log(P(D|\theta)) = \sum_{t=2}^{t=T}\sum_{r,i,j\neq R} d(q_t = r, q_{t-1} = i, q_{t-2} = j)\log m_{rij} + \sum_{i,j,t}\log(1 - \sum_{r\neq R} m_{rij}) + S$$

Και με την παραγώγιση:

$$\frac{\partial \log\big(P(D|\theta)\big)}{\partial m_{rij}} = 0 \Leftrightarrow \frac{m'_{rij}}{m_{rij}} - \frac{m'_{Rij}}{1 - \sum_{r\neq R} m_{rij}} = 0 \Leftrightarrow m_{rij} = \frac{m'_{rij}}{m'_{Rij}}\left(1 - \sum_{r\neq R} m_{rij}\right) \Leftrightarrow$$

$$\sum_r m_{rij} = \sum_r \frac{m'_{rij}}{m'_{Rij}}\left(1 - \sum_{r\neq R} m_{rij}\right) \Leftrightarrow 1 = \sum_r \frac{m'_{rij}}{m'_{Rij}} m_{Rij} \Leftrightarrow \frac{m_{Rij}}{m'_{Rij}}\sum_r m'_{rij} = 1 \Leftrightarrow m_{Rij} = \frac{m'_{Rij}}{\sum_r m'_{rij}} \Leftrightarrow$$

Όπου, με m'_{rij} συμβολίζεται το πλήθος των μεταβάσεων στην κατάσταση r.

Επομένως, για την ακολουθία D ισχύει:

	$\mathbf{q_{t-1} = 1}$		$\mathbf{q_{t-2} = 2}$	
	$\mathbf{m_{rij} = 1}$	$\mathbf{m_{rij} = 2}$	$\mathbf{m_{rij} = 1}$	$\mathbf{m_{rij} = 2}$
$\mathbf{q_{t-1} = 1}$	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{1}{2}$	$\frac{1}{2}$
$\mathbf{q_{t-2} = 2}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$

Και, για τις πιθανότητες μετάβασης από κενή κατάσταση στις καταστάσεις 1 ή 2 δίνεται ότι:

$$p(q_0 = 1) = p(q_0 = 2) = 0.5$$

Και επίσης,

$$m_{11} = p(q_{t-1} = 1, q_t = 1) = 1$$

$$m_{12} = p(q_{t-1} = 1, q_t = 2) = 0$$

$$m_{21} = p(q_{t-1} = 2, q_t = 1) = 0.5$$

$$m_{22} = p(q_{t-1} = 2, q_t = 2) = 0.5$$

3. Εδώ, η σειρά από καταστάσεις θα είναι D' = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2).

Ζητείται ο υπολογισμός της πιθανότητας της D', δηλαδή:

$$p(q_1 = 1, q_2 = 1, q_3 = 1, \dots, q_{10} = 1, q_{11} = 2 | q_0 = 1)$$

Από τα προηγούμενα ερωτήματα είναι γνωστό ότι λόγω της ύπαρξης ΜΜ – 2, η μετάβαση εξαρτάται μόνο από τις δύο προηγούμενες καταστάσεις. Έτσι,

$$p(D) = p(q_1 \dots q_T; m) = p(q_T | q_{T-1} q_{T-2}; m) \cdot \dots \cdot p(q_2 | q_1 q_0; m) \cdot p(q_1 | q_0; m) \cdot p(q_0) \Leftrightarrow$$

$$p(D|\theta) \Leftrightarrow p(q_0) p(q_1|q_0) \prod_{t=2}^{t=T} m_{q_{t-2} q_{t-1} q_t}$$

Συνεπώς, δεδομένου ότι βρίσκεται κανείς στην κατάσταση 1, θα ισχύει:

$$p(D|q_0 = 1) = p(q_0) p(q_1|q_0) \prod_{t=2}^{t=10} m_{q_{t-2} q_{t-1} q_t} = \frac{1}{2} \cdot m_{11} \cdot m_{111}^9 \cdot m_{112} = \frac{1}{2} \cdot 1 \cdot \left(\frac{3}{5}\right)^9 \cdot \frac{2}{5} = 0.002$$

4. Έστω ότι ως δεδομένα θεωρείται το μοντέλο ΜΜ – 1 οπότε η κάθε μετάβαση εξαρτάται μόνο από την αμέσως προηγούμενη.

(α') Ομοίως με το ερώτημα 2, προκύπτει ο πίνακας μετάβασης:

$\mathbf{m_{ij}}$	$\mathbf{q_t = 1}$	$\mathbf{q_t = 2}$
$\mathbf{q_{t-1} = 1}$	$\frac{5}{9}$	$\frac{4}{9}$
$\mathbf{q_{t-2} = 2}$	$\frac{1}{3}$	$\frac{2}{3}$

(β') Η πιθανότητα θα είναι:

$$p(D|q_0 = 1) = \prod_{t=2}^{t=T} m_{q_{t-1} q_t} = m_{11}^{10} \cdot m_{12} = \left(\frac{5}{9}\right)^{10} \cdot \frac{4}{9} = 0.00124$$

Συνεπώς, παρατηρείται ότι η τιμή αυτή είναι μικρότερη από αυτή του ερωτήματος (3).

