Naïve Bayes

Aaryan

CO21BTECH11001

Naïve Bayes is a classification learning algorithm used for the problems in which features are discrete valued.

For problems where features are not discrete valued i.e. they are continuous real valued vectors, they are recoded into quartiles, such that values less than the 25th percentile are assigned a 1, 25th to 50th a 2, 50th to 75th a 3 and greater than the 75th percentile a 4.

Let number of features of dataset = n

Let number of sets of features = m

Data consists of matrices X and y where i^{th} column of X represents the i^{th} feature of dataset and i^{th} element of y represents the value of variable dependent on set of features listed in i^{th} row of X.

Since it is a generative algorithm, we want to model P(X|y) and P(y)

<u>Naïve Bayes Assumption</u> –

We assume that all features $X_i s$ are conditionally independent given *y.*

Mathematically, we can write –

$$P(X_1, X_2, \dots, X_n|y) = P(X_1|y). P(X_2, y) \dots . P(X_n, y)$$

$$P(X_1, X_2, \dots, X_n|y) = \prod_{i=1}^{n} P(X_i|y)$$

Parameters –

$$\phi_{j|y=1} = P\left(x_j = 1 \middle| y = 1\right)$$
$$\phi_{j|y=0} = P\left(x_j = 1 \middle| y = 0\right)$$
$$\phi_y = P(y = 1)$$

We define a Log-Likelihood function as follows

$$l\left(\phi_y, \phi_{j|y}\right) = \log\left(\prod_{i=1}^{m} P\left(X^{(i)}, y_i; \phi_y, \phi_{j|y}\right)\right)$$

To maximize Log-Likelihood function, we differentiate $l\left(\phi_y, \phi_{j|y}\right)$ and put it's derivative equal to 0, we get –

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y_i = 1\}}{m}$$

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\left\{X_j^{(i)} = 1, y_i = 1\right\}}{\sum_{i=1}^{m} 1\{y_i = 1\}}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\left\{X_j^{(i)} = 1, y_i = 0\right\}}{\sum_{i=1}^{m} 1\{y_i = 0\}}$$

By applying Laplace Smoothing, we get –

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\left\{X_j^{(i)} = 1, y_i = 1\right\} + 1}{\sum_{i=1}^{m} 1\{y_i = 1\} + 2}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\left\{X_j^{(i)} = 1, y_i = 0\right\} + 1}{\sum_{i=1}^{m} 1\{y_i = 0\} + 2}$$

Therefore, we can make the prediction for new set of features x by calculating the following probability –

$$P(y=1|x) = \frac{P(x|y=1)P(y=1)}{P(x|y=1)P(y=1) + P(x|y=0)P(y=0)}$$

Where –

$$P(x|y=1) = \prod_{j=1}^{n} \phi_{j|y=1}$$

$$P(x|y=0) = \prod_{j=1}^{n} \phi_{j|y=0}$$

$$P(y=1) = \phi_y$$

$$P(y=0) = 1 - \phi_y$$

If $P(y=1|x) \geq 0.5$, then y=1

Else y=0

Questions –

1. What is the property of features used in Naïve Bayes Algorithm?
   Ans. They are discrete. For example, they may be Binary({0,1}).

2. What is the basic principle of Naïve Bayes Algorithm?
   Ans. Naive Bayes estimates probabilities based on the class frequencies of each (unique) feature in the training data.

3. What is the assumption that Naïve Bayes Algorithm makes?
   Ans. It assumes that all features are conditionally independent given y i.e. the probability of occurring of one feature given y is independent of probability of occurring of some other feature given y.

4. Why is Laplace Smoothing used?

Ans. In general, the probability of a new value of discrete feature which has not occurred in the training data is not 0. Therefore, to fill that gap, Laplace Smoothing is done.

5. What are the fields of machine learning where Naïve Bayes Algorithm is generally used?
   Ans. It is generally used in Spam filtering, Recommendation systems etc.