

## ASSIGNMENT - 2

**AIM :** Implement Naive Bayes to predict the work type for a person with following parameters, age: 30, Qualification: MTech, Experience: 8.

### OBJECTIVE :

- To understand basic concept of naive bayes classifier.
- To implement naive bayes classifier to predict work type for a person with given attributes.

### SOFTWARE REQUIREMENTS :

- Linux Operating System
- Java Compiler
- Eclipse IDE

### MATHEMATICAL MODEL :

Consider a following set theory notations related to a program. The mathematical model M for Naive Bayes classifier is given as below,

$$M = \{S, S_0, A, G\}$$

Where,

S=State space.i.e All prior probabilities to calculate probability of X being a part of class 'c'

S<sub>0</sub>= Initial State.i.e Training set of tuple

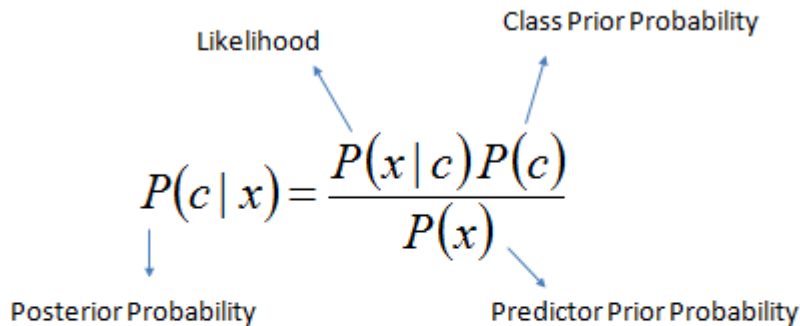
A=Set of Actions/Operators.i.e with given dataset predicting the work type for a person with give parameters.

G=Goal state.In this case predicting accurate work type for a person.

### THEORY :

#### Naive Bayes Classifier :

The Naive Bayes classifier is a simple probabilistic classifier which is based on Bayes theorem with strong and naive independence assumptions. It is one



$$P(c | \mathbf{X}) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Fig : Bayes Rule

of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, language detection and sentiment detection.

You can use Naive Bayes when you have limited resources in terms of CPU and Memory. Moreover when the training time is a crucial factor, Naive Bayes comes handy since it can be trained very quickly.

Let  $X$  be a data tuple. In Bayesian terms,  $X$  is considered evidence. As usual, it is described by measurements made on a set of  $n$  attributes. Let  $H$  be some hypothesis, such as that the data tuple  $X$  belongs to a specified class  $C$ . For classification problems, we want to determine  $P(H|X)$ , the probability that the hypothesis  $H$  holds given the evidence or observed data tuple  $X$ .

The Bayes Naive classifier selects the most likely classification  $V_{nb}$  given the attribute values  $a_1, a_2, \dots, a_n$ . This results in:

$$V_{nb} = \text{argmax}_{v_j \in V} P(v_j) \prod P(a_i | v_j)$$

We generally estimate  $P(a_i | v_j)$  using m-estimates:

$$P(a_i | v_j) = \mathbf{f}(\mathbf{x}) = \frac{nc + mp}{n + m}$$

where:

$n$  = the number of training examples for which  $v = v_j$

$nc$  = number of examples for which  $v = v_j$  and  $a = a_i$

$p$  = a priori estimate for  $P(a_i|v_j)$   
 $m$  = the equivalent sample size

### Naive bayes : Car Theft Example

Attributes are Color, Type, Origin, and the subject, stolen can be either yes or no.

#### 1. data set

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

2. **Training example** We want to classify a Red Domestic SUV. Note there is no example of a Red Domestic SUV in our data set. Looking back at equation (2) we can see how to compute this. We need to calculate the probabilities.