Άσκηση 2.7 (Linear Discriminant Analysis)

Στο μάθημα είδαμε ότι η Linear Discriminant Analysis (LDA) βασίζεται στην ανάστροφη σχέση των μητρών (πινάκων) S_W και S_B:

$$S_W = \sum_{i=1}^{|Classes|} \mathbb{E}_{x|x\in w_i}[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^T]$$

$$S_B = \sum_{i=1}^{|Classes|} P(w_i)(\overrightarrow{\mu_i}-\vec{\mu})(\overrightarrow{\mu_i}-\vec{\mu})^T$$

όπου το w_i αναπαριστά μια κλάση με μέση τιμή $\overrightarrow{\mu_i}$, $|Classes|$ είναι το πλήθος των κλάσεων και $\vec{\mu}$ είναι η μέση τιμή όλων των δειγμάτων.

(α) *Δείξτε ότι στην περίπτωση διαχωρισμού δύο κλάσεων w_1 και w_2, ο πίνακας SB μπορεί να γραφτεί στη μορφή $S_B = P(w_1)P(w_2)(\overrightarrow{\mu_2}-\overrightarrow{\mu_1})(\overrightarrow{\mu_2}-\overrightarrow{\mu_1})^T$.*

(β) *Βασιζόμενοι στο υποερώτημα (α), να βρείτε το ιδιοδιάνυσμα του πίνακα $S_W^{-1}S_B$ και την ιδιοτιμή του.*

Λύση:

(α) Είναι γνωστό ότι για |Classes| = 2:

$$\mathrm{S_B} = \mathrm{P(w_1)}(\overrightarrow{\mu_1}-\vec{\mu})(\overrightarrow{\mu_1}-\vec{\mu})^{\mathrm{T}} + \mathrm{P(w_2)}(\overrightarrow{\mu_2}-\vec{\mu})(\overrightarrow{\mu_2}-\vec{\mu})^{\mathrm{T}}$$

Επιπλέον, είναι γνωστό ότι:

$$\vec{\mu} = \mathrm{P(w_1)}\overrightarrow{\mu_1} + \mathrm{P(w_2)}\overrightarrow{\mu_2}$$

$$\mathrm{P(w_1)} + \mathrm{P(w_2)} = 1$$

Επομένως,

$$\mathrm{S_B} = \mathrm{P(w_1)}[\overrightarrow{\mu_1} - \mathrm{P(w_1)}\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}][\overrightarrow{\mu_1} - \mathrm{P(w_1)}\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}]^{\mathrm{T}} +$$

$$+\mathrm{P(w_2)}[\overrightarrow{\mu_2} - \mathrm{P(w_1)}\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}][\overrightarrow{\mu_2} - \mathrm{P(w_1)}\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}]^{\mathrm{T}} =$$

$$= \mathrm{P(w_1)}\big[\big(1-\mathrm{P(w_1)}\big)\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}\big]\big[\big(1-\mathrm{P(w_1)}\big)\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}\big]^{\mathrm{T}} +$$

$$+\mathrm{P(w_2)}[(1-\mathrm{P(w_2)})\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}][(1-\mathrm{P(w_2)})\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}]^{\mathrm{T}} =$$

$$= \mathrm{P(w_1)}[\mathrm{P(w_2)}\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}][\mathrm{P(w_2)}\overrightarrow{\mu_1} - \mathrm{P(w_2)}\overrightarrow{\mu_2}]^{\mathrm{T}} +$$

$$+\mathrm{P(w_2)}[\mathrm{P(w_1)}\overrightarrow{\mu_2} - \mathrm{P(w_1)}\overrightarrow{\mu_1}][\mathrm{P(w_1)}\overrightarrow{\mu_2} - \mathrm{P(w_1)}\overrightarrow{\mu_1}]^{\mathrm{T}} =$$

$$= \mathrm{P(w_1)P(w_2)^2}(\overrightarrow{\mu_1}-\overrightarrow{\mu_2})(\overrightarrow{\mu_1}-\overrightarrow{\mu_2})^{\mathrm{T}} + \mathrm{P(w_2)P(w_1)^2}(\overrightarrow{\mu_2}-\overrightarrow{\mu_1})(\overrightarrow{\mu_2}-\overrightarrow{\mu_1})^{\mathrm{T}} =$$

$$= \mathrm{P(w_1)P(w_2)}(\overrightarrow{\mu_2}-\overrightarrow{\mu_1})(\overrightarrow{\mu_2}-\overrightarrow{\mu_1})^{\mathrm{T}}\big(\mathrm{P(w_2)} + \mathrm{P(w_1)}\big) =$$

$$= \mathrm{P(w_1)P(w_2)}(\overrightarrow{\mu_2}-\overrightarrow{\mu_1})(\overrightarrow{\mu_2}-\overrightarrow{\mu_1})^{\mathrm{T}}$$

(β) 1ος Τρόπος Επίλυσης:

Ο υπολογισμός θα προκύψει από την εξίσωση: $S_W^{-1}S_B w = \lambda w$

Όπου,

$$S_B w = P(w_1)P(w_2)(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})^T w =$$

$$= (\overrightarrow{\mu_2} - \overrightarrow{\mu_1})(P(w_1)P(w_2)(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})^T w) =$$

$$= k(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})$$

Θέτοντας $k = P(w_1)P(w_2)(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})^T w$

Προκύπτει λοιπόν ότι το $S_B w$ έχει την ίδια κατεύθυνση με το διάνυσμα $\overrightarrow{\mu_2} - \overrightarrow{\mu_1}$.

