Naive Bayes Classifier

Marcos Daniel Calderón-Calderón

May 24, 2020

1 Introduction

The Naive Bayesian classifier is based on Bayes' theorem. In this type of classifier, $\mathbf{x} = (x_1, \ldots, x_n)$ represents the feature vector (every x_i is an independient variable). Bayes' theorem provides a way of calculating the posterior probability $P(c_k \mid \mathbf{x})$ (see equation (1.1)).

$$P(c_k \mid \mathbf{x}) = \frac{P(c_k)P(\mathbf{x} \mid c_k)}{P(\mathbf{x})}. \tag{1.1}$$

Using Bayesian probability terminology, equation (1.1) can be written as:

$$Posterior = \frac{Prior \times Likelihood}{Evidence}. \tag{1.2}$$

In equation (1.2), only the numerator is important. The denominator is just a constant that only depends on $\mathbf{x}$. Also, the numerator is equivalent to the joint probability model:

$$P(c_k)P(\mathbf{x} \mid c_k) = P(c_k, x_1, \ldots, x_n) = \\ P(x_1 \mid x_2, \ldots, x_n, c_k)P(x_2 \mid x_3, \ldots, x_n, c_k) \ldots P(x_{n-1} \mid x_n, c_k)P(x_n \mid c_k)P(c_k) \tag{1.3}$$

Now, the "naive" conditional independence assumption is applied: all features in $\mathbf{x}$ are mutually independent. Under this assumption, the joint model can be expressed as:

$$
\begin{aligned}
P(c_k \mid \mathbf{x}) &\propto P(c_k, x_1, \ldots, x_n) \\
&= P(c_k) P(x_1 \mid c_k) P(x_2 \mid c_k) P(x_3 \mid c_k) \cdots P(x_n \mid c_k) \\
&= P(c_k) \prod_{i=1}^{n} P(x_i \mid c_k)
\end{aligned}
\tag{1.4}
$$

Under the independence assumptions, the conditional distribution over the class variable c_k is:

$$
P(c_k \mid \mathbf{x}) = \frac{1}{Z} P(c_k) \prod_{i=1}^{n} P(x_i \mid c_k) \tag{1.5}
$$

the evidence $Z = P(\mathbf{x}) = \sum_k P(c_k) P(\mathbf{x} \mid c_k)$ is a constant scaling factor dependent only on $\mathbf{x}$ (the values of $\mathbf{x}$ are known).

The Naive Bayes classifier combines the model presented in equation (1.5) with a decision rule. One common rule is to choose the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. For this, we find the probability of given set of inputs for all possible values of the class variable $\mathbf{c}$ and pick up the output with maximum probability:

$$
\hat{y} = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}} \; P(c_k) \prod_{i=1}^{n} P(x_i \mid c_k). \tag{1.6}
$$

2 Example

Table 1 shows 14 rows of data with specific weather conditions, in addition, in each row there is a field that indicates whether it is possible to play golf.

Outlook	Temperature	Humidity	Windy	Play
Sunny	High	High	False	No
Sunny	High	High	True	No
Overcast	High	High	False	Yes
Rain	Medium	High	False	Yes
Rain	Low	Normal	False	Yes
Rain	Low	Normal	True	No
Overcast	Low	Normal	True	Yes
Sunny	Medium	High	False	No
Sunny	Low	Normal	False	Yes
Rain	Medium	Normal	False	Yes
Sunny	Medium	Normal	True	Yes
Overcast	Medium	High	True	Yes
Overcast	High	Normal	False	Yes
Rain	Medium	High	True	No

Table 1: Weather conditions for playing golf.

Suppose you have a new feature vector $\mathbf{t}$, and based on table 1, you need to decide if you can play golf or not:

$$\mathbf{t} = (Sunny, \quad High, \quad Normal, \quad False). \tag{2.1}$$

For $\mathbf{t}$ specified in equation (2.1), the probability of playing golf is given applying equation (1.4):

$$P(Yes \mid \mathbf{t}) \propto P(Yes)P(Sunny \mid Yes)P(High \mid Yes)P(Normal \mid Yes)P(False \mid Yes) \tag{2.2}$$

$$P(No \mid \mathbf{t}) \propto P(No)P(Sunny \mid No)P(High \mid No)P(Normal \mid No)P(False \mid No) \tag{2.3}$$

Using the information in table 1, the probabilities for calculating expressions (2.2) and (2.3) are obtained:

$$P(Yes \mid \mathbf{t}) \propto (0.64)(0.22)P(0.22)(0.67)(0.67) = 0.0139 \tag{2.4}$$

$$P(No \mid \mathbf{t}) \propto (0.36)(0.60)(0.80)(0.20)(0.40) = 0.0138 \tag{2.5}$$

Since $P(Yes \mid \mathbf{t}) + P(No \mid \mathbf{t}) = 1$, results presented in equation (2.4) and equation (2.5) can be normalized to obtain a probability distribution:

$$P(Yes \mid \mathbf{t}) = \frac{0.0139}{0.0139 + 0.0138} = 50.18\% \tag{2.6}$$

$$P(No \mid \mathbf{t}) = \frac{0.0138}{0.0139 + 0.0138} = 49.82\% \tag{2.7}$$

In conclusion: $P(Yes \mid \mathbf{t}) > P(No \mid \mathbf{t})$ (by a very small margin of difference) and it is possible to play golf if $\mathbf{t}$ occurs.