Machine Learning - Theoretical exercise 3

Téo Bouvard

January 27, 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 $\chi = \{x_1, \dots, x_N\}$,

$$\mu = \frac{1}{N}\sum_{k=1}^{N} x_k \tag{1}$$

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

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

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

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

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

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 $\implies g_1(x) = g_2(x)$

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

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

$$\left|\Sigma_1\right| = 1 \implies \frac{1}{2}\ln\left|\Sigma_1\right| = 0$$
$$\left|\Sigma_2\right| = 4 \implies \frac{1}{2}\ln\left|\Sigma_1\right| = \ln 2$$

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

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

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

a) Equation (4) can be used as a classifier by considering the inequality between the two discriminant functions. This leads to the following decision rule

$$
\text{decide}\begin{cases}\omega_1 \text{ if } (4)>0\\ \omega_2 \text{ otherwise}\end{cases}
$$

For $x=(2.5\,2.0)^T$ we have $(4)=0.745..$, which means x would be classified as belonging to ω_1. In this case, we did not properly prove that the inequality leading to decision rule was in the right direction. This is done in the following question.

b) To classify x with a Parzen window technique, we are going to make the assumption that $x\in\omega_2$, and show that this leads to a contradiction. First, we write down the inequality between the two probability density functions in the case $x\in\omega_2$.

$$
\begin{gathered}
p_1(x)<p_2(x)\\
\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{1}{V_{N_1}}\phi(u)<\frac{1}{N_2}\sum_{j=1}^{N_2}\frac{1}{V_{N_2}}\phi(u)
\end{gathered}
$$

In our case, $N_1=N_2=N$ so these factors cancel out. Furthermore, $V_N=h_N^2=\frac{h_1^2}{N}$ so this factor cancels out too.

$$\sum_{i=1}^{N} \phi(u) < \sum_{j=1}^{N} \phi(u)$$

$$\sum_{i=1}^{N} \frac{1}{2\pi \left|I\right|^{\frac{1}{2}}} e^{-\frac{1}{2}u^T I^{-1} u} < \sum_{j=1}^{N} \frac{1}{2\pi \left|I\right|^{\frac{1}{2}}} e^{-\frac{1}{2}u^T I^{-1} u}$$

$$\sum_{i=1}^{N} e^{-\frac{1}{2}\|u\|^2} < \sum_{j=1}^{N} e^{-\frac{1}{2}\|u\|^2}$$

$$\sum_{i=1}^{N} e^{-\frac{1}{2}\|\frac{x-x_i}{h_N}\|^2} < \sum_{j=1}^{N} e^{-\frac{1}{2}\|\frac{x-x_j}{h_N}\|^2}$$

$$\sum_{i=1}^{N} e^{-\frac{N}{2h_1^2}\|x-x_i\|^2} < \sum_{j=1}^{N} e^{-\frac{N}{2h_1^2}\|x-x_j\|^2}$$

To compute the numerical values appearing in this equation, we must first compute the squared distances between the test point x and each other data point.

$$\|x - x_{11}\|^2 = \frac{65}{4} \qquad \|x - x_{12}\|^2 = \frac{17}{4} \qquad \|x - x_{13}\|^2 = \frac{145}{4} \qquad \|x - x_{14}\|^2 = \frac{73}{4}$$
$$\|x - x_{21}\|^2 = \frac{73}{4} \qquad \|x - x_{22}\|^2 = \frac{901}{25} \qquad \|x - x_{23}\|^2 = \frac{116}{25} \qquad \|x - x_{24}\|^2 = \frac{89}{4}$$

Which leads to the following inequality at the test point, for $h_1 = 0.5$.

$$1.714 \times 10^{-15} < 7.568 \times 10^{-17}$$

This equation is a contradiction, which means that our hypothesis $x \in \omega_2$ is false. Therefore, $x \in \omega_1$.

c) We can evaluate the previous inequality with $h_1 = 5$

$$1.272 < 1.147$$

Which is also false, proving that $x \in \omega_1$ with this greater window size too. However the ratio of the probability densities has greatly decreased. With $h_1 = 0.5$, we had $\frac{p_2(x)}{p_1(x)} = 0.04$, whereas with $h_1 = 5$ we have $\frac{p_2(x)}{p_1(x)} = 0.9$. This is because increasing the window size leads to a smoother contribution of each sample, and therefore to a greater overlap of the two probability density functions. In other terms, the resulting probability density functions are less 'spiky' at the sample locations when increasing the window size.

d) We can use the squared distances computed in the previous question to build a k-nearest neighbors classifier. In the case $k = 1$, x is classified as ω_1 because the closest data sample x_{12} belongs to χ_1.

e) If we extend the previous classifier to the three closest neighbors, we have x_{12} and x_{11} from χ_1 in first and third position, and x_{23} from χ_2 in second position. That means x will also be classified as belonging to ω_1 because two of the three closest neighbors are from this class.

Problem 3

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

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

Let L be the likelihood function which describes the probability that a a set of samples $\chi = \{x_1, \ldots, x_N\}$ 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 $\mathcal{L}$.

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

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

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

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

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