Επομένως, η λύση μπορεί να προκύψει ως εξής:

$$S_W^{-1}k(\overrightarrow{\mu_2} - \overrightarrow{\mu_1}) = \lambda w$$

$$w = S_W^{-1}(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})$$

Αφού,

$$S_W^{-1}S_B w = S_W^{-1}S_B\left(S_W^{-1}(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})\right) =$$

$$= S_W^{-1}\left(\lambda(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})\right) = \lambda\left(S_W^{-1}(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})\right) \Rightarrow$$

$$S_W^{-1}S_B w = \lambda w$$

Και για την ιδιοτιμή τελικά, $\lambda = P(w_1)P(w_2)(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})^T S_W^{-1}(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})$

2ος Τρόπος Επίλυσης:

Από τον ορισμό του LDA, σκοπός είναι η μεγιστοποίηση του όρου:

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

Οπότε,

$$\frac{\partial J}{\partial w} = 0 \Rightarrow (w^T S_B w)S_w w = (w^T S_w w)S_B w \Rightarrow \frac{w^T S_B w}{w^T S_w w}S_w w = S_B w \Rightarrow \frac{w^T S_B w}{w^T S_w w}w = S_w^{-1}S_B w$$

Δηλαδή, $\lambda = \frac{w^T S_B w}{w^T S_w w}$

Έπειτα, διαγράφοντας όλους τους βαθμωτούς όρους, το w θα είναι:

$$w = P(w_1)P(w_2)(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})^T S_W^{-1}(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})$$

Και, τελικά

$$w = S_W^{-1}(\overrightarrow{\mu_2} - \overrightarrow{\mu_1})$$

Άσκηση 2.8 (ICA)

(α) *Για μία τυχαία μεταβλητή x, με μηδενική μέση τιμή, η κύρτωση ορίζεται ως:*

$$kurt(x) \coloneqq E[x^4] - 3(E[x^2])^2$$

Να δείξετε ότι για δύο στατιστικώς ανεξάρτητες τυχαίες μεταβλητές x και y, με μηδενικές μέσες τιμές, ισχύει η παρακάτω σχέση:

$$kurt(x + y) = kurt(x) + kurt(y)$$

(β1) *Θεωρήστε ότι σας δίνεται ένα σύνολο από N στατιστικώς ανεξάρτητες τυχαίες μεταβλητές s_i, με μηδενικές μέσες τιμές, μοναδιαίες διασπορές, και τιμές a_i για τις κυρτώσεις τους που κυμαίνονται από $-a$ έως a, για κάποια άγνωστη αλλά σταθερή τιμή a. Οι τυχαίες μεταβλητές s_i αναμειγνύονται μέσω σταθερών βαρών w_i ως εξής:*

$$x \coloneqq \sum_{i}^{N} w_i s_i$$

Να προσδιορίσετε τους περιορισμούς που πρέπει να ικανοποιούνται για τα βάρη w_i, ώστε και το μείγμα των τυχαίων μεταβλητών, x, να έχει μοναδιαία διασπορά.

(β2) *Να αποδείξετε ότι η κύρτωση ενός ομοιόμορφα σταθμισμένου μείγματος ($w_i = w_j,\ \forall i, j$) N τυχαίων μεταβλητών συγκλίνει στο μηδέν, καθώς το N απειρίζεται. Θεωρήστε ότι το μείγμα, x, έχει μοναδιαία διασπορά.*

Λύση:

(α) Είναι γνωστό ότι για δυο στατιστικά ανεξάρτητες τυχαίες μεταβλητές x, y με μηδενική μέση τιμή θα ισχύει:

$$\mathbb{E}[xy] = \mathbb{E}[x] \cdot \mathbb{E}[y] \quad \& \quad \mathbb{E}[x] = 0, \mathbb{E}[y] = 0$$

$$\text{kurt}(x + y) = \mathbb{E}[(x + y)^4] - 3(\mathbb{E}[(x + y)^2])^2$$

$$\text{kurt}(x) = \mathbb{E}[x^4] - 3(\mathbb{E}[x^2])^2 \quad \& \quad \text{kurt}(y) = \mathbb{E}[y^4] - 3(\mathbb{E}[y^2])^2$$

Όπου,

$$\mathbb{E}[(x + y)^4] = \mathbb{E}[x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4] =$$

$$= \mathbb{E}[x^4] + 4\mathbb{E}[x^3y] + 6\mathbb{E}[x^2y^2] + 4\mathbb{E}[xy^3] + \mathbb{E}[y^4] =$$

$$= \mathbb{E}[x^4] + 4\mathbb{E}[x^3]\mathbb{E}[y] + 6\mathbb{E}[x^2]\mathbb{E}[y^2] + 4\mathbb{E}[x]\mathbb{E}[y^3] + \mathbb{E}[y^4] =$$

$$= \mathbb{E}[x^4] + 6\mathbb{E}[x^2]\mathbb{E}[y^2] + \mathbb{E}[y^4]$$

Αφού όλες οι δυνάμεις των x, y είναι στατιστικά ανεξάρτητες.

Και,

$$(\mathbb{E}[(x + y)^2])^2 = (\mathbb{E}[x^2 + 2xy + y^2])^2 = (\mathbb{E}[x^2] + 2\mathbb{E}[xy] + \mathbb{E}[y^2])^2 =$$

$$= (\mathbb{E}[x^2] + 2\mathbb{E}[x]\mathbb{E}[y] + \mathbb{E}[y^2])^2 = (\mathbb{E}[x^2] + \mathbb{E}[y^2])^2 = (\mathbb{E}[x^2])^2 + 2\mathbb{E}[x^2]\mathbb{E}[y^2] + (\mathbb{E}[y^2])^2$$

