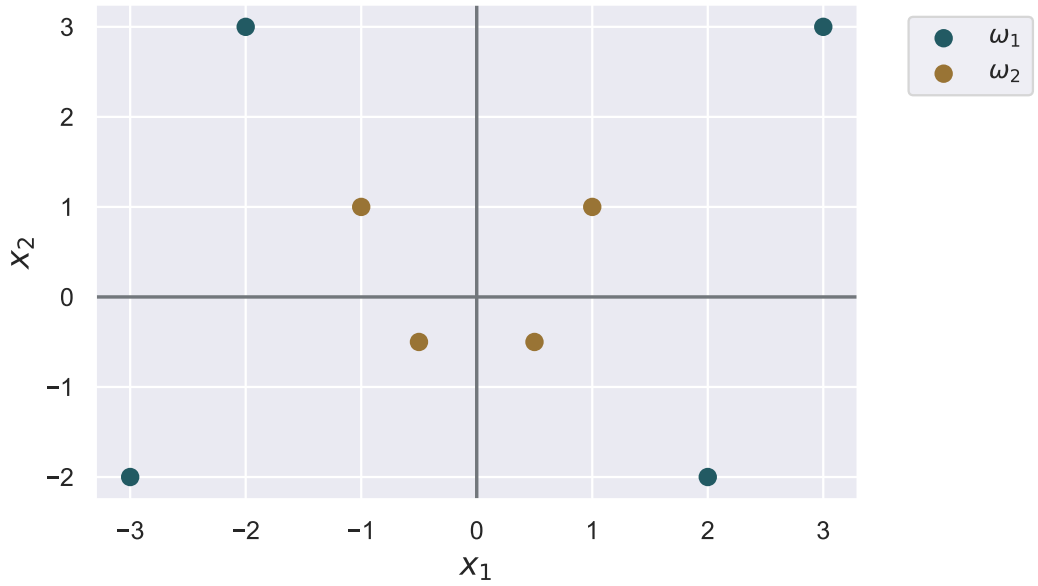
1 Support Vector Machine

1.1 Όπως φαίνεται και από την απεικόνιση του διαγράμματος διασποράς στην Εικόνα 1.1, τα σημεία που αντιστοιχούν στα 8 διανύσματα χαρακτηριστικών $\tilde{\mathbf{x}}_n = [x_1, x_2]^\mathsf{T}$ δεν ανήκουν σε γραμμικά διαχωρίσιμες κλάσεις.



Εικόνα 1.1: Απεικόνιση των 8 σημείων που αντιστοιχούν στα διανύσματα $\tilde{\mathbf{x}}_n$ των δύο κλάσεων.

Για το λόγο αυτό, αξιοποιείται η συνάρτηση $\phi : \mathbb{R}^2 \to \mathbb{R}^4$, η οποία ορίζεται μέσω της σχέσης

$$\phi(\tilde{\mathbf{x}}_n) = \left[1, x_1, x_2, \frac{{x_1}^2 + {x_2}^2 - 5}{3}\right]^\mathsf{T} \equiv \mathbf{y}_n. \tag{1.1}$$

Οι νέες μεταβλητές, $\mathbf{y}_n$, που ορίζονται βάσει της Σχέσης (1.1) παρουσιάζονται στον Πίνακα 1.1. Η απεικόνιση των αρχικών σημείων σε τρεις διαστάσεις (η τέταρτη διάσταση είναι τετριμμένη και προκύπτει απλώς λόγω επαύξησης) μέσω της ϕ καθιστά τις δύο κλάσεις γραμμικά διαχωρίσιμες, μέσω ενός προς προσδιορισμό υπερεπιπέδου.

n	$\tilde{\mathbf{x}}_n$	$\mathbf{y}_n$	Κλάση
1	$[3, 3]^\mathsf{T}$	$[1, 3, 3, 13/3]^\mathsf{T}$	ω_1
2	$[2, -2]^\mathsf{T}$	$[1, 2, -2, 1]^\mathsf{T}$	ω_1
3	$[-3, -2]^\mathsf{T}$	$[1, -3, -2, 8/3]^\mathsf{T}$	ω_1
4	$[-2, 3]^\mathsf{T}$	$[1, -2, 3, 8/3]^\mathsf{T}$	ω_1
5	$[1, 1]^\mathsf{T}$	$[1, 1, 1, -1]^\mathsf{T}$	ω_2
6	$[0.5, -0.5]^\mathsf{T}$	$[1, 0.5, -0.5, -1.5]^\mathsf{T}$	ω_2
7	$[-0.5, -0.5]^\mathsf{T}$	$[1, -0.5, -0.5, -1.5]^\mathsf{T}$	ω_2
8	$[-1, 1]^\mathsf{T}$	$[1, -1, 1, -1]^\mathsf{T}$	ω_2

Πίνακας 1.1: Αντιστοίχιση των $\tilde{\mathbf{x}}_n$ σε $\mathbf{y}_n$ μέσω της απεικόνισης ϕ.

1.2 Στόχος είναι ο προσδιορισμός του προαναφερθέντος υπερεπιπέδου μέσω ενός διανύσματος βάρους, $\mathbf{w} = [w_0, w_1, w_2, w_3]^\top$, τέτοιο, ώστε η συνάρτηση $g(\mathbf{y}_n) = \mathbf{w}^\top \mathbf{y}_n$ να αποτελεί συνάρτηση διαχωρισμού. Με άλλα λόγια, θα πρέπει η συνθήκη $z_n g(\mathbf{y}_n) \geq 1$ να ισχύει για κάθε $n = 1, \dots, 8$. Η πρόσθετη συνθήκη που διαφοροποιεί το πρόβλημα SVM από άλλα προβλήματα διακρινουσών μεθόδων, είναι η μεγιστοποίηση του περιθωρίου ταξινόμησης, β, το οποίο υπεισέρχεται στη σχέση

$$\frac{z_n g(\mathbf{y}_n)}{\|\tilde{\mathbf{w}}\|_2} \geq \beta, \tag{1.2}$$

όπου $\tilde{\mathbf{w}} = [w_1, w_2, w_3]$. Η δεύτερη αυτή συνθήκη ισοδυναμεί με την ελαχιστοποίηση της ποσότητας $\|\tilde{\mathbf{w}}\|_2$. Έτσι, ο προσδιορισμός του $\mathbf{w}$ ανάγεται στην ελαχιστοποίηση της Λαγκρανζιανής

$$\tilde{\mathcal{L}}(\mathbf{w}, \alpha_1, \dots, \alpha_8) = \frac{1}{2}\tilde{\mathbf{w}}^\top \tilde{\mathbf{w}} - \sum_{n=1}^{8} \alpha_n \left[z_n \mathbf{w}^\top \mathbf{y}_n - 1 \right] \tag{1.3}$$

ως προς το $\tilde{\mathbf{w}}$ και παράλληλα τη μεγιστοποίησή της ως προς τους πολλαπλασιαστές Lagrange $\alpha_1, \dots, \alpha_8 \geq 0$. Εάν $\alpha_j > 0$ για κάποιο $\mathbf{y}_j$, τότε (και μόνο τότε) το σημείο αυτό αντιστοιχεί σε διάνυσμα υποστήριξης [1], καθώς η συνεισφορά του στην εύρεση του υπερεπιπέδου διαχωρισμού μέσω της Σχέσης (1.3) είναι μη μηδενική. Αξιοποιώντας τις συνθήκες Karush-Kuhn-Tucker (ΚΚΤ), το πρόβλημα ανάγεται [2] στην επίλυση των εξισώσεων Euler-Lagrange για την

$$\tilde{\mathcal{L}}'(\alpha_1, \dots, \alpha_8) = \sum_{n=1}^{8} \alpha_n - \frac{1}{2} \sum_{n=1}^{8} \sum_{k=1}^{8} \alpha_n \alpha_k z_n z_k \mathbf{y}_n^\top \mathbf{y}_k, \tag{1.4}$$

με τους πολλαπλασιαστές Lagrange να υπόκεινται επιπλέον στην απαίτηση

$$\sum_{n=1}^{8} z_n \alpha_n = 0. \tag{1.5}$$

Εισάγοντας έναν πρόσθετο πολλαπλασιαστή Lagrange, λ, προκειμένου η απαίτηση (1.5) να ενσωματωθεί στην υπό μελέτη Λαγκρανζιανή, προκύπτει τελικά

$$\mathcal{L}(\alpha_1, \dots, \alpha_8, \lambda) = \sum_{n=1}^{8} \alpha_n - \frac{1}{2} \sum_{n=1}^{8} \sum_{k=1}^{8} \alpha_n \alpha_k z_n z_k \mathbf{y}_n^\top \mathbf{y}_k - \lambda \sum_{n=1}^{8} z_n \alpha_n. \tag{1.6}$$

Από τις εξισώσεις Euler-Lagrange, η $\partial \mathcal{L} / \partial \lambda = 0$ ισοδυναμεί με τη Σχέση (1.5), ενώ οι εξισώσεις $\partial \mathcal{L} / \partial \alpha_q = 0$ παίρνουν τη μορφή

$$\begin{aligned} 0 &= \sum_{n=1}^{8} \delta_{nq} - \frac{1}{2} \sum_{n=1}^{8} \sum_{k=1}^{8} \left(\delta_{nq} \alpha_k + \alpha_n \delta_{kq} \right) z_n z_k \mathbf{y}_n^\top \mathbf{y}_k - \lambda \sum_{n=1}^{8} z_n \delta_{nq} \\ &= 1 - \frac{1}{2} \sum_{n=1}^{8} 2 \alpha_n z_n z_q \mathbf{y}_n^\top \mathbf{y}_q - z_q \lambda \Leftrightarrow z_q \sum_{n=1}^{8} \alpha_n z_n \mathbf{y}_n^\top \mathbf{y}_q + z_q \lambda = 1 \\ &\Leftrightarrow \sum_{n=1}^{8} \alpha_n z_n \mathbf{y}_n^\top \mathbf{y}_q + \lambda = z_q, \end{aligned} \tag{1.7}$$

όπου στην τελευταία ισοδυναμία αξιοποιήθηκε το γεγονός πως ${z_q}^2 = 1$. Αντικαθιστώντας τις τιμές του Πίνακα 1.1, προκύπτει - συμπεριλαμβανομένης της (1.3) - ένα γραμμικό σύστημα 9 εξισώσεων, το οποίο μπορεί να γραφεί στη μορφή

$$\tilde{\mathbf{A}} \cdot \tilde{\mathbf{X}} = \tilde{\mathbf{Z}}, \tag{1.8}$$

όπου $\tilde{\mathbf{Z}} \equiv [z_1, \dots, z_8, 0]^\top$, $\tilde{\mathbf{X}} = [\alpha_1, \dots, \alpha_8, \lambda]^\top$ και

$$\tilde{\mathbf{A}} = \begin{pmatrix} -340/9 & -16/3 & 22/9 & -140/9 & 8/3 & -11/2 & -17/2 & -10/3 & 1 \\ -16/3 & -10 & -5/3 & 19/3 & 0 & 3/2 & -1/2 & 4 & 1 \\ 22/9 & -5/3 & -190/9 & -73/9 & -20/3 & -7/2 & -1/2 & -2/3 & 1 \\ -140/9 & 19/3 & -73/9 & -190/9 & -2/3 & -11/2 & -7/2 & 10/3 & 1 \\ -8/3 & 0 & 20/3 & 2/3 & 4 & 5/2 & 3/2 & 2 & 1 \\ 11/2 & -3/2 & 7/2 & 11/2 & 5/2 & 15/4 & 13/4 & 3/2 & 1 \\ 17/2 & 1/2 & 1/2 & 7/2 & 3/1 & 13/4 & 15/4 & 5/2 & 1 \\ 10/3 & 4 & 2/3 & -10/3 & 2 & 3/2 & 5/2 & 4 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}. \tag{1.9}$$

Έτσι, εφόσον ο πίνακας $\tilde{\mathbf{A}}$ είναι αντιστρέψιμος, οι πολλαπλασιαστές Lagrange θα δίνονται από τη σχέση $\tilde{\mathbf{X}} = \tilde{\mathbf{A}}^{-1} \cdot \tilde{\mathbf{Z}}$. Το ζήτημα που προκύπτει στο σημείο αυτό είναι πως η τάξη (rank) του $\tilde{\mathbf{A}}$ είναι 5, επομένως δεν αντιστρέφεται. Το γεγονός αυτό υποδεικνύει πως από τα 8 διαθέσιμα σημεία, θα πρέπει μόνο τα 4 να χρησιμοποιηθούν στην εκπαίδευση του SVM, θέτοντας τους πολλαπλασιαστές Lagrange που αντιστοιχούν στα υπόλοιπα ίσους με μηδέν. Με τον τρόπο αυτό, το πρόβλημα ανάγεται στην αντιστροφή ενός 5×5 πίνακα. Επειδή τα τέσσερα σημεία για τα οποία οι πολλαπλασιαστές Lagrange τίθενται εξ αρχής ίσοι με μηδέν εξ ορισμού δε θα μπορούν να αντιστοιχούν σε διανύσματα υποστήριξης [1], είναι χρήσιμο ανάμεσα σε αυτά να βρίσκονται τα σημεία που μπορεί κανείς με σχετική βεβαιότητα να αποφανθεί πως απέχουν αρκετά από το βέλτιστο υπερεπίπεδο. Για το σκοπό αυτό δεν απαιτείται η απεικόνιση των σημείων $\mathbf{y}_n$ σε έναν τρισδιάστατο χώρο και η εκ των προτέρων γνώση του υπερεπιπέδου. Αντιθέτως, εκφράζοντας την $g(\mathbf{y}_n)$ συναρτήσει των x_1, x_2 αντί για τα y_0, y_1, y_2, y_3, προκύπτει

$$g(x_1, x_2) = \mathbf{w}^\top \phi(\tilde{\mathbf{x}}) = \frac{w_3}{3}{x_1}^2 + w_1 x_1 + \frac{w_3}{3}{x_2}^2 + w_2 x_2 + w_0 - \frac{5w_3}{3}. \tag{1.10}$$

Η εξίσωση $g(x_1, x_2) = 0$ αντιστοιχεί στην εξίσωση της επιφάνειας (γραμμής) διαχωρισμού του αρχικού προβλήματος. Στην προκείμενη περίπτωση είναι η

$$\left(x_1 + \frac{3w_1}{2w_3}\right)^2 + \left(x_2 + \frac{3w_2}{2w_3}\right)^2 = 5 + \frac{9\left({w_1}^2 + {w_2}^2\right)}{4{w_3}^2} - \frac{3w_0}{w_3} \tag{1.11}$$

και αντιστοιχεί στην εξίσωση ενός κύκλου, αφού $w_3 \neq 0$[1]. Βάσει αυτού, μπορεί κανείς να συμπεράνει από την Εικόνα 1.1 πως τα σημεία $\tilde{\mathbf{x}}_1$ και $\tilde{\mathbf{x}}_3$ δε μπορούν έτσι κι αλλιώς να αντιστοιχούν σε διανύσματα υποστήριξης. Για τα υπόλοιπα σημεία, ένα αντίστοιχο συμπέρασμα δε μπορεί να εξαχθεί με μια απλή οπτική εποπτεία, επομένως για λόγους συμμετρίας επιλέγεται να απορριφθούν δύο σημεία της κλάσης ω_2 από την εκπαίδευση του SVM, τα $\tilde{\mathbf{x}}_6$ και $\tilde{\mathbf{x}}_7$. Θέτοντας, λοιπόν, $\alpha_1 = \alpha_3 = \alpha_6 = \alpha_7 = 0$, το πρόβλημα τώρα ανάγεται στην επίλυση του γραμμικού συστήματος 5 εξισώσεων

[1] Η περίπτωση $w_3 = 0$ αναγάγει τη γραμμή σε ευθεία και ήδη είναι γνωστό πως τα αρχικά σημεία ανήκουν σε μη γραμμικά διαχωρίσιμες κλάσεις, επομένως είναι βέβαιο πως $w_3 \neq 0$

$$\mathbf{A}\cdot\mathbf{X}=\mathbf{Z}, \tag{1.12}$$

όπου $\mathbf{Z}\equiv[z_2,z_4,z_5,z_8,0]^\top$, $\mathbf{X}=[\alpha_2,\alpha_4,\alpha_5,\alpha_8,\lambda]^\top$ και

$$\mathbf{A}=\begin{pmatrix} -10 & {}^{19}/_{3} & 0 & -4 & 1\\ {}^{19}/_{3} & {}^{-190}/_{9} & {}^{-2}/_{3} & {}^{10}/_{3} & 1\\ 0 & {}^{2}/_{3} & 4 & 2 & 1\\ 4 & {}^{-10}/_{3} & 2 & 4 & 1\\ -1 & -1 & 1 & 1 & 0 \end{pmatrix}. \tag{1.13}$$

Τώρα, ο $\mathbf{A}$ είναι αντιστρέψιμος και μέσω μιας απλής μεθόδου (π.χ. Gauss-Jordan) προκύπτει πως

$$\mathbf{A}^{-1}=\begin{pmatrix} {}^{-157}/_{2025} & {}^{4}/_{675} & {}^{161}/_{1350} & {}^{-193}/_{4050} & {}^{-17}/_{45}\\ {}^{4}/_{675} & {}^{-13}/_{225} & {}^{-17}/_{450} & {}^{121}/_{1360} & {}^{-1}/_{15}\\ {}^{-161}/_{1350} & {}^{17}/_{450} & {}^{403}/_{900} & {}^{-989}/_{2700} & {}^{-1}/_{30}\\ {}^{193}/_{4050} & {}^{-121}/_{1350} & {}^{-989}/_{2700} & {}^{3307}/_{8100} & {}^{53}/_{90}\\ {}^{17}/_{45} & {}^{1}/_{15} & {}^{-1}/_{30} & {}^{53}/_{90} & -1 \end{pmatrix}. \tag{1.14}$$

Ως εκ τούτου, οι λύσεις $\mathbf{X}=\mathbf{A}^{-1}\cdot\mathbf{Z}$ είναι οι

$$\alpha_2=\frac{58}{405},\quad \alpha_4=\frac{14}{135},\quad \alpha_5=\frac{22}{135},\quad \alpha_8=\frac{34}{405},\quad \lambda=\frac{1}{9}. \tag{1.15}$$

Καμία από τις προκύπτουσες λύσεις δεν είναι αρνητική, επομένως όλες είναι δεκτές. Συνοψίζοντας, προκύπτει πως πράγματι τα σημεία με $n=2,4,5,8$ αντιστοιχούν σε διανύσματα υποστήριξης, αφού για αυτά ισχύει $\alpha_n>0$. Αντίθετα, τα σημεία με $n=1,3,6,7$ δεν αντιστοιχούν σε διανύσματα υποστήριξης, αφού οι αντίστοιχοι πολλαπλασιαστές Lagrange είναι μηδενικοί.

1.3 Δεδομένων των συντελεστών Lagrange, το διάνυσμα $\tilde{\mathbf{w}}$ υπολογίζεται μέσω της

$$\tilde{\mathbf{w}}=\sum_{n=1}^{8}\alpha_n z_n\tilde{\mathbf{y}}_n, \tag{1.16}$$

όπου το $\tilde{\mathbf{y}}_n$ ορίζεται μέσω της $\mathbf{y}_n=[1,\tilde{\mathbf{y}}_n]^\top$. Προκύπτει, έτσι

$$\tilde{\mathbf{w}}=\left[0,\frac{2}{9},-\frac{2}{3}\right]^\top. \tag{1.17}$$

Για τον προσδιορισμό του w_0, αξιοποιείται η σχέση

$$z_n\mathbf{w}^\top\mathbf{y}_n=1, \tag{1.18}$$

η οποία επαληθεύεται για τα σημεία που αντιστοιχούν σε διανύσματα υποστήριξης. Αντικαθιστώντας οποιοδήποτε εκ των $\mathbf{y}_2,\mathbf{y}_4,\mathbf{y}_5$ και $\mathbf{y}_8$ στη Σχέση (1.18) προκύπτει για το w_0 η ίδια τιμή: $w_0=1/9$. Ισχύει, λοιπόν

$$\mathbf{w}\equiv[w_0,w_1,w_2,w_3]^\top=\left[\frac{1}{9},0,\frac{2}{9},-\frac{2}{3}\right]^\top, \tag{1.19}$$

με βάση το οποίο μπορεί να υπολογιστεί το γινόμενο $z_n\mathbf{w}^\top\mathbf{y}_n$ για κάθε $n = 1, \dots, 8$. Τα αποτελέσματα συνοψίζονται στον Πίνακα 1.2. Η αρχική υπόθεση πως τα σημεία με $n = 1$ και $n = 3$ δε θα μπορούσαν ούτως ή άλλως να αντιστοιχούν σε διανύσματα υποστήριξης επαληθεύεται, αφού για αυτά ισχύει πως $z_1\mathbf{w}^\top\mathbf{y}_1 = z_3\mathbf{w}^\top\mathbf{y}_3 > 1$. Από την άλλη, φαίνεται πως τα σημεία με $n = 6$ και $n = 7$ ανήκουν στο όριο ταξινόμησης, επομένως μια διαφορετική αρχική επιλογή θα μπορούσε να οδηγήσει στο να θεωρηθούν διανύσματα υποστήριξης, παρότι στην προκείμενη περίπτωση κάτι τέτοιο δεν ισχύει[2]. Σε κάθε περίπτωση, επαληθεύεται η συνθήκη $z_n\mathbf{w}^\top\mathbf{y}_n \geq 1$, για κάθε $n = 1, \dots, 8$.

n	1	2	3	4	5	6	7	8
$z_n\mathbf{w}^\top\mathbf{y}_n$	$\frac{19}{9}$	1	$\frac{19}{9}$	1	1	1	1	1
Διάνυσμα Υποστήριξης;	×	✓	×	✓	✓	×	×	✓

Πίνακας 1.2: Υπολογισμός του $z_n\mathbf{w}^\top\mathbf{y}_n$ για κάθε $n = 1, \dots, 8$.

