1. Exercise

To show: $R^* = \frac{1}{2} - \frac{1}{4}\int |f_0(x) - f_1(x)|dx$

Prerequisite information:

- For two functions: $min\big(f(x), g(x)\big) = \frac{f(x)+g(x)-|f(x)-g(x)}{2}|$

Solution:

$$
\begin{aligned}
\eta(x) = \mathbb{P}(Y = 1|X = x) &= \frac{\mathbb{P}(X = x|Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X)} \\
&= \frac{f_1(x)q_1}{f_1(x)q_1 + f_2(x)q_2} \\
&= \frac{f_1(x)\frac{1}{2}}{f_0(x)\frac{1}{2} + f_1(x)\frac{1}{2}} \\
&= \frac{f_1(x)}{f_0(x) + f_1(x)}
\end{aligned}
$$

In the same way we can express $\big(1 - \eta(x)\big)$ in the following way:

$$
= \big(1 - \eta(x)\big) = \frac{f_0(x)}{f_0(x) + f_1(x)}
$$

We proceed with the definition of the Bayes Risk we start and substitute $\eta(x) and \big(1 - \eta(x)\big)$:

$$
\begin{aligned}
R^* &= \mathbb{E}\big[min(\eta(x), 1 - \eta(x)\big] \\
&= \frac{1}{2}\mathbb{E}\left[\frac{f_1(x)}{f_0(x) + f_1(x)} + \frac{f_0(x)}{f_0(x) + f_1(x)} - \left|\frac{f_1(x)}{f_0(x) + f_0(x)} - \frac{f_1(x)}{f_0(x) + f_1(x)}\right|\right] \\
&= \frac{1}{2}\mathbb{E}\left[1 - \left|\frac{f_1(x) - f_0(x)}{f_0(x) + f_0(x)}\right|\right] \\
&= \frac{1}{2} - \frac{1}{2}\mathbb{E}\left[\frac{|f_1(x) - f_0(x)|}{f_0(x) + f_0(x)}\right]
\end{aligned}
$$

By integrating with respect to the marginal probability of X and use the fact that both priors are one half, which leads to:

$$
\begin{aligned}
&= \frac{1}{2} - \frac{1}{2}\int \frac{|f_1(x) - f_0(x)|}{2\mathbb{P}(X)}\mathbb{P}(X)\ dx \\
&= \frac{1}{2} - \frac{1}{2}\int \frac{|f_1(x) - f_0(x)|}{2}\ dx \\
&= \frac{1}{2} - \frac{1}{4}\int |f_1(x) - f_0(x)|dx
\end{aligned}
$$

2. Exercise

To show: Determination of the Bayes with linear cases of the Bayes decision

Solution:

We start with the definition of the bayes classifier:

$$g^*(x) = \begin{cases} 1, \text{if } \eta(x) > \frac{1}{2} \\ \\ 0, \text{otherwise} \end{cases}$$

After some basic manipulations we substitute $\eta(x)$ and the Bayes classifier becomes:

$$g^*(x) = \begin{cases} 1, \text{ if } f_1(x)q_1 > f_0(x)q_0 \\ \\ 0, \text{ otherwise} \end{cases}$$

Next we show the case where its linear. In doing so, we first take logarithms of both sides:

$$g^*(x) = \begin{cases} 1, \text{ if } \ln(f_1(x)q_1) > \ln(f_0(x)q_0) \\ \\ 0, \text{ otherwise} \end{cases}$$

We replace the class conditional probabilities on both sides of the euqation:

$$\ln(\sqrt{(2\pi)^d|\Sigma_1|}) - \frac{1}{2}(x-m_1)^T\Sigma_1^{-1}(x-m_1) + \ln(q_1) > \ln(\sqrt{(2\pi)^d|\Sigma_0|}) - \frac{1}{2}(x-m_0)^T\Sigma_0^{-1}(x-m_0) + \ln(q_0)$$

In general we can indentify **two cases**, where Bayes decision becomes **linear**. The first case is given if the covariance matrix is diagonal with constant standard deviation ($\Sigma_i = \sigma^2 I$). After expanding the term in parenthesis there is only one quadratic term in x on both sides. However, since they both are not dependend on an i related term, the can be canceld and both terms become linear in x.

$$\ln(\sqrt{(2\pi)^d\sigma^2}) - \frac{1}{2\sigma^2}(x-m_1)^T(x-m_1) + \ln(q_1) > \ln(\sqrt{(2\pi)^d\sigma^2}) - \frac{1}{2\sigma^2}(x-m_0)^T(x-m_0) + \ln(q_0)$$

The second case is given, if both covariances are equal ($\Sigma = \Sigma_0 = \Sigma_1$). Again, after expanding the term in parenthesis the quadratic term in x can be canceled from both sides of the inequality.

$$\ln(\sqrt{(2\pi)^d|\Sigma|}) - \frac{1}{2}(x-m_1)^T\Sigma^{-1}(x-m_1) + \ln(q_1) > \ln(\sqrt{(2\pi)^d|\Sigma|}) - \frac{1}{2}(x-m_0)^T\Sigma^{-1}(x-m_0) + \ln(q_0)$$

3. Exercise

To show: Predictor function minimizing the expected loss $\mathbb{E}\big[\ell(f(X),Y)\big]$, for $\ell(y,y') = (y-y')^2$

Solution:

Let $f(x) = y'$

$$\begin{aligned}
\mathbb{E}\big[\ell(f(X),Y)\big] &= \mathbb{E}\Big[\big(Y - f(X)\big)^2\Big|X = x\Big] \\
&= \int (y-y')^2 f_{Y|X}(y|x)dy
\end{aligned}$$

We take the derivative with respect to y' and hence we get:

$$-2\int (y-y') f_{Y|X}(y|x)dy = 0$$

Splitting the integral into two parts and bringing each of them to one side leads to:

$$\begin{aligned}
&= \int y f_{Y|X}(y|x)dy - \int y' f_{Y|X}(y|x)dy = 0 \\
\int y' f_{Y|X}(y|x)dy &= \int y f_{Y|X}(y|x)dy \\
y' \int f_{Y|X}(y|x)dy &= \int y f_{Y|X}(y|x)dy
\end{aligned}$$

Finally, we notice that the conditional probability of y on the left hand side integrates to one. Furthermore, we notice that the right hand side is equal to the conditonal expectation. Hence, we found the final solution:

$$y' = f(x) = \mathbb{E}\big[Y\big|X = x\big]$$

[1]

[1] Showing that the second derivative is greater than zero is omitted at this place.

4. Exercise

To show: Minimizing the absolute loss function: $\ell(y, y') = |y - y'|$ with $\phi(y|x)$

Solution:

Likewise the last exercise we compute the conditional expected value:

$$\begin{aligned} R = \mathbb{E}\big[\ell(y, y')\big] &= \mathbb{E}\big[|y - y'|\big|X\big] \\ &= \int_{-\infty}^{y} (y - y')\phi(y|x)dy + \int_{y}^{\infty} -(y - y')\phi(y|x)dy \end{aligned}$$

By taking the first derivative, the function simplyfies as follows:

$$\begin{aligned} \frac{\partial R}{\partial y} &= -\int_{-\infty}^{y} \phi(y|x)dy + \int_{y}^{\infty} \phi(y|x)dy = 0 \\ \int_{-\infty}^{y} \phi(y|x)dy &= \int_{y}^{\infty} \phi(y|x)dy \end{aligned}$$

Finally, we notice that this quantity is minimized if y is the median.[2]

[2]Showing that the second derivative is greater than zero is omitted at this place.

5. Exercise

To show: For the K-Nearest-Neighbor rule; Show that $\lim_{n\to\infty} \|X_{(k)} - X\| = 0$, in probability

Solution:

Assume that the distance between X and its nearest neighbhour is greater than epsilon:

$$\|X_{(k)} - X\| > \epsilon \tag{1}$$

, where epsilon is a positiv number ($\epsilon > 0$). Furthermore, we define a sphere $S_{(X,\epsilon)}$, which is has its center at the X and the radius around it at ϵ. Moreover, we notice the that the meassure $\mu(S_{(X,\epsilon))})$ is a quantity greater than zero. **Iff** (1) is true the following statement holds:

$$\sum_{i=0}^{n} \mathbf{1}_{X_i \in S_{(X,\epsilon)}} < k \tag{2}$$

The indicator function counts 1, each time a point is in the sphere and obviously this number has to be smaller than k (as long as (1) is true). Dividing both sides of (2) by n leads to:

$$\frac{1}{n}\sum_{i=0}^{n} \mathbf{1}_{X_i \in S_{(X,\epsilon)}} < \frac{k}{n}$$

For increasing n, the left hand side converges to a value greater than zero (by SLN), which is the meassure $\mu(S_{(X,\epsilon))})$. However the righthand goes to zero (for k fixed). Hence in the limit this induces a **contradiction** of (1) for n goes to infinity.

6. Exercise

To show: Expected risk of the 1-nearest neighbor classifier is greater than 1/4, but $R^* = 0$

Solution:

Since the Bayes Risk is zero (R^*=0), we first conclude that $\eta(x) = \{0, 1\}$, which implies that both classes are separable. For **infinite** sample size the expected risk ($\mathbb{E}\big[R(g_n)\big]$) will be equal to the Bayes Risk and hence 0 (said to be "universally consistent").
The risk in general of some abitrary classifier can be between zero and one.
For this exercise we now consider m disjoint intervalls (and two classes in each intervall with zero and one) and n data points.
If there is **at least** one data point in each intervall and class, another entering data point data point will be classified correctly. However assuming m to be much greater than n implies a possible risk due to the fact that some intervalls might not be covered and the nearest neighbour classifier assigns the wrong label. Therefore, recalling g_n for the nearest neighbour:

$$g_n(x) = Y_1(x)$$

where $Y_1(x)$ is the label of the nearest neighbour. Now assume there are n data points already in m intervalls. We assume that a new data point is uniformly popping up in one intervall. Since this is random the risk itself becomes a random variable (between zero and one). More detailed, the classifier classifies 0 or 1 and this is either wrong or right. Hence it expected value can to be assumed $\frac{1}{2}$ and hence:

$$\mathbb{E}\big[R(g_n)\big] = \frac{1}{2} > \frac{1}{4}$$

In extension there can be shown that for any classifier the expected risk is:

$$\mathbb{E}\big[R(g_n)\big] \geq \frac{1}{2} - \epsilon$$

Where ϵ is a small number, which implies the same conclusion

[3]

[3] c.f: A L. Devroye, L. Györfi, G. Lugosi (1996): Probabilistic Theory of Pattern Recognition, TH.7.1, page 124