

Μηχανική Μάθηση

1^η Σειρά Ασκήσεων

Ιωάννης Τσαντήλας, 03120883

el20883@mail.ntua.gr

Σημείωση: Όπου χρησιμοποιήθηκε ργθση, υπάρχει το αντίστοιχο footnote.

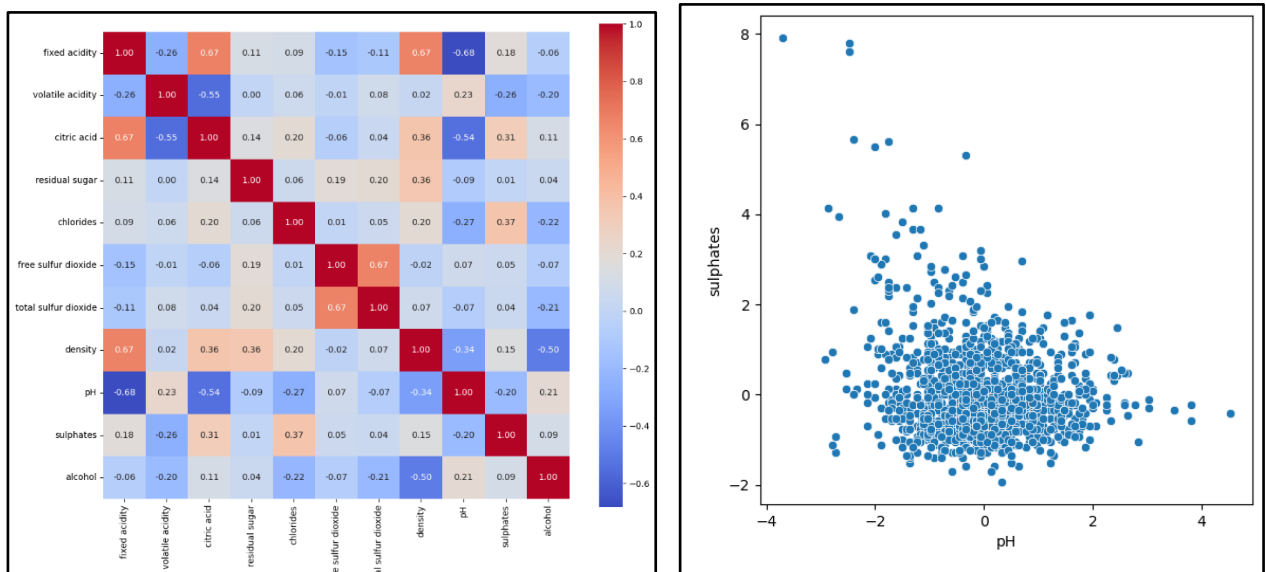
Contents

Άσκηση 1.1 (Linear and Ridge Regression)	1
Άσκηση 1.2 (Multivariate Gaussian distribution)	4
Άσκηση 1.3 (Bayes Classifier)	6
Άσκηση 1.4 (Perceptron – Multilayer Perceptron)	9
Άσκηση 1.5 Support Vector Machines – Kernels	10
Άσκηση 1.6 (Decision Trees)	11

Άσκηση 1.1 (Linear and Ridge Regression)¹

Ερώτημα (α)

Ακολουθούμε τα εξής βήματα: φορτώνουμε τα δεδομένα, αλλάζουμε το σωστό dellimeter (από ; σε ,), κανονικοποιούμε τα δεδομένα, υπολογίζουμε τον κανονικοποιημένο συντελεστή συσχέτισης και τον πίνακα συσχέτισης και τυπώνουμε τα heatmap και scatter plot:



Ερώτημα (β)

Χωρίζουμε τα δεδομένα σε εκπαίδευση και επαλήθευση, και εφαρμόζουμε την εξίσωση για γραμμική παλινδρόμηση, με X να είναι το input και y το output:

$$w = (X^T X)^{-1} X^T y$$

¹ Κάθε ερώτημα έχει το αντίστοιχο αρχείο του: "Ex_1_1a.py", "Ex_1_1b.py", "Ex_1_1c.py", "Ex_1_1d.py", "Ex_1_1e.py"

```

fixed acidity: 45.31493213424589
volatile acidity: 0.08103428345937097
citric acid: -1.946048336563198
residual sugar: -1.4132303174592855
chlorides: 0.001843873893414505
free sulfur dioxide: -0.002661056202755141
total sulfur dioxide: 0.018532432778375983
density: -0.0067354112741709925
pH: -39.30918274741026
sulphates: -0.7599285236475555
alcohol: 0.16109830270496905

```

Ερώτημα (γ)

Η παλινδρόμηση Ridge εισάγει έναν όρο κανονικοποίησης στη συνάρτηση κόστους, ο οποίος συμβάλλει στην αποφυγή υπερβολικής προσαρμογής. Η εξίσωση για τον υπολογισμό των βαρών στην παλινδρόμηση Ridge, με X το input, y το output, λ παράμετρος κανονικοποίησης και I ο πίνακας ταυτότητας, είναι:

$$w = (X^T X + \lambda I)^{-1} X^T y$$

Παρατηρούμε πως όσο αυξάνετε το λ , η απόλυτη τιμή των βαρών μειώνεται, το οποίο αιτιολογείται από την Ridge, αφού μειώνει τα βάρη για να αποφύγει την υπερβολική προσαρμογή.

For $\lambda=10$:

```

Bias (Intercept): 5.2500
fixed acidity: 0.0779
volatile acidity: -0.2577
citric acid: -0.1889
residual sugar: 0.0076
chlorides: 0.0016
free sulfur dioxide: 0.1238
total sulfur dioxide: -0.1774
density: -0.0382
pH: -0.0837
sulphates: 0.0413
alcohol: 0.1240

```

For $\lambda=100$:

```

Bias (Intercept): 5.2500
fixed acidity: 0.0309
volatile acidity: -0.1020
citric acid: -0.0462
residual sugar: 0.0033
chlorides: -0.0031
free sulfur dioxide: 0.0200
total sulfur dioxide: -0.0572
density: -0.0099
pH: -0.0290
sulphates: 0.0186
alcohol: 0.0425

```

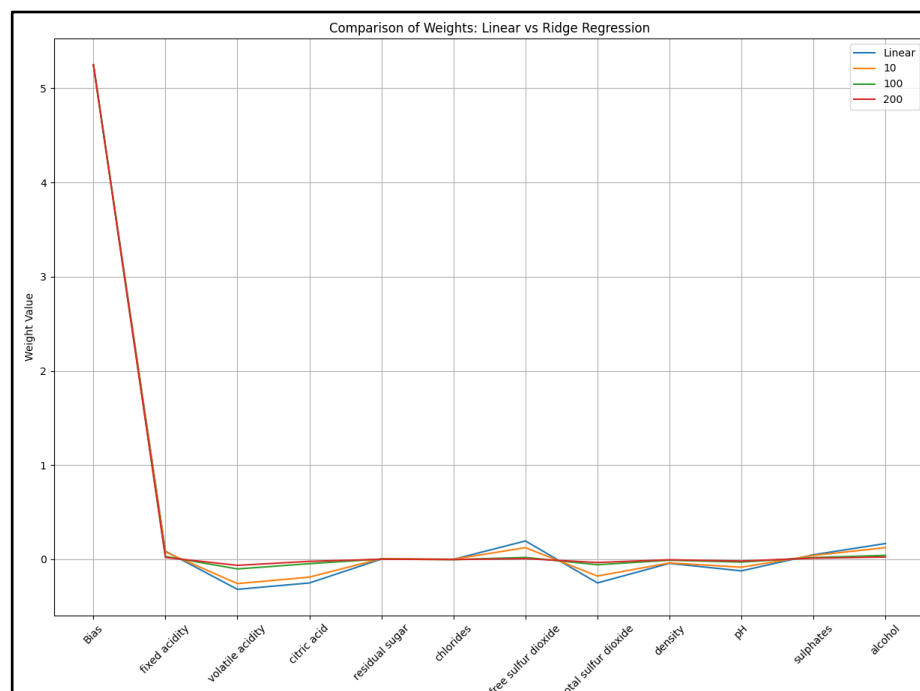
For $\lambda=200$:

```

Bias (Intercept): 5.2500
fixed acidity: 0.0199
volatile acidity: -0.0637
citric acid: -0.0217
residual sugar: 0.0014
chlorides: -0.0023
free sulfur dioxide: 0.0076
total sulfur dioxide: -0.0338
density: -0.0044
pH: -0.0179
sulphates: 0.0122
alcohol: 0.0248

