Navie Bayes

Jerry Peng

1 Navie Bayes

Naive Bayes is a set of supervised learning algorithms based on Bayes' theorem, used for classification problems. The "naive" in Naive Bayes comes from its assumption that features are mutually independent. Although this assumption rarely holds in real-world applications, Naive Bayes performs well in many practical problems, especially in text classification.
The formula is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

In classification problems, the composition of the aforementioned definition is as follows:

- $P(A|B)$ is posterior probability: It represents the probability of event A occurring given the feature B.
- $P(B|A)$ is likelihood: It represents the probability of feature B occurring given the category A.
- $P(A)$ is the prior probability, representing the probability of category A occurring.
- $P(B)$ is the evidence or marginal probability, representing the probability of feature B occurring.

In Naive Bayes classification, what we are looking for is the posterior probability of each category given feature X. Then, the sample is assigned to the category with the highest posterior probability. For example, suppose there is a feature set $X = \{x_1, x_2, \ldots, x_n\}$, then, the independence assumption of Naive Bayes is:

$$P(X|A) = P(x_1|A) \times P(x_2|A) \times \cdots \times P(x_n|A)$$

1.1 Example: Spam Email Classifier

If your email system wants to filter out spam, it can use the Naive Bayes classifier to determine whether the email is a regular email (ham) or spam (spam).

Training Set We have the following email content and their corresponding classifications:

- "Win big prizes for free" → "spam"
- "Free shopping voucher" → "spam"
- "Meeting at 10 o'clock tomorrow" → "ham"
- "Project meeting is tomorrow" → "ham"

Candidate "Free meeting tomorrow"

First, we need to calculate prior probabilities:

$$P(spam) = \frac{2}{4} = 0.5$$

$$P(ham) = \frac{2}{4} = 0.5$$

Then, we can calculate the likelihood:

- For "free":

$$P(free|spam) = \frac{2}{2} = 1$$

$$P(free|ham) = \frac{0}{2} = 0$$

- For "tomorrow":

$$P(tomorrow|spam) = \frac{0}{2} = 0$$

$$P(tomorrow|ham) = \frac{2}{2} = 1$$

- For "meeting":

$$P(meeting|spam) = \frac{0}{2} = 0$$

$$P(meeting|ham) = \frac{2}{2} = 1$$

Finally, we can calculate posterior probabilities:

$$\begin{aligned} &P(spam|free, tomorrow, meeting) \propto P(spam) \times P(free|spam) \times P(meeting|spam) \times P(tomorrow|spam) \\ &= 0.5 \times 1 \times 0 \times 0 \\ &= 0 \end{aligned} \tag{2}$$

$$\begin{aligned} &P(ham|free, tomorrow, meeting) \propto P(ham) \times P(free|ham) \times P(meeting|ham) \times P(tomorrow|ham) \\ &= 0.5 \times 1 \times 0 \times 0 \\ &= 0 \end{aligned} \tag{3}$$

From the above calculations, we can see that the probabilities for both are 0, which is a special case. This is because some conditional probabilities are 0, leading to the entire posterior probability being 0. This problem is typically addressed by using Laplace smoothing to avoid situations where the conditional probability is 0.

1.2 Laplace Smoothing

Laplace smoothing (also known as add-one smoothing) is a technique to prevent conditional probabilities from being 0 in Naive Bayes. Its main purpose is to address the issue of zero probabilities when calculating probabilities. This situation is especially common in text classification tasks, for example, when a word has never appeared in a category in the training data. The basic idea of Laplace smoothing is to add a positive value (usually 1) to the frequency of each word to avoid situations where the probability is 0. When using Laplace smoothing, the formula for calculating conditional probability is:

$$P(w|c) = \frac{count(w, c) + 1}{count(c) + |V|} \tag{4}$$

where:

- $P(w|c)$ is the conditional probability of word w given category c.
- $count(w, c)$ is the occurrence count of word w in category c.
- $count(c)$ is the total number of words in category c.
- $|V|$ is the vocabulary size in the training data.

In this way, even if a word has never appeared in a certain category, its conditional probability will not be 0. This is because the numerator (word frequency) is at least 1, and the denominator is the total number of words plus the vocabulary size (ensuring the denominator is greater than the numerator, so the probability value is between 0 and 1).

Let's take the spam classifier mentioned above as an example:

First, we need to calculate prior probabilities:

$$P(spam) = \frac{2}{4} = 0.5$$

$$P(ham) = \frac{2}{4} = 0.5$$

Then, we can calculate the likelihood:

- For "free":

$$P(free|spam) = \frac{2+1}{2+3} = \frac{3}{5}$$

$$P(free|ham) = \frac{0+1}{2+3} = \frac{1}{5}$$

- For "tomorrow":

$$P(tomorrow|spam) = \frac{0+1}{2+3} = \frac{1}{5}$$

$$P(tomorrow|ham) = \frac{2+1}{2+3} = \frac{3}{5}$$

- For "meeting":

$$P(meeting|spam) = \frac{0+1}{2+3} = \frac{1}{5}$$

$$P(meeting|ham) = \frac{2+1}{2+3} = \frac{3}{5}$$

Finally, we can calculate posterior probabilities:

$$\begin{aligned}
&P(spam|free, tomorrow, meeting) \propto P(spam) \times P(free|spam) \times P(meeting|spam) \times P(tomorrow|spam) \\
&= 0.5 \times \frac{3}{5} \times \frac{1}{5} \times \frac{1}{5} \\
&= 0.012
\end{aligned} \tag{5}$$

$$\begin{aligned}
&P(ham|free, tomorrow, meeting) \propto P(ham) \times P(free|ham) \times P(meeting|ham) \times P(tomorrow|ham) \\
&= 0.5 \times \frac{1}{5} \times \frac{3}{5} \times \frac{3}{5} \\
&= 0.036
\end{aligned} \tag{6}$$

Therefore, "Free meeting tomorrow" is classified as "ham".

References