

Machine Learning - Theoretical exercise 3

Téo Bouvard

January 26, 2020

Problem 1

- a) Assuming an gaussian distribution $X \sim \mathcal{N}(\mu, \Sigma)$ the maximum likelihood method states that for a set of measurements $\chi = \{x_1, \dots, x_N\}$,

$$\mu = \frac{1}{N} \sum_{k=1}^N x_k \quad (1)$$

$$\Sigma = \frac{1}{N} \sum_{k=1}^N (x_k - \mu)(x_k - \mu)^T \quad (2)$$

We first estimate the mean vectors of the two distributions using (1)

$$\begin{aligned} \mu_1 &= \frac{1}{4} \left(\begin{bmatrix} 2 \\ 6 \end{bmatrix} + \begin{bmatrix} 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 3 \\ 8 \end{bmatrix} + \begin{bmatrix} 4 \\ 6 \end{bmatrix} \right) = \frac{1}{4} \begin{bmatrix} 12 \\ 24 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \\ \mu_2 &= \frac{1}{4} \left(\begin{bmatrix} 1 \\ -2 \end{bmatrix} + \begin{bmatrix} 2.7 \\ -4 \end{bmatrix} + \begin{bmatrix} 3.3 \\ 0 \end{bmatrix} + \begin{bmatrix} 5 \\ -2 \end{bmatrix} \right) = \frac{1}{4} \begin{bmatrix} 12 \\ -8 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \end{aligned}$$

We use the estimated mean vectors to compute the covariance matrices according to (2)

$$\begin{aligned} \Sigma_1 &= \frac{1}{4} \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) = \frac{1}{4} \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix} \\ \Sigma_2 &= \frac{1}{4} \left(\begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0.09 & 0.6 \\ 0.6 & 4 \end{bmatrix} + \begin{bmatrix} 0.09 & 0.6 \\ 0.6 & 4 \end{bmatrix} + \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix} \right) = \frac{1}{4} \begin{bmatrix} 8.18 & 1.2 \\ 1.2 & 8.18 \end{bmatrix} = \begin{bmatrix} 2.045 & 0.3 \\ 0.3 & 2 \end{bmatrix} \end{aligned}$$

- b) We use the log discriminant function to compute the decision boundary. Let $x = (x_1 \ x_2)^T$ be on the decision boundary between the two distributions $\implies g_1(x) = g_2(x)$

$$-\frac{1}{2} \ln |\Sigma_1| - \frac{1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) = -\frac{1}{2} \ln |\Sigma_2| - \frac{1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \quad (3)$$

We use the following properties of the covariances matrices Σ_1 and Σ_2 to simplify this equation.

$$\begin{aligned} |\Sigma_1| &= 1 \implies \frac{1}{2} \ln |\Sigma_1| = 0 \\ |\Sigma_2| &= 4 \implies \frac{1}{2} \ln |\Sigma_2| = \ln 2 \end{aligned}$$

We then compute the two remaining terms independently.

$$\begin{aligned}\frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1) &= \frac{1}{2} \begin{bmatrix} x_1 - 3 & x_2 - 6 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1 - 3 \\ x_2 - 6 \end{bmatrix} \\ &= \frac{1}{2} \left((2(x_1 - 3))^2 + \frac{1}{2}(x_2 - 6)^2 \right)\end{aligned}$$

And

$$\begin{aligned}\frac{1}{2}(x - \mu_2)^T \Sigma_2^{-1}(x - \mu_2) &= \frac{1}{2} \begin{bmatrix} x_1 - 3 & x_2 + 2 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{3}{40} \\ -\frac{3}{40} & \frac{409}{800} \end{bmatrix} \begin{bmatrix} x_1 - 3 \\ x_2 + 2 \end{bmatrix} \\ &= \frac{1}{2} \left(\frac{1}{2}(x_1 - 3)^2 - \frac{3}{20}(x_1 - 3)(x_2 + 2) + \frac{409}{800}(x_2 + 2)^2 \right)\end{aligned}$$

Which gives us the final decision boundary equation

$$-\frac{3}{4}(x_1 - 3)^2 - \frac{1}{4}(x_2 - 6)^2 + \frac{409}{1600}(x_2 + 2)^2 - \frac{3}{40}(x_1 - 3)(x_2 + 2) + \ln 2 = 0$$

This equation describes a hyperbola whose upper part is slightly tilted towards the left. This is due to the samples of χ_2 not describing a circle, and thus orienting the decision border sideways.

