1. Na raspolaganju nam je skup za učenje $\mathcal{D}$ koji se sastoji od deset primjera iz $\mathcal{X} = \mathbb{R}^4$:

i	x_1	x_2	x_3	x_4
1	−1.28	−1.51	−1.65	−2.29
2	−1.83	−0.99	−2.51	−0.72
3	3.24	0.11	2.15	−1.67
4	−2.47	−1.95	−0.65	−2.61
5	−1.07	0.15	−0.40	−2.31
6	−1.58	0.06	0.26	−0.55
7	−1.00	−0.95	−0.63	−2.12
8	−0.53	−0.67	−1.40	−1.65
9	0.50	−0.91	1.31	−1.80
10	0.70	0.11	−2.04	−1.83

(a) Izračunajte ML-procjenu kovarijacijske matrice $\hat{\boldsymbol{\Sigma}}$.

Procjenitelji najveće izglednosti za srednju vrijednost i kovarijacijsku matricu:

$$\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{ML}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{ML}})^{\mathrm{T}}$$

Izračunati procjenitelji:

$$\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \begin{bmatrix} -0.5320 & -0.6550 & -0.5560 & -1.7550 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \begin{bmatrix} 2.4418 & 0.6105 & 1.2935 & 0.0102 \\ 0.6105 & 0.5002 & 0.2914 & 0.1758 \\ 1.2935 & 0.2914 & 1.9438 & -0.0264 \\ 0.0102 & 0.1758 & -0.0264 & 0.4008 \end{bmatrix}$$

(b) Izračunajte Mahalanobisovu udaljenost između točaka $\mathbf{x}^{(1)} = (1, 1, 1, 1)$ i $\mathbf{x}^{(2)} = (1, 1, -1, -1)$. Je li ta udaljenost jednaka euklidskoj udaljenosti? Zašto? Je li to dobro ili loše?

Mahalanobisova udaljenost između točaka $\mathbf{x^{(1)}}$ i $\mathbf{x^{(2)}}$:

$$\Delta = \sqrt{(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})^T}$$

$$\Delta = \quad 4.0301$$

Euklidska udaljenost između točaka $\mathbf{x^{(1)}}$ i $\mathbf{x^{(2)}}$:

$$d(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \sqrt{(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})^T}$$

$$d(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \quad 2.8284$$

Euklidska udaljenost nije jednaka Mahalanobisovoj udaljenosti, jer Mahalanobisova uzima u obzir razlike u varijanci između pojedinih dimenzija te korelacije između varijabli.
(Bili bi jednaki kad bi $\boldsymbol{\Sigma} = \mathbf{I}$)

To je zapravo jako dobro, jer bi nam udaljenosti trebale govoriti koliko su dvije točke su dvije točke 'slične' jedna drugoj, uzimajući u obzir varijance i korelacije, a nama su sličnosti jako važne kad radimo klasifikatore.

Npr, kad bi uspoređivali dvije točke koje su si blizu (u euklidskom pogledu) u jednoj dimenziji, ali ta dimenzija ima veliku varijancu, te dvije točke bi si trebale biti manje slične od dvije točke koje su si blizu u dimenziji koja ima manju varijancu.

(c) Kako možemo utvrditi jesu li dvije varijable statistički nezavisne? Pokušajte to napraviti za varijable x_1 i x_2. Možemo li definitivno utvrditi da su varijable zavisne ili da su nezavisne? Obrazložite odgovor.

Definicija za nezavisnost varijabli:

Dvije slučajne varijable X i Y su **(stohastički) nezavisne** akko za sve intervale A i B, $A, B \subseteq \mathbb{R}$, vrijedi

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)\,.$$

No budući da mi ne znamo funkcije razdiobe i intervale varijabli, trebamo naći neki drugi način za odrediti statističku zavisnost. Recimo, koeficijent korelacije:

Koeficijent korelacije upućuje na mjeru linearne zavisnosti među varijablama X i Y.

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Koeficijent korelacije možemo lagano odrediti iz kovarijacijske matrice. Na našem primjeru to će biti:

$$\hat{\rho}_{X_1,X_2} = \frac{\mathrm{Cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}} = 0.5524$$

Varijable x1 i x2 su zavisne. (Zaključeno iz pozitivne korelacije)

Ne možemo definitivno tvrditi da su dvije varijable zavisne niti nezavisne. Zbog dva problema:

(1) Koeficijent korelacije pokazuje samo linearnu zavisnost, varijable mogu biti nelinearno zavisne, uz koeficijent korelacije 0.

(2) Ovo je samo procjena, procjena varira ovisno o izabranim uzorcima, može se dogoditi da smo dobili uzorke koje unose takvu pogrešku u naše procjene da donesemo krivi zaključak o zavisnosti varijabli.