Naive Bayes with Quantitative Predictors

Problem Setup

Our goal is to predict a response—in this case, a categorical response—using quantative predictors. Let's assume that the response variable is labeled C for category, and there are k possibilities so we can have $C = 1$ through $C = k$ as the possible response values. Assume that X_1 through X_p are the p quantitative predictors.

We want to predict the conditional probabilities of $C = 1$ through $C = k$ given $X_1, \ldots, X_p$.

Bayes' Theorem

Using Bayes' Theorem, we can write

$$P(C = 1|X_1, \ldots, X_p) = \frac{P(X_1, \ldots, X_p|C = 1)P(C = 1)}{P(X_1, \ldots, X_p|C = 1)P(C = 1) + \cdots + P(X_1, \ldots, X_p|C = k)P(C = k)}.$$

The value on the left side is what we want to calculate, so we'll need to evaluate the terms on the right side.

First, there are the prior probabilities to evaluate, that is, the values $P(C = 1)$ through $P(C = k)$. In the lab, task #4 shows how to use the sample proportions for these probabilities.

Second, we have to evaluate probabilities like $P(X_1, \ldots, X_p|C = 1)$. We'll make two simplifying assumptions to evaluate these probabilities:

Simplifying Assumption #1

We will assume that when we condition on the value of C, the X_1 through X_p values are independent. This is a dubious assumption—it's why we call this method "naive Bayes"—but it simplifies the calculations. Using this independence assumption gives

$$P(X_1, \ldots, X_p|C = 1) = P(X_1|C = 1) \times P(X_2|C = 1) \times \cdots \times P(X_p|C = 1).$$

Simplifying Assumption #2

For this lab, we will also assume that for each value of C, the quantitative predictors are all normally distributed. This assumption may also be dubious, but in some problems it's not a bad approximation.

For the purposes of this assumption, we'll need to find the means and standard deviations of all quantitative variables for all training set observations in class $C = 1$. For instance, let's say that the X_1 values for those measurements with $C = 1$ have sample mean $\hat{\mu}_1$ and sample standard deviation S_1. Then we will use the formula for the normal curve for $P(X_1|C = 1)$, as follows:

$$P(X_1|C = 1) = \frac{1}{S_1\sqrt{2\pi}} \exp\left\{\frac{-1}{2S_1^2}(X_1 - \hat{\mu}_1)^2\right\}.$$

In this formula, X_1 is the value of a single observation whose class C we are trying to predict based on its $X_1, \ldots, X_p$ values. For the lab, Tasks 5 and 6 show how to use a function from the scipy.stats library in Python to make the calculations.