1.4 Βάσει της Σχέσης (1.2) και της απαίτησης $z_n\mathbf{w}^\top\mathbf{y}_n \geq 1$ για κάθε n, το περιθώριο, β, της ταξινόμησης προκύπτει αντίστροφο του μέτρου του $\tilde{\mathbf{w}}$, επομένως ισχύει

$$\beta = \frac{1}{\|\tilde{\mathbf{w}}\|_2} = \frac{9\sqrt{10}}{20}. \tag{1.20}$$

Σημειώνεται πως εδώ το όριο της ταξινόμησης έχει θεωρηθεί ως η απόσταση ενός σημείου που αντιστοιχεί σε διάνυσμα υποστήριξης από το αντίστοιχο υπερεπίπεδο, αν και στη βιβλιογραφία συναντάται (σπανιότερα) και ως το άθροισμα των αποστάσεων μεταξύ δύο σημείων που αντιστοιχούν σε διανύσματα υποστήριξης διαφορετικών κλάσεων από το αντίστοιχο υπερεπίπεδο [δηλαδή το διπλάσιο του αποτελέσματος της Σχέσης (1.20)].

1.5 Σε ό,τι αφορά τη συνάρτηση διαχωρισμού στον αρχικό χώρο, αυτή προκύπτει με απλή αντικατάσταση του αποτελέσματος της Σχέσης (1.17) στην (1.10). Ισχύει

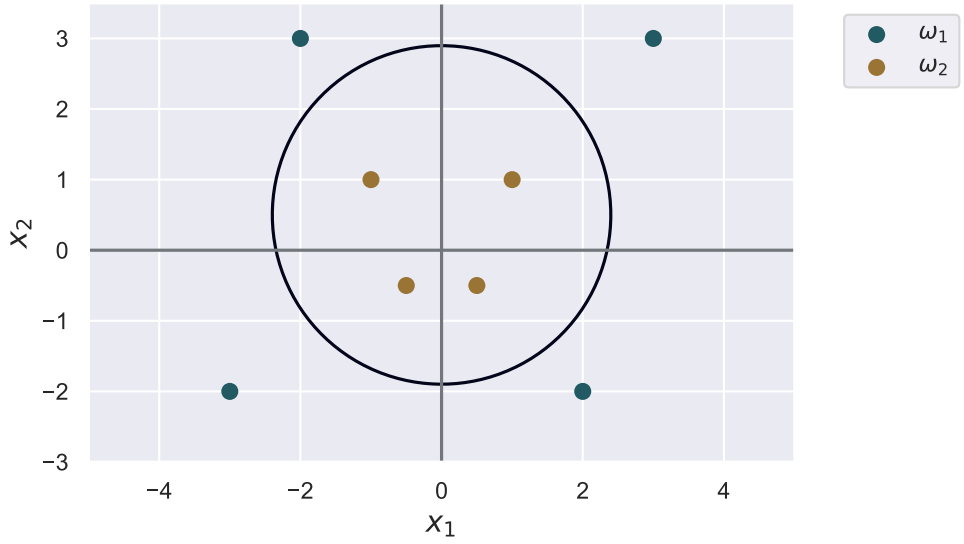
$$g(x_1, x_2) = -\frac{2}{9}\left({x_1}^2 + {x_2}^2 - x_2 - 5.5\right). \tag{1.21}$$

Ο κύκλος που αντιστοιχεί στην εξίσωση $g(x_1, x_2) = 0$ προκύπτει αντικαθιστώντας την τιμή του $\mathbf{w}$ στη Σχέση (1.11) και περιγράφεται από την εξίσωση

$${x_1}^2 + \left(x_2 - \frac{1}{2}\right)^2 = \frac{23}{4}, \tag{1.22}$$

είναι δηλαδή ο κύκλος με κέντρο $K(0, 0.5)$ και ακτίνα $\rho = 0.5\sqrt{23}$. Στο σχήμα της Εικόνας 1.2 απεικονίζεται ο συγκεκριμένος κύκλος μαζί με τα αρχικά σημεία, καθιστώντας έτσι ξεκάθαρο το διαχωρισμό των δύο κλάσεων.

[2] Τονίζεται στο σημείο αυτό πως η συνθήκη προκειμένου ένα σημείο να αντιστοιχεί σε διάνυσμα υποστήριξης δεν είναι να ανήκει στο όριο που θέτει η Σχέση (1.18), αλλά να έχει μη μηδενικό πολλαπλασιαστή Lagrange, ούτως ώστε να έχει συμμετάσχει στην «εκπαίδευση» του ταξινομητή. Το γεγονός αυτό σημαίνει πως μια διαφορετική αρχική επιλογή θα μπορούσε κάλλιστα να έχει οδηγήσει σε διαφορετικά διανύσματα υποστήριξης στο σημείο αυτό, όμως η επιφάνεια διαχωρισμού θα παρέμενε ίδια, αφού η λύση για το διάνυσμα βάρους, $\mathbf{w}$, είναι μοναδική.



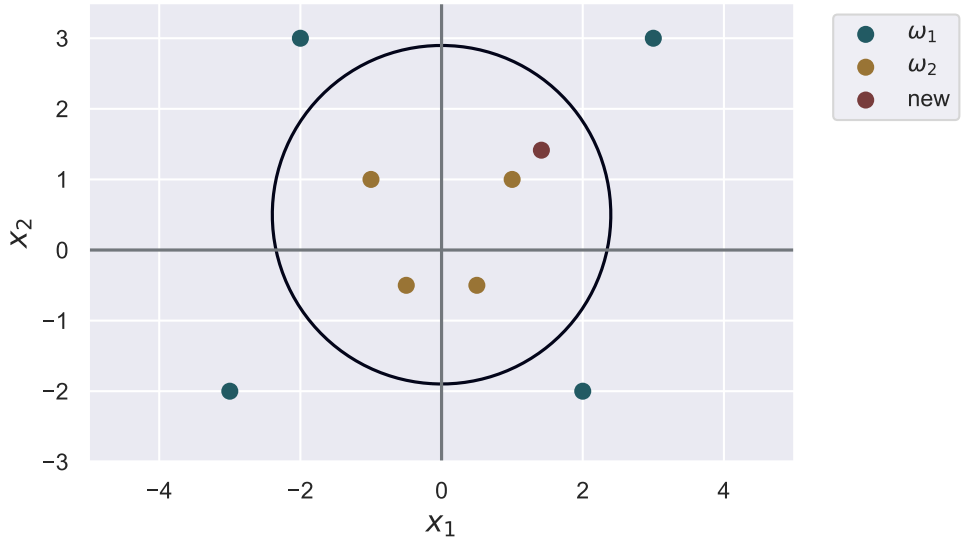
Εικόνα 1.2: Απεικόνιση των αρχικών σημείων και του αντίστοιχου κύκλου διαχωρισμού.

1.6 Όπως ήδη αναλύθηκε, εφόσον οι πολλαπλασιαστές Lagrange των σημείων με $n = 2, 4, 5, 8$ είναι μη μηδενικοί, αυτά είναι και τα σημεία που αντιστοιχούν στα διανύσματα υποστήριξης.

1.7 Σε ό,τι αφορά το σημείο $(-2, 3)$, αυτό είναι ένα από τα αρχικά σημεία, το οποίο είναι γνωστό πως ανήκει στην κλάση ω_1. Από την άλλη, το σημείο $\left(\sqrt{2}, \sqrt{2}\right)$ είναι ένα νέο σημείο, το οποίο ταξινομείται στην κλάση ω_2, αφού

$$\sqrt{2}^2 + \left(\sqrt{2} - \frac{1}{2}\right)^2 = \frac{9}{4} - \sqrt{2} < \rho^2. \tag{1.23}$$

Προσθέτοντας το σημείο $\left(\sqrt{2}, \sqrt{2}\right)$ στο σχήμα της Εικόνας 1.2 προκύπτει η Εικόνα 1.3, στην οποία φαίνεται και οπτικά η ταξινόμηση του σημείου αυτού στην κλάση ω_2, αφού βρίσκεται στο εσωτερικό του κύκλου που ορίζεται από τη Σχέση (1.22).



Εικόνα 1.3: Ταξινόμηση του νέου σημείου, $\left(\sqrt{2}, \sqrt{2}\right)$.

2 Hidden Markov Model

Το υπό μελέτη HMM αποτελείται από τρεις καταστάσεις $q = 1, 2, 3$ και δύο είδη παρατηρήσεων, τις O = H και O = T. Ο πίνακας μεταβάσεων **A** λαμβάνει την τιμή $A_{ij} = 1/3, \forall i, j$, ενώ οι a-priori πιθανότητες είναι επίσης κοινές για κάθε κατάσταση, με $\boldsymbol{\pi} = [1/3, 1/3, 1/3]^{\mathsf{T}}$. Τέλος, οι πιθανότητες των παρατηρήσεων δίνονται από τον πίνακα εκπομπής **b**, για τον οποίο ισχύουν

$$\begin{aligned} b_1(H) = 0.5, \quad b_2(H) = 0.75, \quad b_3(H) = 0.25 \\ b_1(T) = 0.5, \quad b_2(T) = 0.25, \quad b_3(T) = 0.75 \end{aligned} \tag{2.1}$$

2.1 Με βάση τον άνω συμβολισμό, ο αλγόριθμος Forward για την ακολουθία $\mathbf{O} = (\mathrm{H}, \mathrm{T}, \mathrm{H})$ αρχικοποιείται ως

$$\alpha_1(j) = \pi_j b_j(\mathrm{O}_1), \ \ 1 \leq j \leq 3. \tag{2.2}$$

Δεδομένων των α_1, η αναδρομή γίνεται βάσει της σχέσης

$$\alpha_t(j) = b_j(\mathrm{O}_t) \sum_{i=1}^{3} A_{ij} \alpha_{t-1}(i), \ \ 1 \leq j \leq 3, \ \ 1 < t \leq 3. \tag{2.3}$$

Τέλος, η ζητούμενη πιθανότητα υπολογίζεται ως

$$P(\mathbf{O}|\lambda) = \sum_{i=1}^{3} \alpha_3(i). \tag{2.4}$$

Αντικαθιστώντας τις δοθείσες τιμές, η αρχικοποίηση δίνει

$$\alpha_1(1) = \frac{1}{6}, \quad \alpha_1(2) = \frac{1}{4}, \quad \alpha_1(3) = \frac{1}{12}. \tag{2.5}$$

Κατά τον πρώτο κύκλο της αναδρομής, και παρατηρώντας ότι το στοιχείο πίνακα A_{ij} είναι κοινό για κάθε i, j και ως εκ τούτου μπορεί να βγει από το άθροισμα, προκύπτουν

$$\alpha_2(1) = \frac{1}{12}, \quad \alpha_2(2) = \frac{1}{24}, \quad \alpha_2(3) = \frac{1}{8}. \tag{2.6}$$

Αντίστοιχα, κατά το δεύτερο κύκλο παίρνει κανείς

$$\alpha_3(1) = \frac{1}{24}, \quad \alpha_3(2) = \frac{1}{16}, \quad \alpha_3(3) = \frac{1}{48}, \tag{2.7}$$

συνεπώς η ζητούμενη πιθανότητα ισούται τελικά με

$$P(\mathbf{O}|\lambda) = \alpha_3(1) + \alpha_3(2) + \alpha_3(3) = \frac{1}{8}. \tag{2.8}$$

2.2 Ο αλγόριθμος Backward αρχικοποιείται με τετριμμένο τρόπο ως

$$\beta_3(j) = 1, \ \ 1 \leq j \leq 3, \tag{2.9}$$

ενώ η αντίστοιχη αναδρομική σχέση είναι η

$$\beta_t(j) = \sum_{i=1}^{3} A_{ji} b_i(O_{t+1})\, \beta_{t+1}(i), \;\; 1 \le j \le 3, \;\; 1 \le t < 3. \tag{2.10}$$

Σημειώνεται εδώ πως κατά τον πρώτο κύκλο της αναδρομής τα β_3 πρακτικά βγαίνουν εκτός αθροίσματος, αφού είναι ίσα για κάθε $j = 1, 2, 3$. Επιπλέον, το στοιχείο πίνακα A_{ij} παραμένει κοινό για κάθε συνδυασμό των i, j, επομένως και αυτό μπορεί να βγει από το άθροισμα. Τελικά, ο υπολογισμός των νέων β ανάγεται στην πράξη

$$\beta_2(j) = \frac{1}{3} \cdot 1 \sum_{i=1}^{3} b_i(O_3), \tag{2.11}$$

η οποία θα ξαναδώσει κοινές τιμές για τα β_2, με αποτέλεσμα να αναμένει κανείς κοινές τιμές και για τα β_1[3]. Γίνεται, λοιπόν, εμφανές, πως ο αλγόριθμος Backward είναι πολύ πιο απλός από τον Forward για το μοντέλο λ. Έχοντας τελικά φτάσει στα β_1, η ζητούμενη πιθανότητα υπολογίζεται ως

$$P(\mathbf{O}|\lambda) = \sum_{j=1}^{3} \pi_j b_j(O_1)\, \beta_1(j) \tag{2.12}$$

Αντικαθιστώντας τις δοθείσες τιμές για τις παραμέτρους, ο πρώτος κύκλος της αναδρομής δίνει

$$\beta_2(1) = \frac{1}{2}, \quad \beta_2(2) = \frac{1}{2}, \quad \beta_2(3) = \frac{1}{2}, \tag{2.13}$$

ενώ ο δεύτερος δίνει

$$\beta_1(1) = \frac{1}{4}, \quad \beta_1(2) = \frac{1}{4}, \quad \beta_1(3) = \frac{1}{4}. \tag{2.14}$$

Τελικά, η πιθανότητα υπολογίζεται ίση με

$$P(\mathbf{O}|\lambda) = \frac{1}{3} \cdot \frac{1}{4} \sum_{j=1}^{3} b_j(O_1) = \frac{1}{8}, \tag{2.15}$$

το οποίο, όπως είναι αναμενόμενο, ταυτίζεται με το αποτέλεσμα του αλγορίθμου Forward.

2.3 Στα πλαίσια προσδιορισμού της πιθανότερης σειράς καταστάσεων δεδομένης της ακολουθίας **O**, αξιοποιείται η αποκωδικοποίηση Viterbi. Συγκεκριμένα, η διαδικασία είναι σχεδόν ταυτόσημη με αυτή του αλγορίθμου Forward, απλώς τα αθροίσματα αντικαθίστανται από μεγιστοποιήσεις. Με άλλα λόγια, η αρχικοποίηση γίνεται ως

$$v_1(j) = \pi_j b_j(\mathrm{O}_1), \;\; 1 \le j \le 3. \tag{2.16}$$

[3] Θα μπορούσε κανείς, μάλιστα, να γενικεύσει λέγοντας πως στο συγκεκριμένο πρόβλημα η πιθανότητα οποιασδήποτε ακολουθίας μήκους N θα αντιστοιχεί σε πιθανότητα ίση με $(1/2)^N$, καθώς $\sum_{i=1}^{3} b_i(O_t) = 3/2$ για κάθε t, από όπου προκύπτει πως $\beta_t = \beta_{t+1}/2$, λόγω της Σχέσης (2.10).

Σε ό,τι αφορά την αναδρομή, αυτή γενικά εκφράζεται ως

$$v_t(j) = b_j(\mathrm{O}_t) \max_{i=1}^{3} \left[A_{ij}\, v_{t-1}(i)\right], \;\; 1 \le j \le 3, \;\; 1 < t \le 3. \tag{2.17}$$

Στη συγκεκριμένη περίπτωση, όμως, εφόσον τα A_{ij} είναι κοινά για κάθε συνδυασμό i, j και ίσα με $1/3$, η παραπάνω σχέση ισοδυναμεί με την

$$v_t(j) = \frac{1}{3}\, b_j(\mathrm{O}_t) \max_{i=1}^{3} [v_{t-1}(i)] \propto \max_{i=1}^{3} \left[b_j(\mathrm{O}_{t-1})\right]. \tag{2.18}$$

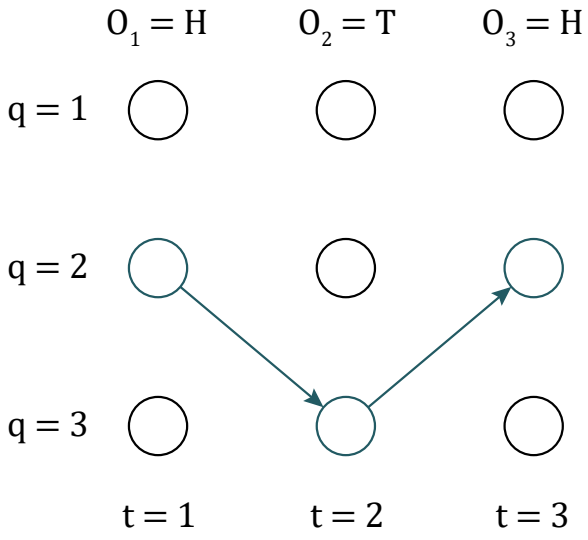
Το γεγονός αυτό σημαίνει πως στην προκείμενη περίπτωση ο αλγόριθμος Viterbi απλοποιείται σημαντικά, αφού δεν απαιτείται πρώτα ο υπολογισμός της μέγιστης πιθανότητας μέσω των αναδρομών αυτών και κατόπιν ο προσδιορισμός της ακολουθίας των καταστάσεων που οδήγησε σε αυτήν. Αντίθετα, η Σχέση (2.18) υποδεικνύει πως η πιθανότερη κατάσταση του κύκλου t γίνεται αυτομάτως γνωστή κατά τον κύκλο $t+1$, διότι το πρόβλημα έχει αναχθεί στον προσδιορισμό της κατάστασης q που μεγιστοποιεί το στοιχείο $b_q(\mathrm{O}_t)$. Με λίγα λόγια, η Σχέση (2.18) σε συνδυασμό με τις Σχέσεις (2.1) καθιστούν το πρόβλημα ισοδύναμο με την αντιστοίχιση

$$\mathrm{H} \leftrightarrow 2, \quad \mathrm{T} \leftrightarrow 3. \tag{2.19}$$

Έτσι, η πιο πιθανή ακολουθία καταστάσεων που αντιστοιχεί στις παρατηρήσεις $\mathbf{O}$ θα είναι η

$$2 \to 3 \to 2,$$

όπως απεικονίζεται και στο Trellis της Εικόνας 2.1.



Εικόνα 2.1: Trellis όπου με μπλε απεικονίζεται η πιθανότερη ακολουθία καταστάσεων βάσει της ακολουθίας παρατηρήσεων $\mathbf{O}$ = (H, T, H).

Σε ό,τι αφορά την τελική πιθανότητα, αυτή θα ισούται με

$$p(q_1 = 2, q_2 = 3, q_3 = 2|\mathbf{O}) = \left(\frac{1}{3}\right)^3 \cdot 0.75^3 = \frac{1}{64}, \tag{2.20}$$

όπου ο παράγοντας 1/3 προκύπτει από την αρχικοποίηση λόγω των a-priori πιθανοτήτων, ο παράγοντας $(1/3)^2$ προκύπτει από τους δύο κύκλους της αναδρομής, ενώ οι παράγοντες 0.75 προκύπτουν ως οι τιμές των $b_2(\mathrm{H})$ και $b_3(\mathrm{T})$. Μπορεί, μάλιστα, κανείς να γενικεύσει λέγοντας πως για μια ακολουθία παρατηρήσεων μήκους N, η μέγιστη πιθανότητα που θα προκύπτει από τον αλγόριθμο Viterbi θα ισούται με

$$p\left(\hat{Q}|\mathbf{O}\right) = \left(\frac{1}{3}\right)^N \left(\frac{3}{4}\right)^N = \left(\frac{1}{2}\right)^{2N} \tag{2.21}$$

και φυσικά θα αντιστοιχεί στην ακολουθία καταστάσεων που θα δίνεται από τις αντιστοιχίες της Σχέσης (2.19). Καθαρά για λόγους πληρότητας, παρατίθενται οι αναλυτικοί υπολογισμοί των v_t για κάθε κατάσταση. Η αρχικοποίηση δίνει

$$v_1(1) = \frac{1}{6}, \quad v_1(2) = \frac{1}{4}, \quad v_1(3) = \frac{1}{12}, \tag{2.22}$$

συνεπώς για τον πρώτο κύκλο του αλγορίθμου χρησιμοποιείται σε κάθε περίπτωση η $v_1(2)$. Βάσει αυτής, για τον πρώτο κύκλο προκύπτουν

$$v_2(1) = \frac{1}{24}, \quad v_2(2) = \frac{1}{48}, \quad v_2(3) = \frac{1}{16}. \tag{2.23}$$

Όπως ήταν αναμενόμενο, για το δεύτερο κύκλο χρησιμοποιείται σε κάθε περίπτωση η $v_2(3)$, παίρνοντας έτσι

$$v_3(1) = \frac{1}{94}, \quad v_3(2) = \frac{1}{64}, \quad v_3(3) = \frac{1}{192} \tag{2.24}$$

και επαληθεύοντας όσα αναλύθηκαν παραπάνω.