```

Ερώτημα (δ)



Ερώτημα (ε)

Θυμίζουμε πως:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Με y_i η πραγματική τιμή, \hat{y}_i η προβλεπόμενη τιμή από το μοντέλο και n το πλήθος των παρατηρήσεων.

```
RMSE for Linear Regression on training set: 0.5650814403773798
RMSE for Linear Regression on test set: 0.597839593404858

RMSE for Ridge Regression on training and test sets for each lambda:
For lambda=10:
  - Training set: 0.5702608388054518
  - Test set: 0.5311592226199388

For lambda=100:
  - Training set: 0.6131819183852613
  - Test set: 0.4899314154781816

For lambda=200:
  - Training set: 0.6275839015699406
  - Test set: 0.5047256022504643
```

Παρατηρούμε πως για $\lambda=100$ έχω το χαμηλότερο RMSE για την εκπαίδευση, το οποίο υποδηλώνει πως δεν πρέπει να το αυξάνουμε άσκοπα, αλλά να βρούμε μία χρυσή τομή για να πετύχουμε το καλύτερο RMSE.

Άσκηση 1.2 (Multivariate Gaussian distribution)

Ερώτημα (α)

Η μέση τιμή της υπό συνθήκης $\mu_{x_1|x_2}$ είναι η αναμενόμενη τιμή της x_1 , θεωρώντας πως $x_2 = \alpha$. Γνωρίζουμε από την κοινή κατανομή Gauss ότι οι x_1 και x_2 είναι γραμμικά εξαρτημένες. Η γραμμική διακύμανση της x_1 στη x_2 μας δίνει την καλύτερη γραμμική πρόβλεψη του x_1 με βάση το x_2 . Σε όρους της αναδρομικής εξίσωσης, έχουμε:

$$x_1 = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2) + \epsilon$$

Όπου το ϵ αντιπροσωπεύει τον όρο σφάλματος ο οποίος κατανέμεται κανονικά με μηδενική μέση τιμή και διακύμανση $\sigma_{x_1|x_2}^2$. Θέτοντας $x_2 = \alpha$, εξαλείφουμε τον όρο σφάλματος για να λάβουμε τον υπό συνθήκη μέσο όρο:

$$\mu_{x_1|x_2} = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(\alpha - \mu_2)$$

Η διακύμανση του x_1 που δεν εξηγείται από το x_2 , ή η υπό συνθήκη διακύμανση $\sigma_{x_1|x_2}^2$, είναι το μέρος του σ_1^2 που δεν εξηγείται από τη συνδιακύμανση σ_{12}^2 . Αυτό μπορεί να προκύψει από τον τύπο για τη διακύμανση μιας υπό συνθήκη κατανομής σε μια διμεταβλητή κανονική:

$$\sigma_{x_1|x_2}^2 = \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}$$

Αυτός ο τύπος αντιπροσωπεύει τη διακύμανση του όρου σφάλματος ϵ στην παραπάνω αναδρομική εξίσωση.

Ερώτημα (β)

Για να βρούμε την υπό συνθήκη κατανομή του (x_1, x_2) δεδομένου $x_3=1$, χωρίζουμε το μέσο διάνυσμα και τον πίνακα συνδιακύμανσης σε δύο μέρη: ένα για το (x_1, x_2) και ένα για το x_3 :

$$\mu = [\mu_{123}] = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 2 & 1 \\ 0.5 & 1 & 3 \end{bmatrix}$$

Όπου ο Σ_{11} είναι ο πίνακας συνδιακύμανσης των (x_1, x_2) , Σ_{22} είναι η διασπορά της x_3 και ο $\Sigma_{21} = \Sigma_{12}^T$ είναι ο πίνακας συνδιακύμανσης των (x_1, x_2) με την x_3 . Το υπό συνθήκη μέσο διάνυσμα $\mu_{(x_1, x_2)|x_3}$ και ο υπό συνθήκη πίνακας συνδιακύμανσης $\Sigma_{(x_1, x_2)|x_3}$ δίνονται από:

$$\mu_{(x_1, x_2)|x_3} = \mu_{12} + \Sigma_{12,3} \Sigma_{22}^{-1} (x_3 - \mu_3)$$

$$\Sigma_{(x_1, x_2)|x_3} = \Sigma_{11} - \Sigma_{12,3} \Sigma_{22}^{-1} \Sigma_{21}$$

Για $x_3 = 1$:

$$\mu_{(x_1, x_2)|x_3=1} = \begin{bmatrix} -2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \cdot \frac{1}{3} \cdot (1 - 2) = \begin{bmatrix} -\frac{7}{6} \\ \frac{1}{3} \end{bmatrix}$$

$$\Sigma_{(x_1, x_2)|x_3=1} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 2 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \cdot \frac{1}{3} \cdot [0.5 \quad 1] = \begin{bmatrix} \frac{11}{12} & \frac{57}{90} \\ \frac{57}{90} & \frac{5}{3} \end{bmatrix}$$

Ερώτημα (γ)

Ομοίως:

$$\mu = [\mu_{123}] = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 2 & 1 \\ 0.5 & 1 & 3 \end{bmatrix}$$

Όπου ο Σ_{11} είναι ο πίνακας συνδιακύμανσης των (x_1, x_3) , Σ_{22} είναι η διασπορά της x_2 και ο $\Sigma_{21} = \Sigma_{12}^T$ είναι ο πίνακας συνδιακύμανσης των (x_1, x_3) με την x_2 . Το υπό συνθήκη μέσο διάνυσμα $\mu_{(x_1, x_3)|x_2}$ και ο υπό συνθήκη πίνακας συνδιακύμανσης $\Sigma_{(x_1, x_3)|x_2}$ δίνονται από:

$$\mu_{(x_1, x_3)|x_2} = \mu_{13} + \Sigma_{13,2} \Sigma_{22}^{-1} (x_2 - \mu_2)$$

