Naïve Bayes Classifier

Prerequisite

- Basic Concepts of probability, and probability distribution function
- Problem of Classification
- Evaluation Metrics for classification accuracy

Objectives(Lerner will able to understand and explain)

- Sample space, event, mutually exclusive events, dependent events.
- Conditional probability and baye's theorem.
- Naïve Baye's Classifier and Naïve assumption.
- Gaussian Naïve Baye's Classifier
- Pros and Cons of Naïve baye's Classifier.

Naïve Bayes classifier works on the concept of probability and has a wide range of applications like spam filtering, sentiment analysis, document classification etc. The principle of naïve bayes classifier is based on the work of Thomas Bayes (1702-1761) of Bayes Theorem for conditional probability.

Before moving towards the heavy concepts of probability, let us first go through with the terminology and some important concepts.

Sample Space and Event-

A **sample space** is defined as the collection of all possible outcomes of a random experiment. For example, consider a random experiment of tossing a coin twice, then the sample space is defined as:

$$sample\ space = \{TT,\ TH,\ HT,\ HH\}$$

An **event** is defined as a subset of possible outcomes of any random experiment. For the above example let the event of getting exactly one Head on a coin is:

$$E(getting\ exactly\ one\ head) = \{HT\ TH\}$$

Based on the above, **Probability** of an event is defined as the ratio of cardinality of the event set with respect to the sample space set.

$$Probabilty = \frac{|Event|}{|Sample\ Space|}$$

$$Prob(getting\ exactly\ one\ heads) = \frac{|\{HH\}|}{|\{TT, TH, HT, HH\}|} = \frac{1}{4} = 0.25$$

$$Prob(getting\ atleast\ one\ head) = \frac{|\{HT, TH, HH\}|}{|\{TT, TH, HT, HH\}|} = \frac{3}{4} = 0.75$$

Mutually Exclusive Event-

Two events are said to be mutually exclusive if the occurrence of one event precludes the occurrence of another or the two events cannot occur at the same time. For example, in an experiment of tossing a coin the event of getting head or tail are mutually exclusive as the two events cannot occur together.

$$P(head) = \frac{1}{2}$$
$$P(tail) = \frac{1}{2}$$
$$P(head \cap tail) = \emptyset$$

Independent Event-

The two events are said to be independent if the probability of one event occurrence does not affect the probability of occurrence of other events. Let A and B are two independent events then:

$$P(A \cap B) = P(A) * P(B)$$

Dependent Events or Conditional Probability-

The two events are said to be dependent if they affect each other or the probability of one event change with respect to other events. Conditional probability is defined for dependent events and read as the probability of event A given B(B has already occurred).

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

or the probability of event B given A is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad \because \quad P(A \cap B) = P(B \cap A)$$

Baye’s Theorem-

Baye’s theorem has basically described the probability of an event which is actually based on the preceding values of the event. Baye's theorem is the extended version of conditional probability. With conditional probability, we know that the probability of event A given B is:

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

and the probability of event B given A is:

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

Using the two equations:

$$P(A|B) = \frac{P\left(\frac{B}{A}\right)*P(A)}{P(B)} \qquad known\ as\ baye's\ theorem.$$

Baye’s theorem provides a way to calculate posterior probability P(A|B) from P(A), P(B) and P(B|A).

Naïve Baye’s Classifier-

The Naïve Baye's Classifier works on the principle of baye's theorem with a naïve assumption that the presence of a particular feature in a class is completely unrelated to the presence of other features(Input features are independent from each other). The naïve baye's classifier is expressed as:

$$P(c|x) = \frac{P(x|c)*P(c)}{P(x)}$$

Where, P(c|x) is the posterior probability of class c for given predictor feature x.

P(c) is the prior probability of class c.

