Exercise 7

To show:

Let the joint distribution of (X,Y) be such that X is uniform on the interval [0, 1], and for all $x \in [0,1]$, $\eta(x) = x$. Determine the prior probabilities $\mathbf{P}\{Y = 0\}$, $\mathbf{P}\{Y = 1\}$ and the class-conditional densities $f(x|Y = 0)$ and $f(x|Y = 1)$.
Calculate R^*, R_{1-NN} , and R_{3-NN} (i.e., the Bayes risk and the asymptotic risk of the 1-, and 3-nearest neighbor rules).

Solution:

A)

Firstly, compute the prior as by marginalising $\eta(x)$. Notice that for a uniform random variable (here X) the density is one if defined on interval [0, 1] and zero otherwise. Hence:

$$p(x) = \begin{cases} \frac{1}{1-0} & \text{for } 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

Using that fact we get:

$$\boldsymbol{P}(Y = 1) = \int \boldsymbol{P}(Y = 1|X)\boldsymbol{P}(X)dx = \int_0^1 x \; dx = \frac{1}{2}$$

By computing the complement of the previous equation, one obtains the other prior:

$$\boldsymbol{P}(Y = 0) = 1 - \boldsymbol{P}(Y = 1) = \frac{1}{2}$$

B)

Aplying Bayes Theorem to comppute the the class conditional probabilities including the results from above:

$$\boldsymbol{P}(Y = 1|X) = \frac{\boldsymbol{P}(X|Y = 1)\boldsymbol{P}(Y = 1)}{\boldsymbol{P}(X)}$$
$$f(x|Y = 1) = \boldsymbol{P}(X|Y = 1) = \frac{\boldsymbol{P}(Y = 1|X)\boldsymbol{P}(X)}{\boldsymbol{P}(Y = 1)} = 2x$$

Using the marginal probability of x we compute f_0

$$\begin{aligned}
\boldsymbol{P}(X) &= \boldsymbol{P}(Y=1)\boldsymbol{P}(X|Y=1) + \boldsymbol{P}(Y=0)\boldsymbol{P}(X|Y=0) \\
&= \frac{1}{2}\boldsymbol{P}(X|Y=1) + \frac{1}{2}\boldsymbol{P}(X|Y=0) = 1 \\
f(x|Y=0) &= \boldsymbol{P}(X|Y=0) = 2x-2
\end{aligned}$$

C)

Finally compute the risks as follows. Since the priors are both one half use the equation from the last homework:

$$\begin{aligned}
R^* &= \frac{1}{2} - \frac{1}{4}\int |f_0(x) - f_1(x)|dx \\
&= \frac{1}{2} - \frac{1}{4}\int_0^1 |2-4x|dx \\
&= \frac{1}{4}
\end{aligned}$$

Next, compute the risk of the first nearest neighbhor as follows:

$$\begin{aligned}
R^{1-NN} &= \mathbb{E}\big[2\eta(x)(1-\eta(x)\big] \\
&= 2\mathbb{E}[x(1-x)] \\
&= 2\mathbb{E}[x-x^2] \\
&= 2\int_0^1 x - x^2 dx \\
&= \frac{1}{3}
\end{aligned}$$

Finally compute the risk of the 3-Nearest-Neighbhor classifier. Following the binomial expansion the risk for the 3-Nearest-Neighbhor classifier becomes:

$$\begin{aligned}
R^{3-NN} &= \mathbb{E}\big[\eta(x)^3(1-\eta(x)) + 6\eta(x)^2(1-\eta(x))^2 + \eta(x)(1-\eta(x))^3\big] \\
&= \mathbb{E}\big[4x^4 - 8x^3 + 3x^2 + x\big] \\
&= \int_0^1 4x^4 - 8x^3 + 3x^2 + x \; dx \\
&= \frac{3}{10}
\end{aligned}$$

Exercise 8

To show:

Let $X_1, \ldots, X_n$ be independent random variables taking values in $[0,1]$. Denote $m = \mathbb{E}\Big[\sum_{i=1}^n X_i\Big]$. Prove that for any $t \geq m$,

$$\boldsymbol{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq (\frac{m}{t})^t e^{t-m}$$

Hint: Use Chernov's bounding technique. Use the fact that by convexity of $e^{\lambda x}$, $e^{\lambda x} \leq xe^{\lambda} + (1-x)$.