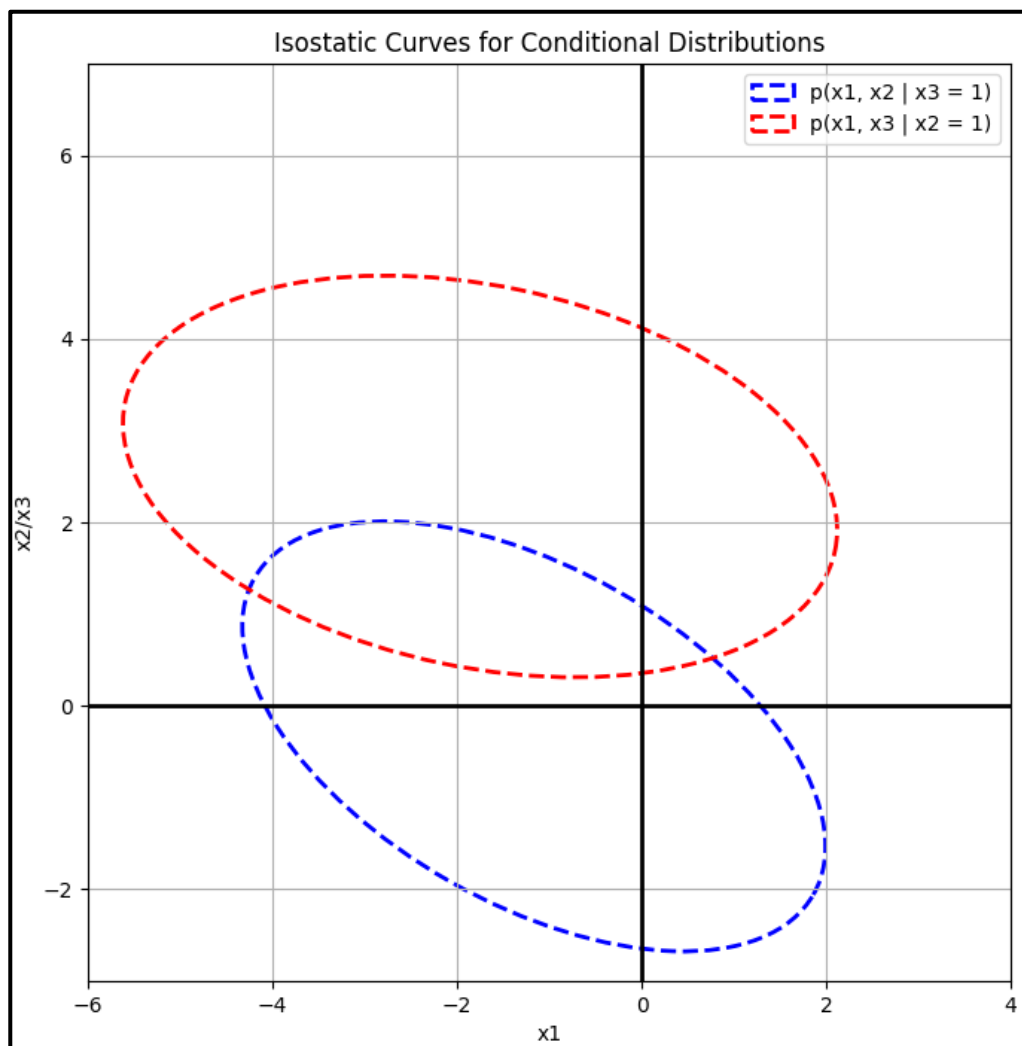
$$\Sigma_{(x_1, x_3)|x_2} = \Sigma_{11} - \Sigma_{13,2} \Sigma_{22}^{-1} \Sigma_{21}$$

Για $x_2 = 1$:

$$\mu_{(x_1, x_3)|x_2=1} = \begin{bmatrix} -2 \\ 2 \end{bmatrix} + \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \cdot \frac{1}{2} \cdot (1 - 0) = \begin{bmatrix} -1.75 \\ 2.5 \end{bmatrix}$$

$$\Sigma_{(x_1, x_3)|x_2=1} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 3 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \cdot \frac{1}{2} \cdot [0.8 \quad 0.5] = \begin{bmatrix} 0.8 & 0.375 \\ 0.1 & 0.05 \end{bmatrix}$$

Ερώτημα (δ)²



Ερώτημα (α)

Αφού $\Sigma = I$, αυτό σημαίνει ότι οι μεταβλητές είναι ανεξάρτητες και έχουν την ίδια διακύμανση. Σε έναν Bayesian ταξινομητή δύο κλάσεων, το όριο απόφασης μεταξύ των δύο κλάσεων μπορεί να βρεθεί θέτοντας τις εκ των υστέρων πιθανότητες τους ίσες μεταξύ τους, πράγμα που απλοποιείται στη σύγκριση των διακριτικών συναρτήσεων τους. Για δύο Gaussian κατανομές με ίσους πίνακες συνδιακύμανσης, η συνάρτηση διάκρισης δίνεται από:

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i) + \ln P(\omega_i)$$

Όπου μ_i είναι ο μέσος όρος της i -οστής κλάσης, Σ είναι ο πίνακας συνδιακύμανσης και $P(\omega_i)$ είναι η εκ των προτέρων πιθανότητα της κλάσης ω_i . Δεδομένου ότι δεν έχουμε πληροφορίες σχετικά με τις προκαταλήψεις, θα υποθέσουμε ότι είναι ίσες, επομένως ο όρος $\ln P(\omega_i)$ μπορεί να αγνοηθεί για το όριο απόφασης. Το όριο απόφασης βρίσκεται τότε όταν $g_1(x) = g_2(x)$, το οποίο απλοποιείται, δεδομένου ότι ο πίνακας συνδιακύμανσης είναι ο πίνακας ταυτότητας (άρα $\Sigma^{-1} = I$), σε:

$$\begin{aligned}(x - \mu_1)^T (x - \mu_1) &= (x - \mu_2)^T (x - \mu_2) \rightarrow \\ x^T x - 2\mu_1^T x + \mu_1^T \mu_1 &= x^T x - 2\mu_2^T x + \mu_2^T \mu_2 \rightarrow \\ 2(\mu_1^T - \mu_2^T) &= \mu_2^T \mu_2 - \mu_1^T \mu_1\end{aligned}$$

Και από τα δεδομένα του προβλήματος:

$$-8x_1 - 2x_2 + 1 = 0$$

Εδώ, x_1 και x_2 είναι οι συντεταγμένες στο δισδιάστατο χώρο χαρακτηριστικών R^2 . Αυτή η γραμμή θα διαχωρίσει τις δύο κλάσεις με βάση τον κανόνα απόφασης του Bayes.

Ερώτημα (β)

Τώρα έχουμε τον κοινό πίνακα συνδιακύμανσης $\hat{\Sigma}$ και για τις δύο κλάσεις, ο οποίος δεν είναι ο πίνακας ταυτότητας. Ο πίνακας συνδιακύμανσης είναι:

$$\hat{\Sigma} = \begin{bmatrix} 1 & -0.6 \\ -0.6 & 1 \end{bmatrix}$$

Για δύο κλάσεις με κοινό πίνακα συνδιακύμανσης, το όριο απόφασης μπορεί ακόμα να βρεθεί θέτοντας τις συναρτήσεις διάκρισης τους ίσες μεταξύ τους. Οι συναρτήσεις διάκρισης για Gaussian κατανομές με κοινό αλλά μη ταυτόσημο πίνακα συνδιακύμανσης είναι:

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \hat{\Sigma}^{-1}(x - \mu_i) + \ln P(\omega_i)$$

Δεδομένου ότι οι προκαταλήψεις $P(\omega_i)$ είναι ίσες, ο όρος $\ln P(\omega_i)$ δεν επηρεάζει το όριο της απόφασης και μπορεί να αγνοηθεί. Το όριο απόφασης βρίσκεται τότε όταν $g_1(x) = g_2(x)$, το οποίο απλοποιείται ως εξής:

$$\begin{aligned}(x - \mu_1)^T \hat{\Sigma}^{-1}(x - \mu_1) &= (x - \mu_2)^T \hat{\Sigma}^{-1}(x - \mu_2) \rightarrow \\ x^T \hat{\Sigma}^{-1} x - 2\mu_1^T \hat{\Sigma}^{-1} x + \mu_1^T \hat{\Sigma}^{-1} \mu_1 &= x^T \hat{\Sigma}^{-1} x - 2\mu_2^T \hat{\Sigma}^{-1} x + \mu_2^T \hat{\Sigma}^{-1} \mu_2 \rightarrow \\ -2\mu_1^T \hat{\Sigma}^{-1} x + \mu_1^T \hat{\Sigma}^{-1} \mu_1 &= -2\mu_2^T \hat{\Sigma}^{-1} x + \mu_2^T \hat{\Sigma}^{-1} \mu_2 \rightarrow \\ 2(\mu_1^T \hat{\Sigma}^{-1} - \mu_2^T \hat{\Sigma}^{-1})x &= \mu_2^T \hat{\Sigma}^{-1} \mu_2 - \mu_1^T \hat{\Sigma}^{-1} \mu_1\end{aligned}$$

Και από τα δεδομένα του προβλήματος³:

³ "Ex_1_3b_eq.py"

$$-7.1875x_1 - 5.3125x_2 + 2.65625 = 0$$

Ερώτημα (γ)

Εδώ, το λ_{ij} αντιπροσωπεύει το κόστος λανθασμένης ταξινόμησης ενός αντικειμένου από την κλάση i ως κλάση j . Στο πλαίσιο της Bayes θεωρίας αποφάσεων, αυτό σημαίνει ότι εξετάζουμε τώρα μια συνάρτηση απωλειών στη διαδικασία λήψης αποφάσεων. Για δύο κλάσεις, ο κανόνας απόφασης γίνεται:

- Ανάθεσε x στην κλάση ω_1 , αν $\lambda_{21}P(\omega_2|x) < \lambda_{12}P(\omega_1|x)$.
- Διαφορετικά, αναθέστε x στην κλάση ω_2 .

