Μηχανική Μάθηση

7ο Εξάμηνο 2022 – 2023

Assignment 1 – Solutions

Ζαρίφης Στέλιος – el20435

Email: el20435@mail.ntua.gr

Contents

Άσκηση 1.1

Προεργασία

Όπως υποδεικνύεται στην εκφώνηση, κανονικοποιούμε τα χαρακτηριστικά x_1, x_2 αφαιρώντας τη μέση τιμή τους και ύστερα διαιρώντας με την τυπική απόκλιση:

```
df['athletes'] -= df['athletes'].mean()
df['events'] -= df['events'].mean()
df['medals'] -= df['medals'].mean()

df['athletes'] /= df['athletes'].std()
df['events'] /= df['events'].std()
df['medals'] /= df['medals'].std()
```

Ερώτημα α

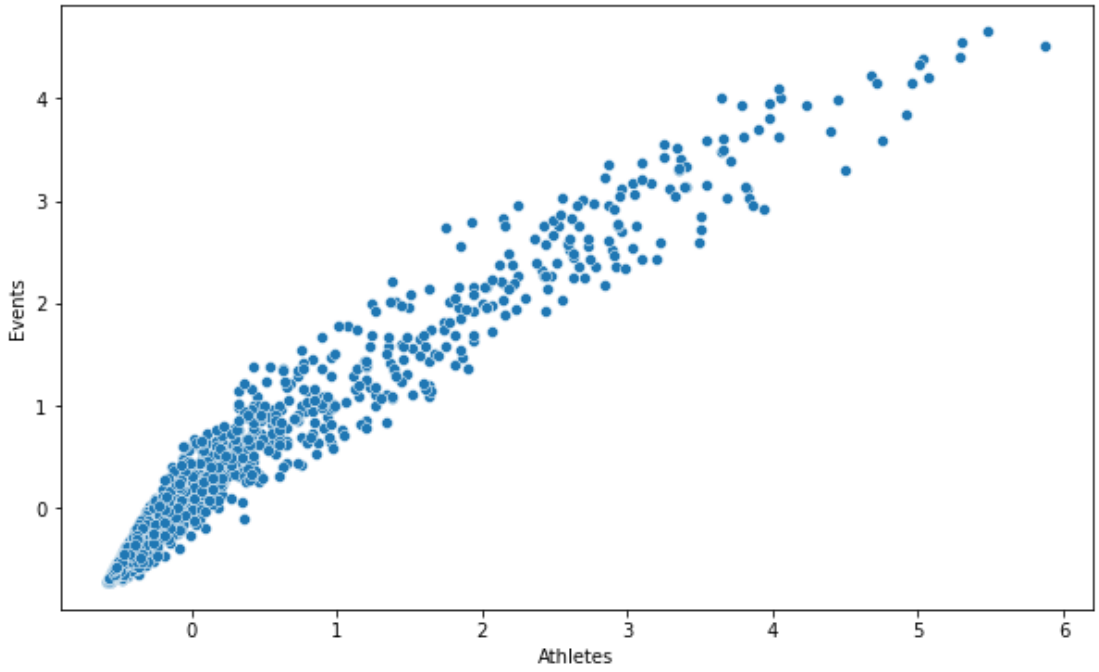
Υπολογίζουμε το συντελεστή συσχέτισης των μεταβλητών:

```
r12 = np.sum(df['athletes']*df['events']/df.count()[0])
```

Βρήκαμε:

$$r_{12} \cong 0.9761$$

Η συσχέτιση είναι πολύ μεγάλη. Όπως μπορούμε να δούμε και στο scatterplot, βλέπουμε μεγάλη ομοιότητα στις μεταβολές των x_1, x_2. Δηλαδή, κατά τον τρόπο που αυξάνεται η x_1 τείνει να αυξηθεί και η x_2 και αντίστροφα.



Ερώτημα β

Κάνουμε ένα shuffling στο dataset διότι οι εγγραφές είναι sorted και κρατάμε τις στήλες που μας ενδιαφέρουν, δηλαδή τις x_1, x_2, y. Δημιουργούμε τον πίνακα X προσθέτοντας μια στήλη με 1 στην αρχή και χωρίζουμε το dataset σε train και test set.

Είμαστε πλέον έτοιμοι για τη γραμμική παλινδρόμηση. Από τη μέθοδο ελαχιστοποίησης του αθροίσματος των τετραγωνικών σφαλμάτων έχουμε λύση:

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

Και είναι:

$$\boldsymbol{y} = \begin{bmatrix} y[0] \\ y[1] \\ \vdots \\ y[n] \end{bmatrix} \text{ και } X = \begin{bmatrix} 1 & x_1[0] & x_2[0] \\ 1 & x_1[1] & x_2[1] \\ \vdots & \vdots & \vdots \\ 1 & x_1[n] & x_2[n] \end{bmatrix}$$

Παραθέτουμε τον κώδικα στο ερώτημα (γ) ο οποίος μας δίνει βάρη για τη γραμμική παλινδρόμηση:

```
Linear Regression:
[-6.68917288e-05  1.91616987e+00 -1.10051469e+00]
```

Δηλαδή $w_{lin-Reg} = [-6.68917288e-05\,, 1.91616987e+00, -1.10051469e+00]$

Ερώτημα γ

Αντίστοιχα, για τη ridge παλινδρόμηση έχουμε λύση:

$$\boldsymbol{w} = (X^T X + \lambda I_3)^{-1} X^T \boldsymbol{y}$$

Και παίρνουμε βάρη:

```
Ridge Regression with λ = 1:
[-8.87750278e-05  1.87798835e+00 -1.06260686e+00]
Ridge Regression with λ = 10:
[-2.40248701e-04  1.60475646e+00 -7.91798893e-01]
Ridge Regression with λ = 100:
[-5.14075807e-04  8.15541622e-01 -2.52442502e-02]
```

Δηλαδή

$$w_{ridge-Reg}^{\lambda=1} = [-8.87750278e-05\ \ 1.87798835e+00\ \ -1.06260686e+00]$$

$$w_{ridge-Reg}^{\lambda=10} = [-2.40248701e-04, 1.60475646e+00, -7.91798893e-01]$$

$$w_{ridge-Reg}^{\lambda=100} = [-5.14075807e-04, 8.15541622e-01, -2.52442502e-02]$$

Ο κώδικας που χρησιμοποιήθηκε είναι:

```
Xt = X.transpose()
XtX = np.matmul(Xt, X)
XtXinv = np.linalg.inv(XtX)
Xty = np.matmul(Xt, y)
I = np.identity(3)
XtXplusIinv = np.linalg.inv(XtX + I)
XtXplus10Iinv = np.linalg.inv(XtX + 10*I)
XtXplus100Iinv = np.linalg.inv(XtX + 100*I)

w = {'Linear Regression'                   : np.matmul(XtXinv, Xty),
     'Ridge Regression with \u03bb = 1'    : np.matmul(XtXplusIinv,
Xty),
     'Ridge Regression with \u03bb = 10'   : np.matmul(XtXplus10Iinv,
Xty),
     'Ridge Regression with \u03bb = 100' : np.matmul(XtXplus100Iinv,
Xty)}

for key, val in w.items():
    print(key + ":")
    print(val)
```

Ερώτημα δ

Από τα βάρη που βρήκαμε στα ερωτήματα (β) και (γ), κάνουμε προβλέψεις στα train και test set και χρησιμοποιώντας το target των train και test set εκτιμάμε τα σφάλματα εκπαίδευσης και επαλήθευσης $RMSE$:

```
pred_train = {}
for reg, ws in w.items():
    pred_train[reg] = np.matmul(ws.transpose(), X.transpose())

pred_test = {}
for reg, ws in w.items():
    pred_test[reg] = np.matmul(ws.transpose(), X_test.transpose())



RMSE = {}
for reg, pred in pred_train.items():
    RMSE[reg] = np.sqrt(np.sum((pred -
y.transpose())**2)/X.count()[0])

RMSE = pd.DataFrame.from_dict(RMSE, orient = 'index', columns =
['RMSE'])
RMSE.sort_values(by = ['RMSE'])

RMSE_test = {}
for reg, pred in pred_test.items():
    RMSE_test[reg] = np.sqrt(np.sum((pred -
y_test.transpose())**2)/X_test.count()[0])

RMSE_test = pd.DataFrame.from_dict(RMSE_test, orient = 'index',
columns = ['RMSE_test'])
RMSE_test.sort_values(by = ['RMSE_test'])
```

Και λαμβάνουμε τα εξής αποτελέσματα:

	RMSE_train		RMSE_test
Linear Regression	0.492617	Linear Regression	0.485715
Ridge Regression with λ = 1	0.492687	Ridge Regression with λ = 1	0.485807
Ridge Regression with λ = 10	0.497283	Ridge Regression with λ = 10	0.486216
Ridge Regression with λ = 100	0.54779	Ridge Regression with λ = 100	0.520963

Επιλέγουμε $\lambda = 1$ ή $\lambda = 10$ αφού για αυτές τις τιμές το $RMSE$ είναι μικρότερο (η διαφορά τους είναι αμελητέα).

Συμπέρασμα: Γενικά, η αύξηση της τιμής του λ αυξάνει την ισχύ κανονικοποίησης του μοντέλου, γεγονός που μπορεί να οδηγήσει σε μείωση του overfitting και να βελτιώσει την απόδοση γενίκευσης του μοντέλου. Ωστόσο, εάν η τιμή του λ είναι πολύ μεγάλη, μπορεί επίσης να οδηγήσει σε underfitting, όπου το μοντέλο δεν είναι σε θέση να καταγράψει τα υποκείμενα μοτίβα στα δεδομένα.

Ο συνολικός κώδικας που χρησιμοποιήθηκε, παρατίθεται εδώ:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv('olympic_teams.csv')

df['athletes'] -= df['athletes'].mean()
df['events'] -= df['events'].mean()
df['medals'] -= df['medals'].mean()

df['athletes'] /= df['athletes'].std()
df['events'] /= df['events'].std()
df['medals'] /= df['medals'].std()

r12 = np.sum(df['athletes']*df['events']/df.count()[0])

Xt = X.transpose()
XtX = np.matmul(Xt, X)
XtXinv = np.linalg.inv(XtX)
Xty = np.matmul(Xt, y)
I = np.identity(3)
XtXplusIinv = np.linalg.inv(XtX + I)
XtXplus10Iinv = np.linalg.inv(XtX + 10*I)
XtXplus100Iinv = np.linalg.inv(XtX + 100*I)

w = {'Linear Regression'                   : np.matmul(XtXinv, Xty),
     'Ridge Regression with \u03bb = 1'    : np.matmul(XtXplusIinv,
Xty),
     'Ridge Regression with \u03bb = 10'   : np.matmul(XtXplus10Iinv,
Xty),
     'Ridge Regression with \u03bb = 100' : np.matmul(XtXplus100Iinv,
Xty)}

for key, val in w.items():
```

```
    print(key + ":")
    print(val)

pred_train = {}
for reg, ws in w.items():
    pred_train[reg] = np.matmul(ws.transpose(), X.transpose())

pred_test = {}
for reg, ws in w.items():
    pred_test[reg] = np.matmul(ws.transpose(), X_test.transpose())

RMSE = {}
for reg, pred in pred_train.items():
    RMSE[reg] = np.sqrt(np.sum((pred -
y.transpose())**2)/X.count()[0])

RMSE = pd.DataFrame.from_dict(RMSE, orient = 'index', columns =
['RMSE'])
RMSE.sort_values(by = ['RMSE'])

RMSE_test = {}
for reg, pred in pred_test.items():
    RMSE_test[reg] = np.sqrt(np.sum((pred -
y_test.transpose())**2)/X_test.count()[0])