Solution:

Use the definition from class and choose t' to be $t' = t - m$. Use MGF on both sides of the inequality

$$\boldsymbol{P}\left(\sum_{i=1}^n X_i - m \geq t'\right) = \boldsymbol{P}\left(\sum_{i=1}^n X_i \geq t\right) = \boldsymbol{P}\left(e^{\lambda\left(\sum_{i=1}^n X_i\right)} \geq e^{\lambda t}\right)$$

Apply **Markow's Inequality** to obtain the following:

$$\leq \frac{\mathbb{E}\left[e^{\lambda(\sum_{i=1}^n X_i)}\right]}{e^{\lambda t}}$$

Using **convexity** and simplyfing the expression leads to:

$$\begin{aligned}
&\leq e^{-\lambda t}\left(\mathbb{E}\big[\sum_{i=1}^n X_i e^{\lambda}\big] + \mathbb{E}\big[1 - \sum_{i=1}^n X_i\big]\right) \\
&= e^{-\lambda t}\left(e^{\lambda}\mathbb{E}\big[\sum_{i=1}^n X_i\big] + 1 - \mathbb{E}\big[\sum_{i=1}^n X_i\big]\right) \\
&= e^{-\lambda t}\big(e^{\lambda} m + 1 - m\big)
\end{aligned}$$

Choose λ to be $\lambda = \ln\left(\frac{t}{m}\right)$ and plug it into the last expression to obtain:

$$\begin{aligned}
&= e^{-\ln\left(\frac{t}{m}\right)t}\big(e^{\ln\left(\frac{t}{m}\right)} m + 1 - m\big) \\
&= \left(\frac{t}{m}\right)^{-t}\left(\frac{t}{m} m + 1 - m\right) \\
&= \left(\frac{m}{t}\right)^{t}\big(1 + (t - m)\big)
\end{aligned}$$

By the definition of the exponential function, $e^x \geq 1 + x$ holds and hence we showed:

$$\leq \left(\frac{m}{t}\right)^{t} e^{(t-m)}$$

Exercise 9

To show:

Let R_{k-NN} denote the asymptotic risk of the k-nearest neighbor clasifier, where k is an odd positive integer. Use the expression of R_{k-NN} found in class to show that

$$R_{k-NN} - R^* \leq \sup_{p\in[0,1/2]} (1-2p)\boldsymbol{P}\big(Bin(k,p) > k/2\big)$$

Use Hoeffding's inequality to deduce from this that

$$R_{k-NN} - R^* \leq \frac{1}{\sqrt{ke}}$$

Solution:

Start again by the definition derived in class:

$$R_{k-NN} = R^* + \mathbb{E}\left[|2\eta(x)-1|\boldsymbol{P}\left(\mathrm{Bin}\big(k, \min[\eta(x), 1-\eta(x)]\big) > \frac{k}{2}\middle| X\right)\right]$$

Set p as the mean of the binomial and therefore $p = min\big(\eta(x), 1-\eta(x)\big)$. Consider two cases, where $\eta(x)$ is either smaller or greater than $1/2$. For both cases it holds: $|2\eta(x)-1| = 1-2p$. Notice that this true, no matter what x is and hence:

$$= R^* + (1-2p)\boldsymbol{P}\left(\mathrm{Bin}\big(k,p\big) > \frac{k}{2}\right)$$

Notice that the last is expression is naturally upper bounded by its supremum and p takes value within zero and a half. From this follows the first soluton

$$\leq R^* + \sup_{p\in[0,1/2]} (1-2p)\boldsymbol{P}\left(\mathrm{Bin}\big(k,p\big) > \frac{k}{2}\right)$$

For the second part substract the mean of the binomial on both sides of the inequality and divide k:

$$R_{k-NN} - R^* \leq \sup_{p\in[0,1/2]} (1-2p)\boldsymbol{P}\left(\frac{\mathrm{Bin}(k,p)-kp}{k} > \frac{1}{k}\left(\frac{k}{2} - kp\right)\right)$$
$$R_{k-NN} - R^* \leq \sup_{p\in[0,1/2]} (1-2p)\boldsymbol{P}\left(\frac{\mathrm{Bin}(k,p)-kp}{k} > \frac{1}{2} - p\right)$$

Apply **Hoeffding's Inequality** as an upper bound for the last expression:

$$\leq \sup_{p\in[0,1/2]} (1-2p)e^{-2k(\frac{1}{2}-p)^2}$$

Set u as $u = 1 - 2p$ with $u \in [0, 1]$ and plug it into the last expression

$$= \sup_{u \in [0,1]} u e^{-\frac{k}{2} u^2}$$

The final step is achieved by maximizing the right hand side of the last quantity. By taking the derivative and solving for u one gets:

$$\begin{aligned} \frac{\partial}{\partial u} &= e^{-\frac{k}{2} u^2} - k u^2 e^{-\frac{k}{2} u^2} = 0 \\ u &= \frac{1}{\sqrt{k}} \end{aligned}$$

Substitute that in the former equation to get:

$$\begin{aligned} &= \frac{1}{\sqrt{k}} e^{-\frac{k}{2} (\frac{1}{\sqrt{k}})2} = \frac{1}{\sqrt{k}} e^{-\frac{1}{2}} \\ &= \frac{1}{\sqrt{ke}} \end{aligned}$$

Exercise 10

To show:

(RADEMACHER AVERAGES.) Let A be a bounded subset of $\mathbb{R}^n$. Define the Rademacher average.