- c) In order to match more closely a parabolic decision boundary, one should gather more data samples in χ_1 and χ_2 , which would smooth out irregularities in the samples.

Problem 2

Problem 3

Let x be a l -dimensional random vector following a multivariate gaussian distribution described by the probability density function p such that

$$p(x) = \frac{1}{(2\pi)^{\frac{l}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Let L be the likelihood function which describes the probability that a set of samples $\chi = \{x_1, \dots, x_N\}$ was generated by the parameters (μ, θ) . As our goal is to find the maximum of L , we use its natural logarithm in order to make differentiation easier. This does not change the location of the maximum as the natural logarithm is monotonically increasing over $]0, +\infty[$ and therefore over the image domain of p . In the following, we denote this natural logarithm of the likelihood function as \mathcal{L} .

$$\begin{aligned}\mathcal{L}(\mu, \theta) &= \ln p(\chi; \mu, \theta) \\ &= \ln \prod_{k=1}^N p(x_k; \mu, \theta) \\ &= \sum_{k=1}^N \ln p(x_k; \mu, \theta) \\ &= \sum_{k=1}^N \left[-\frac{l}{2} \ln 2\pi - \frac{1}{2} |\Sigma| - \frac{1}{2} (x_k - \mu)^T \Sigma^{-1} (x_k - \mu) \right] \\ &= -\frac{Nl}{2} \ln 2\pi - \frac{N}{2} |\Sigma| - \frac{1}{2} \sum_{k=1}^N (x_k - \mu)^T \Sigma^{-1} (x_k - \mu)\end{aligned}$$

In order to find the maximum likelihood estimate of μ , denoted $\hat{\mu}$ in the following, we find the root of $\frac{\partial \mathcal{L}}{\partial \mu}$.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu} &= -\frac{1}{2} \frac{\partial N \ln 2\pi}{\partial \mu} - \frac{1}{2} \frac{\partial |\Sigma|}{\partial \mu} - \frac{1}{2} \frac{\partial \sum_{k=1}^N (x_k - \mu)^T \Sigma^{-1} (x_k - \mu)}{\partial \mu} \\ &= -\frac{1}{2} \sum_{k=1}^N \frac{\partial (x_k - \mu)^T \Sigma^{-1} (x_k - \mu)}{\partial \mu}\end{aligned}$$

We use the matrix differentiation identity $\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^T \mathbf{A}$, with $\mathbf{x} = (x_k - \mu)$ and $\mathbf{A} = \Sigma^{-1}$ which holds if \mathbf{A} is symmetric and \mathbf{A} is not a function of \mathbf{x} . In this case, Σ^{-1} is a covariance matrix and is therefore symmetric, furthermore Σ is not a function of μ , so we can safely use this identity.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu} &= -\frac{1}{2} \sum_{k=1}^N 2(x_k - \mu)^T \Sigma^{-1} \\ &= -\Sigma^{-1} \sum_{k=1}^N (x_k - \mu)^T\end{aligned}$$

We now find the root of this equation to find $\hat{\mu}$.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu} &= 0 \\ \Sigma^{-1} \sum_{k=1}^N (x_k - \hat{\mu})^T &= 0\end{aligned}$$

If we consider the case where the covariance matrix is not equivalent to a null matrix, the previous equality becomes

$$\begin{aligned}\sum_{k=1}^N (x_k - \hat{\mu})^T &= 0 \\ \sum_{k=1}^N \hat{\mu} &= \sum_{k=1}^N x_k \\ N\hat{\mu} &= \sum_{k=1}^N x_k \\ \hat{\mu} &= \frac{1}{N} \sum_{k=1}^N x_k\end{aligned}$$