Naive Bayes
Epoch IIT Hyderabad

Chakka Surya Saketh
AI22BTECH11005

Introduction

The Naive Bayes algorithm is a popular and straightforward classification technique based on Bayes' theorem. It is widely used in various machine learning applications due to its simplicity, efficiency, and ability to handle high-dimensional data. This model is easy to build and is mostly used for large datasets. It is a probabilistic machine learning model that is used for classification problems. The core of the classifier depends on the Bayes theorem.
It is called Naive because of the assumption that 2 variables are independent when they may not be. In a real-world scenario, there is hardly any situation where the features are independent.

Implementation

Bayes theorem provides a way of computing posterior probability $\Pr(c|x)$ from $\Pr(c), \Pr(x) and \Pr(x|c)$. Look at the equation below:

$$\Pr(c/x) = \frac{\Pr(x/c)\Pr(c)}{\Pr(x)}$$

When there are multiple X variables, we simplify it by assuming that X's are independent, so- For n number of X, the formula becomes Naive Bayes:

$$P\big(Y = k|X1, X2....Xn\big) = \frac{\mathrm{P(Y)}\prod_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{P(Xi|Y)}}{P(X1) * P(X2).... * P(Xn)}$$

1) Convert the data into a frequency table.

2) Then create a likelihood table by finding the probabilities.

3) Use Naive Bayesian equation to calculate the posterior probability of each class

Using these steps classify the data into classes. This algorithm is mostly used in text classification (nlp) and with problems having multiple classes.

Gaussian Naive Bayes

Gaussian Naive Bayes is used when we assume all the continuous variables associated with each feature to be distributed according to Gaussian Distribution. Gaussian Distribution is also called Normal distribution. The conditional probability changes here since we have different values now. Also, the (PDF) probability density function of a normal distribution is given by:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

We can use this formula to compute the probability of likelihoods if our data is continuous.

Applications

1) Text Classification: Naive Bayes is widely used in natural language processing tasks, such as spam filtering, sentiment analysis, and document categorization.
2) Medical Diagnosis: Naive Bayes finds applications in medical diagnosis, such as disease classification and prediction based on patient symptoms.
3) Image Recognition: In image recognition, Naive Bayes is used for tasks like object recognition and facial expression classification.

Conclusion

Gaussian Naive Bayes is a powerful classification algorithm for continuous data that is widely used in various machine learning applications. By assuming that the features follow a Gaussian distribution and are conditionally independent given the class, the algorithm efficiently and effectively classifies data points into different categories. Despite its assumptions, Gaussian Naive Bayes is a valuable tool in the machine learning toolkit, especially for real-world problems with continuous data.