RMSE_test = pd.DataFrame.from_dict(RMSE_test, orient = 'index',
columns = ['RMSE_test'])
RMSE_test.sort_values(by = ['RMSE_test'])
```

Και μπορεί να βρεθεί στο link επειδή υλοποιήθηκε σε Python Jupyter μέσω Google Colab.

Άσκηση 1.2

Ερώτημα α

Θεωρούμε την τυχαία μεταβλητή x που ακολουθεί Gaussian Distribution με παραμέτρους μ, σ^2: $x \sim \mathcal{N}(\mu, \sigma^2)$.

Η συνάρτηση κατανομής πιθανότητας της x είναι:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

Θα αποδείξουμε ότι η μεταβλητότητα της x ισούται με την παράμετρο σ^2. Έχουμε:

$$\sigma_x^2 = Var[x] = \mathbb{E}[(x-\mu)^2] = \int_{\Re} (x-\mu)^2 f_x(x) dx =$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

Αλλάζουμε μεταβλητή:

$$y = \frac{x-\mu}{\sigma\sqrt{2}}, dx = \sigma\sqrt{2} dy$$

Και καταλήγουμε:

$$\sigma_x^2 = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2} dy$$

Αλλάζουμε για μία ακόμα φορά μεταβλητή:

$$y^2 = x, dy = d\left(\sqrt{x}\right) = \frac{1}{2\sqrt{x}} dx$$

Και τελικά είναι:

$$\sigma_x^2 = \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} x e^{-x} \frac{1}{2\sqrt{x}} dx$$

Παρατηρούμε πως η συνάρτηση του ολοκληρώματος έχει άρτια συμμετρία οπότε το ολοκλήρωμα ισούται με 2 φορές το ολοκλήρωμα στην περιοχή $\mathbb{R}^+$ και μπορούμε να γράψουμε:

$$\sigma_x^2 = \frac{2\sigma^2}{\sqrt{\pi}} 2 \int_0^{\infty} x e^{-x} \frac{1}{2\sqrt{x}} dx = \frac{2\sigma^2}{\sqrt{\pi}} \int_0^{\infty} x^{1/2} e^{-x} dx \quad (1)$$

Τέλος, θεωρούμε τη συνάρτηση Γ η οποία ορίζεται ως εξής:

$$\Gamma(x) = \int_0^{\infty} z^{x-1} e^{-z} dz$$

Και έχει την ιδιότητα:

$$\Gamma(\alpha+1)=\alpha\Gamma(\alpha), \Gamma(1/2)=\sqrt{\pi}$$

Από την (1) βλέπουμε ότι:

$$\sigma_x^2=\frac{2\sigma^2}{\sqrt{\pi}}\int_0^\infty x^{1/2}e^{-x}dx=\frac{2\sigma^2}{\sqrt{\pi}}\int_0^\infty x^{3/2-1}e^{-x}dx=\frac{2\sigma^2}{\sqrt{\pi}}\Gamma(3/2)=\frac{2\sigma^2}{\sqrt{\pi}}\frac{\Gamma(1/2)}{2}=\frac{2\sigma^2}{\sqrt{\pi}}\frac{\sqrt{\pi}}{2}$$

Άρα:

$$\boxed{\sigma_x^2=\sigma^2}$$

Ερώτημα β

Θεωρούμε τυχαίο διάνυσμα $x\in\mathbb{R}^2$ που ακολουθεί την κανονική κατανομή, $x\sim\mathcal{N}(x|\boldsymbol{\mu},\Sigma)$ με

$$\boldsymbol{\mu}=\begin{bmatrix}1\\1\end{bmatrix}, \Sigma=\begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}$$

Το $\boldsymbol{x}=(x_1,x_2)$ θα έχει από κοινού συνάρτηση κατανομής πιθανότητας:

$$f_x(x_1,x_2)=f_x(\boldsymbol{x})=\frac{1}{2\pi|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Οι ζητούμενες ισοσταθμικές καμπύλες της κατανομής του x είναι οι Γεωμετρικοί Τόποι των σημείων του επιπέδου (x_1,x_2) για τα οποία η από κοινού συνάρτηση κατανομής πιθανότητας έχει σταθερή τιμή. Δηλαδή εκείνα τα $\boldsymbol{x}=(x_1,x_2)$ για τα οποία:

$$f_x(\boldsymbol{x})=\bar{c}\Rightarrow\frac{1}{2\pi|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]=\bar{c}\Rightarrow(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=c\ (1)$$

Θα διαγωνιοποιήσουμε τον Σ:

$$\Sigma=U^T\Lambda U$$

Επειδή ο Σ είναι συμμετρικός (και άρα έχει πραγματικές ιδιοτιμές), επιλέγουμε $U=[\vec{u}_1,\vec{u}_2]$ όπου $\vec{u}_1,\vec{u}_2$ τα ιδιοδιανύσματα του Σ που σχηματίζουν μια ορθοκανονική βάση και συνεπώς $\Lambda=diag\{\lambda_1,\lambda_2\}$. Θεωρούμε επίσης ότι ο Σ είναι αντιστρέψιμος (αλλιώς δεν ορίζεται η από κοινού συνάρτηση κατανομής πιθανότητας).

Ας υπολογίσουμε τα ορθοκανονικά ιδιοδιανύσματα του Σ:

Ιδιοτιμές του Σ:

$$\det(\lambda I - \Sigma) = 0 \Rightarrow \begin{vmatrix} \lambda - 2 & -1 \\ -1 & \lambda - 2 \end{vmatrix} = 0 \Rightarrow \lambda^2 - 4\lambda + 3 = 0 \Rightarrow (\lambda - 1)(\lambda - 3) = 0$$

$$\lambda_i\{\Sigma\} = \{1, 3\}$$

Ιδιοδιανύσματα του Σ:

Για $\lambda = 1$:

$$\Sigma\vec{u}_1 = \lambda\vec{u}_1 \Rightarrow \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = 1\begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} \Rightarrow \begin{cases} 2u_{11} + u_{12} = u_{11} \\ u_{11} + 2u_{12} = u_{12} \end{cases} \Rightarrow \begin{cases} u_{11} = -u_{12} \\ u_{11} = -u_{12} \end{cases}$$

Επιλέγουμε (αυθαίρετα) $u_{11} = 1/\sqrt{2}$ άρα $u_{12} = -1/\sqrt{2}$.

Η επιλογή έγινε ώστε να καταλήξουμε σε ορθοκανονική βάση γιατί με τέτοιο $\vec{u}_1$ είναι $|\vec{u}_1| = 1$, όπως πρέπει σε μια τέτοια βάση.

Για $\lambda = 3$:

$$\Sigma\vec{u}_2 = \lambda\vec{u}_2 \Rightarrow \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} = 3\begin{bmatrix} u_{21} \\ u_{22} \end{bmatrix} \Rightarrow \begin{cases} 2u_{21} + u_{22} = 3u_{21} \\ u_{21} + 2u_{22} = 3u_{22} \end{cases} \Rightarrow \begin{cases} u_{21} = u_{22} \\ u_{21} = u_{22} \end{cases}$$

Επιλέγουμε με όμοιο τρόπο $u_{21} = 1/\sqrt{2}$ άρα $u_{22} = 1/\sqrt{2}$.

Δείξαμε λοιπόν ότι:

$$\Sigma = U^T \Lambda U = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

Και η (1) γίνεται:

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c \Rightarrow (\boldsymbol{x} - \boldsymbol{\mu})^T (U^T \Lambda U)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c \Rightarrow$$

$$\Rightarrow (\boldsymbol{x} - \boldsymbol{\mu})^T U^{-1} \Lambda^{-1} (U^T)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c \ (2)$$

Όμως φροντίσαμε ο U να είναι ορθοκανονικός, συνεπώς έχει την ιδιότητα $UU^T = U^T U = I$.

Συνεπώς από τη (2) έχουμε:

$$(\boldsymbol{x} - \boldsymbol{\mu})^T U^T \Lambda^{-1} U(\boldsymbol{x} - \boldsymbol{\mu}) = c \Rightarrow [U(\boldsymbol{x} - \boldsymbol{\mu})]^T \Lambda^{-1} [U(\boldsymbol{x} - \boldsymbol{\mu})] = c \ (3)$$

Ορίζουμε $\boldsymbol{y} = U(\boldsymbol{x} - \boldsymbol{\mu})$ και η μορφή της (3) απλοποιείται σε:

$$\boldsymbol{y}^T \Lambda^{-1} \boldsymbol{y} = c = [y_1 \quad y_2] \frac{1}{\lambda_1 \lambda_2}\begin{bmatrix} \lambda_2 & 0 \\ 0 & \lambda_1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = c \Rightarrow \frac{1}{\lambda_1 \lambda_2}(\lambda_2 y_1^2 + \lambda_1 y_2^2) = c \Rightarrow$$

$$\Rightarrow \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = c \ (4)$$

Στα αλήθεια δε χρειαζόταν αυτή η ανάλυση καθώς η μορφή των ισοσταθμικών καμπυλών εξαγάγεται εύκολα με απλή αντικατάσταση των τιμών των $\boldsymbol{\mu}, \Sigma$ στη σχέση:

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c \Rightarrow \cdots \Rightarrow x_1^2 + x_1x_2 + x_2^2 - 2x_1 - 2x_2 + 1 = c/2$$

Αλλά αξίζει να παρουσιάσουμε μια πιο γεωμετρική ερμηνεία:

Από την (4) βλέπουμε ότι στο επίπεδο (y_1, y_2) που ορίζουν τα components του $\boldsymbol{y}$ έχουμε μια έλλειψη με μήκος $2\sqrt{\lambda_1 c}$ και πλάτος $2\sqrt{\lambda_2 c}$, όπως φαίνεται και εδώ:

$$\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = c \Rightarrow \frac{y_1^2}{\left(\sqrt{\lambda_1 c}\right)^2} + \frac{y_2^2}{\left(\sqrt{\lambda_2 c}\right)^2} = 1$$

Όμως μας ενδιαφέρει να εκφράσουμε τον Γεωμετρικό Τόπο στο επίπεδο (x_1, x_2). Θα εξετάσουμε πώς από το επίπεδο (x_1, x_2) πήγαμε στο (y_1, y_2). Ο μετασχηματισμός από τα προηγούμενα βήματα είναι:

$$\boldsymbol{y} = U(\boldsymbol{x}-\boldsymbol{\mu})$$