Οπότε,

$$\mathrm{kurt}(x+y) = \mathbb{E}[(x+y)^4] - 3(\mathbb{E}[(x+y)^2])^2 =$$
$$= \mathbb{E}[x^4] + 6\mathbb{E}[x^2]\mathbb{E}[y^2] + \mathbb{E}[y^4] - 3[(\mathbb{E}[x^2])^2 + 2\mathbb{E}[x^2]\mathbb{E}[y^2] + (\mathbb{E}[y^2])^2] =$$
$$= \mathbb{E}[x^4] + 6\mathbb{E}[x^2]\mathbb{E}[y^2] + \mathbb{E}[y^4] - 3(\mathbb{E}[x^2])^2 - 6\mathbb{E}[x^2]\mathbb{E}[y^2] - 3(\mathbb{E}[y^2])^2 =$$
$$= \mathbb{E}[x^4] + \mathbb{E}[y^4] - 3(\mathbb{E}[x^2])^2 - 3(\mathbb{E}[y^2])^2 =$$
$$= [\mathbb{E}[x^4] - 3(\mathbb{E}[x^2])^2] + [\mathbb{E}[y^4] - 3(\mathbb{E}[y^2])^2] =$$
$$= \mathrm{kurt}(x) + \mathrm{kurt}(y)$$

(β1) Ισχύει ότι είναι επιθυμητό το μείγμα τυχαίων μεταβλητών x να έχει μοναδιαία διασπορά, οπότε:

$$\forall i: \mathrm{Var}[s_i] = 1$$

<u>1ος Τρόπος Επίλυσης</u>:

$$V[x] = V\left[\sum_i^N w_i s_i\right] = \sum_i^N V[w_i s_i] = \sum_i^N w_i^2 V[s_i] \xrightarrow{V[s_i]=1} \sum_i^N w_i^2 = 1$$

<u>2ος Τρόπος Επίλυσης</u>:

$$\mathrm{Var}[x] = 1 \Leftrightarrow \mathbb{E}[(x - \mathbb{E}[x])^2] = 1 \Leftrightarrow \mathbb{E}[x^2] - (\mathbb{E}[x])^2 = 1 \Leftrightarrow$$
$$\Leftrightarrow \mathbb{E}\left[\left(\sum_i^N w_i s_i\right)^2\right] - \left(\mathbb{E}\left[\sum_i^N w_i s_i\right]\right)^2 = 1 \Leftrightarrow$$
$$\Leftrightarrow \mathbb{E}\left[\left(\sum_i^N w_i s_i\right)\left(\sum_j^N w_j s_j\right)\right] - \left(\sum_i^N w_i \mathbb{E}[s_i]\right)\left(\sum_j^N w_j \mathbb{E}[s_j]\right) = 1 \Leftrightarrow$$
$$\Leftrightarrow \mathbb{E}\left[\sum_i^N \sum_j^N w_i w_j s_i s_j\right] - \sum_i^N \sum_j^N w_i w_j \mathbb{E}[s_i]\,\mathbb{E}[s_j] = 1 \Leftrightarrow$$
$$\Leftrightarrow \sum_i^N \sum_j^N w_i w_j \mathbb{E}[s_i s_j] - \sum_i^N \sum_j^N w_i w_j \mathbb{E}[s_i]\,\mathbb{E}[s_j] = 1 \Leftrightarrow$$
$$\Leftrightarrow \sum_i^N w_i^2 [\mathbb{E}[s_i^2] - (\mathbb{E}[s_i])^2] + \sum_i^N \sum_j^N w_i w_j (\mathbb{E}[s_i]\,\mathbb{E}[s_j] - \mathbb{E}[s_i]\mathbb{E}[s_j]) = 1 \Leftrightarrow$$
$$\Leftrightarrow \sum_i^N w_i^2 \mathrm{Var}[s_i] + 0 = 1 \xLeftrightarrow{\mathrm{Var}[s_i]=1} \sum_i^N w_i^2 = 1$$

(β2) 1ος Τρόπος Επίλυσης: Παρατηρείται ότι:

$$\mathrm{kurt}(w \cdot x) = \mathbb{E}[(w \cdot x)^4] - 3\mathbb{E}[(w \cdot x)^2]^2 = w^4[\mathbb{E}[x^4] - 3(\mathbb{E}[x^2])^2] = w^4 \cdot \mathrm{kurt}(x)$$

Επίσης, από το (α) ισχύει:

$$\mathrm{kurt}\left(\sum_i^N w_i s_i\right) = \sum_i^N \mathrm{kurt}(w_i s_i)$$

Οπότε,

$$\mathrm{kurt}(x) = \sum_i^N \mathrm{kurt}(w_i s_i) = \sum_i^N w_i^4 \cdot \mathrm{kurt}(s_i) = w^4 \sum_i^N \mathrm{kurt}(s_i) = w^4 \sum_i^N a_i$$

Άρα,

$$-a \le a_i \le a \Rightarrow -a \cdot N \le \sum_i^N a_i \le a \cdot N \Rightarrow -a \cdot N \cdot w^4 \le w^4 \sum_i^N a_i \le a \cdot N \cdot w^4 \qquad (1)$$

Ακόμη, από (β1) ισχύει:

$$V[x] = 1 \Rightarrow \sum_i^N w_i^2 = 1 \Rightarrow N \cdot w^2 = 1 \Rightarrow w^2 = \frac{1}{N} \Rightarrow w^4 = \frac{1}{N^2} \Rightarrow w = \pm\sqrt{\frac{1}{N}}$$

Έτσι, η (1) γίνεται:

$$-\frac{\alpha}{N} \le \mathrm{kurt}(x) \le \frac{a}{N}$$

Και από το Κριτήριο Παρεμβολής: $\lim_{N\to\infty} \mathrm{kurt}(x) = 0$

2ος Τρόπος Επίλυσης:

Από (β1) ισχύει:

$$x = \sqrt{\frac{1}{N}} \sum_i^N s_i$$

Άρα,

$$|\mathrm{kurt}(x)| = \left|\mathrm{kurt}\left(\sqrt{\frac{1}{N}} \sum_i^N s_i\right)\right| = \left|\left(\sqrt{\frac{1}{N}}\right)^4 \mathrm{kurt}\left(\sum_i^N s_i\right)\right| = \frac{1}{N^2}\left|\sum_i^N \mathrm{kurt}(s_i)\right|$$

Οπότε,

$$\frac{1}{N^2}\left|\sum_i^N \mathrm{kurt}(s_i)\right| \le \frac{1}{N^2}\sum_i^N |\mathrm{kurt}(s_i)| \le \frac{\alpha}{N^2}\sum_i^N 1 \xrightarrow{-\alpha \le \mathrm{kurt}(s_i) \le a} \frac{1}{N^2}\sum_i^N |\mathrm{kurt}(s_i)| \le \frac{\alpha}{N^2} N = \frac{a}{N}$$

Έτσι, προκύπτει ότι: $\lim_{N\to\infty} \mathrm{kurt}(x) = 0$

Άσκηση 2.9 (Logistic Regression)

Θεωρήστε το πρόβλημα logistic regression για ένα σύνολο δεδομένων $\{\varphi_n, t_n\}$*, όπου* $t_n \in \{0, 1\}$ *και* $\varphi_n = \varphi(x_n)$ *είναι οι κατηγορίες και οι συναρτήσεις βάσης, αντίστοιχα, για δείγματα* $n = \{1, 2, \ldots, N\}$*. Η συνάρτηση σφάλματος E(w), η οποία αναφέρεται συνήθως και ως crossentropy, ορίζεται ως:*

$$E(w) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

όπου w είναι το διάνυσμα βαρών, $y_n = \sigma(w^T \varphi_n)$ *η έξοδος του μοντέλου logistic regression στο διάνυσμα εισόδου* x_n*, και σ(a) =* $\frac{1}{1+e^{-a}}$ *η logistic sigmoid συνάρτηση.*

(α) *Να δείξετε ότι για ένα γραμμικώς διαχωρίσιμο σύνολο δεδομένων, η λύση μέγιστης πιθανοφάνειας για το μοντέλο logistic regression αντιστοιχεί στην εύρεση ενός διανύσματος w, για το οποίο η επιφάνεια απόφασης* $w^T \varphi(x) = 0$ *διαχωρίζει τις κλάσεις, απειρίζοντας ταυτόχρονα το μέτρο του διανύσματος w.*

(β) *Η Hessian μήτρα για το logistic regression δίνεται από τη σχέση:*

$$H = [\Phi^T R \Phi]$$

όπου Φ ο πίνακας των χαρακτηριστικών και R είναι ένας διαγώνιος πίνακας με στοιχεία $y_n(1 - y_n)$*. Να δείξετε ότι η Hessian μήτρα H είναι θετικώς ορισμένη. Ως εκ τούτου, δείξτε ότι η συνάρτηση σφάλματος είναι κυρτή συνάρτηση του w και ότι έχει μοναδικό ελάχιστο.*

(γ) *Να γράψετε κώδικα που θα υλοποιεί τον iterative reweighted least squares (IWLS) αλγόριθμο για logistic regression. Χρησιμοποιώντας τον αλγόριθμο αυτό, να υπολογίσετε και να σχεδιάσετε τα διαχωριστικά επίπεδα απόφασης που αντιστοιχούν στο σύνολο δεδομένων του προβλήματος τριών κλάσεων που έχει ανεβεί στο mycourses (αρχείο MLR.data). Οι δύο πρώτες στήλες περιλαμβάνουν τα διανύσματα χαρακτηριστικών, ενώ η τρίτη την κλάση. Συγκρίνετε τα αποτελέσματα με εκείνα που θα προέκυπταν εάν εφαρμοζόταν ταξινόμηση με βάση τα ελάχιστα τετράγωνα, σχεδιάζοντας τα αντίστοιχα διαχωριστικά επίπεδα απόφασης.*

Σημείωση: Λεπτομέρειες από θεωρία του logistic regression και IWLS μπορούν να βρεθούν στις σχετικές διαφάνειες του μαθήματος και τις ενότητες 4.3.2, 4.3.3, 4.3.4 από [3].

Λύση:

(α) 1ος Τρόπος Επίλυσης:

Για ένα γραμμικώς διαχωρίσιμο σύνολο δεδομένων, παρατηρείται ότι η λύση μέγιστης πιθανοφάνειας για το μοντέλο Logistic Regression αντιστοιχεί στην περίπτωση που σ = 0.5. Δηλαδή,

$$\frac{1}{1 + e^{-w^T \varphi(x)}} = \frac{1}{2} \Rightarrow w^T \varphi(x) = 0$$

Με αυτή την επιλογή, διαχωρίζεται η επιφάνεια απόφασης $w^T \varphi(x) = 0$ σε:

$$\text{αν για τα δείγμα είναι } t_n = 1, \text{τότε } w^T \varphi(x_n) > 0, \text{για τα θετικά δείγματα}$$

$$\text{αν για τα δείγματα είναι } t_n = 0, \text{τότε } w^T \varphi(x_n) < 0, \text{για τα αρνητικά δείγματα}$$

Μάλιστα, αυξάνοντας το μέτρο του διανύσματος w μέχρι το άπειρο, οι εξισώσεις δεν παύουν να ισχύουν.