$$R_n(A) = \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right|$$

where $\sigma_1, \ldots, \sigma_n$ are independent random variables with $\mathbf{P}\{\sigma_i = 1\} = \mathbf{P}\{\sigma_i = -1\} = 1/2$ and $a_1, \ldots, a_n$ are the components of the vector a. Let $A, B \subset \mathbb{R}^n$ be bounded sets and let $c \in \mathbb{R}$ be a constant. Prove the following "structural" results:

$$R_n(A \cup B) \leq R_n(A) + R_n(B), \qquad Rn(c \cdot A) = |c| R_n(A), \qquad R_n(A \oplus B) \leq R_n(A) + R_n(B)$$

where $c \cdot A = \{ca : a \in A\}$ and $A \oplus B = \{a + b : a \in A, b \in B\}$. Moreover, if $absconv(A) = \{\sum_{j=1}^{N} c_j a^{(j)} : N \in \mathbb{N}, \sum_{j=1}^{N} |c_j| \leq 1, a^{(j)} \in A\}$ is the absolute convex hull of A,then

$$R_n(A) = R_n(absconv(A))$$

Solution:

A)

The first case follows simply from the union bound.

$$\begin{aligned}
R_n(A \cup B) &= \mathbb{E} \sup_{d \in A \cup B} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i d_i \right| \\
&\leq \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right| + \mathbb{E} \sup_{b \in B} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i b_i \right| \\
&= R_n(A) + R_n(B)
\end{aligned}$$

B)

In the second case we can just pull out the constant term, since it neither depends on i nor the expected value:

$$\begin{aligned}
R_n(c \cdot A) &= \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i c a_i \right| \\
&= |c| \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right|
\end{aligned}$$

C)

Start by writing it out:

$$R_n(A \oplus B) = \mathbb{E} \sup_{a \in A, b \in B} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i (a_i + b_i) \right|$$

This can be bounded using the **Triangle Inequality**:

$$\begin{aligned} &\leq \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right| + \mathbb{E} \sup_{b \in B} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i b_i \right| \\ &= R_n(A) + R_n(B) \end{aligned}$$

D)

First, notice that A is subset of the absolute convex hull and hence $A \subset absconv(A)$ and hence conclude:

$$R_n(A) \leq R_n(absconv(A))$$

In the following in order to show that they are actually equal, one have to show also that $R_n(absconv(A)) \leq R_n(A)$. Hence, choose a convex combination of A (with $\sum_{i=1}^{N} |c_i| \leq 1$).

$$\begin{aligned} R_n\left(\sum_{i=1}^{N} |c_i| A \right) &\leq \sum_{i=1}^{N} R_n\big(|c_i| A\big) \\ &= \sum_{i=1}^{N} |c_i| R_n\big(A\big) \end{aligned}$$

Notice, that this convext combination of A is smaller or euqal to the set itself. Following that:

$$\leq R_n\big(A\big)$$

Using that property it turnes out:

$$R_n(A) \leq R_n(absconv(A)) \leq R_n(A) \qquad \Longrightarrow \qquad R_n(A) = R_n(absconv(A))$$

Question 11

To show:

A half plane is a set of the form $H_{a,b,c} = \{(x, y) \in \mathbb{R}^2 : ax + bx \geq c\}$ for some real numbers a, b, c. Determine the $n-$the shatter coefficient of the classes

$$\mathcal{A}_0 = \{H_{a,b,0} : a, b \in \mathbb{R}\} \quad \textit{and} \quad \mathcal{A}_0 = \{H_{a,b,c} : a, b, c \in \mathbb{R}\}$$

Solution:

Following Corollary 13.1.[1] for classes of half spaces having the form $\{x : ax \geq b\}$ it holds that $V_{\mathcal{A}}$ is $V_{\mathcal{A}} = d + 1$. Furthermore, the shatter coefficiant becomes:

$$S(\mathcal{A}, n) = \sum_{i=0}^{d} \binom{n-1}{i}$$

This follows by applying the theorem of Cover (1965,outlined on page 237) and choosing basis function, which are simply the coordinates of the space $\phi_1(x) = x^{(1)}, \ldots, \phi_d(x) = x^{(d)}, \phi_{d+1}(x) = 1$. Consider now the explicit cases of the exercise. In the first case the constant is zero and hence:

$$V_{\mathcal{A}_0} = 2$$

Applying the formular of the shatter coefficient leads to:

$$S(\mathcal{A}_0, n) = 2 \sum_{i=0}^{1} \binom{n-1}{i} = 2\big(1 + (n-1)\big) = 2n$$

Notice for the second case we have a constant and hence the VC-Dimension grows (considereing $\phi_{d+1}(x) = 1$). Hence for the second case:

$$V_{\mathcal{A}_1} = d + 1 = 3$$

Again by applying the formula above one obtain the final result:

$$S(\mathcal{A}_0, n) = 2 \sum_{i=0}^{2} \binom{n-1}{i} = 2\left(1 + (n-1) + \frac{1}{2}(n-2)(n-1)\right) = n^2 - n + 2$$

[1] c.f: A L. Devroye, L. Györfi, G. Lugosi (1996): Probabilistic Theory of Pattern Recognition, page 237 et seqq.