Αυτό σημαίνει ότι έχουμε μια μετατόπιση του συστήματος (x_1, x_2) στο σημείο $\boldsymbol{\mu} = (1, 1)$ και ύστερα τον γραμμικό μετασχηματισμό (scaling and rotation) που προκαλεί ο πίνακας U:

Η κάθε στήλη του U σηματοδοτεί πού θα καταλήξουν τα μοναδιαία διανύσματα που κάνουν span το επίπεδο (x_1, x_2) (δηλαδή τα $\hat{\imath} = [1 \quad 0]^T, \hat{\jmath} = [0 \quad 1]^T$). Δηλαδή το $\hat{\imath}$ θα μετασχηματιστεί σε $\vec{u}_1 = \left[1/\sqrt{2} \quad -1/\sqrt{2}\right]^T$ και το $\hat{\jmath}$ σε $\vec{u}_2 = \left[1/\sqrt{2} \quad 1/\sqrt{2}\right]^T$. Συνεπώς, επειδή ο U είναι ορθοκανονικός δεν παρατηρούμε scaling στο μήκος των διανυσμάτων βάσης, αλλά μόνο στροφή και μάλιστα γωνίας θ όπου:

$$U = \begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix} = \begin{bmatrix}1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2}\end{bmatrix} \Rightarrow \begin{cases}\cos\theta = 1/\sqrt{2} \\ \sin\theta = 1/\sqrt{2}\end{cases} \Rightarrow \theta = \pi/4\ rad$$

Είδαμε τελικά ότι ο ζητούμενος Γεωμετρικός Τόπος είναι ελλείψεις μήκους $2\sqrt{\lambda_1 c}$ και πλάτους $2\sqrt{\lambda_2 c}$ στο σύστημα (y_1, y_2) το οποίο σε σχέση με το (x_1, x_2) έχει κέντρο το $\boldsymbol{\mu} = (1, 1)$ και έχει περιστραφεί $\pi/4\ rad$.

Ερώτημα γ

Θεωρούμε τις ανεξάρτητες παρατηρήσεις $x_1, x_2, \dots, x_N \in \mathbb{R}^l$ που ακολουθούν κανονική κατανομή $\mathcal{N}(\boldsymbol{x_n}|, \Sigma)$, δηλαδή άγνωστης παραμέτρου $\boldsymbol{\mu}$. Για το ζητούμενο ξεκινάμε με τον ορισμό του εκτιμητή μεγίστης πιθανοφάνειας:

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{ML}} = \arg\max_{\mu} p(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N}; \boldsymbol{\mu})$$

Επειδή όμως θεωρήσαμε στατιστικά ανεξάρτητες παραηρήσεις είναι:

$$p(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N}; \boldsymbol{\mu}) = \prod_{n=1}^{N} p(\boldsymbol{x_n}; \boldsymbol{\mu})$$

Ενώ ισχύει λόγω $\mathcal{N}(\boldsymbol{x_n}|, \Sigma)$ ότι:

$$p(\boldsymbol{x_n}; \boldsymbol{\mu}) = \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right], \forall n = 1, 2, \dots, N$$

Λόγω του εκθετικού, μπορούμε να λογαριθμήσουμε την p και να αναζητήσουμε τη μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας (αφού $\ln(\cdot)$ γνησίως αύξουσα):

$$L(\boldsymbol{\mu}) = \ln \prod_{n=1}^{N} p(\boldsymbol{x_n}; \boldsymbol{\mu}) = \ln \prod_{n=1}^{N} \frac{1}{(2\pi)^{l/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] =$$

$$= -\frac{N}{2}\ln\left((2\pi)^l|\Sigma|\right) - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

Για τον εκτιμητή μεγίστης πιθανοφάνειας έχουμε μεγιστοποίηση της $p(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N}; \boldsymbol{\mu})$ και συνεπώς της $L(\boldsymbol{\mu})$. Άρα, στο μέγιστο από το θεώρημα του Fermat έχουμε μηδενισμό της παραγώγου της:

$$\frac{\partial L(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = 0 \Rightarrow \frac{\partial}{\partial \boldsymbol{\mu}}\left(\sum_{n=1}^{N}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) = 0 \ (1)$$

Λόγω ιδιότητας:

$$\frac{\partial \boldsymbol{u}^T X \boldsymbol{u}}{\partial \boldsymbol{u}} = (X + X^T)\boldsymbol{u}$$

Έχουμε από την (1):

$$\frac{\partial L(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = 0 \Rightarrow \sum_{n=1}^{N}(\Sigma^{-1} + (\Sigma^{-1})^T)(\boldsymbol{x}-\boldsymbol{\mu}) = 0 \Rightarrow \sum_{n=1}^{N}(\boldsymbol{x}-\boldsymbol{\mu}) = 0 \Rightarrow \sum_{n=1}^{N}\boldsymbol{x} = \sum_{n=1}^{N}\boldsymbol{\mu} \Rightarrow$$

$$\Rightarrow N\boldsymbol{\mu} = \sum_{n=1}^{N} \boldsymbol{x} \Rightarrow \boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x} \ (2)$$

Είδαμε λοιπόν ότι όταν μεγιστοποιείται η $L(\boldsymbol{\mu})$ για το $\boldsymbol{\mu}$ ισχύει η σχέση (2) οπότε:

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{ML}} = \arg\max_{\mu} p(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_N}; \boldsymbol{\mu}) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}$$

Και δείξαμε το ζητούμενο: ο εκτιμητής μεγίστης πιθανοφάνειας $\widehat{\boldsymbol{\mu}}_{\boldsymbol{ML}}$ της παραμέτρου μ ισούται με τον δειγματικό μέσο.

Άσκηση 1.3

Ερώτημα 1

Η τιμή κατωφλίου x_t χωρίζει τις δύο κλάσεις απόφασης και ελαχιστοποιεί το σφάλμα απόφασης του Bayes Classifier. Το σφάλμα ορίζεται ως η πιθανότητα λανθασμένης ταξινόμησης του x στην κλάση ω_1 συν την πιθανότητα λανθασμένης ταξινόμησης του x στην κλάση ω_2. Όπως φαίνεται στο διάγραμμα, σφάλμα έχουμε στις σκιαγραφημένες περιοχές, όταν δηλαδή

Είναι προφανές ότι το κατώφλι απόφασης είναι αδύνατον να είναι μικρότερο του 0 ή μεγαλύτερο του 3. Αυτό συμβαίνει διότι οι κατανομές είναι όμοιες, άρα αν το κατώφλι βρίσκεται στον υπόλοιπο χώρο η πιθανότητα σφάλματος θα είναι πάντα μεγαλύτερη από την περίπτωση που βρίσκεται στο $[0, 3]$ (γιατί προστίθεται νέο εμβαδόν).

$$P_e = P(x \in \Re_1, x \in \omega_2) + P(x \in \Re_2, x \in \omega_1) =$$
$$= P(\omega_2) \int_{\Re_1} p(x|\omega_2)dx + P(\omega_1) \int_{\Re_2} p(x|\omega_1)dx$$

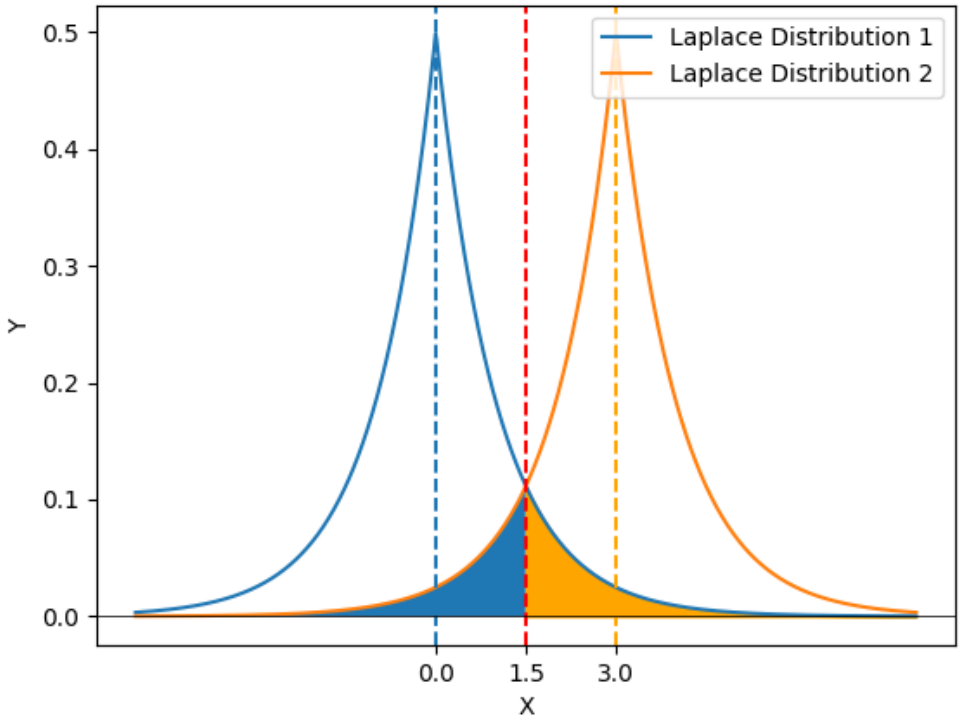
Επειδή $0 < x_t < 3$, μπορούμε να γράψουμε:

$$P_e = \frac{1}{2}\int_{\Re_1} \frac{1}{2\alpha}\exp\left(-\frac{|x-\mu_2|}{a}\right)dx + \frac{1}{2}\int_{\Re_2} \frac{1}{2\alpha}\exp\left(-\frac{|x-\mu_1|}{a}\right)dx =$$
$$= \frac{1}{2}\int_{-\infty}^{x_t} \frac{1}{2}e^{-|x-3|}dx + \frac{1}{2}\int_{x_t}^{\infty} \frac{1}{2}e^{-|x|}dx = \frac{1}{4}\int_{-\infty}^{x_t} e^{x-3}dx + \frac{1}{4}\int_{x_t}^{\infty} e^{-x}dx =$$
$$= \frac{1}{4}[e^{x-3}]|_{-\infty}^{x_t} - \frac{1}{4}[e^{-x}]|_{x_t}^{\infty} = \frac{1}{4}e^{x_t-3} + \frac{1}{4}e^{-x_t}$$

Ο ταξινομητής ελαχιστοποιεί το σφάλμα απόφασης. Αναζητούμε ακρότατα της ανωτέρω συνάρτησης:

$$\frac{dP_e}{dx_t} = 0 \Rightarrow \frac{1}{4}e^{x_t-3} - \frac{1}{4}e^{-x_t} = 0 \Rightarrow e^{x_t-3} = e^{-x_t} \xrightarrow[\gamma\nu.\alpha\acute{\upsilon}\xi ov\sigma\alpha]{\exp(\cdot)} x_t - 3 = -x_t \Rightarrow x_t = 3/2$$

Έχουμε ακρότατο (μοναδικό ελάχιστο) για $x_t = 3/2$. Συνεπώς αυτό είναι και το κατώφλι απόφασης.


Laplace Distribution 1
Laplace Distribution 2
0.5
0.4
0.3
0.2
0.1
0.0
Y
0.0
1.5
3.0
X

Ερώτημα 2

Αποδίδοντας διαφορετική βαρύτητα στα 2 σφάλματα, ορίζουμε το μέσο ρίσκο, το οποίο ο Bayes Classifier θα προσπαθήσει να ελαχιστοποιήσει.

Ρίσκο για την κλάση ω_1:

$$r_1 = \lambda_{11}\int_{\Re_1} p(x|\omega_1)dx + \lambda_{12}\int_{\Re_2} p(x|\omega_1)dx$$

Ρίσκο για την κλάση ω_2:

$$r_2 = \lambda_{21}\int_{\Re_1} p(x|\omega_2)dx + \lambda_{22}\int_{\Re_2} p(x|\omega_2)dx$$

Με $\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 1/2, \lambda_{21} = 1$. Το μέσο ρίσκο είναι:

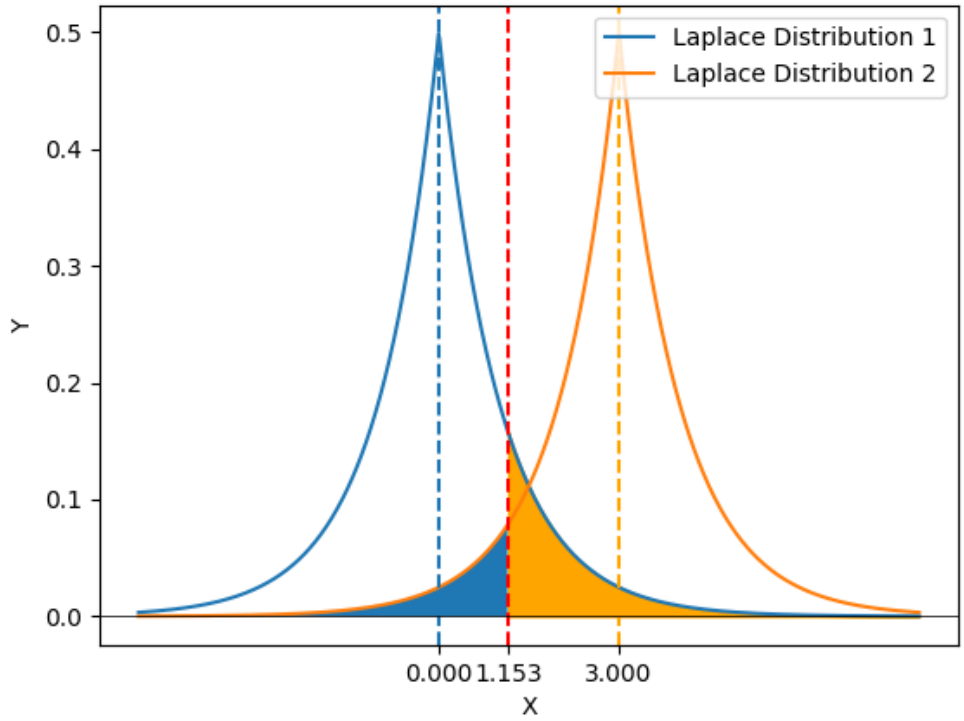
$$\begin{aligned} r &= P(\omega_1)r_1 + P(\omega_2)r_2 = \frac{1}{2}\frac{1}{2}\int_{\Re_2} p(x|\omega_1)dx + \frac{1}{2}\int_{\Re_1} p(x|\omega_2)dx = \\ &= \frac{1}{4}\int_{x_r}^{\infty}\frac{1}{2\alpha}\exp\left(-\frac{|x-\mu_1|}{a}\right)dx + \frac{1}{2}\int_{-\infty}^{x_r}\frac{1}{2\alpha}\exp\left(-\frac{|x-\mu_2|}{a}\right)dx = \\ &= \frac{1}{4}\int_{x_r}^{\infty}\frac{1}{2}e^{-|x|}dx + \frac{1}{2}\int_{-\infty}^{x_r}\frac{1}{2}e^{-|x-3|}dx = \frac{1}{8}\int_{x_r}^{\infty}e^{-x}dx + \frac{1}{4}\int_{-\infty}^{x_r}e^{x-3}dx = \\ &= -\frac{1}{8}[e^{-x}]|_r^{\infty} + \frac{1}{4}[e^{x-3}]|_{-\infty}^{x_r} = \frac{1}{8}e^{-x_r} + \frac{1}{4}e^{x_r-3} \end{aligned}$$

