Naive Bayes
Epoch IIT Hyderabad

Ankit Saha
AI21BTECH11004

19 Jul 2022

1. Introduction

Naive Bayes is a supervised machine learning algorithm used for classification. It is based on the Bayes' theorem in probability theory. It is called "naive" because these models are based on independence assumptions between features. Naive Bayes is widely used in text processing becaues of its efficiency and simplicity.

2. Mathematical Formulation

Suppose we are given m labelled samples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)$ where $x_i \in \mathbb{R}^n, y_i \in \{1, 2, \ldots, c\}\ \forall i \in [m]$. Here y_i denotes the category of $\mathbf{x}_i$ among c possible categories. From this, we can find the likelihoods of the features of the input vectors for every category, i.e.,

$$\Pr\left(x_{ij} \mid y_i = k\right) \quad \forall i \in [m]\ \forall j \in [n]\ \forall k \in [c] \tag{2.1}$$

Given a test sample $\mathbf{x}_t$, our goal is to assign a category to it. This is usually done by the *maximum a posteriori* (MAP) rule. We find the category that is the most probable to be assigned to a given input.

$$\hat{y} = \arg\max_{k \in [c]} \Pr\left(y_i = k \mid \mathbf{x}_t\right) \tag{2.2}$$

According to Bayes' rule,

$$\Pr\left(y_i = k \mid \mathbf{x}_t\right) = \frac{\Pr\left(\mathbf{x}_t \mid y_i = k\right) \Pr\left(y_i = k\right)}{\Pr\left(\mathbf{x}_t\right)} \tag{2.3}$$

By our "naive" assumption that all features are independent, we have

$$\Pr\left(\mathbf{x}_t \mid y_i = k\right) = \Pr\left(x_{t1} \mid y_i = k\right) \Pr\left(x_{t2} \mid y_i = k\right) \cdots \Pr\left(x_{tn} \mid y_i = k\right) \tag{2.4}$$

$$= \prod_{j=1}^{n} \Pr\left(x_{tj} \mid y_i = k\right) \tag{2.5}$$

This can be found easily as we already know the individual probability values. Now, notice that the denominator $\Pr\left(\mathbf{x}_t\right)$ is independent of k. Therefore, the optimization problem reduces to

$$\hat{y} = \arg\max_{k \in [c]} \Pr\left(y_i = k\right) \prod_{j=1}^{n} \Pr\left(x_{tj} \mid y_i = k\right) \tag{2.6}$$

$\Pr\left(y_i = k\right)$ is known as the prior probability and is usually chosen to be proportional to the number of samples that have the label k

$$\Pr\left(y_i = k\right) = \frac{n(y_i = k)}{m} \tag{2.7}$$

3. Questions

i) Define conditional probability and give a mathematical expression for it.
ii) Prove Bayes' rule.
iii) Give a practical application of the naive Bayes classifier.
iv) How would one evaluate the performance of the model?
v) Explain the zero-frequency problem and propose a solution to fix it.

4. Answers

i) Conditional probability of $A \mid B$ is defined as the probability of event A happening given that event B has already happened.

$$\Pr(AB) = \Pr(B)\Pr(A \mid B) \tag{4.1}$$

$$\implies \Pr(A \mid B) = \frac{\Pr(AB)}{\Pr(B)} \tag{4.2}$$

ii) Similarly, we have

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)} \tag{4.3}$$

On dividing the equations, we get

$$\frac{\Pr(A \mid B)}{\Pr(B \mid A)} = \frac{\Pr(A)}{\Pr(B)} \tag{4.4}$$

$$\therefore \Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \tag{4.5}$$

iii) Naive Bayes classifiers are used in spam detection. With a given e-mail as input, a naive Bayes model cateogrizes it into one of two categories: not spam or spam.
iv) Just like most other classification models, confusion matrices can be used as an evaluation metric for naive Bayes classifiers. A confusion matrix tracks the number of true positives, true negatives, false positives and false negatives (for binary classification; for c classes, we get a $c \times c$ confusion matrix) and computes the appropriate ratios of these quantities to quanitfy the performance of the model.
v) When the frequency of a feature is zero in one of the categories, i.e., $\Pr\left(x_{ij} \mid y_i = k\right) = 0$, then the posterior probability will be zero every time irrespective of other features if this feature is present in the sample. This is known as the zero-frequency problem. A simple fix is to increase the frequencies of every feature in every category by a fixed small positive number α so that there are no zero probabilities at all.