Με άλλα λόγια, για τον συγκεκριμένο διαχωρισμό με LR, για την εύρεση των βαρών θα πρέπει να ελαχιστοποιηθεί το cross-entropy. Έτσι,

$$\vec{\nabla}E(w) = 0 \Rightarrow \sum_{n=1}^{N}(y_n - t_n)\varphi_n = 0 \Rightarrow (y_n - t_n)\varphi_n = 0 \Rightarrow \left(\frac{1}{1+e^{-w^T\varphi_n}} - t_n\right)\varphi_n = 0$$

Όμως, $\varphi_n \neq 0$, οπότε,

$$\frac{1}{1+e^{-w^T\varphi_n}} = t_n \Rightarrow 1 - t_n = t_n e^{-w^T\varphi_n} \Rightarrow -w^T\varphi_n + \ln t_n = \ln(1-t_n) \Rightarrow$$

$$\Rightarrow w^T\varphi_n = \ln\frac{t_n}{1-t_n}$$

Όμως, $t_n \in \{0, 1\}$ οπότε $\ln\frac{t_n}{1-t_n} = \pm\ln(0)$.

Άρα θα πρέπει το $w^T\varphi$ να απειρίζεται.

Όμως, $w^T\varphi = |w| \cdot |\varphi| \cdot \cos\theta$, όπου cosθ φραγμένο ως γωνία και $|\varphi|$ φραγμένο ως είσοδος.

Συνεπώς, το $w^T\varphi$ απειρίζεται όταν απειρίζεται το μέτρο του διανύσματος w.

<u>2ος Τρόπος Επίλυσης</u>:

Έστω, $y_i = p(t_i = 1|x_i; w)$

Οπότε, $p(t_i|x_i; w) = y_i^{t_i}(1-y_i)^{1-t_i}$

Έτσι, η μέγιστη πιθανοφάνεια υπολογίζεται ως εξής:

$$w_{ML} = \arg\max_w \prod_{i=1}^{N} p(t_i|x_i; w) = \arg\max_w \sum_{i=1}^{N} \ln p(t_i|x_i; w) =$$

$$= \arg\max_w \sum_{i=1}^{N} t_i \ln y_i + (1-t_i)\ln(1-y_i)$$

Μάλιστα, η παραπάνω σχέση υπολογίζει και την ελαχιστοποίηση της δεδομένης συνάρτησης σφάλματος E(w).

Εφόσον το πρόβλημα είναι γραμμικά διαχωρίσιμο, θα υπάρχει w που να διαχωρίζει επιτυχώς τα δείγματα και μάλιστα, για $w^T\varphi(x) = 0$ ως διαχωριστική επιφάνεια, προκύπτει $p(t_i|x_i; w) = 0.5$.

Επιπλέον, με δεδομένο πως $\forall\varphi_i$ θα έχει το σωστό πρόσημο, η σχέση μεγιστοποιείται στο 0 για κάποιο w με το μέτρο να απειρίζεται ώστε $y_i = 1$ αν $t_i = 1$ και αντίστοιχα για $t_i = 0$.

(β) 1ος Τρόπος Επίλυσης:

Η Hessian μήτρα δίνεται από την σχέση: $H = [\Phi^T R \Phi]$. Έστω, ένα διάνυσμα $\vec{r} \neq 0$.

Για να δειχθεί ότι η Hessian μήτρα είναι θετικώς ορισμένη, θα πρέπει πάντα να ισχύει: $r^T H r > 0$

Οπότε, $r^T H r = r^T \Phi^T R \Phi r$, όπου, Φ είναι ο πίνακας των χαρακτηριστικών και R ένας διαγώνιος πίνακας με στοιχεία $y_n(1 - y_n)$.

Για τον πίνακα R ισχύει: $0 < y_n < 1 \Rightarrow 0 < y_n(1 - y_n) < 1$

Επομένως, φαίνεται πως όλα τα στοιχεία του πίνακα R είναι θετικά. Άρα, και το $r^T H r > 0$. Συνεπώς, η Hessian μήτρα H είναι θετικά ορισμένη. Επιπλέον, από τα παραπάνω προκύπτει ότι η συνάρτηση σφάλματος θα είναι κυρτή.

2ος Τρόπος Επίλυσης:

Για την Hessian μήτρα ισχύει ότι:

$$H = \sum_{n=1}^{N} y_n(1 - y_n)\varphi_n \varphi_n^T$$

Για $x \in \mathbb{R}^n, x \neq 0$, ισχύει ότι:

$$x^T H x = \sum_{n=1}^{N} y_n(1 - y_n)(x^T \varphi_n)(\varphi_n^T x) = \sum_{n=1}^{N} y_n(1 - y_n)(x^T \varphi_n)(x^T \varphi_n)^T = \sum_{n=1}^{N} y_n(1 - y_n)k_n^2$$

Επίσης, επειδή $0 < y_n < 1$, για $k_n = x^T \varphi_n = 0, \forall n$ τότε θα ισχύει ότι $x \neq 0$. Δηλαδή, όλα τα διανύσματα βάσης είναι κάθετα στον x, το οποίο είναι άτοπο. Επομένως, $\exists n$ τέτοιο ώστε $k_n = 0$. Οπότε, η Hessian μήτρα είναι θετικά ορισμένη, $x^T H x > 0$.