Ο ταξινομητής ελαχιστοποιεί το σφάλμα απόφασης. Αναζητούμε ακρότατα της τελευταίας συνάρτησης:

$$\begin{aligned} &\frac{dr}{dx_r} = 0 \Rightarrow -\frac{1}{8}e^{-x_r} + \frac{1}{4}e^{x_r-3} = 0 \Rightarrow \frac{1}{2}e^{-x_r} = e^{x_r-3} \Rightarrow \\ &\Rightarrow e^{-\ln 2}e^{-x_r} = e^{x_r-3} \xrightarrow[\gamma\nu.\alpha\acute{\upsilon}\xi ουσα]{\exp(\cdot)} -x_r - \ln 2 = x_r - 3 \Rightarrow x_r = \frac{3-\ln 2}{2} \end{aligned}$$

Έχουμε ακρότατο (μοναδικό ελάχιστο) για $x_r = (3-\ln 2)/2$. Συνεπώς αυτό είναι και το κατώφλι απόφασης.

Όπως φαίνεται και στο γράφημα, το ρίσκο για την κλάση ω_1 είναι μικρότερο από εκείνο της κλάσης ω_2, άρα μας είναι σημαντικότερο να μη γίνει λάθος ταξινόμηση στην κλάση ω_2 από το να μη γίνει λάθος ταξινόμηση στην κλάση ω_1. Έτσι, η μετατόπιση του κατωφλίου προς τη μέση τιμή της κατανομής της κλάσης ω_1 μας οδηγεί να κάνουμε πιο συχνά λάθος ταξινόμηση στην κλάση που μας είναι περισσότερο ασήμαντη.



Ερώτημα 3

Υποερώτημα i

$$\Sigma = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}$$

Έχουμε:

$$\boldsymbol{w} = \boldsymbol{\mu_1} - \boldsymbol{\mu_2} = \begin{bmatrix} 2 \\ -2 \end{bmatrix} - \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

$$\boldsymbol{x_0} = \frac{1}{2}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) - \sigma^2 \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right)\frac{\boldsymbol{\mu_1} - \boldsymbol{\mu_2}}{\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|^2} = \frac{1}{2}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$$

Άρα το υπερεπίπεδο απόφασης είναι:

$$\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x_0}) = 0 \Rightarrow 3(x - 0.5) - 4y = 0 \Rightarrow y = \frac{3}{4}x - \frac{3}{8}$$

Υποερώτημα ii

$$\Sigma = \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$$

Έχουμε:

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) = \begin{bmatrix} 3/4 & -1/4 \\ -1/4 & 3/4 \end{bmatrix}\begin{bmatrix} 3 \\ -4 \end{bmatrix} = \begin{bmatrix} 13/4 \\ -15/4 \end{bmatrix}$$

$$\boldsymbol{x_0} = \frac{1}{2}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) - \sigma^2 \ln\left(\frac{P(\omega_1)}{P(\omega_2)}\right)\frac{\boldsymbol{\mu_1} - \boldsymbol{\mu_2}}{\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|^2_{\Sigma^{-1}}} = \frac{1}{2}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$$

Άρα το υπερεπίπεδο απόφασης είναι:

$$\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x_0}) = 0 \Rightarrow 13(x - 0.5) - 15y = 0 \Rightarrow y = \frac{13}{15}x - \frac{13}{30}$$

Ερώτημα 4

Περίπτωση i

Για την περίπτωση i, αφού ο πίνακας Σ είναι διαγώνιος με ίδια διαγώνια στοιχεία, αρκεί ένας ταξινομητής ελάχιστης ευκλείδειας απόστασης (η απόσταση Mahalanobis εκφυλίζεται σε ευκλείδεια):

Απόσταση σημείου από την κλάση 1:

$$d_1 = \|\hat{\boldsymbol{x}} - \mu_1\|_2 = \sqrt{(4-2)^2 + \big(3-(-2)\big)^2} = \sqrt{29}$$

Απόσταση σημείου από την κλάση 2:

$$d_2 = \|\hat{\boldsymbol{x}} - \mu_2\|_2 = \sqrt{\big(4-(-1)\big)^2 + (3-2)^2} = \sqrt{26}$$

Αφού $d_2 < d_1$, το σημείο θα ταξινομηθεί στην κλάση 2.

Περίπτωση ii

Για την περίπτωση ii, αφού ο πίνακας Σ δεν είναι διαγώνιος, θα χρησιμοποιηθεί ταξινομητής ελάχιστης Mahalanobis απόστασης:

Απόσταση σημείου από την κλάση 1:

$$d_1 = \|\hat{\boldsymbol{x}} - \mu_1\|_{\Sigma^{-1}} = \sqrt{(\hat{\boldsymbol{x}} - \boldsymbol{\mu_1})^T\Sigma^{-1}(\hat{\boldsymbol{x}} - \boldsymbol{\mu_1})} = \sqrt{67/4}$$

Απόσταση σημείου από την κλάση 2:

$$d_2 = \|\hat{\boldsymbol{x}} - \mu_2\|_{\Sigma^{-1}} = \sqrt{(\hat{\boldsymbol{x}} - \boldsymbol{\mu_2})^T\Sigma^{-1}(\hat{\boldsymbol{x}} - \boldsymbol{\mu_2})} = \sqrt{17}$$

Αφού $d_1 < d_2$, το σημείο θα ταξινομηθεί στην κλάση 2.

Άσκηση 1.4

Ερώτημα α

Θεωρούμε συνάρτηση ενεργοποίησης $f(u) = cu$ τόσο για το κρυφό στρώμα όσο και για το στρώμα εξόδου. Θα εκφράσουμε την έξοδο Y με είσοδο $[X_1 \quad X_2]$ χρησιμοποιώντας το εξής notation:

$\varphi^{(i)}(\cdot)$: $activation\ function\ of\ layer\ i\ (here, all\ neurons\ have\ the\ same\ \varphi\ in\ each\ layer)$

$h_i^{(l)}$: $output\ of\ neuron\ i\ in\ layer\ l$

$w_{ij}^{(l)}$: $weight\ from\ neuron\ j\ to\ neuron\ i\ in\ layer\ l$

Έτσι γράφουμε:

$$h_i^{(1)} = \varphi^{(1)}\left(w_{i1}^{(1)}x_1 + w_{i3}^{(1)}x_2\right) \Rightarrow \begin{cases} h_1^{(1)} = \varphi^{(1)}\left(w_{11}^{(1)}x_1 + w_{12}^{(1)}x_2\right) \\ h_2^{(1)} = \varphi^{(1)}\left(w_{21}^{(1)}x_1 + w_{22}^{(1)}x_2\right) \end{cases}$$

$$Y = \varphi^{(2)}\left(w_{11}^{(2)}h_1^{(1)} + w_{12}^{(2)}h_2^{(1)}\right)$$

Από το σχήμα εύκολα μπορούμε να δούμε πως:

$$w_{11}^{(1)} = w_1, w_{12}^{(1)} = w_3, w_{21}^{(1)} = w_2, w_{22}^{(1)} = w_4, w_{11}^{(2)} = w_5, w_{12}^{(2)} = w_6$$
$$\varphi^{(1)} = \varphi^{(2)} = f(\cdot)$$

Άρα:

$$\begin{aligned} Y &= f\left(w_5h_1^{(1)} + w_6h_2^{(1)}\right) = f\big(w_5f(w_1x_1 + w_3x_2) + w_6f(w_2x_1 + w_4x_2)\big) = \\ &= c\big(w_5c(w_1x_1 + w_3x_2) + w_6c(w_2x_1 + w_4x_2)\big) = \\ &= c^2(w_1w_5x_1 + w_3w_5x_2 + w_2w_6x_1 + w_4w_6x_2) = \\ &= f\left(f\big((w_1w_5 + w_2w_6)x_1 + (w_3w_5 + w_4w_6)x_2\big)\right) \end{aligned}$$

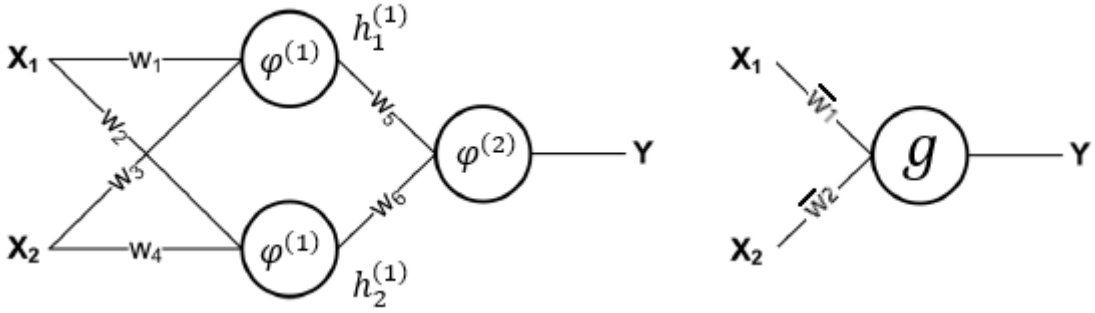
Για το ισοδύναμο Perceptron (χωρίς κρυφό στρώμα) ορίζουμε:

$$Activation\ function: g = f \circ f$$
$$Weights: \overline{w}_1 = w_1w_5 + w_2w_6, \overline{w}_2 = w_3w_5 + w_4w_6$$

Οπότε η σχέση εισόδου – εξόδου γίνεται:

$$Y = g(\overline{w}_1x_1 + \overline{w}_2x_2)$$

Και παραθέτουμε τα σχήματα του αρχικού και του μετασχηματισμένου Perceptron:



Ερώτημα β

Είναι δυνατόν για οποιοδήποτε Multi – Layer Perceptron με γραμμικές συναρτήσεις ενεργοποίησης να βρεθεί ένα ισοδύναμο νευρωνικό δίκτυο χωρίς κρυμμένα στρώματα. Αυτό συμβαίνει διότι η έξοδος κάθε νευρώνα κάθε επιπέδου είναι γραμμικός συνδυασμός των εξόδων (μετά από τη συνάρτηση ενεργοποίησής τους) των νευρώνων του προηγούμενου επιπέδου με συντελεστές τα αντίστοιχα βάρη. Αυτή η έξοδος περνά από τη συνάρτηση ενεργοποίησης του κάθε νευρώνα, η οποία είναι γραμμική.

Βλέπουμε λοιπόν ότι έχουμε σε κάθε επίπεδο γραμμικούς μετασχηματισμούς γραμμικών συνδυασμών εισόδων από προηγούμενους νευρώνες που είναι επίσης γραμμικές. Άρα η έξοδος είναι πάντα γραμμικός συνδυασμός των εισόδων, δηλαδή υπάρχουν συντελεστές (συναρτήσει των αρχικών βαρών και της σταθεράς της συνάρτησης ενεργοποίησης).

Ερώτημα γ

Εργαζόμαστε όμοια με το ερώτημα (α) και η έξοδος του Perceptron θα είναι

$$\begin{aligned} Y &= g\left(w_5 h_1^{(1)} + w_6 h_2^{(1)}\right) = g\big(w_5\sigma(w_1x_1 + w_3x_2) + w_6\sigma(w_2x_1 + w_4x_2)\big) = \\ &= sgn\big(w_5\sigma(w_1x_1 + w_3x_2) + w_6\sigma(w_2x_1 + w_4x_2)\big) = \\ &= sgn\left(w_5\frac{1}{1+\exp(-w_1x_1 - w_3x_2)} + w_6\frac{1}{1+\exp(-w_2x_1 - w_4x_2)}\right) \end{aligned}$$