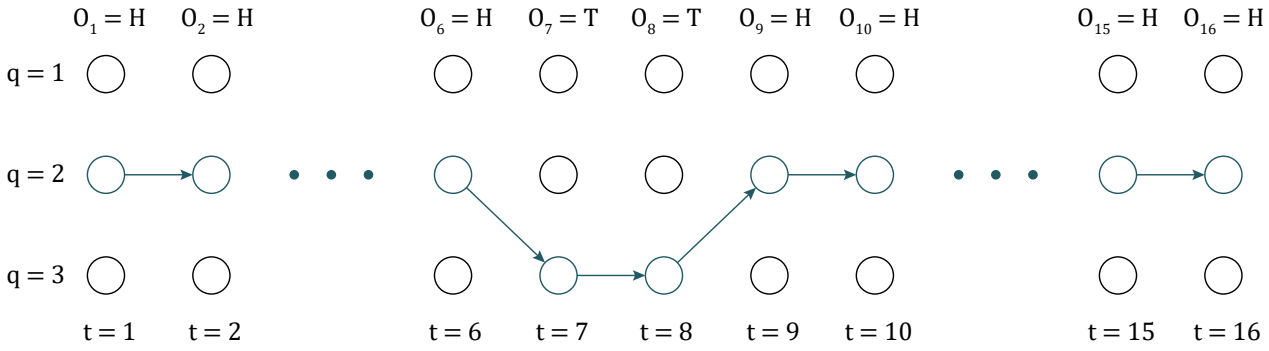
2.4 Σε ό,τι αφορά την εκπαίδευση του μοντέλου χρησιμοποιώντας εκπαίδευση Viterbi με την ακολουθία παρατηρήσεων $\mathbf{O}'$ = (H, H, H, H, H, H, T, T, H, H, H, H, H, H, H, H), το πρώτο βήμα, δηλαδή η αποκωδικοποίηση, είναι τετριμμένο βάσει της παραπάνω ανάλυσης. Το μήκος της ακολουθίας είναι $N = 16$, συνεπώς η πιο πιθανή ακολουθία καταστάσεων (βλ. Trellis της Εικόνας 2.2) είναι η

$$\hat{Q}^{(1)} = 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2,$$

με πιθανότητα $p\left(\hat{Q}^{(1)}|\mathbf{O}'\right) = (1/2)^{32} \simeq 2.3 \cdot 10^{-10}$. Δεδομένων των παραπάνω καταστάσεων, μπορεί κανείς να υπολογίσει εκ νέου τις παραμέτρους του μοντέλου λ, χρησιμοποιώντας τον αλγόριθμο μέγιστης πιθανοφάνειας, ο οποίος εδώ ανάγεται σε απλές καταμετρήσεις της μορφής

$$p(c = k) = \frac{\#\text{ occurrences where } c = k}{\#\text{ occurrences where } c = \star}. \tag{2.25}$$

Σε ό,τι αφορά τον επαναπροσδιορισμό των a-priori πιθανοτήτων, αφού ο αλγόριθμος εκπαίδευσης Viterbi ασχολείται μόνο με την πιθανότερη ακολουθία, οδηγεί σε $\pi_1^{(1)} = \pi_3^{(1)} = 0$ και $\pi_2^{(1)} = 1$. Σε ό,τι έχει να κάνει με τα στοιχεία του πίνακα μετάβασης, οι μόνες μεταβάσεις που παρατηρούνται είναι οι $2 \to 2$, $2 \to 3$, $3 \to 2$ και $3 \to 3$. Έτσι, τα αντίστοιχα στοιχεία πίνακα $A_{ij}^{(1)}$ προκύπτουν ίσα με



Εικόνα 2.2: Trellis όπου με μπλε απεικονίζεται η πιθανότερη ακολουθία καταστάσεων βάσει της ακολουθίας παρατηρήσεων $\mathbf{O}'$.

$$A_{22}^{(1)} = \frac{12}{13}, \quad A_{23}^{(1)} = \frac{1}{13}, \quad A_{32}^{(1)} = \frac{1}{2}, \quad A_{33}^{(1)} = \frac{1}{2}, \tag{2.26}$$

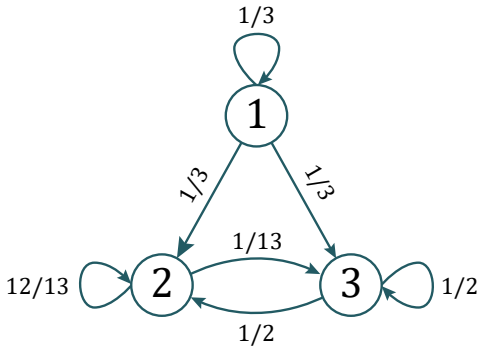
καθώς επίσης $A_{21}^{(1)} = A_{31}^{(1)} = 0$, ενώ τα A_{1j} δεν μπορούν να ανανεωθούν. Ισχύει, δηλαδή:

$$\boldsymbol{\pi}^{(1)} = [0, 1, 0]^{\mathsf{T}} \quad \text{και} \quad \mathbf{A}^{(1)} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 12/13 & 1/13 \\ 0 & 0.5 & 0.5 \end{pmatrix}. \tag{2.27}$$

Σχετικά με τα στοιχεία $b_i\,(O_t)$, λόγω της αντιστοιχίας της Σχέσης (2.19), ισχύουν

$$b_2^{(1)}\,(H) = 1 \quad \text{και} \quad b_3^{(1)}\,(T) = 1, \tag{2.28}$$

με τα υπόλοιπα στοιχεία $b_i^{(1)}\,(O_t)$ να ισούνται με 0. Γίνεται στο σημείο αυτό αντιληπτό πως η κατάσταση $q = 1$ έχει πρακτικά αφαιρεθεί από το μοντέλο, αφού δεν υπάρχει δυνατότητα μετάβασης από άλλη κατάσταση σε αυτήν (βλ. Εικόνα 2.3) και παράλληλα η ακολουθία καταστάσεων δεν μπορεί να ξεκινά από αυτήν λόγω της Σχέσης (2.28) (ακόμα κι αν οι a-priori πιθανότητες ήταν διαφορετικές)[4].



Εικόνα 2.3: Το δίκτυο που αντιστοιχεί στο σύστημα μετά την πρώτη επανάληψη του αλγορίθμου εκπαίδευσης Viterbi.

[4] Ένας τρόπος με τον οποίο θα μπορούσε κανείς να διατηρήσει την κατάσταση αυτή στο μοντέλο θα ήταν μέσω της διαδικασίας pseudo-counting, όπου κατά τον υπολογισμό των πιθανοτήτων με βάση τη Σχέση (2.25) γίνεται η παραδοχή πως κάθε κατάσταση εμφανίζεται τουλάχιστον μία φορά.

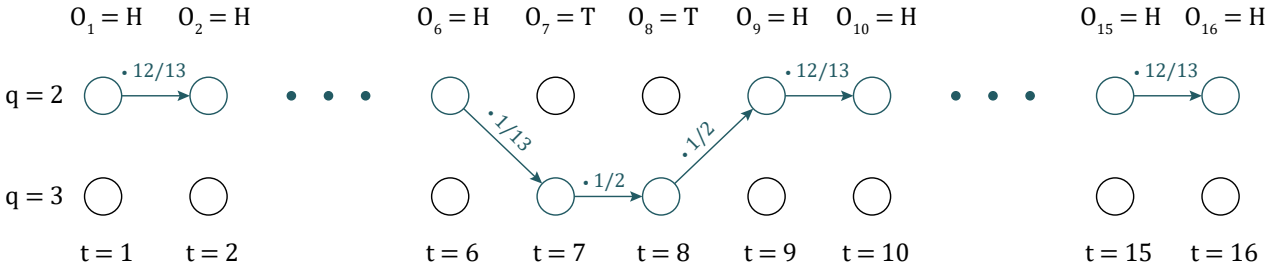
Προχωρώντας στη δεύτερη επανάληψη του αλγορίθμου, η αποκωδικοποίηση Viterbi παραμένει τετριμμένη υπόθεση, παρότι πλέον τα στοιχεία A_{ij} δεν είναι όλα μεταξύ τους ίσα. Ο λόγος είναι η Σχέση (2.28), βάσει της οποίας η αρχικοποίηση του αλγορίθμου δίνεται από τα

$$v_1(1) = 0, \quad v_1(2) = 1, \quad v_1(3) = 0, \tag{2.29}$$

ενώ τα υπόλοιπα βήματα της αναδρομής της Σχέσης (2.17) απλοποιούνται τώρα στη μορφή

$$v_t(j) = \begin{cases} A_{jj} v_{t-1}(j), & \text{εάν } O_t = O_{t+1} \\ A_{ij} v_{t-1}(i), & \text{εάν } O_t \neq O_{t+1} \end{cases}, \quad i \neq j, \ 2 \leq i,j \leq 3, \ 1 < t \leq 16 \tag{2.30}$$

Με άλλα λόγια, η αποκωδικοποίηση ισοδυναμεί, ξανά, με την αντιστοίχιση της Σχέσης (2.19). Οι παρατηρήσεις αυτές συνοψίζονται στο Trellis της Εικόνας 2.4, όπου η κατάσταση $q = 1$ έχει αφαιρεθεί, μιας και δεν υπάρχει τρόπος να επανέλθει στην ανάλυση μετά την πρώτη επανάληψη της εκπαίδευσης Viterbi.



Εικόνα 2.4: Trellis όπου με μπλε απεικονίζεται η πιθανότερη ακολουθία καταστάσεων για τη δεύτερη επανάληψη της εκπαίδευσης Viterbi.

Βάσει αυτών, προκύπτει ξανά πως η πιθανότερη ακολουθία καταστάσεων είναι η

$$\hat{Q}^{(2)} = 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 3 \to 3 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2 \to 2,$$

αυτή τη φορά με πιθανότητα

$$p\left(\hat{Q}^{(2)} | \mathbf{O}'\right) = v_1(2) \cdot \left(\frac{12}{13}\right)^5 \cdot \frac{1}{13} \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{12}{13}\right)^7 = \frac{12^{12}}{4 \cdot 13^{13}} \simeq 7.4 \cdot 10^{-3}. \tag{2.31}$$

Εφόσον η πιθανότερη ακολουθία καταστάσεων προκύπτει ίδια με αυτή που προέκυψε κατά την πρώτη επανάληψη του αλγορίθμου, συμπεραίνει κανείς πως το μοντέλο λ δεν πρόκειται να εκπαιδευτεί περαιτέρω, αφού $\boldsymbol{\pi}^{(2)} = \boldsymbol{\pi}^{(1)}$, $\mathbf{A}^{(2)} = \mathbf{A}^{(1)}$ και $b_q^{(2)}(O_t) = b_q^{(1)}(O_t)$. Ως εκ τούτου, για τον αλγόριθμο εκπαίδευσης Viterbi έχει επέλθει σύγκλιση ήδη από τη δεύτερη επανάληψη.

2.5 Ο αλγόριθμος εκπαίδευσης Forward-Backward έχει τα χαρακτηριστικά κάθε αλγορίθμου EM (Expectation-Maximization) και για το λόγο αυτό αναμένεται να είναι ακριβέστερος ως προς τα αποτελέσματά του, ενώ η σύγκλισή του σε κάποιο μέγιστο είναι εγγυημένη. Από την άλλη, το τρομερό προτέρημα του αλγορίθμου εκπαίδευσης Viterbi, ο οποίος είναι το αντίστοιχο ενός pseudo-EM για HMMs, είναι ο σημαντικά μειωμένος υπολογιστικός χρόνος σε σχέση με τον Forward-Backward, ο οποίος αφού καλείται να επιτελέσει δύο περάσματα στα δεδομένα, ένα για τη διαδικασία Forward και ένα για την Backward. Φυσικά, η ταχύτητα αυτή έρχεται με το κόστος της ακρίβειας, αφού η σύγκλιση του αλγορίθμου εκπαίδευσης Viterbi δεν είναι εγγυημένη, ενώ τα αντίστοιχα αποτελέσματα για τις τελικές πιθανότητες δεν αναμένεται να είναι εξίσου ακριβή με αυτά του Forward-Backward.

3 Classification & Regression Trees

Σκοπός του προβλήματος είναι η αξιοποίηση του δοθέντος πίνακα αξιολογήσεων, με σκοπό τη διερεύνηση των παραμέτρων που επηρεάζουν περισσότερο την ικανοποίηση ενός πελάτη, σε ό,τι αφορά τη χρήση ενός αυτόματου συστήματος διαλόγου για κράτηση εισιτηρίων.

3.1 Για το σκοπό αυτό, κατασκευάζεται ένα δένδρο ταξινόμησης, όπου οι κατηγορίες ω_1 και ω_2 είναι το εάν ο πελάτης έμεινε εν τέλει ικανοποιημένος ή όχι, αντίστοιχα, από το αυτόματο σύστημα διαλόγου. Οι παράμετροι για τη συγκεκριμένη ταξινόμηση είναι τρεις: η ποσοστιαία ακρίβεια ως προς τις λέξεις που αναγνωρίστηκαν επιτυχώς από το σύστημα αναγνώρισης φωνής (`WORD_ACCURACY`), ο χρόνος που διήρκεσε η διάδραση σε λεπτά (`TASK_DURATION`) και το κατά πόσο η κράτηση εν τέλει ολοκληρώθηκε επιτυχώς ή όχι (`TASK_COMPLETION`). Δεδομένου πως το δένδρο ταξινόμησης είναι δυαδικό, όλες οι ερωτήσεις τίθενται σε μορφή τέτοια, ώστε να επιδέχονται απαντήσεις τύπου ναι/όχι (Y/N). Έτσι, ενώ η ερώτηση που αφορά την παράμετρο `TASK_COMPLETION` είναι η προφανής («ολοκληρώθηκε η κράτηση;»), οι ερωτήσεις ως προς τις άλλες δύο παραμέτρους πρέπει να τεθούν με τη μορφή ανισώσεων. Συνολικά, οι 9 ερωτήσεις που μπορούν να τεθούν με βάση τα δεδομένα του πίνακα αξιολογήσεων είναι οι ακόλουθες:

Q1 : Was the task completed ?

Q2 : Was the word accuracy > 95% ? Q3 : Was the word accuracy > 90% ?

Q4 : Was the word accuracy > 85% ? Q5 : Was the word accuracy > 80% ?

Q6 : Was the task duration > 4 min ? Q7 : Was the task duration > 3 min ?

Q8 : Was the task duration > 2 min ? Q9 : Was the task duration > 1 min ?

Σημειώνεται στο σημείο αυτό πως οι ερωτήσεις που αφορούν τις μη κατηγορηματικές παραμέτρους `TASK_DURATION` και `WORD_ACCURACY` θα μπορούσαν να περιλαμβάνουν και ισότητες (π.χ. «ήταν η διάρκεια ακριβώς 4 λεπτά;»), ή και ανισότητες της μορφής «ήταν η διάρκεια μεταξύ 2 και 4 λεπτών;», όμως επιλέχθηκε η μορφή των ερωτήσεων να είναι τέτοια, που να αντιστοιχεί στο παράδειγμα που λύθηκε στις διαλέξεις, όπου τέτοιου είδους ερωτήσεις επίσης απορρίφθηκαν. Τέλος, είναι προφανές πως οι ερωτήσεις που αντιστοιχούν στις Q2-Q9 με ανεστραμμένη φορά της ανισότητας δίνουν την ίδια πληροφορία με τις ερωτήσεις Q2-Q9, καθώς αποτελούν την άρνησή τους (μη συμπεριλαμβανομένης της ισότητας).

3.2 Δεδομένων των ερωτήσεων αυτών και των δεδομένων του πίνακα αξιολογήσεων, η κατασκευή του δυαδικού δένδρου βασίζεται στην επιλογή ερωτήσεων με βάση την ελαχιστοποίηση της εντροπίας σε κάθε ερώτηση. Για το σκοπό αυτό, δεδομένης μιας ερώτησης, Q, τα δεδομένα αρχικά κατηγοριοποιούνται ανάλογα με τις απαντήσεις Y/N. Συγκεκριμένα, τα δεδομένα που αντιστοιχούν σε αποδοχή της ερώτησης Q διαχωρίζονται σε αυτά που ανήκουν στην ω_1 (δηλαδή σε αυτά που αντιστοιχούν σε θετικές αξιολογήσεις) και σε αυτά που ανήκουν στην ω_2 (δηλαδή σε αυτά που αντιστοιχούν σε αρνητικές αξιολογήσεις). Το ίδιο συμβαίνει και για τα δεδομένα που αντιστοιχούν σε άρνηση της ερώτησης Q, δίνοντας έτσι δύο δυάδες της μορφής $(Y_1\ \omega_1, Y_2\ \omega_2)$ για όσα αντιστοιχούν σε απάντηση Y και $(N_1\ \omega_1, N_2\ \omega_2)$ για όσα αντιστοιχούν σε απάντηση N. Με βάση αυτά, υπολογίζεται η εντροπία κάθε κόμβου, σύμφωνα με τις σχέσεις

$$i_\epsilon(\mathrm{Y}) = -\frac{Y_1}{Y_1+Y_2}\log_2\left(\frac{Y_1}{Y_1+Y_2}\right) - \frac{Y_2}{Y_1+Y_2}\log_2\left(\frac{Y_2}{Y_1+Y_2}\right) \tag{3.1}$$

και

$$i_\epsilon\,(\mathrm{N}) = -\frac{N_1}{N_1+N_2}\log_2\left(\frac{N_1}{N_1+N_2}\right) - \frac{N_2}{N_1+N_2}\log_2\left(\frac{N_2}{N_1+N_2}\right). \tag{3.2}$$

Βάσει αυτών, η εντροπία της ερώτησης Q, $i_\epsilon(\mathrm{Q})$, υπολογίζεται σύμφωνα με τη σχέση

$$i_\epsilon\,(\mathrm{Q}) = \frac{Y_1+Y_2}{Y_1+Y_2+N_1+N_2} i_\epsilon\,(\mathrm{Y}) + \frac{N_1+N_2}{Y_1+Y_2+N_1+N_2} i_\epsilon\,(\mathrm{N})\,. \tag{3.3}$$

Πραγματοποιώντας τη διαδικασία αυτή για κάθε ερώτηση Q1-Q9, προκύπτουν τα αποτελέσματα για την εντροπία κάθε ερώτησης που παρατίθενται στον Πίνακα 3.1.

Q	Y	i_ϵ(Y)	N	i_ϵ(N)	i_ϵ(Q)
Q1	$(4\,\omega_1, 1\,\omega_2)$	0.722	$(2\,\omega_1, 2\,\omega_2)$	1.0	0.846
Q2	$(2\,\omega_1, 0\,\omega_2)$	0.0	$(4\,\omega_1, 3\,\omega_2)$	0.985	0.766
Q3	$(3\,\omega_1, 1\,\omega_2)$	0.811	$(3\,\omega_1, 2\,\omega_2)$	0.971	0.9
Q4	$(4\,\omega_1, 1\,\omega_2)$	0.722	$(2\,\omega_1, 2\,\omega_2)$	1.0	0.846
Q5	$(5\,\omega_1, 2\,\omega_2)$	0.863	$(1\,\omega_1, 1\,\omega_2)$	1.0	0.894
Q6	$(1\,\omega_1, 1\,\omega_2)$	1.0	$(5\,\omega_1, 2\,\omega_2)$	0.863	0.894
Q7	$(3\,\omega_1, 1\,\omega_2)$	0.811	$(3\,\omega_1, 2\,\omega_2)$	0.971	0.9
Q8	$(4\,\omega_1, 2\,\omega_2)$	0.918	$(2\,\omega_1, 1\,\omega_2)$	0.918	0.918
Q9	$(5\,\omega_1, 3\,\omega_2)$	0.954	$(1\,\omega_1, 0\,\omega_2)$	0.0	0.848

Πίνακας 3.1: Υπολογισμοί εντροπιών για την πρώτη ερώτηση του CART.

Είναι εμφανές πως η ερώτηση με την ελάχιστη εντροπία είναι η Q2, γι' αυτό εξάλλου έχει σημειωθεί στον Πίνακα 3.1 με μπλε χρώμα, γεγονός που υποδεικνύει πως αυτή πρέπει να είναι η πρώτη ερώτηση του δένδρου. Δεδομένης της Q2, ο κόμβος Y οδηγεί σε κατάσταση μηδενικού impurity, αφού τα δεδομένα ταξινομούνται αποκλειστικά στη μία κατηγορία, ενώ ο κόμβος N έχει μη μηδενικό impurity, γεγονός που σημαίνει πως το δένδρο πρέπει να συνεχιστεί με μια δεύτερη ερώτηση. Έτσι, για τα υπόλοιπα δεδομένα, η διαδικασία επιλογής ερώτησης βάσει ελάχιστης εντροπίας επαναλαμβάνεται, αφαιρώντας όμως πρώτα την ερώτηση Q2. Τα αντίστοιχα αποτελέσματα φαίνονται στον Πίνακα 3.2.

