1 Risk Minimization with Doubt

Suppose we have a classification problem with classes labeled 1, \ldots, c and an additional “doubt” category labeled \( c + 1 \). Let \( f : \mathbb{R}^d \to \{1, \ldots, c + 1\} \) be a decision rule. Define the loss function

\[
L(f(x), y) = \begin{cases} 
0 & \text{if } f(x) = y, \ f(x) \in \{1, \ldots, c\}, \\
\lambda_c & \text{if } f(x) \neq y, \ f(x) \in \{1, \ldots, c\}, \\
\lambda_d & \text{if } f(x) = c + 1
\end{cases}
\]

where \( \lambda_c \geq 0 \) is the loss incurred for making a misclassification and \( \lambda_d \geq 0 \) is the loss incurred for choosing doubt. In words this means the following:

- When you are correct, you should incur no loss.
- When you are incorrect, you should incur some penalty \( \lambda_c \) for making the wrong choice.
- When you are unsure about what to choose, you might want to select a category corresponding to “doubt” and you should incur a penalty \( \lambda_d \).

The risk of classifying a new data point \( x \) as class \( f(x) \in \{1, 2, \ldots, c + 1\} \) is

\[
R(f(x)|x) = \sum_{i=1}^{c+1} L(f(x), i) P(Y = i|x).
\]

(a) Show that the following policy \( f_{opt}(x) \) obtains the minimum risk:

- (R1) Find class \( i \) such that \( P(Y = i|x) \geq P(Y = j|x) \) for all \( j \), meaning you pick the class with the highest probability given \( x \).
- (R2) Choose class \( i \) if \( P(Y = i|x) \geq 1 - \frac{\lambda_d}{\lambda_c} \)
- (R3) Choose doubt otherwise.
(b) How would you modify your optimum decision rule if $\lambda_d = 0$? What happens if $\lambda_d > \lambda_c$? Explain why this is or is not consistent with what one would expect intuitively.
2 The Classical Bias-Variance Tradeoff

Consider a random variable $X$, which has unknown mean $\mu$ and unknown variance $\sigma^2$. Given $n$ iid realizations of training samples $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ from the random variable, we wish to estimate the mean of $X$. We will call our estimate of $\mu$ the random variable $\hat{X}$, which has mean $\hat{\mu}$. There are a few ways we can estimate $\mu$ given the realizations of the $n$ samples:

1. Average the $n$ samples: $\frac{x_1 + x_2 + \ldots + x_n}{n}$.
2. Average the $n$ samples and one sample of 0: $\frac{x_1 + x_2 + \ldots + x_n}{n+1}$.
3. Average the $n$ samples and $n_0$ samples of 0: $\frac{x_1 + x_2 + \ldots + x_n}{n+n_0}$.
4. Ignore the samples: just return 0.

In the parts of this question, we will measure the bias and variance of each of our estimators. The bias is defined as

$$E[\hat{X} - \mu]$$

and the variance is defined as

$$\text{Var}[\hat{X}]$$.

(a) What is the bias of each of the four estimators above?

(b) What is the variance of each of the four estimators above?
(c) Suppose we have constructed an estimator \( \hat{X} \) from some samples of \( X \). We now want to know how well \( \hat{X} \) estimates a new independent sample of \( X \). Denote this new sample by \( X' \). Derive a general expression for \( E[(\hat{X} - X')^2] \) in terms of \( \sigma^2 \) and the bias and variance of the estimator \( \hat{X} \). Similarly, derive an expression for \( E[(\hat{X} - \mu)^2] \). Compare the two expressions and comment on the differences between them.

(d) It is a common mistake to assume that an unbiased estimator is always “best.” Let’s explore this a bit further. Compute \( E[(\hat{X} - \mu)^2] \) for each of the estimators above.

(e) Demonstrate that the four estimators are each just special cases of the third estimator, but with different instantiations of the hyperparameter \( n_0 \).

(f) What happens to bias as \( n_0 \) increases? What happens to variance as \( n_0 \) increases?