Και επιθυμούμε το Y να ικανοποιεί τον πίνακα αληθείας της λογικής συνάρτησης XOR. Άρα:

x_1	x_2	Y
0	0	0
0	1	1
1	0	1
1	1	0

$$Y(x_1 = 0, x_2 = 0) = 0 \Rightarrow w_5 + w_6 < 0$$

$$Y(x_1 = 0, x_2 = 1) = 1 \Rightarrow$$
$$\Rightarrow w_5\frac{1}{1+\exp(-w_3)} + w_6\frac{1}{1+\exp(-w_4)} > 0$$

$$Y(x_1 = 1, x_2 = 0) = 1 \Rightarrow$$
$$\Rightarrow w_5\frac{1}{1+\exp(-w_1)} + w_6\frac{1}{1+\exp(-w_2)} > 0$$

$$Y(x_1 = 1, x_2 = 1) = 0 \Rightarrow$$
$$\Rightarrow w_5\frac{1}{1+\exp(-w_1 - w_3)} + w_6\frac{1}{1+\exp(-w_2 - w_4)} < 0$$

Με αριθμητικές μεθόδους, μπορούμε να προσδιορίσουμε μια τουλάχιστον λύση:

$$w1 = 1.9056159223489448$$
$$w2 = 0.9865108894368007$$
$$w3 = 1.9056159223489448$$
$$w4 = 0.9865108894368007$$
$$w5 = 1.376630930039338$$
$$w6 = -1.5989553304800608$$

Επίσης, μπορούμε να χρησιμοποιήσουμε τις εξής κοντινές «πιο ωραίες» τιμές

$$w1 = 2$$
$$w2 = 1$$
$$w3 = 2$$
$$w4 = 1$$
$$w5 = 1.4$$
$$w6 = -1.6$$

Αλλά αποκλίνουν αρκετά από το αρχικό μοντέλο προσεγγίζει πολύ το συμμετρικό.

Άσκηση 1.5

Ερώτημα α

Έχουμε το μη γραμμικό μετασχηματισμό $\varphi(u) = (u, u^2)$. Συνεπώς, ο πυρήνας που αντιστοιχεί σε αυτόν έχει τη μορφή:

$$K(X_1, X_1') = \varphi(X_1)^T \varphi(X_1') = (X_1, X_1^2)^T \left(X_1', {X_1'}^2\right) = X_1 X_1' + X_1^2 {X_1'}^2$$

Ερώτημα 2

Από την εκφώνηση, χρησιμοποιούμε 2 διανύσματα υποστήριξης: ένα για κάθε κλάση που επιδιώκουμε να διαχωρίσουμε. Επειδή σκοπός μας είναι η ελαχιστοποίηση μιας συνάρτησης κόστους που αυξάνεται με την αύξηση του αθροίσματος των τετραγώνων των αποστάσεων των διανυσμάτων υποστήριξης από την ευθεία, καταλαβαίνουμε ότι η ευθεία απόφασης θα πρέπει να διέρχεται από το μέσο των διανυσμάτων και ταυτόχρονα τα διανύσματα να είναι εκείνα τα 2 σημεία των κλάσεων που απέχουν ελάχιστη απόσταση μεταξύ τους σε σχέση με όλα τα υπόλοιπα ζεύγη.

Μετασχηματίζουμε τα δοθέντα σημεία:

Transformed Feature			Label
$\varphi(x_1) = (-1, 1)$	$\varphi(x_2) = (0, 0)$	$\varphi(x_3) = (1, 1)$	$-$
$\varphi(x_4) = (-3, 9)$	$\varphi(x_5) = (-2, 4)$	$\varphi(x_6) = (3, 9)$	$+$

Βλέπουμε πως τα δύο πλησιέστερα σημεία (που ανήκουν σε διαφορετικές κλάσεις) είναι τα $\varphi(x_1) = (-1, 1)$ και $\varphi(x_5) = (-2, 4)$. Το ευθύγραμμο τμήμα που τα ενώνει έχει την κατεύθυνση του διανύσματος $\varphi(x_5) - \varphi(x_1) = (-2 - (-1), 4 - 1) = (-1, 3)$, δηλαδή κλίση $3/(-1) = -3$.

Ακόμα, όπως είπαμε, η ευθεία απόφασης θα διέρχεται από το μέσο του ευθύγραμμου αυτού τμήματος, δηλαδή το σημείο $\left(\frac{-1-2}{2}, \frac{1+4}{2}\right) = \left(-\frac{3}{2}, \frac{5}{2}\right)$ και για την ελαχιστοποίηση του κόστους θα είναι κάθετη σε αυτό (ελαχιστοποίηση αθροίσματος τετραγώνων αποστάσεων), δηλαδή θα έχει κλίση αντιθετοαντίστροφη της κλίσης του τμήματος ή $1/3$.

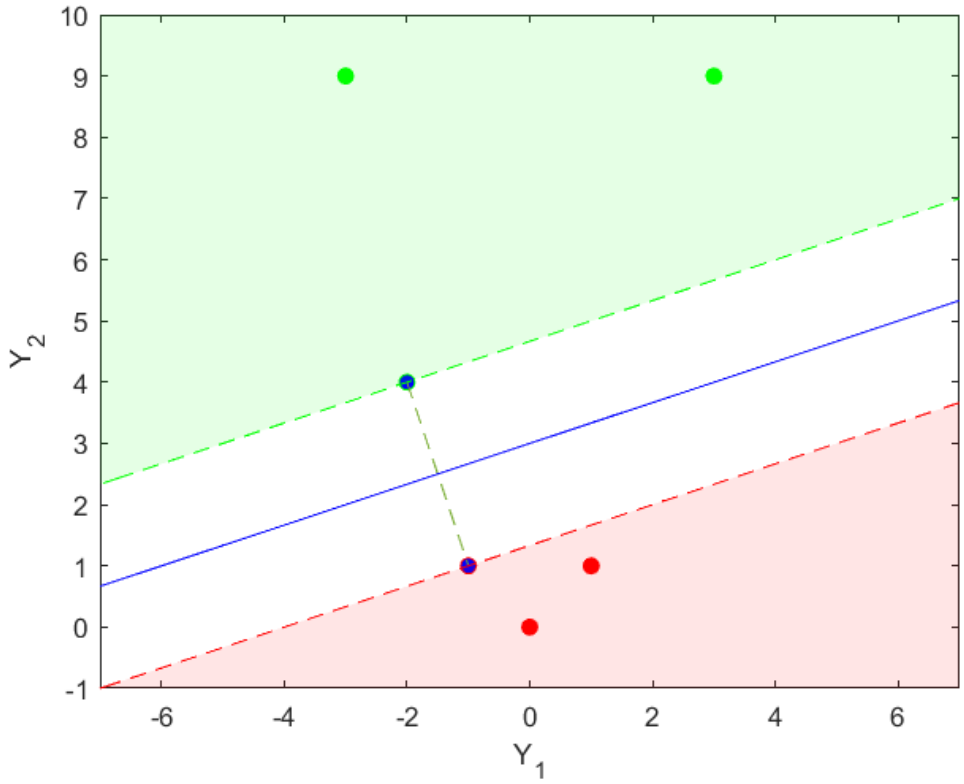
Τέλος, αρκεί να υπολογίσουμε τον σταθερό όρο στην εξίσωση της ευθείας. Αφού διέρχεται από το σημείο $\left(-\frac{3}{2}, \frac{5}{2}\right)$, θα είναι:

$$Y_2 - 5/2 = 1/3\left(Y_1 - (-3/2)\right) \Rightarrow Y_2 = 1/3 Y_1 + 3 \Rightarrow 3Y_2 - Y_1 - 9 = 0$$

Το ζητούμενο πλάτος περιθωρίου είναι το άθροισμα των αποστάσεων των διανυσμάτων υποστήριξης από την ευθεία απόφασης. Επειδή, για μια ακόμα φορά, μιλάμε για την ελαχιστοποίηση του κόστους, οι τελευταίες αποστάσεις είναι ίσες, άρα το άθροισμα είναι η απόσταση των διανυσμάτων υποστήριξης:

$$\gamma = |(-1, 1), (-2, 4)| = \sqrt{\left(-2 - (-1)\right)^2 + (4 - 1)^2} \Rightarrow \boxed{\gamma = \sqrt{10}}$$

Ερώτημα γ



- Τα πράσινα σημεία είναι οι μετασχηματισμοί εκείνων που έχουν labels +
- Τα κόκκινα σημεία είναι οι μετασχηματισμοί εκείνων που έχουν labels -
- Τα σημεία που είναι γεμισμένα με μπλε χρώμα είναι τα support vectors
- Η ευθεία απόφασης είναι η μπλε
- Οι διακεκομμένες γραμμές ορίζουν το περιθώριο απόφασης
- Οι σκιαγραφημένες περιοχές είναι οι περιοχές απόφασης

Ερώτημα δ

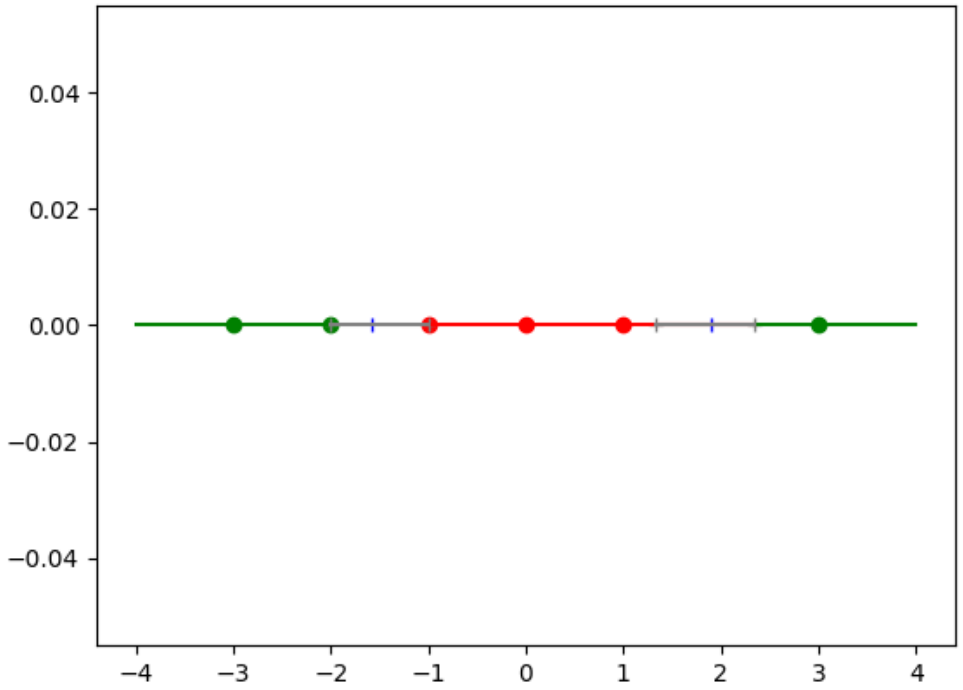
Επανερχόμαστε στο χώρο $\mathbb{R}^1$ ως εξής: $Y_1 = x, Y_2 = x^2$ (αντιστρέφοντας τον μετασχηματισμό). Άρα έχουμε:

$Ευθεία\ Απόφασης$: $3Y_2 - Y_1 - 9 = 0 \Rightarrow 3x^2 - x - 9 = 0 \Rightarrow$

$$\Rightarrow roots = \left\{\frac{1-\sqrt{109}}{6}, \frac{1+\sqrt{109}}{6}\right\}$$

$Ευθεία\ Support\ 1$: $3(Y_2 - 1.5) - (Y_1 + 0.5) - 9 = 0 \Rightarrow 3(x^2 - 1.5) - (x + 0.5) - 9 = 0$
$\Rightarrow roots = \{-2, 7/3\}$
$Ευθεία\ Support\ 2$: $3(Y_2 + 1.5) - (Y_1 - 0.5) - 9 = 0 \Rightarrow 3(x^2 + 1.5) - (x - 0.5) - 9 = 0$
$\Rightarrow roots = \{-1, 4/3\}$

Και σχεδιάζουμε:



- Η κόκκινη περιοχή αντιστοιχεί στα $labels -$
- Η πράσινη περιοχή αντιστοιχεί στα $labels +$
- Η γκρι περιοχή αντιστοιχεί στην περιοχή ανάμεσα στις ευθείες των support vectors
- Τα κόκκινα και πράσινα σημεία είναι τα δεδομένα μας
- Τα γκρι σημάδια προέρχονται από τα support vectors
- Τα μπλε σημάδια προέρχονται από την ευθεία απόφασης

Ερώτημα ε

Μας δίνεται η εξίσωση

$$y(x) = sgn\left(\sum_{n=1}^{|SV|} a_n y_n k(x, u_n) + b\right), SV = \{(-2,4), (-1,1)\}$$

Συγκρίνοντας με την εξίσωση που περιγράφει τη βέλτιστη διαχωριστική επιφάνεια

$$g^*(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 = \sum_{i=1}^{P} \lambda_i d_i \boldsymbol{x_i^T x_i} + w_0, d_1 = +1, d_2 = -1$$

Βλέπουμε πως $y_n = d_n \Rightarrow y_1 = +1, y_2 = -1$

Ο σκοπός είναι η εύρεση των συντελεστών a_n και της σταθεράς b. Από το δυϊκό πρόβλημα τετραγωνικού προγραμματισμού έχουμε τη συνάρτηση κόστους

$$\mathcal{L}^d(\lambda_1, \lambda_2) = \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j d_i d_j \boldsymbol{x_i^T x_j} - \sum_{i=1}^{P}\lambda_i$$

Αντίστοιχα

$$\mathcal{L}^d(a_1, a_2) = \frac{1}{2}\sum_{i=1}^{|SV|}\sum_{j=1}^{|SV|}\alpha_i\alpha_j y_i y_j \boldsymbol{x_i^T x_j} - \sum_{i=1}^{|SV|}a_i$$

Και το πρόβλημα του τετραγωνικού προγραμματισμού γίνεται

$$\min_{a_1,a_2\geq 0}\mathcal{L}^d(a_1, a_2) = \min_{a_1,a_2\geq 0}\left(\sum_{i=1}^{|SV|}\sum_{j=1}^{|SV|}\alpha_i\alpha_j y_i y_j \boldsymbol{x_i^T x_j} - \sum_{i=1}^{|SV|}a_i\right)$$

$$M\varepsilon\ constraints\begin{cases} a_i \geq 0, \forall i \in \{1, 2, \dots, |SV|\} \\ \sum_{n=1}^{|SV|} a_n y_n = 0 \end{cases}$$

Απλοποιούμε τη σχέση

$$\frac{1}{2}\sum_{i=1}^{|SV|}\sum_{j=1}^{|SV|}\alpha_i\alpha_j y_i y_j \boldsymbol{x_i^T x_j} - \sum_{i=1}^{|SV|}a_i =$$
$$= \frac{1}{2}\alpha_1 y_1[\alpha_1 k(u_1, u_1) + \alpha_2 k(u_1, u_2)] + \frac{1}{2}a_2 y_2[\alpha_1 k(u_2, u_1) + \alpha_2 k(u_2, u_2)] - a_1 - a_2$$

Πριν δείξαμε ότι $y_1 = +1, y_2 = -1$. Επιπλέον, $k(u_1, u_2) = k(u_2, u_1)$. Άρα

$$\alpha_1 y_1[\alpha_1 y_1 k(u_1, u_1) + \alpha_2 y_2 k(u_1, u_2)] + a_2 y_2[\alpha_1 y_1 k(u_2, u_1) + \alpha_2 y_2 k(u_2, u_2)] - a_1 - a_2$$
$$= \alpha_1[\alpha_1 k(u_1, u_1) - \alpha_2 k(u_1, u_2)] - a_2[\alpha_1 k(u_2, u_1) - \alpha_2 k(u_2, u_2)] - a_1 - a_2 =$$
$$= \frac{1}{2}a_1^2 k(u_1, u_1) - a_1 a_2 k(u_1, u_2) + \frac{1}{2}a_2^2 k(u_2, u_2) - a_1 - a_2$$

Υπολογίζουμε τις τιμές

$$k(u_1, u_1) = u_1^T u_1 = (-2, 4)^T(-2, 4) = 20$$