Q	Y	i_ϵ(Y)	N	i_ϵ(N)	i_ϵ(Q)
Q1	$(2\,\omega_1, 1\,\omega_2)$	0.918	$(2\,\omega_1, 2\,\omega_2)$	1.0	0.965
Q3	$(1\,\omega_1, 1\,\omega_2)$	1.0	$(3\,\omega_1, 2\,\omega_2)$	0.971	0.979
Q4	$(2\,\omega_1, 1\,\omega_2)$	0.918	$(2\,\omega_1, 2\,\omega_2)$	1.0	0.965
Q5	$(3\,\omega_1, 2\,\omega_2)$	0.971	$(1\,\omega_1, 1\,\omega_2)$	1.0	0.979
Q6	$(1\,\omega_1, 1\,\omega_2)$	1.0	$(3\,\omega_1, 2\,\omega_2)$	0.971	0.979
Q7	$(3\,\omega_1, 1\,\omega_2)$	0.811	$(1\,\omega_1, 2\,\omega_2)$	0.918	0.857
Q8	$(3\,\omega_1, 2\,\omega_2)$	0.971	$(1\,\omega_1, 1\,\omega_2)$	1.0	0.979
Q9	$(3\,\omega_1, 3\,\omega_2)$	1.0	$(1\,\omega_1, 0\,\omega_2)$	0.0	0.857

Πίνακας 3.2: Υπολογισμοί εντροπιών για τη δεύτερη ερώτηση του CART.

Σε αντίθεση με την πρώτη ερώτηση, για τη δεύτερη ερώτηση παρατηρείται ισοβαθμία ως προς την εντροπία, ανάμεσα στις ερωτήσεις Q7 και Q9 (οι δύο ερωτήσεις σημειωμένες με μπλε χρώμα

στον Πίνακα 3.2). Οι ερωτήσεις αυτές δεν είναι ισοδύναμες, αφού η μία οδηγεί σε κατάσταση μηδενικού impurity, ενώ η άλλη σε δύο impure καταστάσεις, χωρίς όμως αυτό να σημαίνει πως η μία υπερτερεί της άλλης, αφού ακόμη δεν είναι γνωστό το τελικό μέγεθος του δένδρου (γεγονός το οποίο παίζει σημαντικό ρόλο). Η πιο ασφαλής προσέγγιση είναι να ληφθούν υπ' όψιν δύο περιπτώσεις: μία όπου η δεύτερη ερώτηση είναι η Q9 (περίπτωση I) και μία όπου η δεύτερη ερώτηση είναι η Q7 (περίπτωση II). Σε ό,τι αφορά την περίπτωση I, ο κόμβος N της Q9 οδηγεί σε pure κατάσταση, ενώ για τον κόμβο Y η διαδικασία επαναλαμβάνεται, έχοντας αφαιρέσει και την ερώτηση Q9 από το σύνολο των ερωτήσεων. Τα σχετικά αποτελέσματα παρατίθενται στον Πίνακα 3.3.

Q	Y	i_ϵ(Y)	N	i_ϵ(N)	i_ϵ(Q)
Q1	$(1\ \omega_1, 1\ \omega_2)$	1.0	$(2\ \omega_1, 2\ \omega_2)$	1.0	1.0
Q3	$(1\ \omega_1, 1\ \omega_2)$	1.0	$(2\ \omega_1, 2\ \omega_2)$	1.0	1.0
Q4	$(2\ \omega_1, 1\ \omega_2)$	0.918	$(1\ \omega_1, 2\ \omega_2)$	0.918	0.918
Q5	$(3\ \omega_1, 2\ \omega_2)$	0.971	$(0\ \omega_1, 1\ \omega_2)$	0.0	0.809
Q6	$(1\ \omega_1, 1\ \omega_2)$	1.0	$(2\ \omega_1, 2\ \omega_2)$	1.0	1.0
Q7	$(3\ \omega_1, 1\ \omega_2)$	0.811	$(0\ \omega_1, 2\ \omega_2)$	0.0	0.541
Q8	$(3\ \omega_1, 2\ \omega_2)$	0.971	$(0\ \omega_1, 1\ \omega_2)$	0.0	0.809

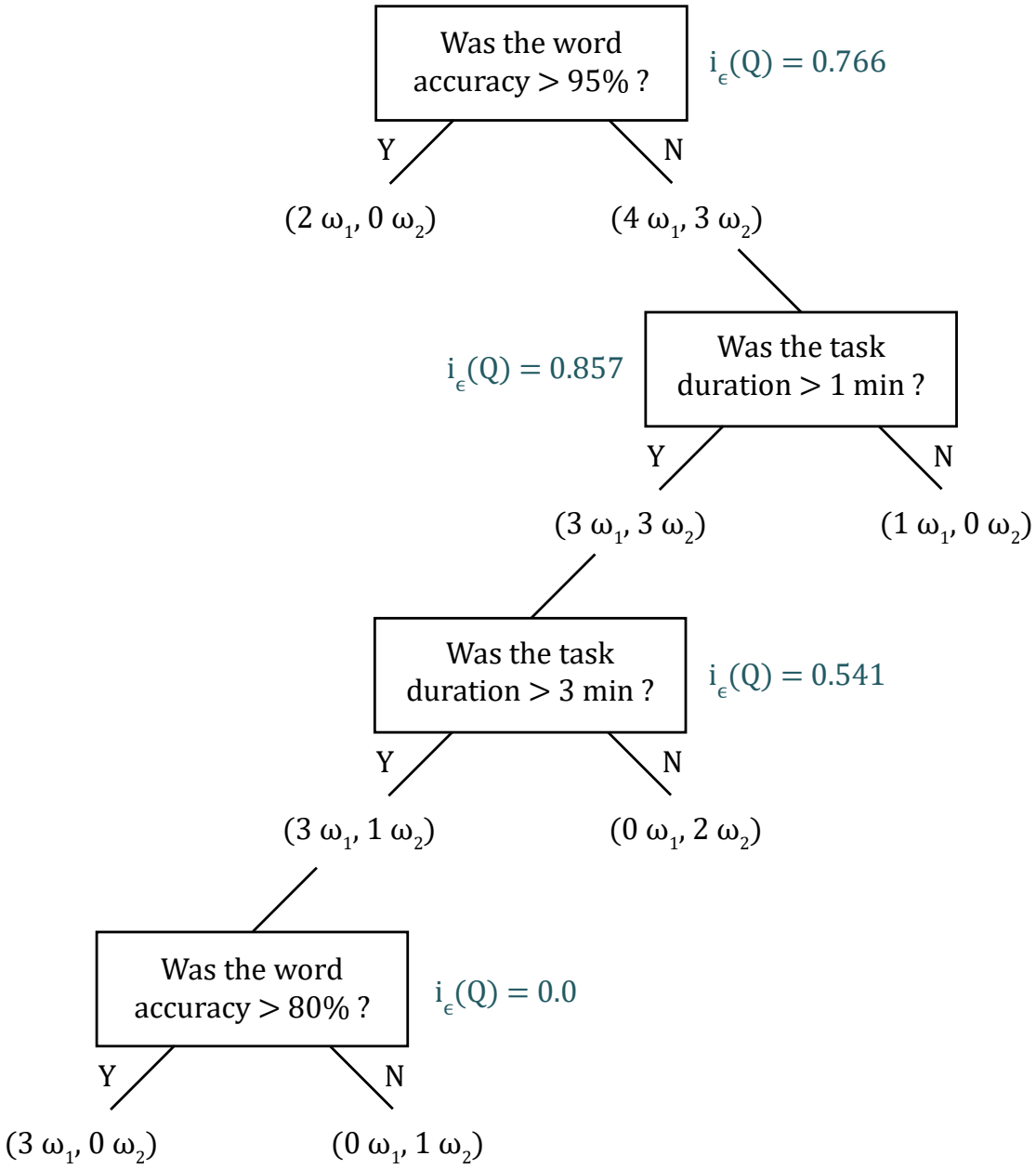
Πίνακας 3.3: Υπολογισμοί εντροπιών για την τρίτη ερώτηση του CART (περίπτωση I).

Ως προς την τρίτη ερώτηση στην περίπτωση I, είναι ξεκάθαρο πως αυτή είναι η Q7, για την οποία ο κόμβος N αντιστοιχεί σε pure κατάσταση, ενώ ο κόμβος Y όχι, πράγμα το οποίο υποδεικνύει πως θα χρειαστεί τουλάχιστον μία ακόμη ερώτηση για την ολοκλήρωση του δένδρου. Αυτό σημαίνει πως η διαδικασία πρέπει να επαναληφθεί για ακόμη μία (τουλάχιστον) φορά, έχοντας πρώτα αφαιρέσει την Q7 και την Q8 από το σύνολο των ερωτήσεων. Ο λόγος για τον οποίο αφαιρείται και η Q8 είναι διότι στον κόμβο Y της Q7 αντιστοιχούν καταχωρήσεις με `TASK_DURATION` > 3 min, οι οποίες προφανώς ικανοποιούν και τη σχέση `TASK_DURATION` > 2 min αυτόματα. Έτσι, προκύπτουν τα αποτελέσματα που παρατίθενται στον Πίνακα 3.4.

Q	Y	i_ϵ(Y)	N	i_ϵ(N)	i_ϵ(Q)
Q1	$(1\ \omega_1, 1\ \omega_2)$	1.0	$(2\ \omega_1, 0\ \omega_2)$	0.0	0.5
Q3	$(1\ \omega_1, 0\ \omega_2)$	0.0	$(2\ \omega_1, 1\ \omega_2)$	0.918	0.689
Q4	$(2\ \omega_1, 0\ \omega_2)$	0.0	$(1\ \omega_1, 1\ \omega_2)$	1.0	0.5
Q5	$(3\ \omega_1, 0\ \omega_2)$	0.0	$(0\ \omega_1, 1\ \omega_2)$	0.0	0.0
Q6	$(1\ \omega_1, 1\ \omega_2)$	1.0	$(2\ \omega_1, 0\ \omega_2)$	0.0	0.5

Πίνακας 3.4: Υπολογισμοί εντροπιών για την τέταρτη ερώτηση του CART (περίπτωση I).

Βάσει των αποτελεσμάτων αυτών προκύπτει πως η τέταρτη ερώτηση είναι η Q5 με μηδενική εντροπία και ως εκ τούτου οδηγεί σε πλήρη κατηγοριοποίηση των εναπομείναντων δεδομένων, ή, με άλλα λόγια, σε δύο καταστάσεις μηδενικού impurity. Αυτό σημαίνει πως με την ερώτηση Q5 το δένδρο για την περίπτωση I ολοκληρώνεται, περιλαμβάνοντας συνολικά 4 ερωτήσεις. Η τελική μορφή του δένδρου απεικονίζεται στην Εικόνα 3.1. Σημειώνεται πως η εντροπία που παρατίθεται με μπλε χρώμα δίπλα από κάθε ερώτηση δεν αντιστοιχεί στην εντροπία της αντίστοιχης κατάστασης, αλλά στην εντροπία που υπολογίστηκε μέσω των Σχέσεων (3.1)-(3.3) για τη συγκεκριμένη ερώτηση (η ελάχιστη εντροπία σε κάθε κύκλο υπολογισμών).



Εικόνα 3.1: Δένδρο απόφασης για την περίπτωση I.

Αξίζει επίσης να αναφερθεί πως η υλοποίηση `DecisionTreeClassifier` της `sklearn` με παράμετρο `criterion='entropy'` καταλήγει επίσης στο συγκεκριμένο δένδρο (και όχι σε αυτό της περίπτωσης II που θα αναλυθεί παρακάτω). Η μόνη διαφορά είναι πως οι ερωτήσεις, Q′, που λαμβάνει υπ' όψιν είναι οι

$$Q1' : \text{Was the task completed ?}$$

$Q2' : \text{Was the word accuracy} \leq 97.5\% \text{ ?}$ $Q3' : \text{Was the word accuracy} \leq 92.5\% \text{ ?}$

$Q4' : \text{Was the word accuracy} \leq 87.5\% \text{ ?}$ $Q5' : \text{Was the word accuracy} \leq 82.5\% \text{ ?}$

$Q6' : \text{Was the task duration} \leq 4.5 \text{ min ?}$ $Q7' : \text{Was the task duration} \leq 3.5 \text{ min ?}$

$Q8' : \text{Was the task duration} \leq 2.5 \text{ min ?}$ $Q9' : \text{Was the task duration} \leq 1.5 \text{ min ?}$

οι οποίες είναι απολύτως ισοδύναμες με τις Q για την κατασκευή του δένδρου, όμως ίσως οδη-

γούν σε ένα πιο εύκολα γενικεύσιμο δέντρο για ταξινόμηση στην περίπτωση εμφάνισης νέων δεδομένων, αφού αφορούν πάντα το μέσο του διαστήματος μεταξύ δύο τύπων παρατηρήσεων.

Προχωρώντας στην περίπτωση II, θεωρώντας πως η δεύτερη ερώτηση είναι η Q7, τόσο ο κόμβος Y, όσο και ο κόμβος N αντιστοιχούν σε impure καταστάσεις. Έτσι, η διαδικασία εύρεσης της κατάλληλης ερώτησης πρέπει να γίνει ξεχωριστά για κάθε κόμβο. Σε ό,τι αφορά τον κόμβο Y, οι υπολογισμοί επαναλαμβάνονται αφαιρώντας τις ερωτήσεις Q8 και Q9, πέραν της Q7, αφού η υπόθεση Q7 είναι ισχυρότερη των Q8 και Q9 και τις περιλαμβάνει. Τα σχετικά αποτελέσματα φαίνονται στον Πίνακα 3.5.

Q	Y	i_ϵ(Y)	N	i_ϵ(N)	i_ϵ(Q)
Q1	(1 ω_1, 1 ω_2)	1.0	(2 ω_1, 0 ω_2)	0.0	0.5
Q3	(1 ω_1, 0 ω_2)	0.0	(2 ω_1, 1 ω_2)	0.918	0.689
Q4	(2 ω_1, 0 ω_2)	0.0	(1 ω_1, 1 ω_2)	1.0	0.5
Q5	(3 ω_1, 0 ω_2)	0.0	(0 ω_1, 1 ω_2)	0.0	0.0
Q6	(1 ω_1, 1 ω_2)	1.0	(2 ω_1, 0 ω_2)	0.0	0.5

Πίνακας 3.5: Υπολογισμοί εντροπιών για την ερώτηση μετά τον κόμβο Y.

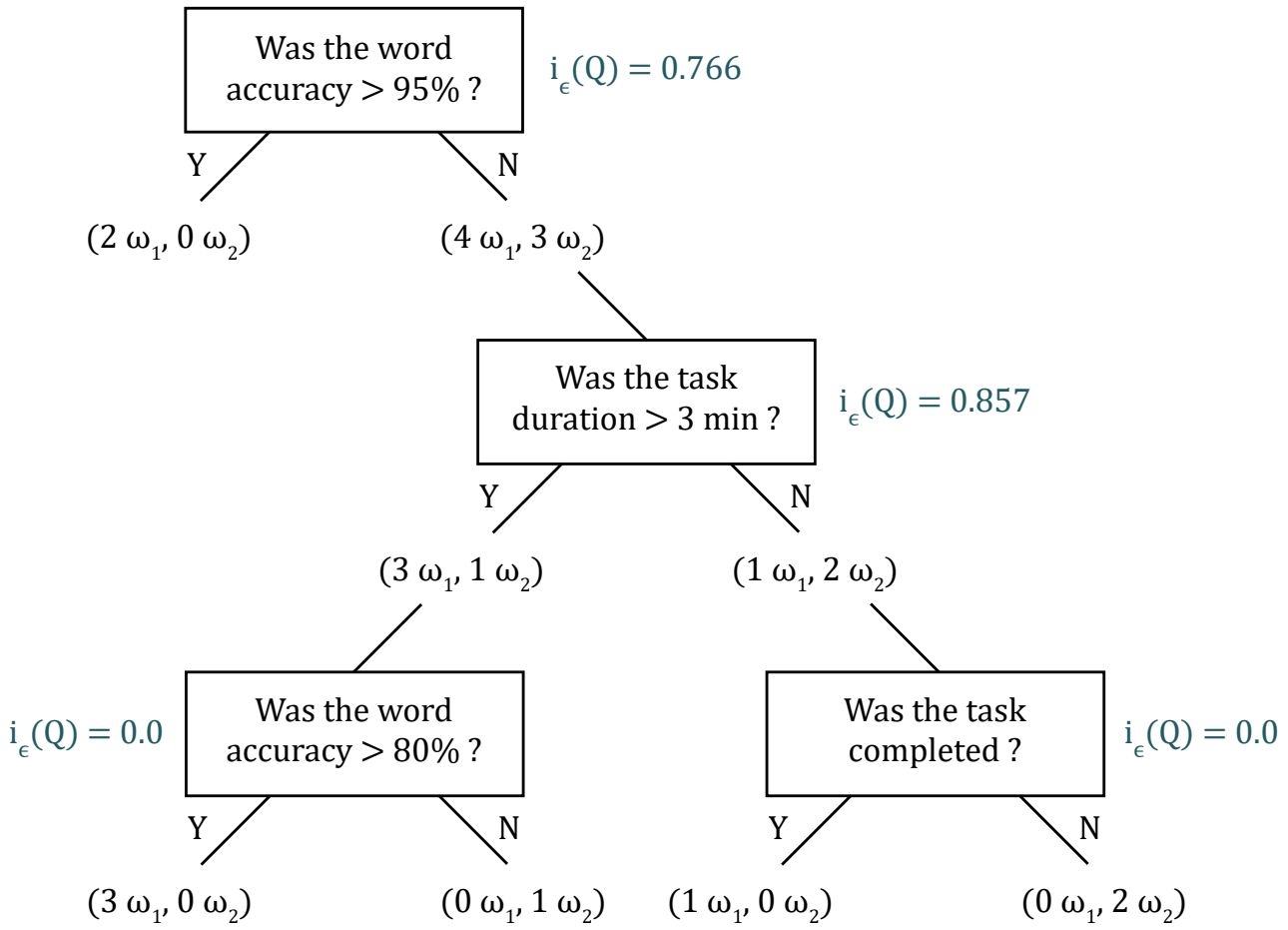
Εύκολα παρατηρεί κανείς πως ο Πίνακας 3.5 είναι ταυτόσημος με τον Πίνακα 3.4, γεγονός αναμενόμενο, αφού προκύπτουν από την ίδια ακολουθία ερωτήσεων. Σε ό,τι αφορά τον κόμβο N μετά από τη δεύτερη ερώτηση της περίπτωσης II, οι σχετικοί υπολογισμοί παρατίθενται στον Πίνακα 3.6, οι οποίοι πραγματοποιήθηκαν αφαιρώντας τις ερωτήσεις Q6 και Q7.

Q	Y	i_ϵ(Y)	N	i_ϵ(N)	i_ϵ(Q)
Q1	(1 ω_1, 0 ω_2)	0.0	(0 ω_1, 2 ω_2)	0.0	0.0
Q3	(0 ω_1, 1 ω_2)	0.0	(1 ω_1, 1 ω_2)	1.0	0.667
Q4	(0 ω_1, 1 ω_2)	0.0	(1 ω_1, 1 ω_2)	1.0	0.667
Q5	(0 ω_1, 2 ω_2)	0.0	(1 ω_1, 0 ω_2)	0.0	0.0
Q8	(0 ω_1, 1 ω_2)	0.0	(1 ω_1, 1 ω_2)	1.0	0.667
Q9	(0 ω_1, 2 ω_2)	0.0	(1 ω_1, 0 ω_2)	0.0	0.0

Πίνακας 3.6: Υπολογισμοί εντροπιών για την ερώτηση μετά τον κόμβο N.

Ενδιαφέρον έχει το γεγονός πως υπάρχουν 3 διαφορετικές ερωτήσεις μηδενικής εντροπίας, που οδηγούν δηλαδή αποκλειστικά σε pure καταστάσεις, ολοκληρώνοντας έτσι την κατασκευή του δένδρου: η πρώτη είναι η Q5, δηλαδή η ερώτηση ελάχιστης εντροπίας που υπολογίστηκε και για τον κόμβο Y. Η δεύτερη είναι η Q9, η οποία ήταν η δεύτερη ερώτηση του δένδρου της περίπτωσης I, καθιστώντας έτσι την περίπτωση II ισοδύναμη με την I ως προς τις ερωτήσεις που τίθενται, με μόνη διαφορά τη σειρά τους. Η τρίτη ερώτηση που μπορεί να ολοκληρώσει το δένδρο της περίπτωσης II είναι η Q1, η οποία δεν έχει εμφανιστεί πουθενά μέχρι το σημείο αυτό. Για λόγους ποικιλίας, επιλέγεται η απεικόνιση του δένδρου της περίπτωσης II να γίνει χρησιμοποιώντας την Q1 ως τελική ερώτηση, με το αντίστοιχο σχήμα φαίνεται στην Εικόνα 3.2.

