1. (Multi-class pattern classification) In the two-class pattern classification problem, we have seen that we can express the excess risk of a classifier \( f \) (that is, the amount by which its risk exceeds the Bayes risk) as

\[
R(f) - R^* = E \left( 1 \left[ f(X) \neq f^*(X) \right] | 2\eta(X) - 1 \right).
\]

It is easy to check that we can write this as

\[
R(f) - R(f^*) = E \left( \max_{y \in Y} P(Y = y|X) - P(Y = f(X)|X) \right).
\]  

(1)

Consider the following multi-class problem: \( X \) is the observation space, \( Y = \{1, \ldots, k\} \) is the outcome space, and \( P \) is a probability distribution on \( X \times Y \). Write \( \eta_y(x) = P(Y = y|X = x) \). The Bayes decision function \( f^* : X \rightarrow Y \) (that is, the classifier that minimizes risk, \( R(f) = P(Y \neq f(X)) \)) is given by

\[
f^*(x) = \arg \max_{y \in Y} \eta_y(x).
\]

(a) Show that, for any \( f : X \rightarrow Y \), (1) is also true in this case.

(b) Given estimates \( \hat{\eta}_y \) of the conditional probability functions \( \eta_y \), the plug-in decision function is

\[
\hat{f}(x) = \arg \max_{y \in Y} \hat{\eta}_y(x).
\]

Derive an upper bound on the excess risk, \( R(\hat{f}) - R^* \), in terms of the function

\[
x \mapsto \max_{y \in Y} |\eta_y(x) - \hat{\eta}_y(x)|.
\]

2. (Perceptron algorithm) Implement the perceptron algorithm. Consider the following two data sets, which are labelled according to a linear threshold function. (Define \( e_i \in \{0, 1\}^d \) as the unit vector with the \( i \)th component equal to 1.)

(a) Choose \( x_i \in \{0, 1\}^{d+1} \) with \( x_i = e_i + e_{d+1}, \ y_i = 1 \) for \( i = 1, \ldots, d \), and \( x_{d+1} = e_{d+1}, \ y_{d+1} = -1 \).

(b) Construct a set \( S_n \subset \{0, 1\}^{2n} \) as follows. Set \( S_1 = \{01, 10, 11\} \), and for \( i \geq 1 \),

\[ S_{i+1} = \{ x01 : x \in S_i \} \cup \{ 11 \ldots 100 \ldots 011 \}. \]

For each element \( b = (b_1 \cdots b_{2n}) \) of \( S_n \), define an associated boolean value \( v_b \in \{0, 1\} \) as

\[
v_b = b_{2n} \land (b_{2n-1} \lor (b_{2n-2} \land (b_{2n-3} \lor (\cdots (b_2 \land b_1) \cdots)))
\]

where \( \land \) denotes conjunction (\( b_1 \land b_2 \) is 1 only for \( b_1 = b_2 = 1 \)) and \( \lor \) denotes disjunction (\( b_1 \lor b_2 \) is 0 only for \( b_1 = b_2 = 0 \)). Finally, set

\[
\{(x_1, y_1), \ldots, (x_{2n+1}, y_{2n+1})\} = \{(b, 1), 2v_b - 1 : b \in S_n\}.
\]

Plot the number of updates made by the perceptron algorithm with these two data sets, as a function of \( n \) for \( d = 2n \). Comment on the difference. Explain why it occurs.

3. (Lower bounds on risk in pattern classification) We say that a class \( F \) of \( \{\pm1\} \)-valued functions defined on \( X \) shatters a set \( \{x_1, \ldots, x_n\} \subseteq X \) if

\[
\{(f(x_1), \ldots, f(x_n)) : f \in F\} = \{\pm1\}^n,
\]

that is, if \( F \) can compute all \( 2^n \) dichotomies of the set. The Vapnik-Chervonenkis dimension of \( F \) is

\[
d_{VC}(F) = \max \{n : F \text{ shatters some } \{x_1, \ldots, x_n\} \subseteq X\}.
\]

We have seen the following minimax lower bound on expected risk for the class of linear threshold functions on \( \mathbb{R}^d \):

1
Theorem 1 For any classification rule $f_n$ and any $n > 1$, there is a probability distribution $P$ on $\mathcal{X} \times \{\pm 1\}$ for which some $f \in F$ has $R(f) = 0$ but

$$\mathbb{E}R(f_n) \geq \left( \frac{\min(n, d) - 1}{2n} \right) \left( 1 - \frac{1}{n} \right)^n.$$  

(a) Show that this result remains true for $F$ an arbitrary set of functions with $d_{VC}(F) \geq d$.

(b) Hence prove minimax lower bounds on expected risk for the following classes of binary-valued functions.

i. **Decision stumps on** $\mathbb{R}^d$,

   $$F = \{ \theta_{H(i,a,s)} : 1 \leq i \leq d, a \in \mathbb{R}, s \in \{\pm 1\} \},$$

   where

   $$\theta_S(x) = \begin{cases} 1 & \text{if } x \in S, \\ -1 & \text{otherwise}, \end{cases}$$

   and $H(i,a,s)$, for $1 \leq i \leq d$, $a \in \mathbb{R}$ and $s \in \{\pm 1\}$, is the halfspace

   $$H(i,a,s) = \{ x \in \mathbb{R}^d : s(x_i - a) > 0 \}.$$  

ii. **Indicators of unions of intervals** in $\mathbb{R}$,

   $$F = \{ \theta_U : U \text{ is a union of up to } k \text{ intervals} \}.$$  

iii. **Indicators of convex sets** in $\mathbb{R}^2$,

   $$F = \{ \theta_S : S \subseteq \mathbb{R}^2 \text{ and } S \text{ convex} \}.$$  

iv. **Decision trees on** $\mathbb{R}^d$ with $m$ nodes. A decision tree is a binary tree $T$ with nodes labeled with decision stumps on $\mathbb{R}^d$. It computes a function $f_T$ defined recursively as follows. Suppose the root of $T$ is a decision stump $s$. If the root has no descendants, $f_T = s$, otherwise if $T$ has left and right subtrees $L$ and $R$, then

   $$f_T(x) = \begin{cases} f_L(x) & \text{if } s(x) = -1, \\ f_R(x) & \text{if } s(x) = 1. \end{cases}$$