$$k(u_1, u_2) = u_1^T u_2 = (-2, 4)^T(-1, 1) = 6$$
$$k(u_1, u_1) = u_1^T u_1 = (-2, 4)^T(-2, 4) = 2$$

Και ο περιορισμός μας δίνει:

$$(+1)a_1 + (-1)a_2 = 0 \Rightarrow a_1 = a_2$$

Άρα η συνάρτηση κόστους γίνεται:

$$\mathcal{L}^d(a_1, a_2) = 10a_1^2 - 6a_1^2 + a_1^2 - a_1 - a_1 = 5\alpha_1^2 - 2\alpha_1$$

Για το πρόβλημα ελαχισσοποίησης είναι:

$$\min_{a_1} \mathcal{L}^d(a_1, a_2) = \min_{a_1} 5\alpha_1^2 - 2\alpha_1$$

Και έχει μοναδική λύση:

$$\alpha_1 = \alpha_2 = \arg\min_{a_1}(5\alpha_1^2 - 2\alpha_1) = 0.2$$

Οι συνθήκες KKT μας λένε:

$$\sum_{n=1}^{|SV|} a_n y_n \varphi(u_n) = a_1\varphi(u_1) - a_2\varphi(u_2) = [-0.4 \quad 0.8] - [-0.2 \quad 0.2] = [-0.2 \quad 0.6]$$

Όμως ισχύει για το κατώφλι:

$$w_0 = b = \frac{1}{|SV|}\sum_{n=1}^{|SV|}\left(\frac{1}{y_n} - w^T\varphi(u_n)\right) =$$
$$= \frac{1}{2}[(1 - (-0.2)(-2) - 0.6 \cdot 4) + (-1 - (-0.2)(-1) - 0.6 \cdot 1)] = -1.8$$

Βρήκαμε λοιπόν $a_1 = a_2 = 0.2, b = -1.8$

Ερώτημα στ

Με την προσθήκη του σημείου $x_7 = 5$ με θετική ετικέτα στο μετασχηματισμένο χώρο καταστάσεων προστίθεται το σημείο $(5, 25)$. Παρατηρούμε ότι το ζεύγος των κοντινότερων μεταξύ τους σημείων διαφορετικών κλάσεων παραμένει το ίδιο με πριν. Άρα, αφού χρησιμοποιούμε ακόμα 2 διανύσματα υποστήριξης, τα οποία είναι τα ίδια με πριν, δε θα αλλάξει η γραμμή απόφασης. Επιπλέον, τα μετασχηματισμένα δεδομένα παραμένουν γραμμικώς διαχωρίσιμα.

Ερώτημα ζ

Θεωρούμε τον μετασχηματισμό

$$\varphi\colon \mathbb{R}^1 \mapsto \mathbb{R}^n, \varphi_n(x) = \left\{e^{-x^2/2}, e^{-x^2/2}x, \frac{e^{-x^2/2}x^2}{\sqrt{2}}, \dots, \frac{e^{-x^2/2}x^i}{\sqrt{i!}}, \dots, \frac{e^{-x^2/2}x^n}{\sqrt{n!}}\right\}, n \to \infty$$

Και έστω $x, y \in \mathbb{R}$ για τα 2 διανύσματα μ, ν

$$\mu = \varphi_\infty(x), \nu = \varphi_\infty(y)$$

Επομένως

$$k(\mu, \nu) = \varphi_\infty(x)^T \varphi_\infty(y) = \mu \cdot \nu = \sum_{i=1}^{\infty} \frac{e^{-x^2/2}x^i}{\sqrt{i!}} \frac{e^{-y^2/2}y^i}{\sqrt{i!}} = e^{-\frac{x^2}{2}-\frac{y^2}{2}} \sum_{i=1}^{\infty} \frac{x^i y^i}{i!}$$

Από τη σειρά Taylor της e^x γνωρίζουμε πως

$$\sum_{i=1}^{\infty} \frac{x^i y^i}{i!} = \sum_{i=1}^{\infty} \frac{(xy)^i}{i!} = e^{xy}$$

Συνεπώς

$$k(x, y) = e^{-\frac{x^2}{2}-\frac{y^2}{2}} e^{xy} = e^{-\frac{1}{2}(x^2-2xy+y^2)} = e^{-\frac{1}{2}(x-y)^2}$$

Ερώτημα η

Ο μετασχηματισμός φ_∞ οδηγεί τα δεδομένα σε ένα χώρο πάρα πολλών διαστάσεων. Τα πάρα πολλά αυτά χαρακτηριστικά που απαιτούνται για την απεικόνιση των σημείων στο νέο χώρο είναι υπεύθυνα για το ότι τα μετασχηματισμένα σημεία θα είναι πολύ αραιά στο χώρο $\mathbb{R}^\infty$. Έτσι, υπερπροσαρμογή είναι πολύ πιθανό να συμβεί.

Άσκηση 1.6

Ερώτημα α

Υποερώτημα 1

Θεωρούμε την τυχαία μεταβλητή X που αντιστοιχεί στα χαρακτηριστικά που θέλουμε να διαχωρίσουμε και και την τυχαία μεταβλητή Y που αντιστοιχεί στο χαρακτηριστικό το οποίο επιλέγουμε.

Ορίζουμε ως εντροπία της τυχαίας μεταβλητής X με $P(X = x) = p(x)$ την ποσότητα

$$H(X) = -\sum_{x} p(x) \log p(x)$$

Επίσης, για την τυχαία μεταβλητή X δεδομένου $Y = y$ η εντροπία εκφράζεται ως

$$H(X|Y = y) = -\sum_{x} p(x|y) \log p(x|y)$$

Έτσι, με $p(x, y)$ να είναι η από κοινού κατανομή των X, Y, για την τυχαία μεταβλητή X δεδομένης της άλλης τυχαίας μεταβλητής Y, η εντροπία θα είναι το άθροισμα για κάθε τιμή που μπορεί να πάρει η Y

$$H(X|Y) = \sum_{y} p(y) H(X|Y = y) = -\sum_{y}\sum_{x} p(y)p(x|y) \log p(x|y) =$$
$$= -\sum_{y}\sum_{x} p(x, y) \log \frac{p(x, y)}{p(y)}$$

Ακόμα, το information gain ενός attribute A αναφορικά με την τυχαία μεταβλητή του target T ορίζεται ως η διαφορά (τυχαιότητα του T πλην τυχαιότητα του A δεδομένου του T)

$$ig(A, T) = H(T) - H(T|A) = -\sum_{t} p(t) \log p(t) + \sum_{a}\sum_{t} p(t, a) \log \frac{p(t, a)}{p(a)}$$

Μπορούμε να εκφράσουμε τον δεύτερο όρο του πρώτου αθροίσματος ως άθροισμα πάνω στο a ως εξής

$$\sum_{t} p(t) \log p(t) = \sum_{t} \log p(t) \sum_{a} p(t, a)$$

Άρα

$$ig(A, T) = \sum_{a}\sum_{t} p(t, a) \log \frac{p(t, a)}{p(a)} - \sum_{t} \log p(t) \sum_{a} p(t, a) =$$

$$= \sum_a \sum_t p(t,a) \log \frac{p(t,a)}{p(a)} - \sum_t \sum_a p(t,a) \log p(t) =$$
$$= \sum_a \sum_t p(t,a) \left(\log \frac{p(t,a)}{p(a)} - \log p(t) \right) =$$
$$= \sum_a \sum_t p(t,a) \left(\log \frac{p(t,a)}{p(a)p(t)} \right) =$$
$$= -\sum_a \sum_t p(t,a) \left(\log \frac{p(a)p(t)}{p(t,a)} \right)$$

Σε αυτό το σημείο θα αξιοποιήσουμε την ανισότητα του Jensen:

$$\mathbb{E}\{\varphi(X)\} \geq \varphi(\mathbb{E}\{X\})$$

Και έχουμε

$$ig(A,T) = -\sum_a \sum_t p(t,a) \left(\log \frac{p(a)p(t)}{p(t,a)} \right) \geq -\log \left(\sum_a \sum_t p(t,a) \frac{p(a)p(t)}{p(t,a)} \right) =$$
$$= -\log \left(\sum_a \sum_t p(a)p(t) \right) =$$
$$= -\log \left(\sum_a p(a) \sum_t p(t) \right) = -\log(1) = 0$$

Αποδείξαμε λοιπόν το ζητούμενο.

Υποερώτημα 2

Για δύο κόμβους $t_i, t_j, i < j$ στο μονοπάτι $p = t_1 t_2 \dots t_i \dots t_j \dots t_n$, ισχύει ότι τα σύνολα $D_m, \forall m > i$ των κόμβων t_m έχουν πληροφορία για το χαρακτηριστικό. Ο κόμβος t_j, επειδή έχει μηδενική αβεβαιότητα για το σε ποια διακλάδωση θα οδηγήσει, σημαίνει ότι ξ τυχαία μεταβλητή t_j έχει $H(t_j) = 0$. Άρα το κέρδος πληροφορίας θα είναι:

$$ig(t_j) = H(R_j) - H(R_j|t_j) = H(R_j) - H(R_j) = 0$$

Ερώτημα β

Outlook	Temperature	Humidity	Windy	Play Tennis
Overcast	4	Normal	TRUE	Yes
Sunny	4	Normal	TRUE	Yes
Rainy	8	Normal	FALSE	Yes
Rainy	22	High	FALSE	Yes
Rainy	22	Normal	FALSE	Yes
Sunny	22	Normal	TRUE	Yes
Overcast	22	High	TRUE	Yes
Overcast	36	High	FALSE	Yes
Overcast	36	Normal	FALSE	Yes
Rainy	8	Normal	TRUE	No
Sunny	22	High	FALSE	No
Rainy	22	High	TRUE	No
Sunny	40	High	FALSE	No
Sunny	40	High	TRUE	No

Ο gini index ποσοτικοποεί το βαθμό «μη καθαρότητας» ενός συνόλου δεδομένων και ορίζεται ως

$$gini(\mathbb{D}) = 1 - \sum_{i \in labels(\mathbb{D})} p_i^2$$

Και το information gain του χαρακτηριστικού X ορίζεται ως

$$ig(X) = gini(Root) - \sum_{v \in values(X)} \frac{|X = v|}{|Root|} gini(X = v)$$

Κατηγορικό χαρακτηριστικό Temperature

Αρχικά αντιμετωπίζουμε το χαρακτηριστικό Temperature ως κατηγορικό. Διακρίνουμε δύο κλάσεις οπότε:

$$gini(Root) = 1 - \left(\frac{5}{14}\right)^2 - \left(\frac{9}{14}\right)^2 = \frac{45}{98}$$

Information Gain για το χαρακτηριστικό Outlook

$gini(Outlook) =$

$$= \frac{5}{14} gini(Outlook = Rainy) + \frac{5}{14} gini(Outlook = Sunny) +$$

$$+\frac{4}{14}gini(Outlook = Overcast) =$$
$$=\frac{5}{14}\left[1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right]+\frac{5}{14}\left[1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right]+\frac{4}{14}\left[1-\left(\frac{4}{4}\right)^2-\left(\frac{0}{4}\right)^2\right]=\frac{12}{35}$$

$$ig(Outlook) = gini(Root) - gini(Outlook) \cong 0.1163$$

Information Gain για το χαρακτηριστικό Temperature

$$gini(Temperature) =$$
$$=\frac{2}{14}gini(Temperature = 4)+\frac{2}{14}gini(Temperature = 8)+$$
$$+\frac{6}{14}gini(Temperature = 22)+\frac{2}{14}gini(Temperature = 36)+$$
$$+\frac{2}{14}gini(Temperature = 40) =$$