3.3 Έχοντας προσδιορίσει πλήρως τα δύο πιθανά δένδρα απόφασης για το συγκεκριμένο πρόβλημα, δεν υπάρχει λόγος το ένα να χαρακτηριστεί ως καλύτερο του άλλου. Και τα δύο απαιτούν συνολικά 4 ερωτήσεις για να ταξινομήσουν πλήρως τα «δεδομένα εκπαίδευσης», με το ένα να το επιτυγχάνει σε 2 διακλαδώσεις, στη μία εκ των οποίων και οι δύο κόμβοι είναι impure, και



Εικόνα 3.2: Δένδρο απόφασης για την περίπτωση II.

το άλλο να το επιτυγχάνει σε 3 διακλαδώσεις, με όλες τους όμως να έχουν πάντοτε έναν τουλάχιστον pure κόμβο. Παρ' όλα αυτά, τα δύο δένδρα επιτρέπουν την εξαγωγή ορισμένων συμπερασμάτων με απόλυτη ασφάλεια. Αρχικά, η πρώτη ερώτηση σε κάθε περίπτωση αφορά την παράμετρο `WORD_ACCURACY`, αφού φαίνεται πως όταν η ακρίβεια είναι άνω του 95%, τότε οι πελάτες μένουν πάντα ικανοποιημένοι. Επιπλέον, η δεύτερη ερώτηση σε κάθε περίπτωση αφορά την παράμετρο `TASK_DURATION`, με διαφοροποιήσεις ως προς τη διάρκεια που εξετάζεται στην ερώτηση. Και στις δύο περιπτώσεις, βέβαια, οι ερωτήσεις που υπεισέρχονται αφορούν ελέγχους για διάρκεια άνω του ενός λεπτού και για διάρκεια άνω των τριών λεπτών. Η παράμετρος `TASK_COMPLETION` υπεισέρχεται μόνο ως τρίτη ερώτηση και μόνο στην περίπτωση II και ακόμα κι εκεί ισοβαθμεί με άλλες δύο ερωτήσεις. Βάσει των παρατηρήσεων αυτών, συμπεραίνει κανείς πως η σημασία των τριών παραμέτρων για την ταξινόμηση είναι

`WORD_ACCURACY` > `TASK_DURATION` > `TASK_COMPLETION`

Έτσι, το βέβαιο συμπέρασμα είναι πως η έμφαση για τη βελτιστοποίηση του συστήματος θα πρέπει να δοθεί στο κομμάτι της αναγνώρισης φωνής και συγκεκριμένα στο ποσοστό των λέξεων που το σύστημα αναγνωρίζει με ακρίβεια.

4 MLP Backpropagation

Βάσει του δοσμένου γράφου, η μεταβλητή εξόδου, $\hat{y}$, δίνεται από τη σχέση

$$\hat{y} = \sigma\left[w_5\sigma\left(w_1x_1 + w_2x_2\right) + w_6\sigma\left(w_3x_3 + w_4x_4\right)\right], \tag{4.1}$$

όπου $\sigma(x)$ είναι η σιγμοειδής συνάρτηση, για την οποία ισχύει

$$\frac{\partial\sigma\left(x\right)}{\partial x} = \sigma\left(x\right)\left[1 - \sigma\left(x\right)\right]. \tag{4.2}$$

Από τη σχέση αυτή προκύπτει $\hat{y} = 0.5867$ και ως εκ τούτου για τη μερική παράγωγο της συνάρτησης L2 Loss ως προς το βάρος w_1 θα ισχύει

$$\frac{\partial L}{\partial w_1} = 2\|y - \hat{y}\|_2\hat{y}\left(1 - \hat{y}\right)w_5\sigma\left(w_1x_1 + w_2x_2\right)\left[1 - \sigma\left(w_1x_1 + w_2x_2\right)\right]x_1 = 0.0008. \tag{4.3}$$

Για να εξαγάγει κανείς το αποτέλεσμα αυτό με έναν τρόπο που να θυμίζει λίγο περισσότερο τη λογική υπολογιστικών βιβλιοθηκών όπως η PyTorch, δηλαδή το σταδιακό υπολογισμό των διάφορων παραμέτρων και των αντίστοιχων παραγώγων τους, θα έγραφε αρχικά

$$s_1 = w_1x_1 + w_2x_2 = 0.99, \quad \text{με} \quad \frac{\partial s_1}{\partial w_1} = x_1 = -0.5 \tag{4.4}$$

και

$$h_1 = \sigma\left(s_1\right) = 0.7291 \quad \text{με} \quad \frac{\partial h_1}{\partial s_1} = h_1\left(1 - h_1\right) = 0.1975. \tag{4.5}$$

Πρόσθετα,

$$s_2 = w_3x_3 + w_4x_4 = 4.86, \quad \text{με} \quad \frac{\partial s_2}{\partial w_1} = 0 \tag{4.6}$$

και

$$h_2 = \sigma\left(s_2\right) = 0.9923. \tag{4.7}$$

Από αυτά, προκύπτει

$$s_3 = w_5h_1 + w_6h_2 = 0.3503, \quad \text{με} \quad \frac{\partial s_3}{\partial h_1} = w_5 = -0.2 \tag{4.8}$$

και επομένως

$$\hat{y} = \sigma\left(s_3\right) = 0.5867, \quad \text{με} \quad \frac{\partial\hat{y}}{\partial s_3} = \hat{y}\left(1 - \hat{y}\right) = 0.2276. \tag{4.9}$$

Τελικά, παίρνει κανείς

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial\hat{y}}\frac{\partial\hat{y}}{\partial w_1} = 2\|y - \hat{y}\|_2\frac{\partial\hat{y}}{\partial s_3}\frac{\partial s_3}{\partial w_1} = 2\|y - \hat{y}\|_2\frac{\partial\hat{y}}{\partial s_3}\frac{\partial s_3}{\partial h_1}\frac{\partial h_1}{\partial w_1} = 2\|y - \hat{y}\|_2\frac{\partial\hat{y}}{\partial s_3}\frac{\partial s_3}{\partial h_1}\frac{\partial h_1}{\partial s_1}\frac{\partial s_1}{\partial w_1}$$

$$= 2(0.5867 - 0.5)\cdot 0.2276\cdot(-0.2)\cdot 0.1975\cdot(-0.5) = 0.0008. \tag{4.10}$$

5 Karhunen-Loève Transform - Principal Component Analysis

5.1 Το γεγονός πως το πρόβλημα έχει λυθεί για $p = 1$ σημαίνει πως έχει αποδειχτεί ότι ο μετασχηματισμός **A** ταυτίζεται με το ιδιοδιάνυσμα $\mathbf{e}_1$ του **R**, ενώ το σφάλμα της προσέγγισης, J, δίνεται από τη σχέση

$$J = \sum_{i=2}^{d} \lambda_i, \tag{5.1}$$

όπου λ_i οι $d-1$ ιδιοτιμές του **R**, οι οποίες έχουν διαταχθεί έτσι, ώστε $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Με άλλα λόγια, το σφάλμα για $p = 1$ δεν είναι παρά το άθροισμα όλων των ιδιοτιμών του **R** πλην της μεγαλύτερης ιδιοτιμής του. Η υπόθεση που γίνεται προκειμένου να προκύψει μια απόδειξη μέσω επαγωγής είναι πως τα συμπεράσματα ισχύουν και για τυχαίο p: επιλέγοντας τα p ιδιοδιανύσματα του **R** ως στήλες του μετασχηματισμού **A** έτσι, ώστε να αντιστοιχούν στις p μεγαλύτερες ιδιοτιμές $\lambda_1, \dots, \lambda_p$, τότε το σφάλμα της προσέγγισης δίνεται από τη σχέση

$$J = \sum_{i=p+1}^{d} \lambda_i. \tag{5.2}$$

Με βάση αυτά, αρκεί να αποδειχθεί πως τα παραπάνω συμπεράσματα ισχύουν και για $p+1$. Αρχικά, θα αποδειχθεί πως ο **R** είναι ερμιτιανός[5], γεγονός το οποίο θα φανεί χρήσιμο και στο ερώτημα **5.2** της άσκησης. Ισχύει:

$$\mathbf{R} = \mathbb{E}\left(\mathbf{x}\mathbf{x}^\dagger\right) \Leftrightarrow \mathbf{R}^\dagger = \left[\mathbb{E}\left(\mathbf{x}\mathbf{x}^\dagger\right)\right]^\dagger = \mathbb{E}\left[\left(\mathbf{x}\mathbf{x}^\dagger\right)^\dagger\right] = \mathbb{E}\left(\mathbf{x}\mathbf{x}^\dagger\right) = \mathbf{R} \Leftrightarrow \mathbf{R}^\dagger = \mathbf{R}. \tag{5.3}$$

Επιστρέφοντας στην απόδειξη, θεωρείται ένα επιπλέον διάνυσμα πέραν των $\mathbf{e}_1, \dots, \mathbf{e}_p$, το **u**, στο οποίο προβάλλονται τα δεδομένα, συνεπώς ισχύει

$$\hat{\mathbf{x}} = \sum_{k=1}^{p} y_k \mathbf{e}_k + y_{p+1}\mathbf{u}. \tag{5.4}$$

Το σφάλμα που διορθώνει η επιπλέον αυτή προβολή είναι βέβαια $\mathbf{u}^\dagger\mathbf{R}\mathbf{u}$, ενώ το **u** θα πρέπει επιπλέον να είναι μοναδιαίο και κάθετο σε όλα τα $\mathbf{e}_1, \dots, \mathbf{e}_p$, χωρίς να γίνεται οποιαδήποτε επιπλέον υπόθεση που να συνδέει το **u** με τον πίνακα συνδιακύμανσης, **R** [εξ ου και η γενική μορφή της (5.4)]. Η Λαγκρανζιανή που πρέπει να στασιμοποιηθεί ώστε να εξασφαλιστεί η μέγιστη διόρθωση και παράλληλα η τήρηση των δύο συνθηκών για το **u** είναι η

$$\mathcal{L} = \mathbf{u}^\dagger\mathbf{R}\mathbf{u} + \alpha\left(1 - \mathbf{u}^\dagger\mathbf{u}\right) + \sum_{i=1}^{p} \beta_i \mathbf{e}_i^\dagger\mathbf{u}, \tag{5.5}$$

όπου πρέπει $\alpha, \beta_i \in \mathbb{R}$, προκειμένου $\mathcal{L} = \mathcal{L}^\dagger$ (ή, ισοδύναμα, προκειμένου η Λαγκρανζιανή να στασιμοποιείται και ως προς $\mathbf{u}^\dagger$), δεδομένης της ερμιτιανότητας του **R**. Οι εξισώσεις Euler-Lagrange $\partial\mathcal{L}/\partial\alpha = 0$ και $\partial\mathcal{L}/\partial\beta_q = 0$ δίνουν απλώς τις συνθήκες κανονικοποίησης και ορθογωνιότητας που επιβλήθηκαν. Από την άλλη, η εξίσωση Euler-Lagrange $\partial\mathcal{L}/\partial\mathbf{u} = 0$ οδηγεί σε

[5] Η ερμιτιανή συζυγία συμβολίζεται εδώ με † αντί για ^{H}, σε συμφωνία με το φορμαλισμό της Κβαντικής Φυσικής.

$$0 = \mathbf{u}^\dagger \mathbf{R} - \alpha \mathbf{u}^\dagger + \sum_{i=1}^{p} \beta_i \mathbf{e}_i^\dagger = \mathbf{u}^\dagger \mathbf{R} \mathbf{e}_q - \alpha \mathbf{u}^\dagger \mathbf{e}_q + \sum_{i=1}^{p} \beta_i \mathbf{e}_i^\dagger \mathbf{e}_q$$

$$= \lambda_q \mathbf{u}^\dagger \mathbf{e}_q - 0 + \sum_{i=1}^{p} \beta_i \delta_{iq} = \beta_q. \tag{5.6}$$

όπου στην πρώτη ισότητα πολλαπλασιάστηκε παντού από δεξιά το $\mathbf{e}_q$, με $q \in \{1, \dots, p\}$, ενώ στις επόμενες αξιοποιήθηκε η ορθογωνιότητα του $\mathbf{u}$ με τα $\mathbf{e}$. Το συμπέρασμα που προκύπτει είναι πως $\beta_q = 0$, για κάθε $q \in \{1, \dots, p\}$, γεγονός που υποδεικνύει πως η εξίσωση Euler-Lagrange $\partial \mathcal{L} / \partial \mathbf{u} = 0$ ισοδυναμεί στην πραγματικότητα με μια εξίσωση ιδιοτιμών για τον $\mathbf{R}$:

$$0 = \mathbf{u}^\dagger \mathbf{R} - \alpha \mathbf{u}^\dagger \Leftrightarrow \mathbf{R}\mathbf{u} = \alpha \mathbf{u}. \tag{5.7}$$

Αυτό με τη σειρά του σημαίνει πως και το $\mathbf{u}$ πρέπει να είναι ιδιοδιάνυσμα του $\mathbf{R}$, με ιδιοτιμή α, η οποία μπορεί να είναι μια από τις εναπομείνασες $\lambda_{p+1}, \dots, \lambda_d$ (αλλιώς το $\mathbf{u}$ θα αντιστοιχούσε σε κάποιο από τα $\mathbf{e}_i$ με $i = 1, \dots, p$ και άρα δε θα ήταν ορθογώνιο με αυτό όπως απαιτήθηκε). Το σφάλμα της προσέγγισης θα δίνεται τώρα από τη Σχέση (5.2), λαμβάνοντας όμως υπ' όψιν και τη διόρθωση που προκύπτει λόγω της επιπλέον αύξησης της διαστατικότητας:

$$J = \sum_{i=p+1}^{d} \lambda_i - \mathbf{u}^\dagger \mathbf{R} \mathbf{u} = \sum_{i=p+1}^{d} \lambda_i - \alpha. \tag{5.8}$$

Η Σχέση (5.8) υποδεικνύει πως το συνολικό σφάλμα ελαχιστοποιείται όταν το α είναι η μέγιστη από τις $\lambda_{p+1}, \dots, \lambda_d$, δηλαδή $\alpha = \lambda_{p+1}$, δεδομένου του πώς έχουν διαταχθεί εξ υποθέσεως οι ιδιοτιμές. Ως εκ τούτου, θα πρέπει $\mathbf{u} = \mathbf{e}_{p+1}$, οπότε η Σχέση (5.4) γράφεται ως

$$\hat{\mathbf{x}} = \sum_{k=1}^{p+1} y_k \mathbf{e}_k, \tag{5.9}$$

ενώ το σφάλμα της προσέγγισης δίνεται τώρα από τη σχέση

$$J = \sum_{i=p+2}^{d} \lambda_i, \tag{5.10}$$

αποδεικνύοντας έτσι πως η υπόθεση ισχύει και για $p + 1$ και ολοκληρώνοντας την απόδειξη μέσω επαγωγής.

5.2 Δεδομένου του συναρτησιακού

$$\tilde{J} = \mathrm{Tr}\left(\mathbf{U}^\dagger \mathbf{R} \mathbf{U}\right) + \mathrm{Tr}\left[\mathbf{H}\left(\mathbb{1} - \mathbf{U}^\dagger \mathbf{U}\right)\right] = \mathrm{Tr}\left(\mathbf{U}^\dagger \mathbf{R} \mathbf{U}\right) - \mathrm{Tr}\left(\mathbf{H}\mathbf{U}^\dagger \mathbf{U}\right) + \mathrm{Tr}\left(\mathbf{H}\right), \tag{5.11}$$

και της ιδιότητας

$$\frac{\partial}{\partial \mathbf{X}} \left[\mathrm{Tr}\left(\mathbf{X}\mathbf{A}\right)\right] = \mathbf{A}^\top, \tag{5.12}$$

η ελαχιστοποίηση του $\tilde{J}$ ως προς $\mathbf{U}$ απαιτεί την επίλυση της $\partial\tilde{J}/\partial\mathbf{U} = 0$, δηλαδή

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mathbf{U}}\left[\mathrm{Tr}\left(\mathbf{U}^{\dagger}\mathbf{R}\mathbf{U}\right)\right] - \frac{\partial}{\partial \mathbf{U}}\left[\mathrm{Tr}\left(\mathbf{H}\mathbf{U}^{\dagger}\mathbf{U}\right)\right] = \frac{\partial}{\partial \mathbf{U}}\left[\mathrm{Tr}\left(\mathbf{U}\mathbf{U}^{\dagger}\mathbf{R}\right)\right] - \frac{\partial}{\partial \mathbf{U}}\left[\mathrm{Tr}\left(\mathbf{U}\mathbf{H}\mathbf{U}^{\dagger}\right)\right] \\ &= \left(\mathbf{U}^{\dagger}\mathbf{R}\right)^{\top} - \left(\mathbf{H}\mathbf{U}^{\dagger}\right)^{\top} \Leftrightarrow \mathbf{U}^{\dagger}\mathbf{R} = \mathbf{H}\mathbf{U}^{\dagger} \end{aligned} \tag{5.13}$$

όπου στη δεύτερη ισότητα χρησιμοποιήθηκε η κυκλικότητα του ίχνους. Δεδομένου πως ο πίνακας $\mathbf{U}$ έχει ως στήλες κάποια διανύσματα $\mathbf{e}_i$, με $i = p+1, \ldots, d$, τα οποία ανήκουν σε ένα σύνολο $\{\mathbf{e}_i\}$ που συνιστά ορθοκανονική βάση, θα ισχύει ότι

$$\mathbf{U}^{\dagger}\mathbf{U} = \mathbb{1}_{(d-p)\times(d-p)}, \tag{5.14}$$

όπου ο $\mathbb{1}_{(d-p)\times(d-p)}$ είναι ο μοναδιαίος πίνακας διάστασης $(d-p)\times(d-p)$. Έτσι, πολλαπλασιάζοντας τη Σχέση (5.13) από δεξιά με $\mathbf{U}$ προκύπτει

$$\mathbf{H} = \mathbf{U}^{\dagger}\mathbf{R}\mathbf{U}. \tag{5.15}$$

Παίρνοντας το ερμιτιανό συζυγές της σχέσης αυτής βρίσκει κανείς πως

$$\mathbf{H}^{\dagger} = \mathbf{U}^{\dagger}\mathbf{R}^{\dagger}\mathbf{U} = \mathbf{U}^{\dagger}\mathbf{R}\mathbf{U} = \mathbf{H}, \tag{5.16}$$

όπου χρησιμοποιήθηκε η ερμιτιανότητα του $\mathbf{R}$, η οποία αποδείχθηκε στο πρώτο σκέλος της άσκησης. Προκύπτει, λοιπόν, πως ο $\mathbf{H}$ είναι εν γένει ερμιτιανός, οπότε η Σχέση (5.13) μπορεί ισοδύναμα να γραφεί ως

$$\mathbf{R}\mathbf{U} = \mathbf{U}\mathbf{H}, \tag{5.17}$$

αποδεικνύοντας έτσι το ζητούμενο. Αξίζει να σημειωθεί πως η υπόθεση πως οι στήλες του $\mathbf{U}$ αποτελούνται από διανύσματα κάποιας ορθοκανονικής βάσης δεν είναι καν απαραίτητη για να αποδειχθεί η Σχέση (5.17), αρκεί να αντικατασταθεί από την υπόθεση πως εάν $\partial\tilde{J}/\partial\mathbf{U} = 0$, τότε και $\partial\tilde{J}/\partial\mathbf{U}^{\dagger} = 0$. Πράγματι, η $\partial\tilde{J}/\partial\mathbf{U}^{\dagger} = 0$ δίνει

$$0 = \frac{\partial}{\partial \mathbf{U}^{\dagger}}\left[\mathrm{Tr}\left(\mathbf{U}^{\dagger}\mathbf{R}\mathbf{U}\right)\right] - \frac{\partial}{\partial \mathbf{U}^{\dagger}}\left[\mathrm{Tr}\left(\mathbf{H}\mathbf{U}^{\dagger}\mathbf{U}\right)\right] = (\mathbf{R}\mathbf{U})^{\top} - (\mathbf{U}\mathbf{H})^{\top} \Leftrightarrow \mathbf{R}\mathbf{U} = \mathbf{U}\mathbf{H}, \tag{5.18}$$

από την οποία μπορεί να προκύψει και η ερμιτιανότητα του $\mathbf{H}$, εάν συνδυαστεί με τη Σχέση (5.13). Σε κάθε περίπτωση, δεδομένης της Σχέσης (5.17), η ελάχιστη τιμή της $\tilde{J}$ θα αντιστοιχεί στην

$$\begin{aligned} \tilde{J}_{\min} &= \mathrm{Tr}\left(\mathbf{U}^{\dagger}\mathbf{U}\mathbf{H}\right) - \mathrm{Tr}\left(\mathbf{H}\mathbf{U}^{\dagger}\mathbf{U}\right) + \mathrm{Tr}\left(\mathbf{H}\right) = \mathrm{Tr}\left(\mathbf{H}\mathbf{U}^{\dagger}\mathbf{U}\right) - \mathrm{Tr}\left(\mathbf{H}\mathbf{U}^{\dagger}\mathbf{U}\right) + \mathrm{Tr}\left(\mathbf{H}\right) \\ &= \mathrm{Tr}\left(\mathbf{H}\right), \end{aligned} \tag{5.19}$$