Δεδομένων των εκ των προτέρων πιθανοτήτων των κλάσεων $P(\omega_1)$ και $P(\omega_2)$ και των υπό συνθήκη πιθανοτήτων $p(x|\omega_1)$ και $p(x|\omega_2)$, ο κανόνας αυτός μπορεί να ξαναγραφεί χρησιμοποιώντας το θεώρημα του Bayes:

- Ανάθεσε x στην κλάση ω_1 , αν $\lambda_{21}p(x|\omega_2)P(\omega_2) < \lambda_{12}p(x|\omega_1)P(\omega_1)$.
- Διαφορετικά, αναθέστε x στην κλάση ω_2 .

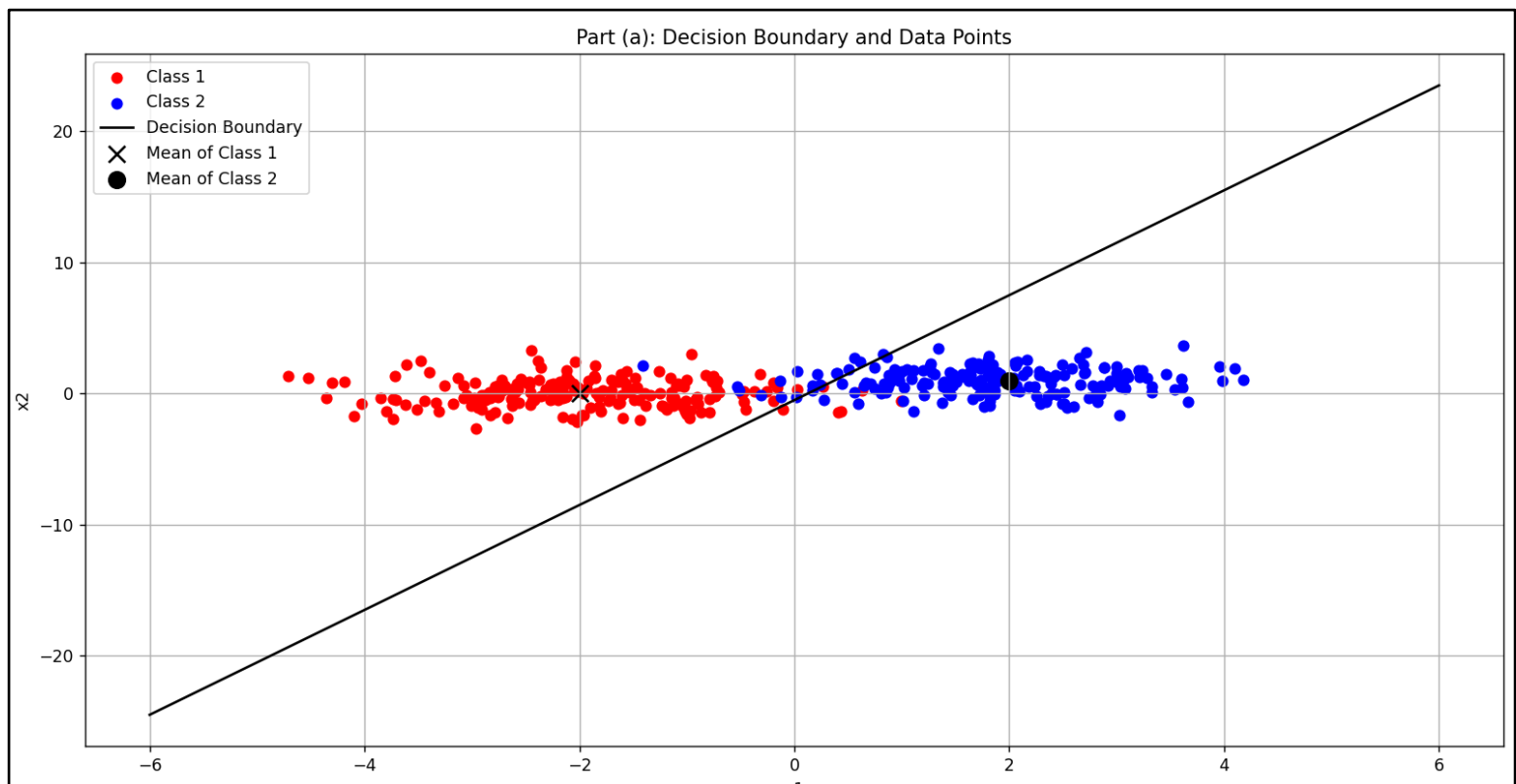
Αφού οι εκ των προτέρων πιθανότητες δεν δίνονται, υποθέτουμε ότι είναι ίσες, δηλαδή $P(\omega_1) = P(\omega_2)$. Άρα, η συνθήκη γίνεται:

$$p(x|\omega_2) < 2 \cdot p(x|\omega_1)$$

Και από τα δεδομένα του προβλήματος⁴:

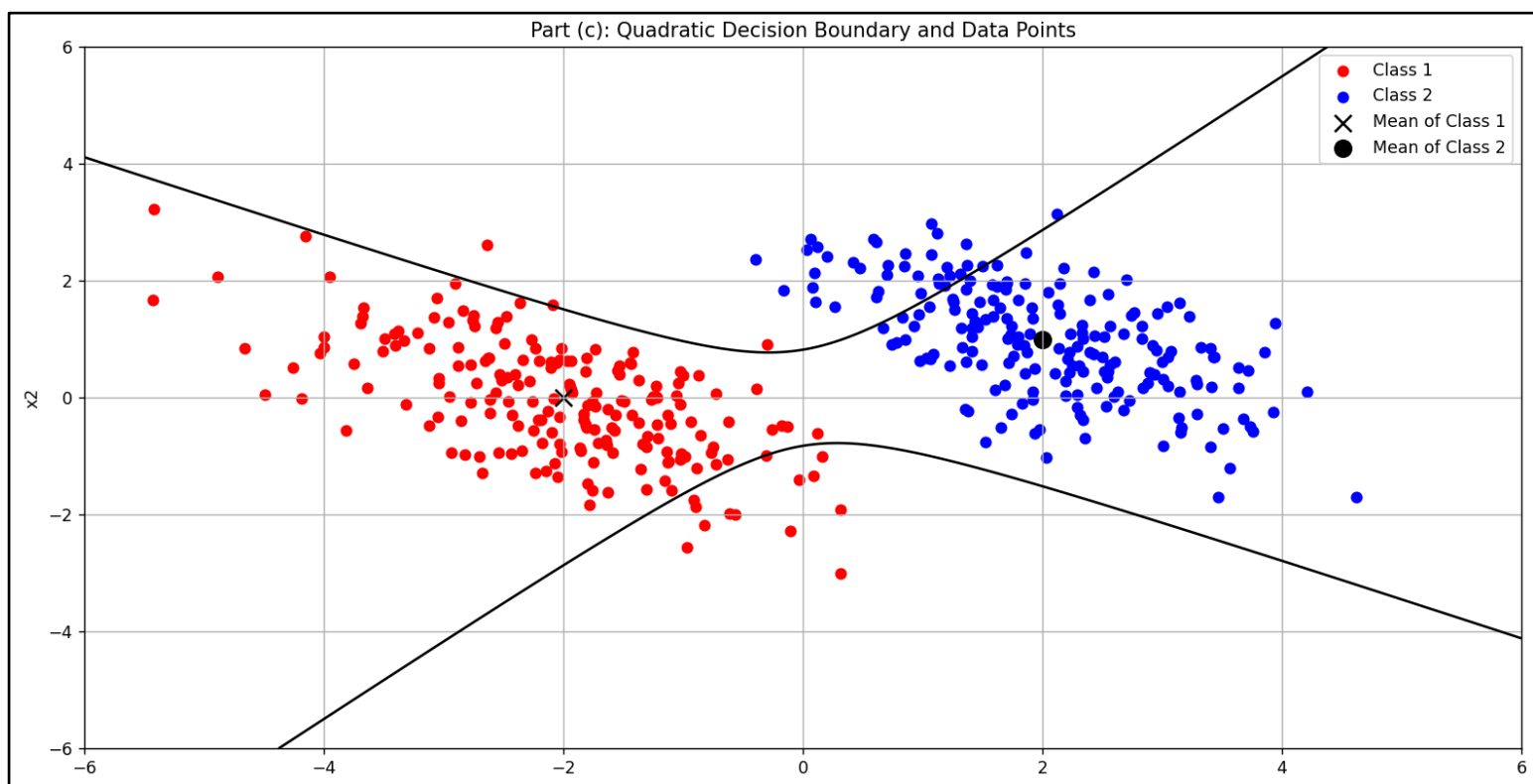
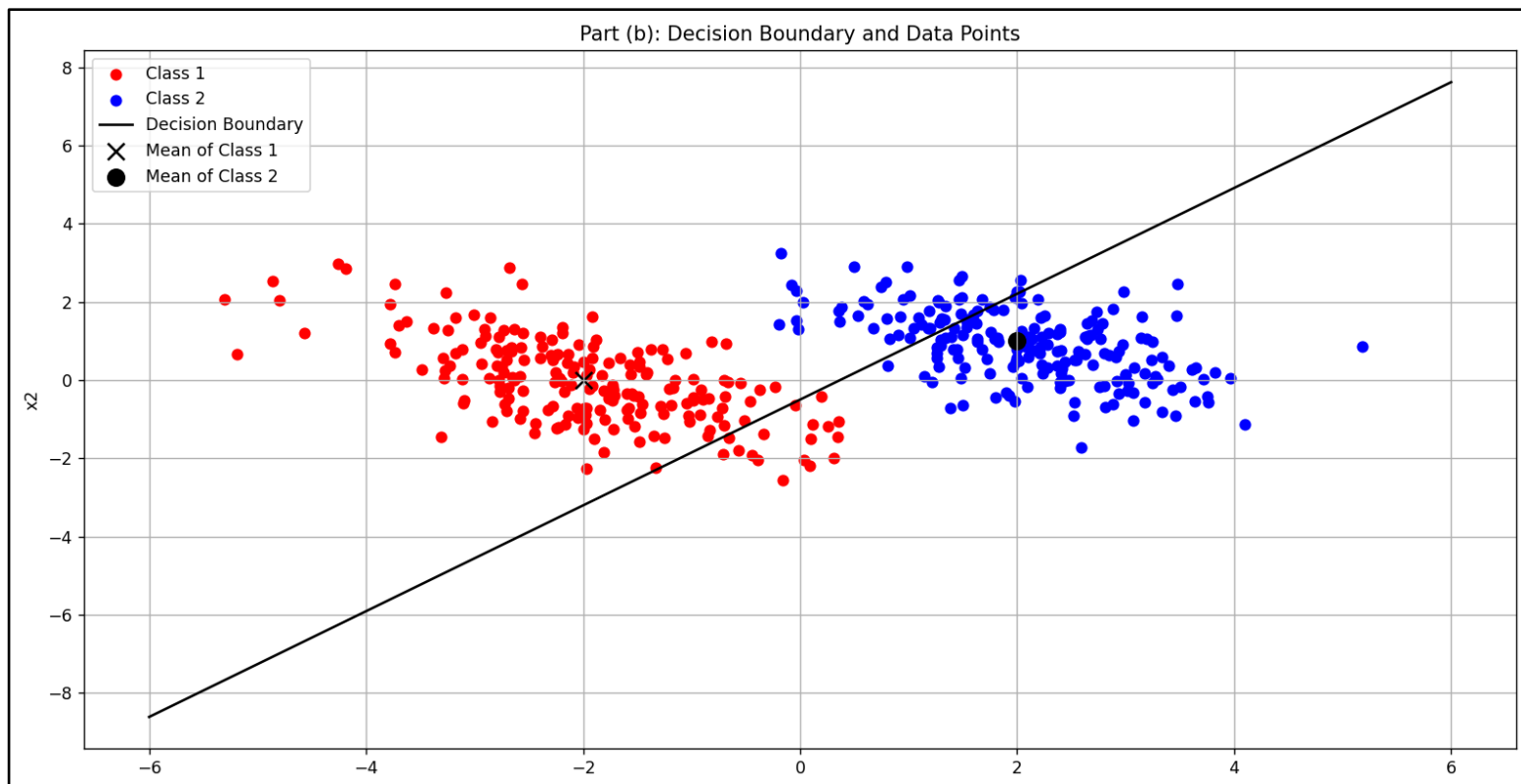
$$3.59375x_1^2 + 2.65625x_1x_2 - 3.9262x_2^2 + 2.65625 = 0$$

Ερώτημα (δ)⁵



⁴ "Ex_1_3c_eq.py"

⁵ "Ex_1_3d_plot_a.py, Ex_1_3d_plot_b.py, Ex_1_3d_plot_c.py"



Σχόλια:

- Σε όλα τα διαγράμματα, όσο μικρότερη είναι η επικάλυψη, τόσο πιο αποτελεσματικό είναι το όριο απόφασης στην ορθή ταξινόμηση των σημείων. Οι θέσεις των κέντρων των κλάσεων (μέσων) σε σχέση με τα όρια απόφασης παρέχουν πληροφορίες για το πόσο καλά αντιπροσωπεύεται κάθε κλάση από το μέσο της και πώς οι εν λόγω μέσοι επηρεάζουν τα όρια απόφασης.
- Συγκρίνοντας τα (α) και (β) δείχνει πώς το σχήμα και ο προσανατολισμός του ορίου απόφασης μπορεί να αλλάξει με βάση τον πίνακα συνδιακύμανσης των κατανομών, ακόμη και όταν οι μέσοι όροι των κλάσεων παραμένουν αμετάβλητοι.
- Το (γ) δείχνει πώς η ενσωμάτωση διαφορετικού κόστους για την εσφαλμένη ταξινόμηση μπορεί να αλλάξει σημαντικά τη φύση του ορίου απόφασης.