$P(\text{Red}|\text{Yes})$ ,  $P(\text{SUV}|\text{Yes})$ ,  $P(\text{Domestic}|\text{Yes})$  ,

$P(\text{Red}|\text{No})$  ,  $P(\text{SUV}|\text{No})$ , and  $P(\text{Domestic}|\text{No})$

and multiply them by  $P(\text{Yes})$  and  $P(\text{No})$  respectively . We can estimate these values using equation (3).

Looking at  $P(\text{Red}|\text{Yes})$ , we have 5 cases where  $v_j = \text{Yes}$  , and in 3 of those cases  $a_i = \text{Red}$ . So for  $P(\text{Red}|\text{Yes})$ ,  $n = 5$  and  $n_c = 3$ . Note that all attribute are binary (two possible values). We are assuming no other information so,  $p = 1 / (\text{number-of-attribute-values}) = 0.5$  for all of our attributes. Our  $m$  value is arbitrary, (We will use  $m = 3$ ) but consistent for all attributes. Now we simply apply equation (3) using the precomputed values of  $n$  ,  $n_c$ ,  $p$ , and  $m$ .

Yes	No
Red n = 5 n.c= 3 p = .5 m = 3	Red n = 5 n.c = 2 p = .5 m = 3
SUV n = 5 n.c= 1 p = .5 m = 3	SUV n = 5 n.c = 3 p = .5 m = 3
Domestic n = 5 n.c= 2 p = .5 m = 3	Domestic n = 5 n.c = 3 p = .5 m = 3

$$P(\text{Red}|\text{Yes})= \mathbf{f}(\mathbf{x}) = \frac{3+3*.5}{5+3} = .56$$

$$P(\text{Red}|\text{No})= \mathbf{f}(\mathbf{x}) = \frac{2+3*.5}{5+3} = .43$$

$$P(\text{SUV}|\text{Yes})= \mathbf{f}(\mathbf{x}) = \frac{1+3*.5}{5+3} = .31$$

$$P(\text{SUV}|\text{No})= \mathbf{f}(\mathbf{x}) = \frac{3+3*.5}{5+3} = .56$$

$$P(\text{Domestic}|\text{Yes})= \mathbf{f}(\mathbf{x}) = \frac{2+3*.5}{5+3} = .43$$

$$P(\text{Domestic}|\text{No})= \mathbf{f}(\mathbf{x}) = \frac{3+3*.5}{5+3} = .56$$

We have  $P(\text{Yes}) = .5$  and  $P(\text{No}) = .5$ , so we can apply equation (2).

For  $v = \text{Yes}$ , we have

$$P(\text{Yes}) * P(\text{Red} | \text{Yes}) * P(\text{SUV} | \text{Yes}) * P(\text{Domestic}|\text{Yes}) \\ = .5 * .56 * .31 * .43 = .037$$

and for  $v = \text{No}$ , we have

$$P(\text{No}) * P(\text{Red} | \text{No}) * P(\text{SUV} | \text{No}) * P(\text{Domestic} | \text{No}) \\ = .5 * .43 * .56 * .56 = .069$$

Since  $0.069 > 0.037$ , our example gets classified as 'NO'

## Types of Probabilities:

### 1. Prior Probabilty

Prior probabilities represent what we originally believed before new evidence is uncovered. New information is used to produce updated probabilities and is a more accurate measure of a potential outcome. It can be represented as,  $P(\mathbf{x})$

### 2. Conditional Probabilty

A conditional probability is the probability of an event, given some other event has already occurred. It can be represented as,  $P(\mathbf{x}|\mathbf{y})$

### 3. Posterior Probabilty

The posterior probability is the probability of event A occurring given that event B has occurred. In other words, Posterior probability is the probability of the parameters given the evidence. It can be represented as,  $P(\mathbf{z}|\mathbf{x})$

## CONCLUSION :

Thus, we have implemented Naive Bayes classifier to predict the work type for a person.

Roll No.	Name of Student	Date of Performance	Date of Submission	Sign.
		/ /2015	/ /2015	