Προφανώς, η επιλογή του $\mathbf{U}$ είναι αυτή που καθορίζει, μέσω της Σχέσης (5.17), τη μορφή του $\mathbf{H}$. Έτσι, οποιοσδήποτε συνδυασμός $\mathbf{U}$ και $\mathbf{H}$ σέβεται τη Σχέση (5.17), δεδομένου του πίνακα συνδιακύμανσης, $\mathbf{R}$, είναι μια αποδεκτή λύση του προβλήματος. Στην ειδική περίπτωση όπου ο $\mathbf{U}$ έχει ως στήλες $d-p$ από τα ιδιοδιανύσματα του $\mathbf{R}$, τότε, εάν ο $\mathbf{R}$ αναπτυχθεί στη βάση $\mathbf{W}$ των ιδιοδιανυσμάτων του, η Σχέση (5.15) επιβάλλει για τον $\mathbf{H}$:

$$\mathbf{H} = \mathbf{U}^{\dagger}\mathbf{W}\Lambda_{\mathbf{R}}\mathbf{W}^{\dagger}\mathbf{U} = \begin{bmatrix} \mathbf{e}_{p+1}^{\dagger} \\ \vdots \\ \mathbf{e}_{d}^{\dagger} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_d \end{bmatrix} \cdot \mathrm{diag}\left[\lambda_1, \dots, \lambda_d\right] \cdot \begin{bmatrix} \mathbf{e}_1^{\dagger} \\ \vdots \\ \mathbf{e}_{d}^{\dagger} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{e}_{p+1} & \cdots & \mathbf{e}_d \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{0}_{(d-p)\times p} & \mathbb{1}_{(d-p)\times(d-p)} \end{bmatrix} \cdot \mathrm{diag}\left[\lambda_1, \dots, \lambda_d\right] \cdot \begin{bmatrix} \mathbb{0}_{p\times(d-p)} \\ \mathbb{1}_{(d-p)\times(d-p)} \end{bmatrix}$$

$$= \mathrm{diag}\left[\lambda_{p+1}, \dots, \lambda_d\right] \tag{5.20}$$

Πράγματι, προκύπτει πως ο **H** πρέπει να είναι ένας διαγώνιος πίνακας με στοιχεία της διαγωνίου τις $d-p$ ιδιοτιμές που αντιστοιχούν στα $d-p$ ιδιοδιανύσματα του **R** τα οποία περιέχει ο **U**. Στην περίπτωση αυτή, θα ισχύει

$$\tilde{J}_{\min} = \mathrm{Tr}\left(\mathbf{H}\right) = \sum_{i=p+1}^{d} \lambda_i. \tag{5.21}$$

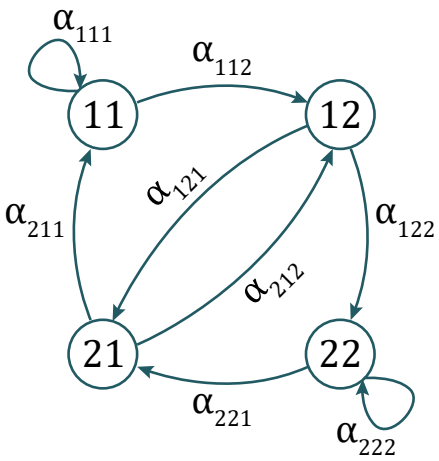
Κατ' επέκταση, εάν οι ιδιοτιμές λ_i του **R** διαταχθούν έτσι, ώστε $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, τότε το αποτέλεσμα της Σχέσης (5.21) ταυτίζεται με αυτό της Σχέσης (5.2). Σημειώνεται πως εδώ ο πίνακας **U** δεν περιλαμβάνει τα p ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές, αλλά όσα έχουν «περισσέψει», αφότου έχει εφαρμοστεί PCA τάξης p. Ισχύει, δηλαδή

$$\mathbf{W} = \begin{bmatrix} \mathbf{A} & \mathbf{U} \end{bmatrix}, \tag{5.22}$$

όπου ο **A** είναι ο μετασχηματισμός των **x** σε **y** της εκφώνησης. Συμπερασματικά, αφού η ειδική περίπτωση στην οποία ο **H** είναι διαγώνιος και περιέχει τις ιδιοτιμές του **R** που αντιστοιχούν στα ιδιοδιανύσματα του **R** που περιέχονται στον **U**, ικανοποιεί τη γενική συνθήκη της Σχέσης (5.17) - η οποία αφορά τυχαίους εν γένει **U** και **H** - είναι μια αποδεκτή λύση.

6 Graphical Models

6.1 Για ένα Μαρκοβιανό Μοντέλο μνήμης 2 (ΜΜ-2) με δύο καταστάσεις, έστω 1 και 2, η κατάσταση q_t δεδομένων των q_{t-1} και q_{t-2} είναι ανεξάρτητη όλων των προηγούμενων. Αυτό σημαίνει πως το Μπεϋζιανό δίκτυο που αντιστοιχεί σε ένα τέτοιο μοντέλο έχει ως καταστάσεις ακολουθίες 2 διαδοχικών καταστάσεων του ΜΜ-2 και οι μεταβάσεις μεταξύ τους πραγματοποιούνται μέσω ενός τανυστή (και όχι πίνακα) μετάβασης με στοιχεία α_{ijk}, όπως φαίνεται στην Εικόνα 6.1.



Εικόνα 6.1: Το Μπεϋζιανό δίκτυο που αντιστοιχεί σε ένα ΜΜ-2 δύο καταστάσεων.

Ο πλήρης προσδιορισμός του μοντέλου ισοδυναμεί με τον προσδιορισμό των α_{111}, α_{121}, α_{212} και α_{222}, αφού τα υπόλοιπα 4 μη ταυτοτικά μηδενικά στοιχεία του τανυστή μετάβασης υπολογίζονται βάσει αυτών, καθώς και τη γνώση των $\pi_i = p\,(q_0 = i)$ και $\pi_{ij} = p\,(q_1 = j | q_0 = i)$.

6.2 Δεδομένης της ακολουθίας

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

τα στοιχεία του τανυστή μετάβασης μπορούν να υπολογιστούν χρησιμοποιώντας τον αλγόριθμο μέγιστης πιθανοφάνειας, ο οποίος στην προκείμενη περίπτωση ισοδυναμεί με απλές καταμετρήσεις της μορφής

$$\alpha_{ijk} = \frac{\#(i, j \to k)}{\#(i, j \to \star)}. \tag{6.1}$$

Έτσι, προκύπτουν

$$\alpha_{111} = \frac{\#(1,1\to 1)}{\#(1,1\to 1)+\#(1,1\to 2)} = \frac{3}{5}, \quad \text{άρα} \quad \alpha_{112} = 1 - \frac{3}{5} = \frac{2}{5}, \tag{6.2}$$

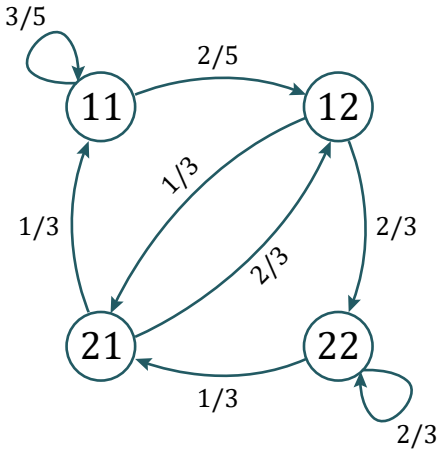
$$\alpha_{121} = \frac{\#(1,2\to 1)}{\#(1,2\to 1)+\#(1,2\to 2)} = \frac{1}{3}, \quad \text{άρα} \quad \alpha_{122} = 1 - \frac{1}{3} = \frac{2}{3}, \tag{6.3}$$

$$\alpha_{212} = \frac{\#(2,1\to 2)}{\#(2,1\to 1)+\#(2,1\to 2)} = \frac{2}{3}, \quad \text{άρα} \quad \alpha_{211} = 1 - \frac{2}{3} = \frac{1}{3} \tag{6.4}$$

και

$$\alpha_{222} = \frac{\#(2,2 \to 2)}{\#(2,2 \to 1) + \#(2,2 \to 2)} = \frac{2}{3}, \quad \text{άρα} \quad \alpha_{221} = 1 - \frac{2}{3} = \frac{1}{3}. \tag{6.5}$$

Το Μπεϋζιανό δίκτυο με τις αντίστοιχες τιμές των υπολογισμένων στοιχείων του πίνακα μετάβασης φαίνεται στην Εικόνα 6.2. Αξίζει να σημειωθεί πως ένα ΜΜ-2 δύο καταστάσεων μπορεί (όπως εξάλλου υποδεικνύει και το δίκτυο της Εικόνας 6.2) να αναχθεί σε ένα ΜΜ-1 τεσσάρων καταστάσεων, επομένως ο **α** μπορεί να λάβει τη συνηθισμένη μορφή πίνακα (με 8 τετριμμένα μηδενικά στοιχεία).

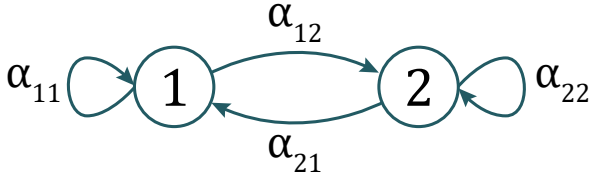


Εικόνα 6.2: Το δίκτυο του ΜΜ-2 με τις τιμές των στοιχείων του πίνακα μετάβασης.

6.3 Δεδομένου ότι $q_0 = 1$, η πιθανότητα, p_1, μιας ακολουθίας με 10 επιπλέον διαδοχικές καταστάσεις $q = 1$ ακολουθούμενες από μια κατάσταση $q = 2$, ισούται με

$$\begin{aligned} p_1 &= p\,(q_1 = 1, q_2 = 1, \dots, q_{10} = 1, q_{11} = 2 | q_0 = 1) \\ &= \frac{p\,(q_0 = 1, q_1 = 1, q_2 = 1, \dots, q_{10} = 1, q_{11} = 2)}{p\,(q_0 = 1)} = \frac{\pi_1 \cdot \pi_{11} \cdot \overbrace{\alpha_{111} \cdot \dots \cdot \alpha_{111}}^{9\ \text{φορές}} \cdot \alpha_{112}}{\pi_1} \\ &= \pi_{11} \cdot {\alpha_{111}}^9 \cdot \alpha_{112} = \underbrace{\pi_{11}}_{=0.25} \cdot \left(\frac{3}{5}\right)^9 \cdot \frac{2}{5} \simeq 1.01 \cdot 10^{-3}. \end{aligned} \tag{6.6}$$

6.4 Στην περίπτωση ενός Μαρκοβιανού Μοντέλου μνήμης 1 (ΜΜ-1), το αντίστοιχο Μπεϋζιανό δίκτυο έχει την αρκετά απλούστερη μορφή που απεικονίζεται στην Εικόνα 6.3.



Εικόνα 6.3: Το Μπεϋζιανό δίκτυο που αντιστοιχεί σε ένα ΜΜ-1 δύο καταστάσεων.

Έχοντας επιστρέψει σε *πίνακα* μετάβασης, τα αντίστοιχα στοιχεία θα δίνονται σύμφωνα με τον κανόνα

$$\alpha_{ij} = \frac{\#(i \to j)}{\#(i \to \star)}, \tag{6.7}$$

ο οποίος, ξανά, ισοδυναμεί με τον αλγόριθμο μέγιστης πιθανοφάνειας. Ισχύει

$$\alpha_{11} = \frac{\#(1 \to 1)}{\#(1 \to 1) + \#(1 \to 2)} = \frac{5}{9}, \quad \text{άρα} \quad \alpha_{12} = 1 - \frac{5}{9} = \frac{4}{9} \tag{6.8}$$

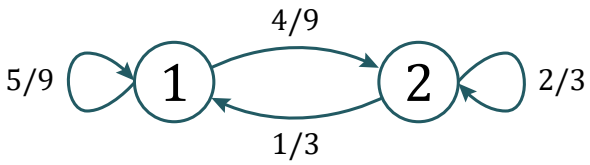
και

$$\alpha_{22} = \frac{\#(2 \to 2)}{\#(2 \to 1) + \#(2 \to 2)} = \frac{2}{3}, \quad \text{άρα} \quad \alpha_{21} = 1 - \frac{2}{3} = \frac{1}{3}. \tag{6.9}$$

Κατά συνέπεια, ο πίνακας μετάβασης θα είναι ο

$$\boldsymbol{\alpha} = \begin{pmatrix} 5/9 & 4/9 \\ 1/3 & 2/3 \end{pmatrix} \tag{6.10}$$

και το αντίστοιχο Μπεϋζιανό δίκτυο αυτό που φαίνεται στην Εικόνα 6.4.



Εικόνα 6.4: Το δίκτυο του ΜΜ-1 με τις τιμές των στοιχείων του πίνακα μετάβασης.

Σε ό,τι αφορά την $p\left(q_1 = 1, q_2 = 1, \dots, q_{10} = 1, q_{11} = 2 | q_0 = 1\right)$, αυτή θα υπολογίζεται τώρα ως

$$\begin{aligned} p_2 &= p\left(q_1 = 1, q_2 = 1, \dots, q_{10} = 1, q_{11} = 2 | q_0 = 1\right) \\ &= \frac{p\left(q_0 = 1, q_1 = 1, q_2 = 1, \dots, q_{10} = 1, q_{11} = 2\right)}{p\left(q_0 = 1\right)} = \frac{\pi_1 \cdot \overbrace{\alpha_{11} \cdot \dots \cdot \alpha_{11}}^{10 \text{ φορές}} \cdot \alpha_{12}}{\pi_1} \\ &= {\alpha_{11}}^{10} \cdot \alpha_{12} = \left(\frac{5}{9}\right)^{10} \cdot \frac{4}{9} \simeq 1.24 \cdot 10^{-3}. \end{aligned} \tag{6.11}$$

Αξίζει να σημειωθεί στο σημείο αυτό πως οι πιθανότητες p_1 του ΜΜ-2 και p_2 του ΜΜ-1 αναμένεται να ανήκουν στην ίδια τάξη μεγέθους εφόσον $\pi_{11} = \mathcal{O}\left(10^{-1}\right)$, το οποίο πράγματι ισχύει[6]. Ο λόγος για αυτό είναι πως οι προβλέψεις για μια δεδομένη ακολουθία δεν μπορεί να είναι δραματικά διαφορετικές ανάλογα με το μοντέλο που επιλέγεται κάθε φορά για την ερμηνεία της, τουλάχιστον εφόσον το εκάστοτε μοντέλο δεν είναι βαθιά προβληματικό (π.χ. να πάσχει σημαντικά από υπερπροσαρμογή).

[6] Βάσει διευκρίνησης που στάλθηκε μέσω email στις 04/01/2022

7 Linear Discriminant Analysis

Στην περίπτωση των 2 κατηγοριών, ω_1 και ω_2, οι σχέσεις για τους $\mathbf{S}_W$ και $\mathbf{S}_B$ είναι

$$\mathbf{S}_W = \sum_{i=1}^{2} \mathbb{E}_{x|x\in\omega_i}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right] \quad \text{και} \quad \mathbf{S}_B = \sum_{i=1}^{2} P(\omega_i)(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^\top, \tag{7.1}$$

αντίστοιχα, όπου

$$\boldsymbol{\mu} = P(\omega_1)\,\boldsymbol{\mu}_1 + P(\omega_2)\,\boldsymbol{\mu}_2. \tag{7.2}$$

7.1 Δεδομένης της Σχέσης (7.2) προκύπτει

$$\begin{aligned}
\mathbf{S}_B &= P(\omega_1)(\boldsymbol{\mu}_1-\boldsymbol{\mu})(\boldsymbol{\mu}_1-\boldsymbol{\mu})^\top + P(\omega_2)(\boldsymbol{\mu}_2-\boldsymbol{\mu})(\boldsymbol{\mu}_2-\boldsymbol{\mu})^\top \\
&= P(\omega_1)\{[1-P(\omega_1)]\,\boldsymbol{\mu}_1 - P(\omega_2)\,\boldsymbol{\mu}_2\}\{[1-P(\omega_1)]\,\boldsymbol{\mu}_1 - P(\omega_2)\,\boldsymbol{\mu}_2\}^\top + \\
&+ P(\omega_2)\{[1-P(\omega_2)]\,\boldsymbol{\mu}_2 - P(\omega_1)\,\boldsymbol{\mu}_1\}\{[1-P(\omega_2)]\,\boldsymbol{\mu}_2 - P(\omega_1)\,\boldsymbol{\mu}_1\}^\top \\
&= P(\omega_1)P^2(\omega_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top + P(\omega_2)P^2(\omega_1)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top \\
&= [P(\omega_1)+P(\omega_2)]\,P(\omega_1)P(\omega_2)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top \\
&= P(\omega_1)P(\omega_2)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top,
\end{aligned} \tag{7.3}$$

αποδεικνύοντας έτσι το ζητούμενο.

7.2 Η εξίσωση ιδιοτιμών για τον πίνακα $\mathbf{S}_W^{-1}\mathbf{S}_B$ είναι η

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{u} = \lambda\mathbf{u}, \tag{7.4}$$

όπου $\mathbf{u}$ είναι το ιδιοδιάνυσμα που αντιστοιχεί στην τιμή λ. Δεδομένης της Σχέσης (7.3), παρατηρεί κανείς πως

$$\begin{aligned}
\mathbf{S}_B\mathbf{u} &= P(\omega_1)P(\omega_2)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\mathbf{u} \\
&= \left[P(\omega_1)P(\omega_2)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\mathbf{u}\right](\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) \Leftrightarrow \\
\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{u} &= \left[P(\omega_1)P(\omega_2)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\mathbf{u}\right]\mathbf{S}_W^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1),
\end{aligned} \tag{7.5}$$

όπου προφανώς η ποσότητα εντός της αγκύλης είναι βαθμωτό μέγεθος. Η Σχέση (7.5) υποδεικνύει πως η επιλογή

$$\mathbf{u} = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) \tag{7.6}$$

είναι πράγματι η κατάλληλη σε ό,τι αφορά το ιδιοδιάνυσμα, με αντίστοιχη ιδιοτιμή το εσωτερικό της αγκύλης, δηλαδή

$$\lambda = P(\omega_1)P(\omega_2)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\mathbf{S}_W^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1). \tag{7.7}$$

8 Independent Component Analysis

8.1 Ορίζοντας την κύρτωση μιας τυχαίας μεταβλητής, x, με μηδενική μέση τιμή ως

$$\mathrm{kurt}\,(x) \coloneqq \mathbb{E}\left[x^4\right] - 3\left(\mathbb{E}\left[x^2\right]\right)^2, \tag{8.1}$$

και λαμβάνοντας υπ' όψιν πως η ανεξαρτησία δύο μεταβλητών x, y σημαίνει πως

$$\mathbb{E}\left[x^\nu\, y^\mu\right] = \mathbb{E}\left[x^\nu\right]\mathbb{E}\left[y^\mu\right], \tag{8.2}$$

βρίσκει κανείς ότι

$$\begin{aligned}\mathbb{E}\left[(x+y)^4\right] &= \mathbb{E}\left[x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4\right] \\ &= \mathbb{E}\left[x^4\right] + 4\mathbb{E}\left[x^3\right]\underbrace{\mathbb{E}\left[y\right]}_{=0} + 6\mathbb{E}\left[x^2\right]\mathbb{E}\left[y^2\right] + 4\underbrace{\mathbb{E}\left[x\right]}_{=0}\mathbb{E}\left[y^3\right] + \mathbb{E}\left[y^4\right] \\ &= \mathbb{E}\left[x^4\right] + \mathbb{E}\left[y^4\right] + 6\mathbb{E}\left[x^2\right]\mathbb{E}\left[y^2\right]\end{aligned} \tag{8.3}$$

καθώς και

$$\begin{aligned}-3\left(\mathbb{E}\left[(x+y)^2\right]\right)^2 &= -3\left(\mathbb{E}\left[x^2 + 2xy + y^2\right]\right)^2 = -3\left(\mathbb{E}\left[x^2\right] + 2\underbrace{\mathbb{E}\left[x\right]\mathbb{E}\left[y\right]}_{=0} + \mathbb{E}\left[y^2\right]\right)^2 \\ &= -3\left(\mathbb{E}\left[x^2\right]\right)^2 - 6\mathbb{E}\left[x^2\right]\mathbb{E}\left[y^2\right] - 3\left(\mathbb{E}\left[y^2\right]\right)^2.\end{aligned} \tag{8.4}$$

Βάσει των Σχέσεων (8.3), (8.4) προκύπτει

$$\begin{aligned}\mathrm{kurt}\,(x+y) &= \mathbb{E}\left[(x+y)^4\right] - 3\left(\mathbb{E}\left[(x+y)^2\right]\right)^2 \\ &= \mathbb{E}\left[x^4\right] + \mathbb{E}\left[y^4\right] + 6\mathbb{E}\left[x^2\right]\mathbb{E}\left[y^2\right] - 3\left(\mathbb{E}\left[x^2\right]\right)^2 - 6\mathbb{E}\left[x^2\right]\mathbb{E}\left[y^2\right] - 3\left(\mathbb{E}\left[y^2\right]\right)^2 \\ &= \mathbb{E}\left[x^4\right] - 3\left(\mathbb{E}\left[x^2\right]\right)^2 + \mathbb{E}\left[y^4\right] - 3\left(\mathbb{E}\left[y^2\right]\right)^2 \\ &= \mathrm{kurt}\,(x) + \mathrm{kurt}\,(y),\end{aligned} \tag{8.5}$$

αποδεικνύοντας έτσι το ζητούμενο.