P(x) is the prior probability of the predictor.

and P(x|c) is the likelihood of the probability of predictor x for given class c.

the above equation is interpreted as:

$$Posterior\ probability = \frac{Prior\ probability*Likelihood}{Evidence}$$

The Bayes rule computes the probability of class c for given x feature. In real life problems, the target class c depends on multiple x variables. So, the formula of bayes rule can be extended for multiple input features like,

$$P(c|x_1, x_2,\ x_3,\ ...x_n) = \frac{P\left(x_1, x_2,\ x_3, ...x_n|c\right)*P(c)}{P\left(x_1,\ x_2, x_3, ...x_n\right)}$$

According to the naïve assumption of input features are independent,

$$P(c|x_1, x_2,\ x_3,\ ...x_n) = \frac{P(c)*P(c)*P(c).......P(x_n|c)*P(c)}{P\left(x_1\right)*P\left(x_2\right)*P\left(x_3\right).......P(x_n)}$$

The Bayesian classifier works on maximum a posterior or MAP decision rule which assigns the label $\hat{y} = c$ for class c is,

$$\hat{y} = argmax\,\{P(c) *\ P(c) *\ P(c).......\ P(x_n|c)\ *\ P(c)\}$$

Working steps-

Step 1- Compute the prior probabilities for given class labels.

Step 2- Compute the Likelihood of evidence with each attribute for each class.

Step 3- Calculate the posterior probabilities using Bayes rule.

Step 4- Select the class which has higher probability for given inputs.

Gaussian Naïve Bayes-

It is easy to compute probabilities for categorical variable x, but what if the variable x is continuous? If we assume that the variable x follows a specific distribution then we can use the probability density function to compute the probability of given value. Gaussian Naïve Baye's Classifier assumes that the continuous values associated with each feature are normally distributed or follow normal distribution.

$$P(c) = \frac{1}{\sqrt{2\pi\sigma^2_{(x_i, c)}}} e^{\frac{-(x_i - \mu_{(x_i, c)})}{2\sigma^2_{(x_i, c)}}}$$

Laplace correction-

In the formula the denominator term comprises the probabilities of evidence. When you model a naïve bayes with many input features then it may be possible that one or some of the probabilities of evidence come as zero. This creates the problem of zero division and fails the total computation. To avoid this situation the variables are increased with a small value 1 in the numerator, so the probability does not become zero. This correction is called Laplace correction.

Example-

Let us consider a problem to predict a CAR Racing Match (yes/no) based on parameters like weather conditions of cloudy scene, Temperature, Moisture on track and turbulent. The historical data of some previous matches are given in the table. Using the given data we need to predict the car race match on the day of sunshine, chill, high and false values of parameter cloudy scene, temperature, moisture and turbulent respectively.

Sno.	Cloudy scene	Temperature	Moisture on track	Turbulent	Car Race
1	Hazy	Heated	High	False	Yes
2	Hazy	Chill	Normal	True	Yes
3	Hazy	Moderate	High	True	Yes
4	Hazy	Heated	Normal	False	Yes
5	Rainfall	Moderate	High	False	Yes
6	Rainfall	Chill	Normal	False	Yes
7	Rainfall	Chill	Normal	True	No
8	Rainfall	Moderate	Normal	False	Yes
9	Rainfall	Moderate	High	True	No
10	Sunshine	Heated	High	False	No
11	Sunshine	Heated	High	True	No
12	Sunshine	Moderate	High	False	No
13	Sunshine	Chill	Normal	False	Yes
14	Sunshine	Moderate	Normal	True	Yes

Let us assume $x_1 = Sunshine,\ x_2 = Chill,\ x_3 = High\ and\ x_4 = False$

Naïve Bayes Classifier works in following steps:

Step 1- Calculate the Prior Probabilities;

$$P(CarRace\ =\ Yes) = \frac{9}{14}\quad and\ \ P(CarRace\ =\ No) = \ \frac{5}{14}$$

