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

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

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 λ_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 λ_d.

In lecture, you saw a definition of risk over the expectation of data points. We can also define the risk of classifying a new individual data point $\mathbf{x}$ as class $f(\mathbf{x}) \in \{1, 2, \ldots, c+1\}$, and reason about what the risk would be for all possible values of $\mathbf{x}$. We define the risk as

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

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

- (**R1**) Find the non-doubt class i such that $P(Y = i|\mathbf{x}) \geq P(Y = j|\mathbf{x})$ for all j, meaning you pick the class with the highest probability given x.
- (**R2**) Choose class i if $P(Y = i|\mathbf{x}) \geq 1 - \frac{\lambda_d}{\lambda_c}$
- (**R3**) Choose doubt otherwise.

Hint: It will first help you to approach the risk function on a case-by-case basis to help simplify the expression. What is the risk if we choose the "doubt" class? What is it if we choose a non-doubt class as our prediction?

In order to prove that $f_{opt}(x)$ minimizes risk, consider proof techniques that show that $f_{opt}(x)$ "stays ahead" of all other policies that *don't* follow these rules. For example, you could take a proof-by-contradiction approach: assume there exists some other policy, say $f'(x)$, that minimizes risk more than $f_{opt}(x)$. What are the scenarios where the predictions made by $f_{opt}(x)$ and $f'(x)$ might differ? In these scenarios, and based on the rules above that $f_{opt}(x)$ follows, why would $f'(x)$ not be able to beat $f_{opt}(x)$ in risk minimization?

(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 μ and unknown variance σ^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 μ the random variable $\hat{X}$, which has mean $\hat{\mu}$. There are a few ways we can estimate μ 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 σ^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?

