CHAPTER 5

The Optimal Bayes Decision Rule

In the previous chapter, we formulated the following pattern recognition problem. There are two classes 0 and 1 with prior probabilities $P(0)$ and $P(1)$. We observe a vector of features $\overline{x}$ of the object that is related to the class of the object through conditional probabilities $P(\overline{x}|0)$ and $P(\overline{x}|1)$ (in the case of discrete-valued features) or conditional densities $p(\overline{x}|0)$ and $p(\overline{x}|1)$ (in case of continuous-valued features). We wish to decide whether the object belongs to class 0 or class 1, and we would like a decision rule that minimizes the probability of error.

With no observations, the obvious way to minimize the probability of error is to choose the class with larger prior probability. Now we would like to find out how to incorporate the fact that we get to observe the feature vector $\overline{x}$. A natural idea is to try to compute the conditional probabilities of the classes, given that we have observed the feature vector $\overline{x}$, namely $P(0|\overline{x})$ and $P(1|\overline{x})$. These give our "updated" assessment of the probability that the unknown object belongs to each of the two classes *after* observing the feature vector $\overline{x}$. If we had these probabilities, then the obvious way to try to minimize the probability of error would be to decide the class with larger conditional probability. But how can we compute the conditional probabilities that we need?

5.1 BAYES THEOREM

Recall the notion of the conditional probability of an event (Chapter 2). If A and B are two events with $P(A) > 0$, then the conditional probability of B given A, denoted by $P(B|A)$, is defined as

$$P(B|A) = \frac{P(B\&A)}{P(A)}. \tag{5.1}$$

An Elementary Introduction to Statistical Learning Theory, First Edition.
Sanjeev Kulkarni and Gilbert Harman.