Για την κυρτότητα ισχύει ότι: $\nabla(w^T \varphi_n) = \varphi_n$

$$\nabla E(w) = -\left(\sum_{n=1}^{N} t_n \nabla \ln \sigma(w^T \varphi_n) + (1 - t_n)\nabla \ln\left(1 - \sigma(w^T \varphi_n)\right)\right) =$$

$$= \left(\sum_{n=1}^{N} (t_n \frac{1}{y_n} y_n(1 - y_n)\varphi_n - (1 - t_n)\frac{1}{1 - y_n} y_n(1 - y_n)\varphi_n \right) =$$

$$= -\left(\sum_{n=1}^{N} (t_n(1 - y_n)\varphi_n - (1 - t_n)y_n \varphi_n\right) = \sum_{n=1}^{N} (y_n - t_n)\varphi_n$$

$$\nabla\left(\nabla E(w)\right) = \sum_{n=1}^{N} \varphi_n \nabla(y_n - t_n) = \sum_{n=1}^{N} y_n(1 - y_n)\varphi_n \varphi_n^T = H$$

Οπότε, αφού η Hessian είναι θετικά ορισμένη και επειδή η συνάρτηση error είναι κυρτή συνάρτηση του w, προκύπτει ότι θα έχει μοναδικό ελάχιστο που είναι και ολικό.

(γ) Για την υλοποίηση χρησιμοποιήθηκε ο αλγόριθμος του Κεφ. 4.3.3 του [3]. Συγκεκριμένα,

- Ο w είναι ένας πίνακας διάστασης 6×1.
- Ο Hessian πίνακας είναι διάστασης #classes · #features × #classes · #features = 6×6 και υπολογίζεται από τον τύπο 4.110:

$$H_{[2j:2(j+1)-1][2k:2(k+1)-1]} = \nabla_{w_k}\nabla_{w_j}E(w_1, w_2, w_3) = -\sum_{n=1}^{N} y_{nk}\left(I_{kj} - y_{nj}\right)\varphi_n\varphi_n^T$$

- Το gradient που χρησιμοποιείται για τις πράξεις είναι κι αυτό ένας πίνακας 6×1 και υπολογίζεται από τον τύπο:

$$\nabla E_{2j:2(j+1)-1} = \sum_{n=1}^{N}\left(y_{nj} - t_{nj}\right)x_n$$

- Το τελικό διάνυσμα βαρών υπολογίζεται από τον αρχικό τύπο:

$$w^{new} = w^{old} - H^{-1}\nabla E$$

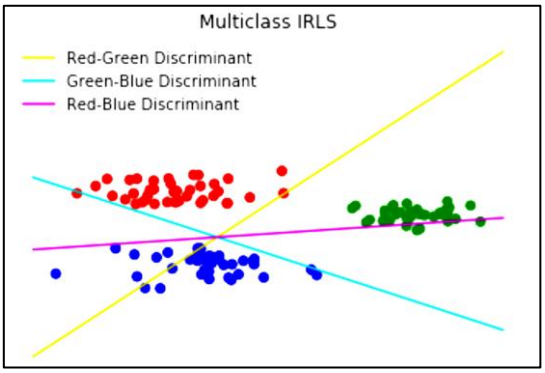
- Οι ευθείες διαχωρισμού υπολογίζονται από τον τύπο:

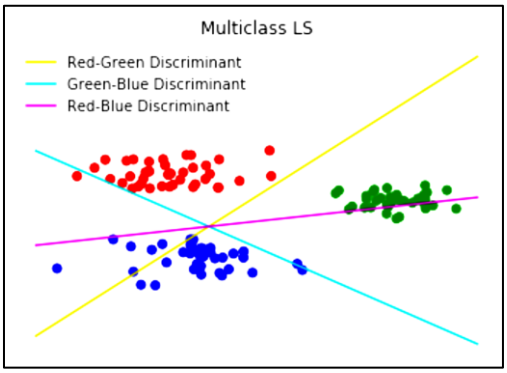
$$\left(w_i - w_j\right)\cdot x = 0$$

Σημειώνεται επίσης ότι το αρχικό διάνυσμα αρχικοποιήθηκε τυχαία με i.i.d. δείγματα κανονικής κατανομής με μ = 0, σ = 0.1.

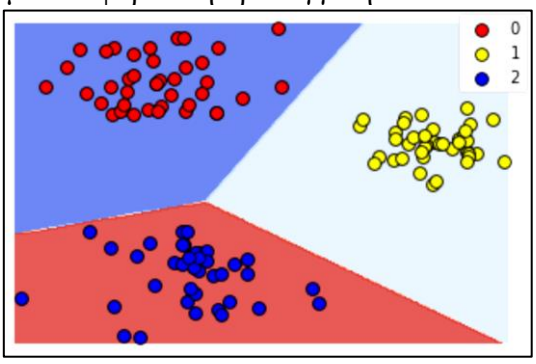
Για την επίτευξη καλύτερων αποτελεσμάτων, αφαιρείται από την είσοδο του softmax η μεγαλύτερη τιμή του διανύσματος για την αποφυγή overflow και προστίθεται στην έξοδο του softmax μια αυθαίρετη τιμή για undeflow και για αποφυγή των μη αντιστρέψιμων Hessians και διαιρέσεις με το μηδέν.

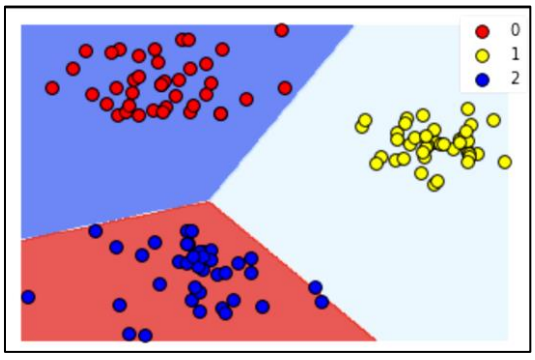
Παρακάτω φαίνεται το αποτέλεσμα για 20 επαναλήψεις.