8.2 Για τις δοθείσες τυχαίες μεταβλητές, s_i, με $i = 1, \dots, N$ ισχύει $\mathbb{E}\left[s_i\right] = 0$ και $\mathbb{E}\left[{s_i}^2\right] = 1$. Δεδομένου πως

$$x = \sum_{i=1}^{N} w_i s_i, \tag{8.6}$$

τότε ισχύει πως

$$\mathbb{E}\left[x\right] = w_i \sum_{i=1}^{N} \mathbb{E}\left[s_i\right] = 0. \tag{8.7}$$

Θεωρώντας πως οι s_i είναι μεταξύ τους ανεξάρτητες, θα ισχύει

$$\mathbb{E}\left[s_i s_j\right] = \mathbb{E}\left[{s_i}^2\right]\delta_{ij} = \delta_{ij}, \tag{8.8}$$

αφού οι μέσες τιμές των s_i είναι εξ υποθέσεως μηδενικές. Με βάση τη Σχέση (8.8), προκύπτει

$$\begin{aligned}\mathrm{Var}\left[x\right] = \mathbb{E}\left[x^2\right] - \underbrace{\left(\mathbb{E}\left[x\right]\right)^2}_{=0} = \mathbb{E}\left[\sum_{i=1}^{N} w_i s_i \sum_{j=1}^{N} w_j s_j\right] = \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j s_i s_j\right] \\ = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \mathbb{E}\left[s_i s_j\right] = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \delta_{ij} = \sum_{i=1}^{N} {w_i}^2.\end{aligned} \tag{8.9}$$

Έτσι, η απαίτηση $\mathrm{Var}\left[x\right] = 1$ ισοδυναμεί με την απαίτηση

$$\sum_{i=1}^{N} {w_i}^2 = 1. \tag{8.10}$$

Θεωρώντας τώρα ένα ομοιόμορφα σταθμισμένο μίγμα N μεταβλητών s_i, θα ισχύει $w_i = w_j = w$ για κάθε i, j. Δεδομένου, επιπλέον, πως το μίγμα έχει μοναδιαία διασπορά, η Σχέση (8.10) δίνει

$$\sum_{i=1}^{N} {w_i}^2 = 1 \Leftrightarrow w^2 N = 1 \Leftrightarrow w = \frac{c}{\sqrt{N}}, \tag{8.11}$$

όπου $c = \pm 1$, αφού και οι δύο λύσεις είναι δεκτές. Για την κύρτωση της x, δεδομένης της Σχέσης (8.5) που αποδείχθηκε στο πρώτο μέρος της άσκησης, ισχύει

$$\begin{aligned}\left|\mathrm{kurt}\left(x\right)\right| = \left|\mathrm{kurt}\left(\frac{c}{\sqrt{N}}\sum_{i=1}^{N} s_i\right)\right| = \frac{c^4}{N^2}\left|\sum_{i=1}^{N}\mathrm{kurt}\left(s_i\right)\right| = \frac{1}{N^2}\left|\sum_{i=1}^{N} a_i\right| \\ \leq \frac{1}{N^2}\sum_{i=1}^{N} a = \frac{aN}{N^2} = \frac{a}{N} \Leftrightarrow -\frac{a}{N} \leq \mathrm{kurt}\left(x\right) \leq \frac{a}{N},\end{aligned} \tag{8.12}$$

αφού a_i είναι η κύρτωση της μεταβλητής s_i, για την οποία ισχύει $-a \leq a_i \leq a$, με $a > 0$. Δεδομένου πως

$$\lim_{N\to\infty}\left(\frac{a}{N}\right) = 0 = \lim_{N\to\infty}\left(-\frac{a}{N}\right), \tag{8.13}$$

το κριτήριο παρεμβολής για τη Σχέση (8.12) υποδεικνύει πως

$$\lim_{N\to\infty}\left[\mathrm{kurt}\left(x\right)\right] = 0, \tag{8.14}$$

αποδεικνύοντας έτσι το ζητούμενο.

9 Logistic Regression

Στα πλαίσια της λογιστικής παλινδρόμησης για ένα σύνολο δεδομένων $\{\phi_n, t_n\}$ με $t_n \in \{0,1\}$ (δυαδική ταξινόμηση), $\phi_n = \phi(\mathbf{x}_n)$ και $n = 1, \dots, N$, η a-posteriori πιθανότητα της μιας κατηγορίας, έστω της $\mathcal{C}_1$, ορίζεται ως

$$p(\mathcal{C}_1|\phi_n) = \sigma\left(\mathbf{w}^\top \phi_n\right) \equiv y_n, \tag{9.1}$$

επομένως $p(\mathcal{C}_2|\phi_n) = 1 - y_n$.

9.1 Σε ένα γραμμικά διαχωρίσιμο σύνολο δεδομένων, τα σημεία n για τα οποία ισχύει $p(\mathcal{C}_1|\phi_n) > p(\mathcal{C}_2|\phi_n)$ θα ταξινομούνται στην κατηγορία $\mathcal{C}_1$, ενώ τα σημεία για τα οποία ισχύει $p(\mathcal{C}_1|\phi_n) < p(\mathcal{C}_2|\phi_n)$ θα ταξινομούνται στην κατηγορία $\mathcal{C}_2$. Έτσι, η επιφάνεια απόφασης θα ορίζεται βάσει των

$$p(\mathcal{C}_1|\phi_n) = p(\mathcal{C}_2|\phi_n) \Leftrightarrow y_n = 1 - y_n \Leftrightarrow \sigma\left(\mathbf{w}^\top \phi_n\right) = 0.5 \Leftrightarrow \sigma\left(\mathbf{w}^\top \phi_n\right) = \sigma(0)$$

$$\Leftrightarrow \mathbf{w}^\top \phi_n = 0, \tag{9.2}$$

όπου στην τελευταία ισοδυναμία αξιοποιείται το γεγονός πως η $\sigma(x)$ είναι 1-1, ενώ η ύπαρξη ενός τέτοιου $\mathbf{w}$ διασφαλίζεται από τη γραμμική διαχωρισιμότητα του συνόλου. Η συνάρτηση πιθανοφάνειας για ένα τέτοιο πρόβλημα μπορεί να γραφεί ως [3]

$$\ell(\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n} \tag{9.3}$$

και επειδή ο φυσικός λογάριθμος είναι γνησίως αύξουσα συνάρτηση του ορίσματός της, η μεγιστοποίηση της πιθανοφάνειας ισοδυναμεί με τη μεγιστοποίηση του λογαρίθμου της, δηλαδή

$$\ln \ell(\mathbf{w}) = \ln \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1-t_n} = \sum_{n=1}^{N} \ln\left[y_n^{t_n} (1 - y_n)^{1-t_n}\right] = \sum_{n=1}^{N} \left[\ln y_n^{t_n} + \ln(1 - y_n)^{1-t_n}\right]$$

$$= \sum_{n=1}^{N} [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)] = -E(\mathbf{w}), \tag{9.4}$$

όπου η $E(\mathbf{w})$ είναι η δοθείσα συνάρτηση σφάλματος (cross-entropy). Προκύπτει, λοιπόν, πως στην προκείμενη περίπτωση η μεγιστοποίηση της πιθανοφάνειας ισοδυναμεί με την ελαχιστοποίηση της συνάρτησης σφάλματος. Για το σκοπό αυτό, γράφει κανείς

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

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

$$= \sum_{n=1}^{N} \frac{t_n - y_n}{y_n (1 - y_n)} y_n (1 - y_n) \phi_n = \sum_{n=1}^{N} (t_n - y_n) \phi_n, \tag{9.5}$$

όπου στην τρίτη ισότητα αξιοποιήθηκε η ιδιότητα

$$\frac{d\sigma(x)}{dx} = \sigma(x)\left[1-\sigma(x)\right] \tag{9.6}$$

της σιγμοειδούς συνάρτησης. Έτσι, η μεγιστοποίηση της πιθανοφάνειας, η οποία ισοδυναμεί με την απαίτηση $-\nabla E(\mathbf{w}) = 0$, βάσει της Σχέσης (9.5) ανάγεται στην εύρεση ενός $\mathbf{w}$ τέτοιου, ώστε

$$\sigma\left(\mathbf{w}^{\top}\phi_n\right) \to \begin{cases} 1, & \text{εάν } t_n = 1 \\ 0, & \text{εάν } t_n = 0 \end{cases}, \ \forall\, n = 1, \dots, N \tag{9.7}$$

και επειδή η σιγμοειδής συνάρτηση είναι 1-1,

$$\mathbf{w}^{\top}\phi_n \to \begin{cases} \infty, & \text{εάν } t_n = 1 \\ -\infty, & \text{εάν } t_n = 0 \end{cases}, \ \forall\, n = 1, \dots, N. \tag{9.8}$$

Συμπεραίνει κανείς πως, προκειμένου η απαίτηση (9.8) να ικανοποιείται για κάθε n, θα πρέπει με βεβαιότητα για το $\mathbf{w}$ να ισχύει

$$\|\mathbf{w}\|_2 \to \infty. \tag{9.9}$$

Αξίζει να σημειωθεί στο σημείο αυτό πως η περίπτωση αυτή ισοδυναμεί με τη χρήση της βηματικής συνάρτησης Heaviside για την ταξινόμηση σημείων, με αποτέλεσμα η a-posteriori πιθανότητα για κάθε σημείο εκπαίδευσης να ισούται με τη μονάδα και συνεπώς το μοντέλο να κινδυνεύει σημαντικά από υπερπροσαρμογή (overfitting).

9.2 Σε ό,τι αφορά τη Hessian μήτρα της cross-entropy στο συγκεκριμένο πρόβλημα, αυτή θα δίνεται ως $\mathbf{H} = \nabla\nabla E(\mathbf{w})$. Ξεκινώντας από τη Σχέση (9.5) και χρησιμοποιώντας ξανά την ιδιότητα (9.6), προκύπτει

$$\mathbf{H} = \nabla \sum_{n=1}^{N} (y_n - t_n)\,\phi_n = \sum_{n=1}^{N} (\nabla y_n)\,\phi_n^{\top} = \sum_{n=1}^{N} y_n\,(1-y_n)\,\phi_n\phi_n^{\top} \equiv \Phi^{\top}\mathbf{R}\Phi, \tag{9.10}$$

όπου $\mathbf{R} = \mathrm{diag}\left[y_n\,(1-y_n)\right]$ και ο Φ είναι ο $N \times M$ πίνακας, του οποίου η n-οστή γραμμή είναι το διάνυσμα $\phi_n^{\top}$, το οποίο είναι ένα M-διάστατο διάνυσμα. Θεωρώντας ένα τυχαίο M-διάσταστο διάνυσμα $\mathbf{v}$, το οποίο δεν ταυτίζεται με το μηδενικό διάνυσμα $\mathbb{0}_M$, ισχύει πως

$$\begin{aligned} \mathbf{v}^{\top}\mathbf{H}\mathbf{v} &= \mathbf{v}^{\top}\left(\sum_{n=1}^{N} y_n\,(1-y_n)\,\phi_n\phi_n^{\top}\right)\mathbf{v} = \sum_{n=1}^{N} y_n\,(1-y_n)\,\mathbf{v}^{\top}\phi_n\phi_n^{\top}\mathbf{v} \\ &= \sum_{n=1}^{N} y_n\,(1-y_n)\left(\phi_n^{\top}\mathbf{v}\right)^{\top}\left(\phi_n^{\top}\mathbf{v}\right) = \sum_{n=1}^{N} y_n\,(1-y_n)\,\|\phi_n^{\top}\mathbf{v}\|_2^2. \end{aligned} \tag{9.11}$$

Σε ό,τι αφορά τις ποσότητες $\|\phi_n^\top \mathbf{v}\|_2^2$, για τις οποίες προφανώς ισχύει $\|\phi_n^\top \mathbf{v}\|_2^2 \geq 0$, έστω πως η ισότητα $\|\phi_n^\top \mathbf{v}\|_2 = 0$ ισχύει για κάθε $n = 1, \dots, N$. Τότε, το διάνυσμα $\mathbf{v}$ ανήκει στον πυρήνα του Φ, επομένως $\dim[\mathrm{Ker}(\Phi)] > 0$. Πρόσθετα, θα πρέπει για την εικόνα του Φ να ισχύει $\dim[\mathrm{Im}(\Phi)] = M$, καθώς εξ υποθέσεως τα διανύσματα $\{\phi_n\}$ συνιστούν βάση σε έναν M-διάστατο χώρο. Συνδυάζοντας τις δύο αυτές παρατηρήσεις, προκύπτει πως

$$\dim[\mathrm{Ker}(\Phi)] + \dim[\mathrm{Im}(\Phi)] > M \Leftrightarrow M > M, \tag{9.12}$$

για το οποίο χρησιμοποιήθηκε το Θεώρημα Διάστασης, οδηγώντας έτσι σε άτοπο. Το συμπέρασμα που προκύπτει από τη μικρή αυτή διερεύνηση είναι πως υπάρχει τουλάχιστον ένα k τέτοιο, ώστε $\|\phi_k^\top \mathbf{v}\|_2 \neq 0$. Επιπλέον, δεδομένου πως το σύνολο τιμών της σιγμοειδούς συνάρτησης είναι το διάστημα $(0, 1)$, θα ισχύει για κάθε n πως $0 < y_n < 1$ και ως εκ τούτου $y_n(1 - y_n) > 0$. Βάσει αυτών, από τη Σχέση (9.11) προκύπτει ότι

$$\mathbf{v}^\top \mathbf{H} \mathbf{v} > 0, \tag{9.13}$$

επομένως η $\mathbf{H}$ είναι πράγματι θετικώς ορισμένη και άρα η $E(\mathbf{w})$ είναι κυρτή ως προς το $\mathbf{w}$ και έχει μοναδικό ελάχιστο, το οποίο προκύπτει από το μηδενισμό της Σχέσης (9.5).

9.3 Ο αλγόριθμος IRLS (Iterative Reweighted Least Squares) είναι μια επαναληπτική διαδικασία που αξιοποιεί τη μέθοδο Newton-Raphson, με σκοπό την ελαχιστοποίηση της cross-entropy και κατ' επέκταση τον προσδιορισμό του διανύσματος βάρους, $\mathbf{w}$. Γράφοντας τη Σχέση (9.5) χρησιμοποιώντας πίνακες, προκύπτει

$$\nabla E(\mathbf{w}) = \Phi^\top (\mathbf{y} - \mathbf{t}), \tag{9.14}$$

όπου $\mathbf{t} = [t_1, \dots, t_N]^\top$ και $\mathbf{y} = [y_1, \dots, y_N]^\top$. Χρησιμοποιώντας την έκφραση αυτή μαζί με τη Σχέση (9.10), η εξίσωση που αφορά στην ανανέωση του διανύσματος βάρους από επανάληψη σε επανάληψη είναι η

$$\mathbf{w}^{\mathrm{new}} = \mathbf{w}^{\mathrm{old}} - \mathbf{H}^{-1} \cdot \nabla E\left(\mathbf{w}^{\mathrm{old}}\right) = \mathbf{w}^{\mathrm{old}} - \left(\Phi^\top \mathbf{R} \Phi\right)^{-1} \Phi^\top (\mathbf{y} - \mathbf{t}). \tag{9.15}$$

Όλη η μέχρι τώρα ανάλυση αφορά το πρόβλημα δυαδικής ταξινόμησης, όμως όλα τα συμπεράσματα μπορούν να επεκταθούν και στην περίπτωση πολλαπλής ταξινόμησης [3], για C κλάσεις. Οι βασικές διαφορές είναι πως η συνάρτηση ενεργοποίησης δε θα είναι η σιγμοειδής, αλλά η softmax, ενώ το $\mathbf{t}$ θα αντιστοιχεί σε έναν $N \times C$ πίνακα, με τη n-οστή γραμμή του να έχει παντού την τιμή 0, εκτός από τη θέση που αντιστοιχεί στην κλάση ταξινόμησης του στοιχείου $\mathbf{x}_n$, η οποία λαμβάνει την τιμή 1. Αντίστοιχα, το $\mathbf{y}$ θα μετατραπεί σε έναν $N \times C$ πίνακα, η n-οστή γραμμή του οποίου θα έχει ως στοιχεία τις τιμές που παίρνει η softmax για την ταξινόμηση του σημείου $\mathbf{x}_n$ σε κάθε κλάση. Τέλος, αντί για 1, θα υπάρχουν πλέον C διανύσματα βαρών, 1 για κάθε κλάση. Έτσι, η αντίστοιχη της Σχέσης (9.5) θα είναι η

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_c) = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n, \tag{9.16}$$

ενώ η αντίστοιχη της Σχέσης (9.10) θα είναι η

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_c) = -\sum_{n=1}^{N} y_{nk} (I_{kj} - y_{nj}) \phi_n \phi_n^\top, \tag{9.17}$$

όπου I είναι ο μοναδιαίος πίνακας. Η αναλογία ως προς το πρόβλημα δυαδικής ταξινόμησης είναι εμφανής, καθώς η Σχέση (9.15) ισχύει αυτούσια, δεδομένων βέβαια των νέων διαστάσεων των πινάκων που υπεισέρχονται σε αυτήν.

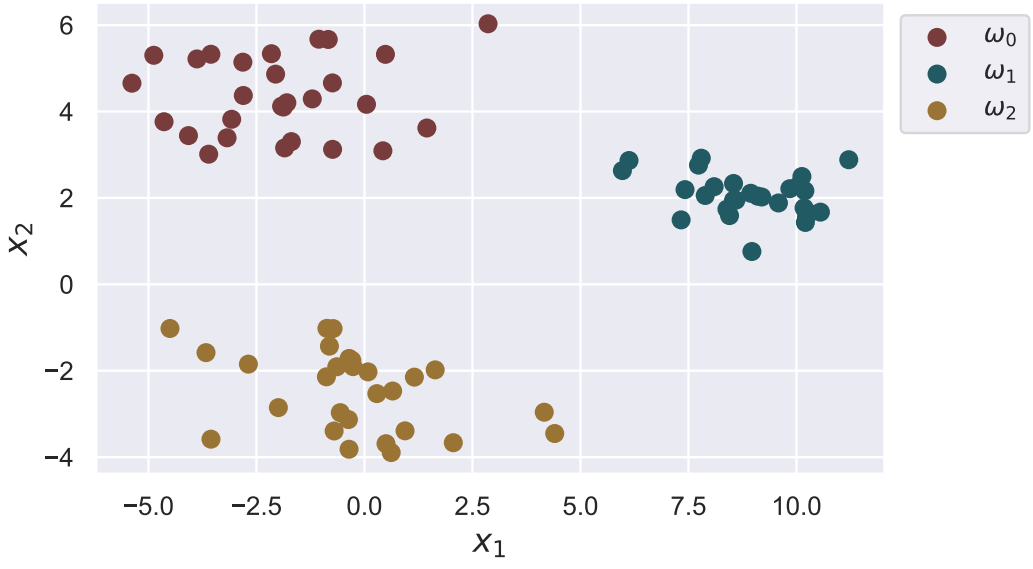
Βάσει της ανάλυσης αυτής, ο αλγόριθμος IRLS υλοποιήθηκε σε Python ως εξής:

```
import numpy as np
from sklearn.base import BaseEstimator, ClassifierMixin

class IWLS(BaseEstimator, ClassifierMixin):
    def __init__(self, C, iters=15, tol=1e-6):
        self.w = None
        self.C = C # Number of classes
        self.tol = tol # convergence tolerance
        self.iters = iters # maximum runs

    # a function to return a matrix with len(X) rows,
    # each having as elements the #classes values of
    # y_c = softmax(w_c*x)
    def softmax(self, X, C):
        # X is a NxM matrix, where N = # of data and M = # of features
        N, M = X.shape
        X = np.asarray(X)
        sft = np.zeros((N,C)) # initialize the N x C matrix Y
        C = self.C

        for i,x in enumerate(X):
            x = np.reshape(x, (M,1)) # reshape into a Mx1 vector

            T = np.array([], dtype='float64')
            for c in range(C):
                wr = np.reshape(self.w[c], (1,M)) # reshape the weight
                t = np.squeeze(np.dot(wr,x)) # w_c*x dot product
                T = np.append(T, t) # append it to T

            # T now has c elements, each corresponding to
            # the dot product w_c*x for category C
            # We use this to calculate this x's softmax
            expons = np.exp(T)
            sfti = expons/expons.sum()
            # Append the calculated softmax values to the sft matrix
            sft[i,:] = np.reshape(sfti, (1,C))

        return sft

    def fit(self, X, y):
        # Same as above
        N, M = X.shape
        X = np.asarray(X)
        y = np.asarray(y)
        inputC = len(np.unique(y))
        C = self.C
        if (inputC != C):
```

```
48              print(f"Input problem: {self.C} =/= {inputC}.")
49
50          Xt = np.transpose(X)
51          Id = np.identity(C)
52
53          T = np.zeros((N,C))
54          for i,c in enumerate(y):
55              T[i,c] = 1 # T matrix, previously t vector
56
57          self.w = np.zeros((C,M)) # initialization of weights
58          nablaE = np.ones((C,M)) # initilization of ∇E(w)
59          Hessian = np.ones((C*M,C*M)) # initialization of Hessian matrix
60
61          current_iter = 0
62          epsilon = 1000.0
63          while current_iter < self.iters and epsilon > self.tol:
64
65              # Calculate the softmax matrix Y
66              sft = self.softmax(X,C)
67
68              # Calculate ∇E(w) as in Bishop eq. 4.96
69              yminusT = sft-T
70              nablaE = np.transpose(np.matmul(Xt,yminusT))
71
72              # Calculate the Hessian
73              for i in range(C):
74                  for j in range(C):
75                      unit = Id[i,j]
76                      R = np.zeros((N,N))
77                      for n in range(N):
78                          R[n,n] = sft[n,i]*(unit-sft[n,j])
79                      XtR = np.matmul(Xt,R)
80                      Hessian[i*M:(i+1)*M,j*M:(j+1)*M] = np.matmul(XtR,X)
81
82              # Update the weights
83              self.w = np.reshape(self.w, (C*M,1))
84              nablaE = np.reshape(nablaE, (C*M,1))
85              # We calculate the pseudoinverse instead
86              # of the inverse, to avoid singularities
87              # due to the x_0 feature
88              Hinv = np.linalg.pinv(Hessian)
89              update = np.matmul(Hinv,nablaE)
90              new_w = self.w - update
91
92              #Switch back to original dimensions
93              nablaE = np.reshape(nablaE, (C,M))
94              self.w = np.reshape(new_w, (C,M))
95
96              current_iter += 1
97              epsilon = max(abs(update))
98
99          if current_iter == self.iters:
```

