Naive Bayes and Linear Classifiers

CS 5350/6350: Machine Learining

This note shows that a binary naive Bayes classifier is a linear classifier.

We will denote inputs by d dimensional vectors, $\mathbf{x} = (x_1, x_2, \cdots, x_d)^T$. (Note that bold face $\mathbf{x}$ denotes the entire feature vector, while each individual feature will be denoted using normal font with the appropriate subscript.) We will assume that our features x_j are all binary (i.e., 0 or 1). Our classifier will predict the label 1 if $P(y = 1|\mathbf{x}) \geq P(y = 0|\mathbf{x})$. Or equivalently,

$$\frac{P(\mathbf{x}|y = 1)P(y = 1)}{P(\mathbf{x}|y = 0)P(y = 0)} \geq 1 \tag{1}$$

By the naive Bayes assumption, we have $P(\mathbf{x}|y) = \prod_{j=0}^{d} P(x_j|y)$. This lets us rewrite the condition for predicting the label 1 from (1) as follows:

$$\frac{P(y = 1)}{P(y = 0)} \cdot \prod_{i=0}^{d} \frac{P(x_j|y = 1)}{P(x_j|y = 0)} \geq 1 \tag{2}$$

To simplify notation, let us denote $P(y = 1)$ by p, $P(x_j = 1|y = 1)$ by a_j and $P(x_j = 1|y = 0)$ by b_j.

Using this notation, we can write

$$P(x_j|y = 1) = a_j^{x_j} \, (1 - a_j)^{(1-x_j)}$$

Note that we can do this because our features are binary and one of x_j or $1 - x_j$ will be zero. Similarly, we can write $P(x_j|y = 0) = b_j^{x_j} \, (1 - b_j)^{(1-x_j)}$.

Using this notation in (2), we get the following equivalent condition for predicting $y = 1$:

$$\frac{p}{1 - p} \cdot \prod_{j=0}^{d} \frac{a_j^{x_j} \, (1 - a_j)^{(1-x_j)}}{b_j^{x_j} \, (1 - b_j)^{(1-x_j)}} \geq 1 \tag{3}$$

Collecting the constants together, we get

$$\left(\frac{p}{1 - p} \prod_{j=0}^{d} \frac{1 - a_j}{1 - b_j} \right) \cdot \prod_{j=0}^{d} \left(\frac{a_j}{b_j} \cdot \frac{1 - b_j}{1 - a_j} \right)^{x_j} \geq 1 \tag{4}$$

Taking log,

$$\log \left(\frac{p}{1 - p} \prod_{j=0}^{d} \frac{1 - a_j}{1 - b_j} \right) + \sum_{j=0}^{d} x_j \log \left(\frac{a_j}{b_j} \cdot \frac{1 - b_j}{1 - a_j} \right) \geq 0 \tag{5}$$

For any input $\mathbf{x}$, the first term in this summation is a constant because it does not have any x_j terms. Let us denote it by $b = \log \left(\frac{p}{1-p} \prod_{j=0}^{d} \frac{1-a_j}{1-b_j} \right)$. Further, let us denote $\log \left(\frac{a_j}{b_j} \cdot \frac{1-b_j}{1-a_j} \right)$ by w_j. Substituting these, we get the familiar expression

$$b + \sum_{j=0}^{d} x_j w_j \geq 0 \tag{6}$$

Recall that we obtained this condition for predicting that the label is 1. This means that our classifier is a linear classifier.

Exercise Suppose the input variables were not binary. This means that $P(x_j|y)$ have to be defined using a probability density functions, one for each value of y and j. Suppose these were Gaussian. Show that the decision boundary is still linear.