Και με μια διαφορετική προσέγγιση:

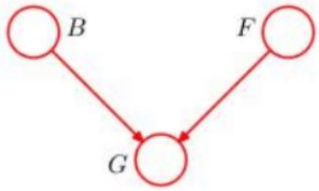




Σχολιασμός:

Οι διαφορές στα παραπάνω αποτελέσματα είναι ελάχιστες αφού, ο IRLS επιτυγχάνει 100 % ακρίβεια ενώ ο LS επιτυγχάνει ακρίβεια 99 %.

Άσκηση 2.10 (Conditional Independence)



Σχήμα 1: Μπαταρία (Battery B), ντεπόζιτο (fuel tank F) και δείκτης (gauge G) ενός αυτοκινήτου

Θεωρείστε το σύστημα καυσίμων ενός αυτοκινήτου, όπως φαίνεται στο Σχήμα 1, με γνωστές τις παρακάτω πιθανότητες:

$$p(B = 1) = 0.95$$

$$p(F = 1) = 0.8$$

$$p(G = 1|B = 1, F = 1) = 0.95$$

$$p(G = 1|B = 1, F = 0) = 0.3$$

$$p(G = 1|B = 0, F = 1) = 0.25$$

$$p(G = 1|B = 0, F = 0) = 0.2$$

Υποθέστε ότι αντί να παρατηρείτε την κατάσταση του δείκτη καυσίμων G απευθείας, ο δείκτης διαβάζεται από τον οδηγό D και μας μεταφέρει την τιμή. Οι δύο πιθανές καταστάσεις είναι είτε ότι διαβάστηκε ο δείκτης ως γεμάτος (D = 1) είτε ως άδειος (D = 0). Ο οδηγός είναι σχετικά αναξιόπιστος, όπως εκφράζεται από τις ακόλουθες πιθανότητες:

$$p(D = 1|G = 1) = 0.8$$

$$p(D = 0|G = 0) = 0.8$$

Σύμφωνα με τον οδηγό, ο μετρητής καυσίμων φαίνεται άδειος, δηλαδή παρατηρούμε ότι D = 0.

1. *Υπολογίστε την πιθανότητα το ντεπόζιτο F να είναι άδειο, με δεδομένη μόνο αυτή την παρατήρηση.*

2. *Υπολογίστε την πιθανότητα το ντεπόζιτο F να είναι άδειο, εάν επιπλέον η μπαταρία B είναι άδεια. Η δεύτερη πιθανότητα είναι μεγαλύτερη ή μικρότερη από την πρώτη· Σχολιάστε το αποτέλεσμα και εξηγήστε γιατί.*

Λύση:

1. Μετά από τον υπολογισμό των πιθανοτήτων, προκύπτουν τα εξής αποτελέσματα:

p(B)		
B	0	1
	0.05	0.95

p(F)		
F	0	1
	0.2	0.8

p(G|BF)		
BF \ G	0	1
00	0.8	0.2
01	0.75	0.25
10	0.7	0.3
11	0.05	0.095

p(D|G)		
G \ D	0	1
0	0.8	0.2
1	0.2	0.8

B	**F**	**G**	**D**	**p(B)**	**p(F)**	**p(D|G)**	**p(G|BF)**	**p(B, F, G, D)**
0	0	0	0	0.05	0.2	0.8	0.8	0.0064
0	0	1	0	0.05	0.2	0.2	0.8	0.004
0	1	0	0	0.05	0.8	0.8	0.75	0.0240
0	1	1	0	0.05	0.8	0.2	0.25	0.0020
1	0	0	0	0.95	0.2	0.8	0.7	0.1064
1	0	1	0	0.95	0.2	0.2	0.3	0.0114
1	1	0	0	0.95	0.8	0.8	0.05	0.0304
1	1	1	0	0.95	0.8	0.8	0.95	0.1444

Ισχύει ότι:

$$p(B, F, G, D) = p(D|G)\cdot p(G|BF)\cdot p(B)\cdot p(F)$$

Και

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

$$p(D = 0) = 0.325$$

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

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

Συνεπώς,

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

2. Μετά από τον υπολογισμό των πιθανοτήτων, προκύπτουν τα εξής αποτελέσματα:

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

$$p(D = 0|B = 0) = 0.033$$

Επομένως,

$$p(F = 0|D = 0, B = 0) = \frac{p(F = 0)p(B = 0)\sum_G p(D = 0|G)p(G|F = 0, B = 0)}{p(D = 0, B = 0)} = 0.207$$

Σχολιασμός:

Παρατηρείται ότι η δεύτερη πιθανότητα είναι μικρότερη. Αυτό οφείλεται στο explaining away.

Resources

[1] Γ. Καραγιάννης και Γ. Σταϊνχάουερ, Αναγνώριση Προτύπων και Μάθηση Μηχανών, ΕΜΠ, 2001.

[2] R. O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, Wiley, 2001.

[3] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[4] S. Theodoridis and K. Koutroumbas, Pattern Recognition, 4th Edition Academic Pres, Elsevier, 2009. Ελληνική μετάφραση: απόδοση-επιμέλεια-πρόλογος ελληνικής έκδοσης Α. Πικράκης, Κ. Κουτρουμπάς, Θ. Γιαννακόπουλος, Επιστημονικές Εκδόσεις Π.Χ. Πασχαλίδης-Broken Hill Publishers LTD, 2012.

[5] Nielsen, M.A., Neural Networks and Deep Learning, in Determination Press, 2015.

[6] Σημειώσεις Μαθήματος & Παλαιότερο υλικό