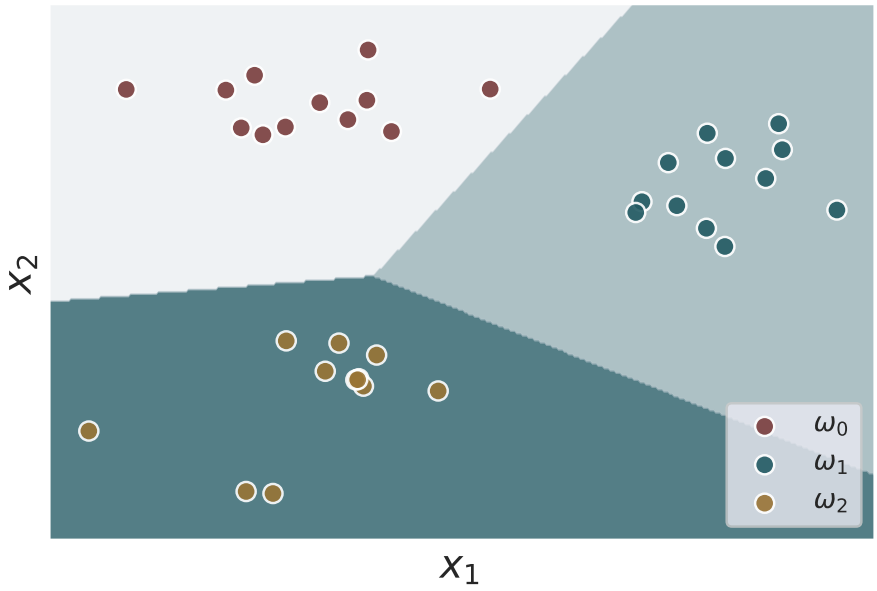
```
100             print(f"Fitting finished after {current_iter} iterations.")
101         else:
102             print(f"Fitting finished, convergence achieved.")
103 
104         return self
105 
106     def predict(self,X):
107         sft = self.softmax(X,self.C)
108         y_preds = np.argmax(sft,axis=1)
109 
110         return y_preds
111 
112     def score(self,X,y):
113         y_preds = self.predict(X)
114         hits = ((y_preds-y)==0).sum()
115 
116         return hits/len(y)
```

Η κλάση γράφηκε στα πρότυπα ενός ταξινομητή της `sklearn`, προκειμένου να μπορεί να αξιοποιηθεί αρμονικά με άλλες μεθόδους και συναρτήσεις της βιβλιοθήκης αυτής. Το πρόβλημα στο οποίο έγινε χρήση της ήταν ένα πρόβλημα ταξινόμησης 2-διάστατων χαρακτηριστικών σε τρεις κλάσεις, ω_0, ω_1 και ω_2. Από τα διαθέσιμα δεδομένα, το 30% δεν αξιοποιήθηκε στη διαδικασία εκπαίδευσης του ταξινομητή λογιστικής παλινδρόμησης, προκειμένου να χρησιμοποιηθούν ως δεδομένα αξιολόγησης. Τα σημεία που αντιστοιχούν στο υπόλοιπο 70% των δεδομένων εκπαίδευσης απεικονίζονται στο διάγραμμα διασποράς της Εικόνας 9.1, όπου ο χρωματικός κώδικας αντιστοιχεί στις διαφορετικές κλάσεις.



Εικόνα 9.1: Δεδομένα εκπαίδευσης του ταξινομητή λογιστικής παλινδρόμησης.

Γίνεται εμφανές πως οι κλάσεις είναι ανά δύο γραμμικά διαχωρίσιμες. Εκπαιδεύοντας ένα μοντέλο της κλάσης που παρατέθηκε παραπάνω, μετά από 15 μόλις επαναλήψεις του αλγορίθμου IRLS, αυτό επιτυγχάνει ακρίβεια 100% στην ταξινόμηση των δεδομένων αξιολόγησης. Οι επιφάνειες (ευθείες) απόφασης μαζί με τα σημεία και τις κλάσεις στις οποίες αυτά ταξινομούνται φαίνονται στο διάγραμμα της Εικόνας 9.2.



Εικόνα 9.2: Ταξινόμηση των δεδομένων αξιολόγησης και απεικόνιση επιφανειών απόφασης.

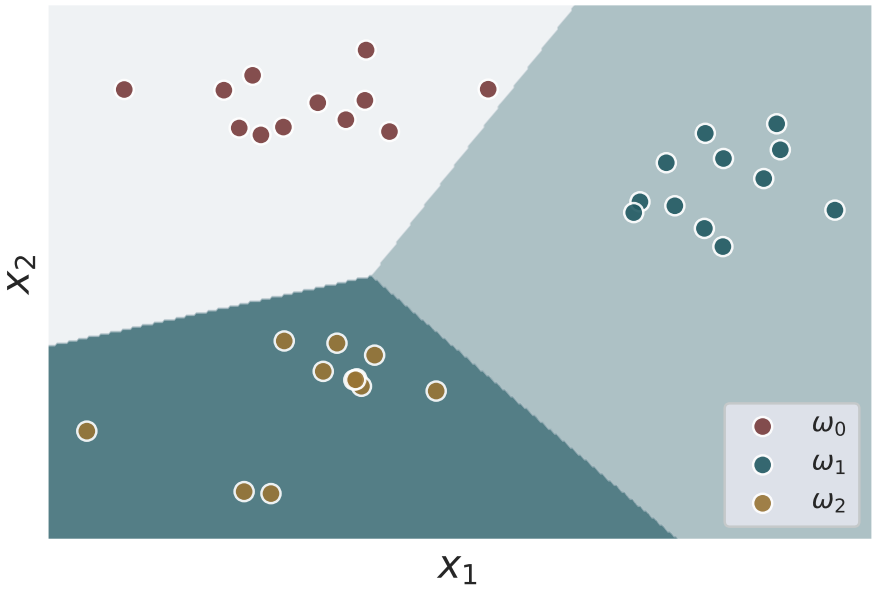
Γίνεται εύκολα αντιληπτό πως, παρά τη γραμμική διαχωρισιμότητα των κλάσεων ανά δύο, το μοντέλο λογιστικής παλινδρόμησης που κατασκευάστηκε δεν προχωρά σε υπερπροσαρμογή (διότι δεν εκπαιδεύτηκε για πολλές επαναλήψεις) και κατ' επέκταση απειρισμό των σχετικών διανυσμάτων βαρών, αντιθέτως πραγματοποιεί με απόλυτη ακρίβεια την ταξινόμηση των δεδομένων αξιολόγησης και ορίζει επιφάνειες απόφασης με σχετικά καλά περιθώρια.

Η ίδια διαδικασία ταξινόμησης μπορεί να πραγματοποιηθεί και μέσω γραμμικής παλινδρόμησης (μέθοδος ελαχίστων τετραγώνων). Στα ίδια πλαίσια, υλοποιείται σε Python μια σχετική κλάση με όνομα `LRMSE`, όπως φαίνεται παρακάτω:

```
class LRMSE(BaseEstimator, ClassifierMixin):
    def __init__(self, C):
        self.w = None
        self.C = C

    def fit(self, X, y):

        C = self.C
        N, M = X.shape
        X = np.asarray(X)
        Xt = np.transpose(X)

        T = np.zeros((N,C))
        for i,c in enumerate(y):
            T[i,c] = 1

        XtX = np.matmul(Xt,X)
        new_w = np.matmul(np.linalg.pinv(XtX), np.matmul(Xt,T))
        self.w = new_w

        return self

```

```
23      def predict(self, X):
24          X = np.asarray(X)
25          ys = np.matmul(X,self.w)
26          y_preds = np.argmax(ys,axis=1)
27
28          return y_preds
29
30      def score(self, X, y):
31          y_preds = self.predict(X)
32          hits = ((y_preds-y)==0).sum()
33
34          return hits/len(y)
```

Πραγματοποιώντας τον ίδιο ακριβώς διαχωρισμό στα δεδομένα εκπαίδευσης και αξιολόγησης, ο ταξινομητής που βασίζεται στην παραπάνω κλάση επιτυγχάνει και αυτός 100% ακρίβεια στην ταξινόμηση των δεδομένων αξιολόγησης, αφότου εκπαιδευτεί. Τα αντίστοιχα αποτελέσματα, μαζί με τις επιφάνειες διαχωρισμού, απεικονίζονται στο σχήμα της Εικόνας 9.3.

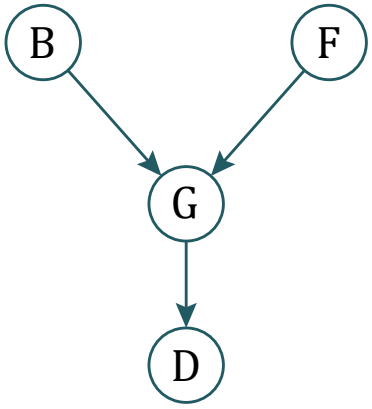


Εικόνα 9.3: Ταξινόμηση των δεδομένων αξιολόγησης μέσω γραμμικής παλινδρόμησης και απεικόνιση επιφανειών απόφασης.

Αξίζει να σημειωθεί πως, παρότι και σε αυτήν την περίπτωση η ταξινόμηση των σημείων αξιολόγησης γίνεται με απόλυτη επιτυχία, οι επιφάνειες διαχωρισμού δεν έχουν εξίσου καλά περιθώρια, γεγονός που φαίνεται από το πόσο κοντά στα αντίστοιχα σύνορα είναι ορισμένα σημεία των κλάσεων ω_0 και ω_2.

10 Conditional Independence

Το δίκτυο που μελετάται στη συγκεκριμένη άσκηση σχετίζεται με το σύστημα καυσίμων ενός αυτοκινήτου και συγκεκριμένα την κατάσταση του δείκτη καυσίμων, G, όπως αυτός παρατηρείται από τον οδηγό, D. Συγκεκριμένα, η κατάσταση $G = 1$ αντιστοιχεί σε ένδειξη του δείκτη για γεμάτο ντεπόζιτο (και αντίστοιχα η $G = 0$ για άδειο), ενώ η κατάσταση D αντιστοιχεί σε παρατήρηση του οδηγού ως προς την ένδειξη του δείκτη για γεμάτο ($D = 1$) ή άδειο ($D = 0$) ντεπόζιτο. Το δίκτυο απεικονίζεται στην Εικόνα 10.1.



Εικόνα 10.1: Γράφος του δικτύου που αφορά τη μπαταρία (B), το ντεπόζιτο (F), το δείκτη (G) και την παρατήρηση του οδηγού (D).

Όπως γίνεται εύκολα αντιληπτό, τα B και F είναι ενδεχόμενα εν γένει ανεξάρτητα, αλλά όχι ανεξάρτητα *δεδομένου του* G (ή, λόγω κληρονομικότητας, του D). Σε ό,τι αφορά τις a-priori πιθανότητες σχετικά με το εάν η μπαταρία είναι φορτισμένη ($B = 1$) ή αφόρτιστη ($B = 0$) και το εάν το ντεπόζιτο είναι γεμάτο ($F = 1$) ή όχι ($F = 0$), αυτές δίνονται στους Πίνακες 10.1. Με μπλε χρώμα έχουν σημειωθεί οι τιμές που δε δίνονται από την εκφώνηση, αλλά μπορούν να εξαχθούν με τετριμμένο τρόπο.

B	0	1
$p(B)$	0.05	0.95

F	0	1
$p(F)$	0.2	0.8

Πίνακες 10.1: A-priori πιθανότητες (αριστερά) για τη μπαταρία και (δεξιά) για το ντεπόζιτο.

Από την άλλη, στους Πίνακες 10.2 παρατίθενται οι δεσμευμένες πιθανότητες $p\,(G|B,F)$ και $p\,(D|G)$, ακολουθώντας την ίδια χρωματική σύμβαση.

(B,F) \ G	0	1
$(0,0)$	0.8	0.2
$(0,1)$	0.75	0.25
$(1,0)$	0.7	0.3
$(1,1)$	0.05	0.95

G \ D	0	1
0	0.8	0.2
1	0.2	0.8

Πίνακες 10.2: Δεσμευμένες πιθανότητες (αριστερά) για το δείκτη και (δεξιά) για τον οδηγό.

10.1 Το πρώτο ζητούμενο αντιστοιχεί στον υπολογισμό της πιθανότητας $p_1 = p\,(F = 0|D = 0)$, για την οποία ισχύει:

$$\begin{aligned} p_1 &= \frac{p\,(D = 0|F = 0) \cdot p\,(F = 0)}{p\,(D = 0)} = \frac{p\,(F = 0)}{p\,(D = 0)} \sum_G p\,(D = 0, G|F = 0) \\ &= \frac{p\,(F = 0)}{p\,(D = 0)} \sum_G p\,(D = 0|G, F = 0) \cdot p\,(G|F = 0) \\ &= \frac{p\,(F = 0)}{p\,(D = 0)} \sum_G p\,(D = 0|G) \cdot p\,(G|F = 0), \end{aligned} \tag{10.1}$$

όπου στην πρώτη ισότητα χρησιμοποιήθηκε ο κανόνας Bayes, στην τρίτη ο ορισμός της δεσμευμένης πιθανότητας και στην τέταρτη αξιοποιήθηκε το γεγονός πως δεδομένης της πληροφορίας για την G, η πληροφορία για την F περιττεύει σε ό,τι αφορά τον προσδιορισμό της δεσμευμένης πιθανότητας σχετικά με την παρατήρηση του οδηγού, D. Γενικά, ισχύει

$$p\,(D|G, B) = p\,(D|G, F) = p\,(D|G)\,, \tag{10.2}$$

γεγονός το οποίο θα αξιοποιηθεί και παρακάτω. Προκειμένου η Σχέση (10.1) να υπολογιστεί, θα πρέπει πρώτα να υπολογιστούν οι επί μέρους όροι που υπεισέρχονται σε αυτήν. Για τον υπολογισμό της $p\,(D = 0)$ απαιτείται πρώτα ο υπολογισμός της $p\,(G = 0)$, για την οποία ισχύει

$$\begin{aligned} p\,(G = 0) &= \sum_{B,F} p\,(G = 0, B, F) = \sum_{B,F} p\,(G = 0|B, F)\, p\,(B|F)\, p\,(F) \\ &= \sum_{B,F} p\,(G = 0|B, F)\, p\,(B)\, p\,(F) = 0.008 + 0.133 + 0.03 + 0.038 = 0.209, \end{aligned} \tag{10.3}$$

όπου στην τρίτη ισότητα χρησιμοποιήθηκε απλώς η ανεξαρτησία των B και F. Προφανώς, το αποτέλεσμα της Σχέσης (10.3) οδηγεί στο $p\,(G = 1) = 1 - p\,(G = 0) = 0.791$. Βάσει αυτών προκύπτει πως

$$p\,(D = 0) = \sum_G p\,(D = 0, G) = \sum_G p\,(D = 0|G)\, p\,(G) = 0.1672 + 0.1582 = 0.3254 \tag{10.4}$$

και προφανώς $p\,(D = 1) = 1 - p\,(D = 0) = 0.6746$. Προχωρώντας, υπολογίζονται οι $p\,(G|F = 0)$ που εμφανίζονται στη Σχέση (10.1):

$$\begin{aligned} p\,(G = 0|F = 0) &= \sum_B p\,(G = 0, B|F = 0) = \sum_B p\,(G = 0|B, F = 0)\, p\,(B|F = 0) \\ &= \sum_B p\,(G = 0|B, F = 0)\, p\,(B) = 0.04 + 0.665 = 0.705 \end{aligned} \tag{10.5}$$

και επομένως $p\,(G = 1|F = 0) = 1 - p\,(G = 0|F = 0) = 0.295$. Αντικαθιστώντας τα παραπάνω αποτελέσματα στη Σχέση (10.1) προκύπτει

$$p_1 = \frac{0.2}{0.3254}\,(0.8 \cdot 0.705 + 0.2 \cdot 0.295) \simeq 0.3829. \tag{10.6}$$

10.2 Σε ό,τι αφορά το δεύτερο ζητούμενο, αυτό αντιστοιχεί στον υπολογισμό της πιθανότητας $p_2 = p\,(F = 0|B = 0, D = 0)$, για την οποία ισχύει

$$\begin{aligned} p_2 &= \frac{p\,(D = 0, B = 0, F = 0)}{p\,(D = 0, B = 0)} = \frac{1}{p\,(D = 0, B = 0)} \sum_G p\,(D = 0, G, B = 0, F = 0) \\ &= \frac{1}{p\,(D = 0, B = 0)} \sum_G p\,(D = 0|G, B = 0, F = 0)\, p\,(G, B = 0, F = 0) \\ &= \frac{1}{p\,(D = 0, B = 0)} \sum_G p\,(D = 0|G)\, p\,(G, B = 0, F = 0) \\ &= \frac{p\,(B = 0)\, p\,(F = 0)}{p\,(D = 0, B = 0)} \sum_G p\,(D = 0|G)\, p\,(G|B = 0, F = 0). \end{aligned} \tag{10.7}$$

Γίνεται εμφανές πως ο προσδιορισμός της p_2 προϋποθέτει τον υπολογισμό της $p\,(D = 0, B = 0)$, ο οποίος με τη σειρά του προϋποθέτει τον υπολογισμό της $p\,(G = 0|B = 0)$. Ισχύει, λοιπόν

$$\begin{aligned} p\,(G = 0|B = 0) &= \sum_F p\,(G = 0, F|B = 0) = \sum_F p\,(G = 0|B = 0, F)\, p\,(F|B = 0) \\ &= \sum_F p\,(G = 0|B = 0, F)\, p\,(F) = 0.16 + 0.6 = 0.76 \end{aligned} \tag{10.8}$$

και ως εκ τούτου ισχύει επίσης $p\,(G = 1|B = 0) = 1 - p\,(G = 0|B = 0) = 0.24$. Έτσι, βρίσκει κανείς πως

$$\begin{aligned} p\,(D = 0, B = 0) &= p\,(D = 0|B = 0)\, p\,(B = 0) = p\,(B = 0) \sum_G p\,(D = 0, G|B = 0) \\ &= p\,(B = 0) \sum_G p\,(D = 0|G, B = 0)\, p\,(G|B = 0) \\ &= p\,(B = 0) \sum_G p\,(D = 0|G)\, p\,(G|B = 0) = 0.05\,(0.608 + 0.048) \\ &= 0.0328. \end{aligned} \tag{10.9}$$

Έτσι, προκύπτει τελικά πως

$$p_2 = \frac{0.05 \cdot 0.2}{0.0328}\,(0.8 \cdot 0.8 + 0.2 \cdot 0.2) = \frac{17}{82} \simeq 0.2073. \tag{10.10}$$

Ο λόγος για τον οποίο προκύπτει $p_1 > p_2$ (μάλιστα, η p_1 είναι σχεδόν διπλάσια της p_2) είναι το λεγόμενο «explaining away» [3] και συγκεκριμένα το ότι η επιπλέον πληροφορία για το γεγονός πως η μπαταρία είναι αφόρτιστη μειώνει την πιθανότητα να είναι άδειο και το ντεπόζιτο. Ο λόγος για αυτό είναι πως από μόνο του το γεγονός πως η μπαταρία είναι αφόρτιστη (το οποίο από μόνο του είναι ένα σχετικά σπάνιο γεγονός, δεδομένων των a-priori πιθανοτήτων) αρκεί για να εξηγήσει την παρατήρηση του οδηγού, επομένως η συνύπαρξη ενός επιπλέον σχετικά σπάνιου γεγονότος είναι λιγότερο πιθανή.

Αναφορες

[1] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press; 2nd edition, 2018.

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

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