$$=\frac{2}{14}\left[1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2\right]+\frac{2}{14}\left[1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right]+\frac{6}{14}\left[1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2\right]+$$
$$+\frac{2}{14}\left[1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2\right]+\frac{2}{14}\left[1-\left(\frac{0}{2}\right)^2-\left(\frac{2}{2}\right)^2\right]=\frac{11}{42}$$

$$ig(Temperature) = gini(Root) - gini(Temperature) \cong 0.1973$$

Information Gain για το χαρακτηριστικό Humidity

$$gini(Humidity) =$$
$$=\frac{7}{14}gini(Humidity = Normal)+\frac{7}{14}gini(Humidity = High) =$$
$$=\frac{7}{14}\left[1-\left(\frac{6}{7}\right)^2-\left(\frac{1}{7}\right)^2\right]+\frac{7}{14}\left[1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2\right]=\frac{18}{49}$$

$$ig(Humidity) = gini(Root) - gini(Humidity) \cong 0.0918$$

Information Gain για το χαρακτηριστικό Windy

$$gini(Windy) =$$
$$=\frac{7}{14}gini(Windy = TRUE)+\frac{7}{14}gini(Windy = FALSE) =$$
$$=\frac{7}{14}\left[1-\left(\frac{2}{7}\right)^2-\left(\frac{5}{7}\right)^2\right]+\frac{7}{14}\left[1-\left(\frac{5}{7}\right)^2-\left(\frac{2}{7}\right)^2\right]=\frac{20}{49}$$

$$ig(Windy) = gini(Root) - gini(Windy) \cong 0.0510$$

Μεγαλύτερο information gain έχουμε για το feature Temperature. Άρα το επιλέγουμε ως Root και συνεχίζουμε.

Outlook	Temperature	Humidity	Windy	Play Tennis
Overcast	4	Normal	TRUE	Yes
Sunny	4	Normal	TRUE	Yes
Overcast	36	High	FALSE	Yes
Overcast	36	Normal	FALSE	Yes

Outlook	Temperature	Humidity	Windy	Play Tennis
Sunny	40	High	FALSE	No
Sunny	40	High	TRUE	No

Βλέπουμε πως από τα δεδομένα μας, είναι βέβαιο ότι:

- $Temperature \in \{4, 36\}$ σημαίνει κατάλληλες συνθήκες για να παίξουμε Tennis (4 από 4 εγγραφές)
- $Temperature = 40$ σημαίνει ακατάλληλες συνθήκες για να παίξουμε Tennis (2 από 2 εγγραφές)

Συνεπώς, αφού αυτές οι τιμές για το Temperature οδηγούν σε μοναδικό label, έχουμε 3 κλάδους, έναν για κάθε τιμή με τους κλάδους των $\{4, 36\}$ να οδηγούν σε φύλλο με τιμή Yes και τον κλάδο του 40 να οδηγεί σε φύλλο με τιμή No.

Ας δούμε σε ποιο χαρακτηριστικό πρέπει να οδηγεί ο κλάδος για $Temperature = 22$ ώστε να πετύχουμε βελτιστότητα. Πλέον κοιτάμε μόνο τις εγγραφές για τις οποίες είναι $Temperature = 22$.

$$gini(Temperature = 22\ ) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = \frac{4}{9}$$

Outlook	Temperature	Humidity	Windy	Play Tennis
Rainy	22	High	FALSE	Yes
Rainy	22	Normal	FALSE	Yes
Sunny	22	Normal	TRUE	Yes
Overcast	22	High	TRUE	Yes
Sunny	22	High	FALSE	No
Rainy	22	High	TRUE	No

Information Gain για το χαρακτηριστικό Outlook δεδομένου Root = Temperature

$gini(Outlook) =$

$$= \frac{3}{6} gini(Outlook = Rainy) + \frac{2}{6} gini(Outlook = Sunny) +$$
$$+ \frac{1}{6} gini(Outlook = Overcast) =$$
$$= \frac{3}{6}\left[1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right] + \frac{2}{6}\left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right] + \frac{1}{6}\left[1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right] = \frac{7}{18}$$

$$ig(Outlook) = gini(Temperature = 22\ ) - gini(Outlook) \cong 0.0556$$

Information Gain για το χαρακτηριστικό Humidity δεδομένου Root = Temperature

$$gini(Humidity) =$$

$$= \frac{2}{6} gini(Humidity = Normal) + \frac{4}{6} gini(Humidity = High) =$$
$$= \frac{2}{6}\left[1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2\right] + \frac{4}{6}\left[1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right] = \frac{1}{3}$$

$$ig(Humidity) = gini(Temperature = 22\ ) - gini(Humidity) \cong 0.1111$$

Information Gain για το χαρακτηριστικό Windy δεδομένου Root = Temperature

$$gini(Windy) =$$

$$= \frac{3}{6} gini(Windy = TRUE) + \frac{3}{6} gini(Windy = FALSE) =$$
$$= \frac{3}{6}\left[1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right] + \frac{3}{6}\left[1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right] = \frac{4}{9}$$

$$ig(Windy) = gini(Temperature = 22\ ) - gini(Windy) = 0$$

Μεγαλύτερο information gain έχουμε για το feature Humidity.

Ας δούμε σε ποιο χαρακτηριστικό πρέπει να οδηγεί ο κλάδος για $Temperature = 8$ ώστε να πετύχουμε βελτιστότητα. Πλέον κοιτάμε μόνο τις εγγραφές για τις οποίες είναι $Temperature = 8$.

$$gini(Temperature = 8\ ) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

Outlook	Temperature	Humidity	Windy	Play Tennis
Rainy	8	Normal	FALSE	Yes
Rainy	8	Normal	TRUE	No

Information Gain για το χαρακτηριστικό Outlook δεδομένου Root = Temperature

$$gini(Outlook) = \frac{2}{2} gini(Outlook = Rainy) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$ig(Outlook) = gini(Temperature = 8\ ) - gini(Outlook) = 0$$

Information Gain για το χαρακτηριστικό Humidity δεδομένου Root = Temperature

$$gini(Humidity) = \frac{2}{2} gini(Humidity = Normal) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$ig(Humidity) = gini(Temperature = 8\ ) - gini(Humidity) = 0$$

Information Gain για το χαρακτηριστικό Windy δεδομένου Root = Temperature

$$gini(Windy) =$$

$$= \frac{1}{2} gini(Windy = TRUE) + \frac{1}{2} gini(Windy = FALSE) =$$

$$= \frac{1}{2}\left[1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right] + \frac{1}{2}\left[1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right] = 0$$

$$ig(Windy) = gini(Temperature = 8) - gini(Windy) = 0.5$$

Μεγαλύτερο information gain έχουμε για το feature Windy.

Συνεπώς επιλέγουμε ως παιδί στον κλάδο $Temperature = 22$ το feature $Humidity$ και ως παιδί στον κλάδο $Temperature = 8$ το feature $Windy$ και συνεχίζουμε.

Outlook	Temperature	Humidity	Windy	Play Tennis
Rainy	8	Normal	FALSE	Yes
Rainy	8	Normal	TRUE	No

Outlook	Temperature	Humidity	Windy	Play Tennis
Rainy	22	Normal	FALSE	Yes
Sunny	22	Normal	TRUE	Yes

Βλέπουμε πως από τα δεδομένα μας, είναι βέβαιο ότι:

- $Temperature = 8 \wedge Windy = TRUE$ σημαίνει ακατάλληλες συνθήκες για να παίξουμε Tennis (1 από 1 εγγραφή)
- $Temperature = 8 \wedge Windy = FALSE$ σημαίνει κατάλληλες συνθήκες για να παίξουμε Tennis (1 από 1 εγγραφή)
- $Temperature = 22 \wedge Humidity = Normal$ σημαίνει κατάλληλες συνθήκες για να παίξουμε Tennis (2 από 2 εγγραφές)

Θα έχουμε λοιπόν 2 κλάδους από τον κόμβο $Windy$ να οδηγούν με συνθήκες $TRUE, FALSE$ σε φύλλα με τιμές No, Yes αντίστοιχα και 1 κλάδο από τον κόμβο $Humidity$ να οδηγεί με συνθήκη $Normal$ σε τιμή Yes.

Ας δούμε σε ποιο χαρακτηριστικό πρέπει να οδηγεί ο κλάδος για $Humidity = High$ ώστε να πετύχουμε βελτιστότητα. Πλέον κοιτάμε μόνο τις εγγραφές για τις οποίες είναι $Temperature = 22 \wedge Humidity = High$.

$$gini(Temperature = 22 \wedge Humidity = High) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = \frac{1}{2}$$

Outlook	Temperature	Humidity	Windy	Play Tennis
Rainy	22	High	FALSE	Yes
Overcast	22	High	TRUE	Yes
Sunny	22	High	FALSE	No
Rainy	22	High	TRUE	No

Information Gain για το χαρακτηριστικό Outlook δεδομένου $Temperature = 22 \wedge Humidity = High$

$$gini(Outlook) =$$
$$= \frac{2}{4} gini(Outlook = Rainy) + \frac{1}{4} gini(Outlook = Sunny) +$$
$$+ \frac{1}{4} gini(Outlook = Overcast) =$$