Step 2- Compute the Likelihood of evidence that goes into denominators

$$P(Sunshine\mid Yes) = \ \frac{P(Sunshine\cap Yes)}{P(Yes)} = \frac{2}{9}\ \ and\ P(Sunshine\mid No) = \ \frac{P(Sunshine\ \cap No)}{P(No)} = \frac{3}{5}$$

$$P(Chill\mid Yes) = \ \frac{P(Chill\cap Yes)}{P(Yes)} = \frac{3}{9}\ \ and\ P(Chill\mid No) = \ \frac{P(Chill\ \cap No)}{P(No)} = \frac{1}{5}$$

$$P(High\mid Yes) = \ \frac{P(High\cap Yes)}{P(Yes)} = \frac{3}{9}\ \ and\ P(High\mid No) = \ \frac{P(High\ \cap No)}{P(No)} = \frac{4}{5}$$

and

$$P(False\mid Yes) = \ \frac{P(False\cap Yes)}{P(Yes)} = \frac{6}{9}\ \ and\ P(False\mid No) = \ \frac{P(False\ \cap No)}{P(No)} = \frac{2}{5}$$

Step 3- Calculate the probability using bayes rule:

$$P(\ Yes|\ Sunshine,\ Chill,\ High,\ False) = \ \frac{P(Yes|\ Sunshine, Chill, High, False)*P(Yes)}{P(Sunshine, Chill, High, False)}$$

and

$$P(\ No|\ Sunshine,\ Chill,\ High,\ False) = \ \frac{P(\ No|\ Sunshine, Chill, High, False)*P(No)}{P(Sunshine, Chill, High, False)}$$

Ignoring the denominator term as it is common in both the equations and applying naïve assumption of independent features. The Probabilities are:

$P(Yes\mid Sunshine,\ Chill,\ High,\ False) = \ P(Sunshine\mid Yes) *\ P(Chill\mid Yes) *\ P(High\mid Yes) *\ P(False\mid Yes) *\ P(Yes$

and

$P(No\mid Sunshine,\ Chill,\ High,\ False) = \ P(Sunshine\mid No) *\ P(Chill\mid No) *\ P(High\mid No) *\ P(\ False\mid No) *\ P(No)$

Substituting the values

$$P(Yes\mid Sunshine,\ Chill,\ High,\ False) = \left(\frac{2}{9}\right)\left(\frac{3}{9}\right)\left(\frac{3}{9}\right)\left(\frac{6}{9}\right)\left(\frac{9}{14}\right)\ \ = 0.0105$$

and

$$P(No\mid Sunshine,\ Chill,\ High,\ False) = \left(\frac{3}{5}\right)\left(\frac{1}{5}\right)\left(\frac{4}{5}\right)\left(\frac{2}{5}\right)\left(\frac{5}{14}\right) = \ 0.0137$$

Step 4- Finally the probabilities are:

$$P(\ Yes|\ Sunshine,\ Chill,\ High,\ False) = \ \frac{0.0105}{0.0105+\ 0.0137}\ = \ 0.44$$

and

$$P(\ No|\ Sunshine,\ Chill,\ High,\ False) = \ \frac{0.0137}{0.0105+\ 0.0137} = \ 0.56$$

From the probability values it is clear that if the parameter values are as Cloudy Scene = Sunshine, Temperature = Chill, Moisture on track = High and Turbulent = False, then there is 44% chance to organize a Car race match.

How to improve the efficiency of naïve baye's model-

1. If continuous features are not normally distributed in the dataset then use any transformation method to convert them as normal distribution.
2. Drop the highly correlated.

Evaluation of Naïve Model-

This is a classification model, so all the classification accuracy metrics are used to evaluate the performance of models which are discussed in the Logistic regression sheet.

Pros and Cons of Naïve Baye's Model-

Pros

- Naïve Baye's model is simple to implement and fast in processing.
- Requires few examples in the train set to work with.
- Perform well with noisy data and missing values.

Cons

- Perform poorly if the dataset contains more continuous input features.
- Predictions are based on the assumption of independent features which is almost impossible in real life scenarios.
- Sometimes the estimated probabilities are less reliable.

****