ASSIGNMENT 7

AIM: Implement Naive Bayes for Concurrent/Distributed application. Approach should handle categorical and continuous data

OBJECTIVE:

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

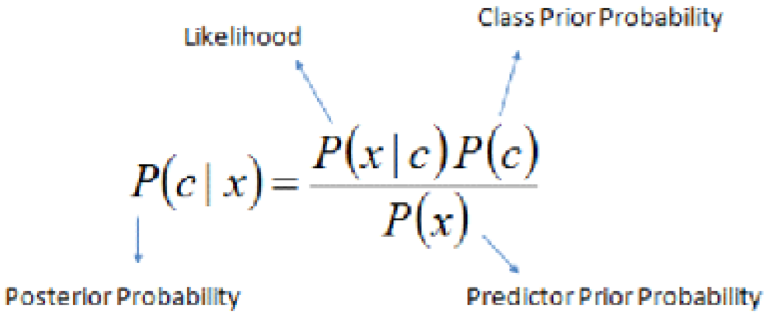
SOFTWARE REQUIREMENTS:

- Linux Operating System
- Java Compiler
- Weka Tool
- 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,So,A,G
Where,
S=State space.i.e All prior probabilities to calculate probability of X being a
part of class c
So= 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 nave independence assumptions. It is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, language detection and sentiment detection.


Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c) P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig : Bayes Rule

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 a1, a2,... a n .This results in:

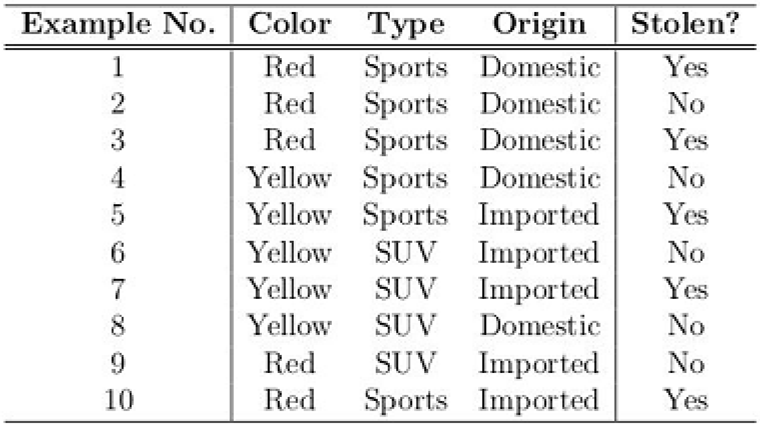
$Vnb = argmaxvjEvP(Vj)P(ai|vj)$

We generally estimate P(ai—vj) using m-estimates:

$P(ai|vj) = f(x) = (nc + mp)/(n + m)$

where:
n = the number of training examples for which v = vj
nc = number of examples for which v = vj and a = ai
2p = a priori estimate for P(ai—vj)
m = the equivalent sample size

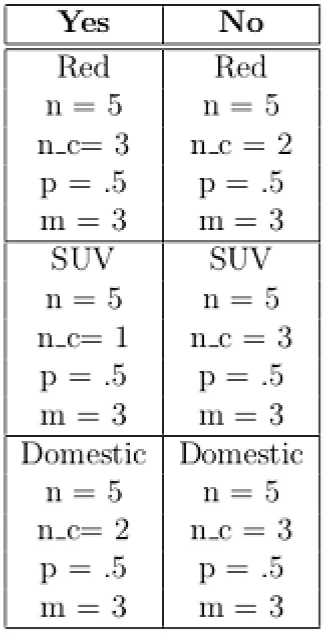
Naive bayes : Car Theft Example
Attributes are Color, Type, Origin, and the subject, stolen can be either yes or no.
1. data set



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(Red—Yes), P(SUV—Yes), P(Domestic—Yes) ,
P(Red—No) , P(SUV—No), and P(Domestic—No)
and multiply them by P(Yes) and P(No) respectively . We can estimate these values using equation (3).
Looking at P(Red—Yes), we have 5 cases where vj = Yes , and in 3 of those cases ai = Red. So for P(Red—Yes), n = 5 and nc = 3. Note that all attribute are binary (two possible values).
We are assuming no other information so, p = 1 / (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 eqauation (3) using the precomputed values of n , nc, 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

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

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

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

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

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

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

We have P(Yes) = .5 and P(No) = .5, so we can apply equation (2).
For v = Yes, we have
P(Yes) * P(Red — Yes) * P(SUV — Yes) * P(Domestic—Yes)
= .5 * .56 * .31 * .43 = .037
and for v = No, we have
P(No) * P(Red — No) * P(SUV — No) * P (Domestic — No)
= .5 * .43 * .56 * .56 = .069
Since 0.069 ¿ 0.037, our example gets classified as NO

Types of Probabilities: 1. Prior Probability :
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(x)
2. Conditional Probability :
A conditional probability is the probability of an event, given some other event has already occured.It can be represented as,P(x—y) 3. Posterior Probability :
The posterior probability is the probability of event A occurring given that event B has occured.In other words,Posterior probability is the probability of the parameters given the evidence.It can be represented as,P(z—x)

CONCLUSION : Thus, we have implemented Naive Bayes for Concurrent/Distributed application

Roll No.	**Name of Student**	**Date of Performance**	**Date of Submission**	**Sign.**
BECOC357	Sunny Shah	28 / 09 / 2017	05 / 10 / 2017	