$$= \frac{2}{4}\left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right] + \frac{1}{4}\left[1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right] + \frac{1}{4}\left[1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2\right] = \frac{1}{4}$$

$$ig(Outlook) = gini(Temperature = 22 \wedge Humidity = High\ ) - gini(Outlook) = 0.25$$

Information Gain για το χαρακτηριστικό Windy δεδομένου $emperature = 22 \wedge Humidity = High$

$$gini(Windy) = \frac{2}{4} gini(Windy = FALSE) + \frac{2}{4} gini(Windy = TRUE) =$$
$$= \frac{2}{4}\left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right] + \frac{2}{4}\left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right] = \frac{1}{2}$$

$ig(Windy) = gini(Temperature = 22 \wedge Humidity = High\ ) - gini(Windy) = 0$

Μεγαλύτερο information gain έχουμε για το feature Outlook.

Outlook	Temperature	Humidity	Windy	Play Tennis
Overcast	22	High	TRUE	Yes

Outlook	Temperature	Humidity	Windy	Play Tennis
Sunny	22	High	FALSE	No

Βλέπουμε πως από τα δεδομένα μας, είναι βέβαιο ότι:

- $Temperature = 22 \wedge Humidity = High \wedge Outlook = Overcast$ σημαίνει κατάλληλες συνθήκες για να παίξουμε Tennis
- $Temperature = 22 \wedge Humidity = High \wedge Outlook = Sunny$ σημαίνει ακατάλληλες συνθήκες για να παίξουμε Tennis

Θα έχουμε λοιπόν 2 κλάδους από τον κόμβο $Outlook$ να οδηγούν με συνθήκες $Overcast, Sunny$ σε φύλλα με τιμές Yes, No αντίστοιχα.

Ας δούμε σε ποιο χαρακτηριστικό πρέπει να οδηγεί ο κλάδος για $Outlook = Rainy$ ώστε να πετύχουμε βελτιστότητα. Πλέον κοιτάμε μόνο τις εγγραφές για τις οποίες είναι $Temperature = 22 \wedge Humidity = High \wedge Outlook = Rainy$.

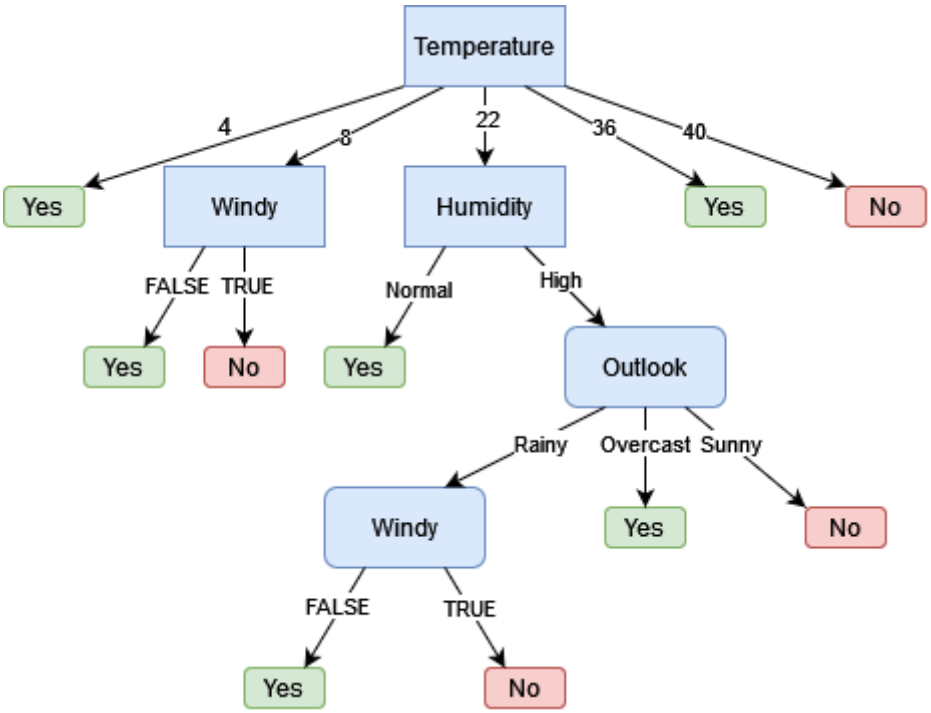
Outlook	Temperature	Humidity	Windy	Play Tennis
Rainy	22	High	FALSE	Yes
Rainy	22	High	TRUE	No

Τετριμμένη περίπτωση: Το μοναδικό χαρακτηριτικό που έμεινε είναι το $Windy$ και σε αυτό πρέπει να οδηγεί ο κλάδος για $Outlook = Rainy$. Βλέπουμε πως από τα δεδομένα μας, είναι βέβαιο ότι:

- $Temperature = 22 \wedge Humidity = High \wedge Outlook = Rainy \wedge Windy = TRUE$ σημαίνει ακατάλληλες συνθήκες για να παίξουμε Tennis
- $Temperature = 22 \wedge Humidity = High \wedge Outlook = Rainy \wedge Windy = FALSE$ σημαίνει κατάλληλες συνθήκες για να παίξουμε Tennis

Θα έχουμε λοιπόν 2 κλάδους από τον κόμβο $Windy$να οδηγούν με συνθήκες $TRUE, FALSE$ σε φύλλα με τιμές No, Yes αντίστοιχα.

Το δέντρο είναι:



Αριθμητικό χαρακτηριστικό Temperature

Θα χρειαστεί να βρεθεί κατάλληλη τιμή κατωφλίου h_0.

Ταξινομούμε όλες τις τιμές θερμοκρασίας $[4, 8, 22, 36, 40]$ και για τα διαδοχικά ζεύξη αυτών βρίσκουμε τις μέσες τιμές $[6, 15, 24, 38]$. Θα αποφανθούμε ποια από αυτές θα επιλέξουμε. Η απάντηση είναι απλή: Εκείνη που μας οδηγεί στο μέγιστο information gain, το οποίο ορίζεται πλέον ως:

$$ig(Temperature) = gini(Root) - \frac{|Temperature \geq h_0|}{|Root|} gini(Temperature \geq h_0) - \\ - \frac{|Temperature < h_0|}{|Root|} gini(Temperature < h_0)$$

$h_0 = 6$

$$ig(Temperature) = gini(Root) - \frac{12}{14} gini(Temperature \geq 6) - \\ - \frac{2}{14} gini(Temperature < 6) = 1 - \left(\frac{7}{12}\right)^2 - \left(\frac{5}{12}\right)^2 + 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = 0.0425$$

$h_0 = 15$

$$ig(Temperature) = gini(Root) - \frac{12}{14} gini(Temperature \geq 15) - \\ - \frac{2}{14} gini(Temperature < 15) = 1 - \left(\frac{7}{12}\right)^2 - \left(\frac{5}{12}\right)^2 + 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = 0.0425$$

$h_0 = 24$

$$ig(Temperature) = gini(Root) - \frac{10}{14} gini(Temperature \geq 24) - \\ - \frac{4}{14} gini(Temperature < 24) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 + 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2 = 0.0163$$

$h_0 = 38$

$$ig(Temperature) = gini(Root) - \frac{2}{14} gini(Temperature \geq 38) - \\ - \frac{12}{14} gini(Temperature < 38) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 + 1 - \left(\frac{9}{12}\right)^2 - \left(\frac{3}{12}\right)^2 = 0.1378$$

Επιλέγουμε $h_0 = 38$.

Με παρόμοια στρατηγική εργαζόμαστε και τώρα:

Για $Temperature \geq h_0$ 2 από 2 εγγραφές έχουν label No. Άρα έχουμε κατευθείαν σύνδεση με φύλλο No.

Εξετάζουμε τις εγγραφές όπου $Temperature < h_0$/

Outlook	Temperature	Humidity	Windy	Play Tennis
Overcast	4	Normal	TRUE	Yes
Sunny	4	Normal	TRUE	Yes
Rainy	8	Normal	FALSE	Yes
Rainy	22	High	FALSE	Yes
Rainy	22	Normal	FALSE	Yes
Sunny	22	Normal	TRUE	Yes
Overcast	22	High	TRUE	Yes
Overcast	36	High	FALSE	Yes
Overcast	36	Normal	FALSE	Yes
Rainy	8	Normal	TRUE	No
Sunny	22	High	FALSE	No
Rainy	22	High	TRUE	No

$$ig(Temperature < h_0) = 1 - \left(\frac{9}{12}\right)^2 - \left(\frac{3}{12}\right)^2 = \frac{3}{8}$$

$$ig(Outlook) = gini(Temperature < h_0) - gini(Outlook) =$$
$$= \frac{3}{12} gini(Outlook = Sunny) + \frac{4}{12} gini(Outlook = Overcast) +$$
$$+ \frac{5}{12} gini(Outlook = Rainy) =$$
$$= \frac{3}{12}\left[1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right] + \frac{4}{12}\left[1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right] + \frac{5}{12}\left[1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right] = 0.0639$$

$$ig(Humidity) = gini(Temperature < h_0) - gini(Humidity) =$$
$$= \frac{7}{12} gini(Humidity = Normal) + \frac{5}{12} gini(Humidity = High) =$$
$$= \frac{7}{12}\left[1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right] + \frac{5}{12}\left[1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right] = 0.0321$$

$$ig(Windy) = gini(Temperature < h_0) - gini(Windy) =$$
$$= \frac{6}{12} gini(Windy = TRUE) + \frac{6}{12} gini(Windy = FALSE) =$$
$$= \frac{6}{12}\left[1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2\right] + \frac{6}{12}\left[1 - \left(\frac{5}{6}\right)^2 - \left(\frac{1}{6}\right)^2\right] = 0.0139$$

Επιλέγουμε το $Outlook$ αφού έχει το μεγαλύτερο information gain.

Παρατηρούμε ότι άπαξ και φτάσουμε στον κόμβο $Outlook$, τότε:

- $Outlook = Overcast$ συνεπάγεται κατάλληλες συνθήκες για Τέννις.

Outlook	Temperature	Humidity	Windy	Play Tennis
Overcast	4	Normal	TRUE	Yes
Overcast	22	High	TRUE	Yes
Overcast	36	High	FALSE	Yes
Overcast	36	Normal	FALSE	Yes

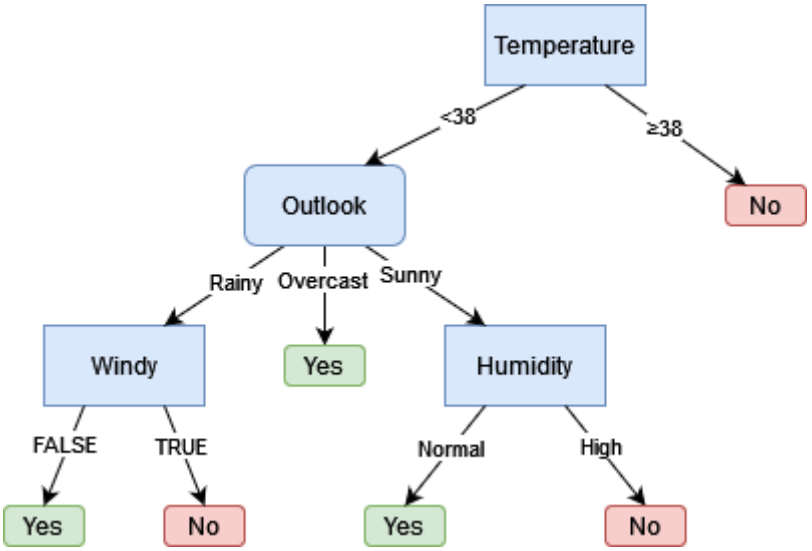
- $Outlook = Rainy$ μόνο ο κόμβος $Windy$ είναι ικανός να προσδιορίσει τις συνθήκες για το Τέννις (μέγιστο information gain).

Outlook	Temperature	Humidity	Windy	Play Tennis
Rainy	8	Normal	FALSE	Yes
Rainy	22	High	FALSE	Yes
Rainy	22	Normal	FALSE	Yes
Rainy	8	Normal	TRUE	No
Rainy	22	High	TRUE	No

- $Outlook = Sunny$ μόνο ο κόμβος $Humidity$ αρκεί να προσδιορίσει τις συνθήκες για το Τέννις (μέγιστο information gain).

Outlook	Temperature	Humidity	Windy	Play Tennis
Sunny	4	Normal	TRUE	Yes
Sunny	22	Normal	TRUE	Yes
Sunny	22	High	FALSE	No

Το δέντρο είναι:



Η διαφορά με το προηγούμενο δέντρο είναι ότι ο αριθμός των κόμβων μειώθηκε. Συνεπώς το νέο μοντέλο είναι πιο αποδοτικό καθώς απαιτούνται λιγότερες συγκρίσεις για την ταξινόμηση, άρα είναι πιο γρήγορο και είναι πιο κατανοητό από εμάς. Συνεπώς, επιλέγουμε αυτό το δέντρο.