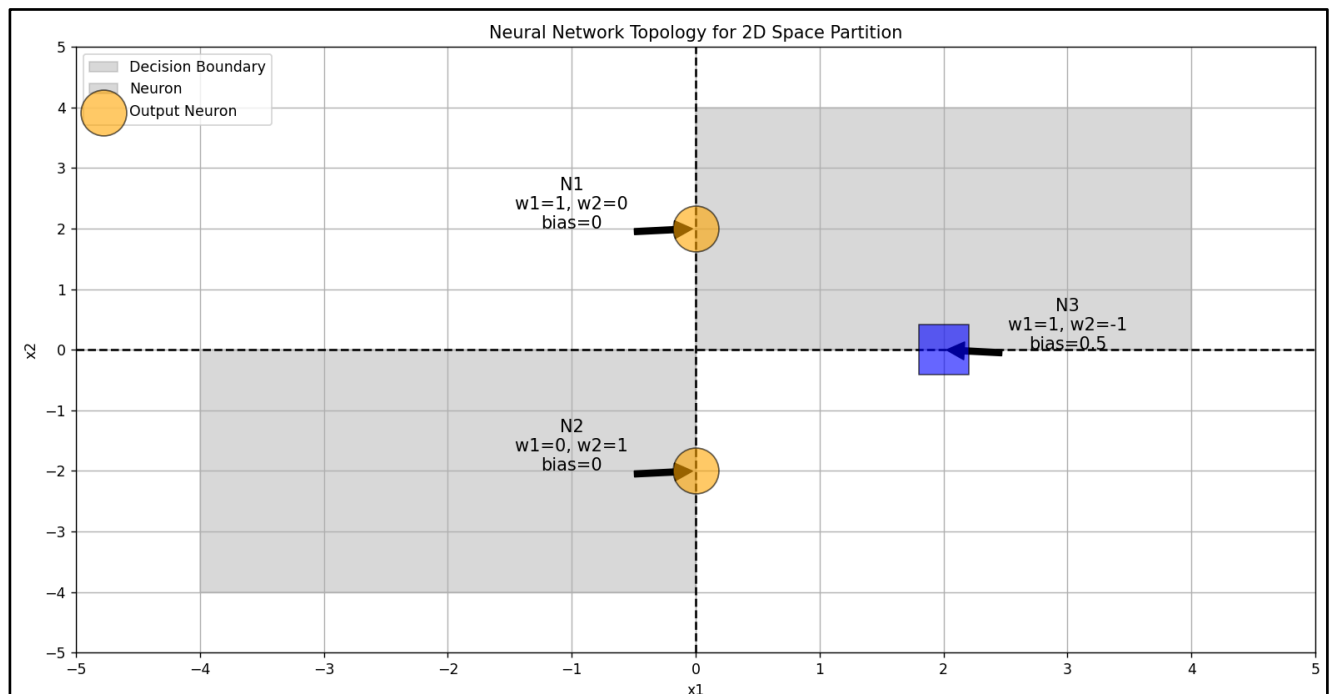
Άσκηση 1.4 (Perceptron – Multilayer Perceptron)

Ερώτημα (α)⁶

Perceptron Training Steps:						
Iteration	Sample (with bias)	Output	Classification	Weight Change	Updated Weights	
1	[1 4 3 6]	1	FP	[-1 -4 -3 -6]	[0 -4 -3 -6]	
2	[1 2 -2 3]	0	FN	[1 2 -2 3]	[1 -2 -5 -3]	
3	[1 1 0 -3]	1	TP	[0 0 0 0]	[1 -2 -5 -3]	
4	[1 4 2 3]	0	TN	[0 0 0 0]	[1 -2 -5 -3]	
5	[1 4 3 6]	0	TN	[0 0 0 0]	[1 -2 -5 -3]	
6	[1 2 -2 3]	0	FN	[1 2 -2 3]	[2 0 -7 0]	
7	[1 1 0 -3]	1	TP	[0 0 0 0]	[2 0 -7 0]	
8	[1 4 2 3]	0	TN	[0 0 0 0]	[2 0 -7 0]	
9	[1 4 3 6]	0	TN	[0 0 0 0]	[2 0 -7 0]	
10	[1 2 -2 3]	1	TP	[0 0 0 0]	[2 0 -7 0]	
11	[1 1 0 -3]	1	TP	[0 0 0 0]	[2 0 -7 0]	
12	[1 4 2 3]	0	TN	[0 0 0 0]	[2 0 -7 0]	

Ερώτημα (β)⁷

- Νεύρων 1 (N_1):** Αυτός ο νευρώνας θα δημιουργήσει ένα κατακόρυφο όριο στο $x_1 = 0$. Έτσι, θέτουμε $w_1 = 1$, $w_2 = 0$ και μια προκατάληψη 0.
- Νεύρων 2 (N_2):** Αυτός ο νευρώνας θα δημιουργήσει ένα οριζόντιο όριο στο σημείο $x_2 = 0$. Έτσι, θέτουμε $w_1 = 0$, $w_2 = 1$ και μια προκατάληψη 0.
- Νεύρων Εξόδου 3 (N_3):** Αυτός ο νευρώνας θα λάβει τις εξόδους από τα N_1 και N_2 . Εάν $N_1 = 1$ ($x_1 > 0$) και η $N_2 = 0$ ($x_2 < 0$), τότε η περιοχή είναι γκριζα και θα πρέπει να χαρακτηριστεί ως θετική. Τα βάρη από το N_1 και το N_2 στο N_3 θα πρέπει να ρυθμιστούν έτσι ώστε η έξοδος να είναι 1 μόνο όταν $N_1 = 1$ και το $N_2 = 0$. Αυτό θα μπορούσε να επιτευχθεί με ένα βάρος 1 από το N_1 , ένα βάρος -1 από το N_2 και μια μεροληψία 0,5 (για να εξασφαλιστεί ότι το ίδιο το όριο ταξινομείται ως θετικό).



⁶ "Ex_1_4a.py"

⁷ "Ex_1_4b.py"

Άσκηση 1.5 Support Vector Machines – Kernels

	SVM	Περιγραφή
α)	$C = 0.1$	Σχετικά ευθεία γραμμή με ανεκτές κάποιες λανθασμένες ταξινομήσεις (δεδομένου ότι το C είναι χαμηλό, πράγμα που σημαίνει ότι η ποινή για λανθασμένη ταξινόμηση είναι χαμηλή).
β)	$C = 10$	Ευθεία γραμμή με λιγότερες λανθασμένες ταξινομήσεις, επειδή ένα υψηλότερο C σημαίνει υψηλότερη ποινή για λανθασμένες ταξινομήσεις.
γ)	$k(u, v) = u^T v + (u^T v)^2$	Πολυωνυμικός πυρήνας βαθμού 2, ο οποίος θα καμπυλώσει το όριο απόφασης. Το όριο θα είναι μια καμπύλη που επιτρέπει κάποιο διαχωρισμό μεταξύ των κλάσεων.
δ)	$k(u, v) = e^{-0.25 \cdot \ u-v\ _2^2}$	Gaussian πυρήνας (RBF) με ορισμένη παράμετρο διασποράς (σχετίζεται με το 0,25 στον εκθέτη). Το όριο περικλείει μια κλάση εξ ολοκλήρου ή καμπυλώνει γύρω από μια κλάση στενά.
ε)	$k(u, v) = e^{-4 \cdot \ u-v\ _2^2}$	Ομοίως με (δ), αλλά με πιο στενή εξάπλωση λόγω του μεγαλύτερου αρνητικού εκθέτη, που οδηγεί σε ένα όριο απόφασης που αγκαλιάζει στενά τα σημεία δεδομένων μιας κλάσης.

Σχήμα	Περιγραφή	SVM
1)	Μη γραμμικό όριο απόφασης που δεν φαίνεται να ταιριάζει πολύ στενά με τα δεδομένα. Αυτό είναι πιο πιθανό να είναι ένας πολυωνυμικός πυρήνας.	(γ)
2)	Καμπύλο όριο που ταιριάζει αρκετά σφιχτά στα δεδομένα, πιθανώς ενδεικτικό ενός πυρήνα RBF με μικρότερη διασπορά (πιο σφιχτή προσαρμογή).	(ε)
3)	Ευθεία γραμμή με πολύ λίγες λανθασμένες ταξινομήσεις, γεγονός που υποδηλώνει υψηλότερη τιμή C για ένα γραμμικό SVM.	(β)
4)	Ευθεία γραμμή, αλλά με μερικές λανθασμένες ταξινομήσεις, γεγονός που υποδηλώνει χαμηλότερη τιμή C για ένα γραμμικό SVM.	(α)
5)	Μη γραμμικό όριο με μέτρια προσαρμογή, το οποίο θα μπορούσε να υποδηλώνει έναν πυρήνα RBF με μεγαλύτερη διασπορά (πιο χαλαρή προσαρμογή).	(δ)
6)	Καμπύλη με στενή προσαρμογή γύρω από μια κλάση, γεγονός που υποδηλώνει έναν πυρήνα RBF με στενή εξάπλωση.	(ε)

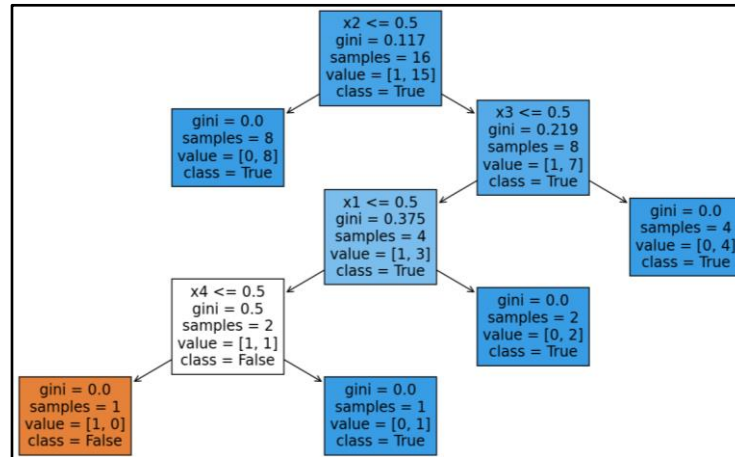
Άσκηση 1.6 (Decision Trees)

Ερώτημα (α)

Υποερώτημα (1): Για το συγκεκριμένο δέντρο, κάθε κόμβος αντιπροσωπεύει ένα από τα x_1, x_2, \dots, x_k και έχει δύο διακλαδώσεις, που αντιπροσωπεύουν το αληθές και το ψευδές του x_i . Αφού το f_n εκφράζεται με η διαζεύξεις, το βάθος του δέντρου στην απλούστερη περίπτωση, θα είναι n , με κάθε επίπεδο να αντιπροσωπεύει μία από τις διαζεύξεις. Τέλος, τα φύλλα αντιπροσωπεύουν την έξοδο του ταξινομητή (αληθής/ ψευδής). Η προσέγγιση στο 1 σημαίνει πως πρέπει να ταξινομεί σωστά τις περιπτώσεις με περιθώριο σφάλματος 1. Επομένως, το δέντρο πρέπει να καλύπτει τα περισσότερα, αλλά όχι απαραίτητα όλα, τα σενάρια που περιγράφονται από τις n διαζεύξεις. Ένα παράδειγμα δέντρου απόφασης, $n=2$:

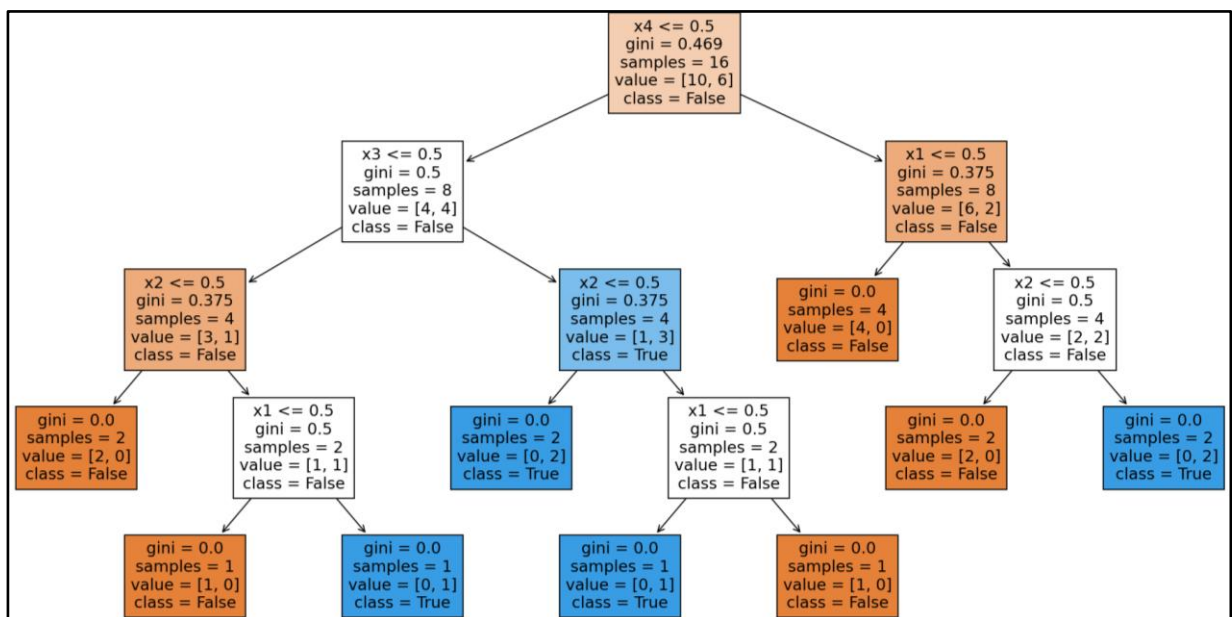
$$f_n = (x_1 \vee \neg x_2) \vee (x_3 \vee x_4)$$

Εδώ, πρώτα ελέγχουμε x_1 ή $\neg x_2$, και στη συνέχεια, με βάση αυτό, ελέγχουμε x_3 ή x_4 . Η απεικόνιση του⁸:



Υποερώτημα (2): Για το συγκεκριμένο δέντρο, κάθε κόμβος θα πρέπει να αναπαριστά μια σύζευξη (πράξη AND) δύο χαρακτηριστικών ή τις αρνήσεις τους. Δεδομένου ότι έχουμε να κάνουμε με αποκλίσεις (πράξεις XOR), πρέπει να διασφαλίσουμε ότι οι διακλαδώσεις του δέντρου αντιπροσωπεύουν σενάρια όπου ακριβώς μία από τις συζυγίες είναι αληθής, αλλά όχι και οι δύο. Ομοίως, τα φύλλα θα αντιπροσώπευαν την τελική ταξινόμηση (αληθής/ ψευδής). Τέλος, η ακρίβεια 1 σημαίνει πως το δέντρο πρέπει να ταξινομεί τέλεια όλες τις πιθανές περιπτώσεις σύμφωνα με την f_n . Ένα παράδειγμα δέντρου απόφασης, $n=2$ ⁹:

$$f_n = (x_1 \wedge x_2) \oplus (x_3 \wedge \neg x_4)$$

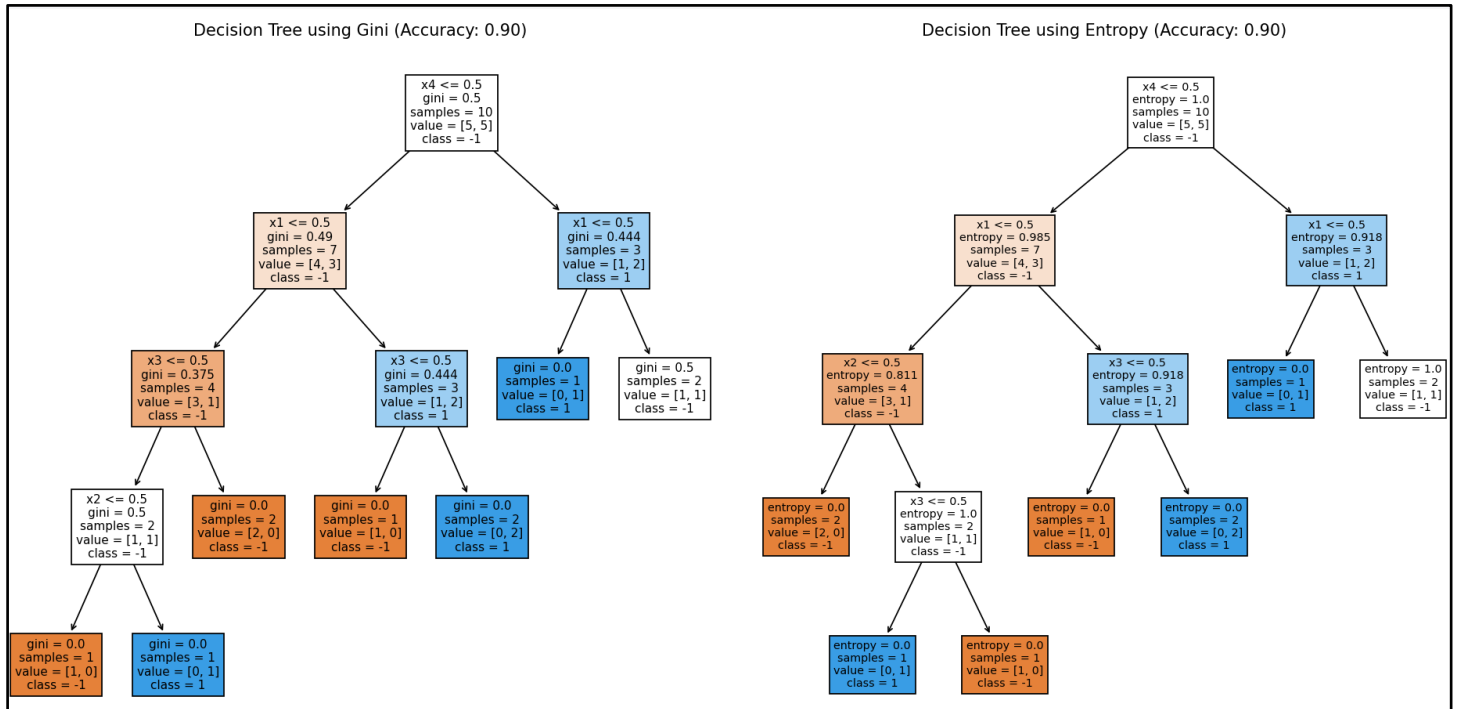


⁸ "Ex_1_6a_1.py"

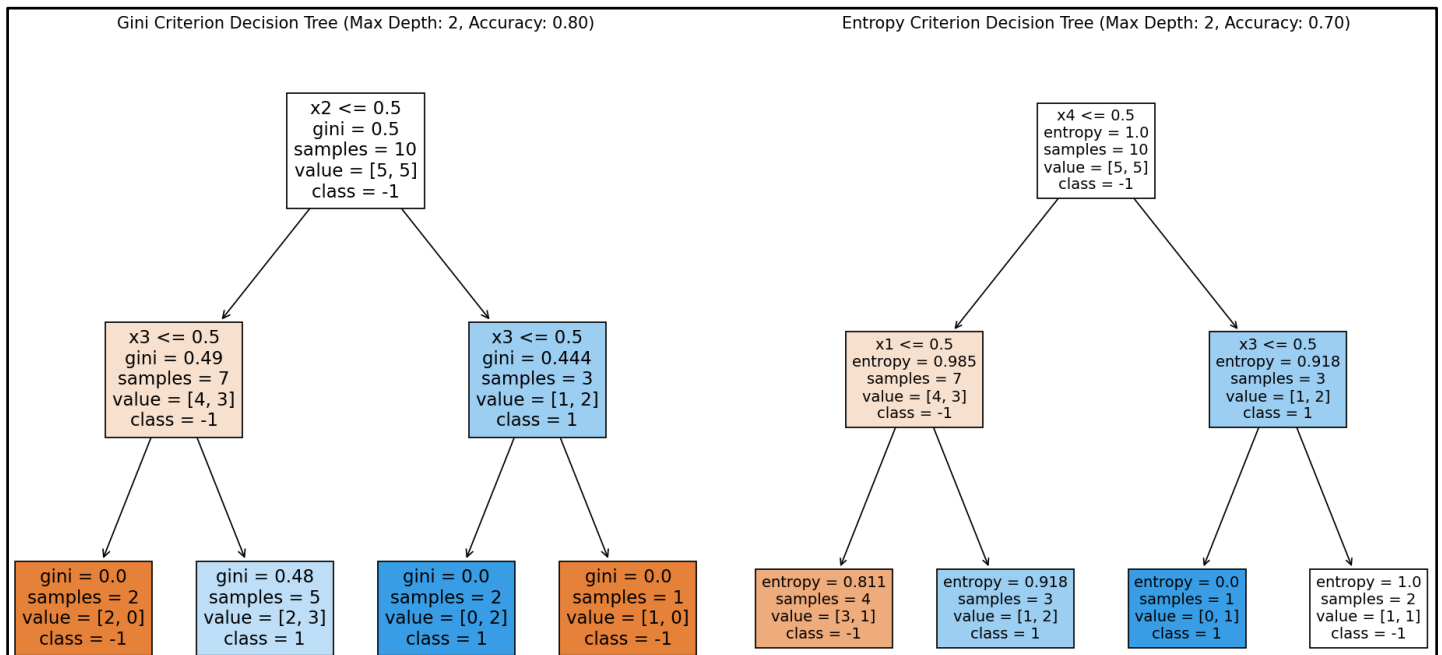
⁹ "Ex_1_6a_2.py"

Ερώτημα (β)

Υποερώτημα (1):¹⁰



Υποερώτημα (2):¹¹



Υποερώτημα (3):

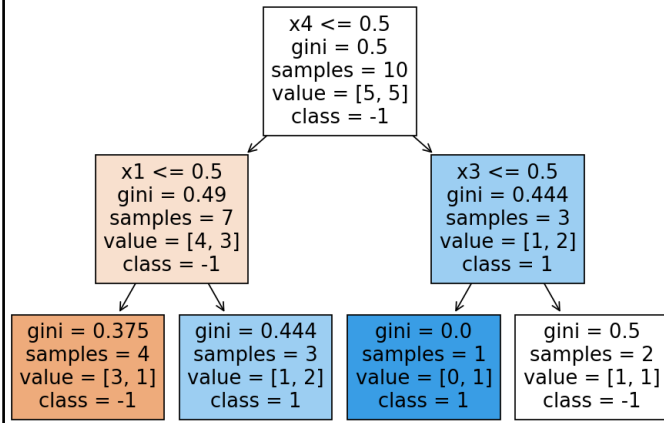
Υπάρχουν διάφορες τεχνικές για να βελτιστοποιήσουμε την αποδοτικότητα. Μεταξύ αυτών, είναι να παραλληλοποιήσουμε τις αποφάσεις σε διάφορα βήματα του δέντρου, ώστε να γίνονται ταυτόχρονα, ενώ μπορούμε να δημιουργήσουμε ένα πιο αποδοτικό dataset, π.χ. σε μορφή binary search tree. Μπορούμε να σταματήσουμε την εκτέλεση ενός βήματος, εάν προσθέτοντας ένα επιπλέον επίπεδο δεν βελτιώνει την. Τέλος, μπορούμε να ελέγχουμε συστηματικά τα βάθη των δέντρων χρησιμοποιώντας αναζήτηση πλέγματος και να διασταυρώνουμε για να ελέγξουμε την απόδοση τους σε κάθε βάθος. Τελικά, έχουμε¹²:

¹⁰ "Ex_1_6b_1.py"

¹¹ "Ex_1_6b_2.py"

¹² "Ex_1_6b_3.py"

Best Decision Tree (Depth: 2, Criterion: Gini, Accuracy: 0.30)



Best Decision Tree (Depth: 2, Criterion: Entropy, Accuracy: 0.30)

