03/05/2021 **Mohd. Aftab Alam**
02013302717

EXPERIMENT 1

Aim :

Study and Implement the Naïve Bayes Learner using WEKA (Breast Cancer Dataset).

Introduction :

The first supervised learning method we introduce is the *multinomial Naïve Bayes or multinomial NB* model, a probabilistic learning method. The probability of a document **d** being in class **c** is computed **as:**

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

where $\mathbf{P(t_k|c)}$ is the conditional probability of term $\mathbf{t_k}$ occurring in a document of class **c**. We interpret $\mathbf{P(t_k|c)}$ as a measure of how much evidence $\mathbf{t_k}$ contributes that **c** is the correct class. **P(c)** is the prior probability of a document occurring in class **c**. If a document's terms do not provide clear evidence for one class versus another, we choose the one that has a higher prior probability. $\mathbf{(t_1, t_2, \ldots, t_{nd})}$ are the tokens in **d** that are part of the vocabulary we use for classification and $\mathbf{n_d}$ is the number of such tokens in **d**. For example, $\mathbf{(t_1, t_2, \ldots, t_{nd})}$ for the one-sentence document Beijing and Taipei join the WTO might be **(Beijing, Taipei, join, WTO)** with $\mathbf{n_d=4}$, if we treat the terms and the as stop words.

In text classification, our goal is to find the *best class* for the document. The best class in NB classification is the most likely or *maximum a posteriori (MAP)* class $\mathbf{c_{map}}$**:**

$$c_{map} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c).$$

We write **Ṕ** for **P** because we do not know the true values of the parameters **P(c)** and $\mathbf{P(t_k|c)}$, but estimate them from the training set.

First discretize the attribute values. By default, Weka's Naïve Bayes classifier assumes that the attributes are normally distributed given the class. You should override this by setting use Supervised Discretization to true using the Generic Object Editor window. This will cause Naïve Bayes to discretize the numeric attributes in the data with a supervised discretization technique. In most practical applications of Naïve Bayes, supervised discretization works better than the default method. It also produces a more comprehensible visualization, which is why we use it here.

Breast Cancer Dataset :

- This dataset includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.
- Number of Instances: 286
- Number of Attributes: 9 + the class attribute

- Attribute Information:
  1. Class: no-recurrence-events, recurrence-events
  2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
  3. menopause: lt40, ge40, premeno.
  4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
  5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35, 36-39.
  6. node-caps: yes, no.
  7. deg-malig: 1, 2, 3.
  8. breast: left, right.
  9. breast-quad: left-up, left-low, right-up, right-low, central.
  10. irradiat: yes, no.

- Missing Attribute Values: (denoted by "?")
  - Attribute Name : Number of instances with missing values
  - node-caps : 8
  - breast-quad : 1

- Class Distribution:
  - no-recurrence-events: 201 instances
  - recurrence-events: 85 instances

Implementation :

```
Classify using Naïve Bayes
=== Run information ===
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     breast-cancer
Instances:    286 Attributes:   10
Age     menopause       tumor-size      inv-nodes       node-caps       deg-malig       breast  breast-quad     irradiat
Class
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
Naive Bayes Classifier
```

Class Attribute	no-recurrence-events (0.7)	recurrence-events (0.3)	
10-19	1.0	1.0	
20-29	2.0	1.0	
30-39	22.0	16.0	
40-49	64.0	28.0	
50-59	72.0	26.0	
60-69	41.0	18.0	
70-79	6.0	2.0	
80-89	1.0	1.0	
90-99	1.0	1.0	
[total]	210.0	94.0	
menopause			
lt40	6.0	3.0	
ge40	95.0	36.0	
premeno	103.0	49.0	
[total]	204.0	88.0	
tumor-size			
0-4	8.0	2.0	
5-9	5.0	1.0	
10-14	28.0	2.0	
15-19	24.0	8.0	
20-24	35.0	17.0	
25-29	37.0	19.0	
30-34	36.0	26.0	
35-39	13.0	8.0	
40-44	17.0	7.0	
45-49	3.0	2.0	
50-54	6.0	4.0	
55-59	1.0	1.0	
[total]	213.0	97.0	
inv-nodes			
0-2	168.0	47.0	
3-5	20.0	18.0	
6-8	8.0	11.0	
9-11	5.0	7.0	
12-14	2.0	3.0	
15-17	4.0	4.0	
18-20	1.0	1.0	
21-23	1.0	1.0	
24-26	1.0	2.0	
27-29	1.0	1.0	
30-32	1.0	1.0	
33-35	1.0	1.0	
36-39	1.0	1.0	
[total]	214.0	98.0	

node-caps		
yes	26.0	32.0
no	172.0	52.0
[total]	198.0	84.0
deg-malig		
	60.0	13.0
	103.0	29.0
	41.0	46.0
[total]	204.0	88.0
breast		
left	104.0	50.0
right	99.0	37.0
[total]	203.0	87.0
breast-quad		
left_up	72.0	27.0
left_low	76.0	36.0
right_up	21.0	14.0
right_low	19.0	7.0
central	18.0	5.0
[total]	206.0	89.0
irradiat		
yes	38.0	32.0
no	165.0	55.0
[total]	203.0	87.0

```
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         205               71.6783 %
Incorrectly Classified Instances        81               28.3217 %
Kappa statistic                          0.2857
Mean absolute error                      0.3272
Root mean squared error                  0.4534
Relative absolute error                 78.2086 %
Root relative squared error             99.1872 %
Total Number of Instances              286

=== Detailed Accuracy By Class ===
TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0.836    0.565    0.778      0.836    0.806      0.288    0.701     0.837     no-recurrence-events
0.435    0.164    0.529      0.435    0.477      0.288    0.701     0.514     recurrence-events

Weighted Avg.    0.717    0.446    0.704      0.717    0.708      0.288    0.701     0.741

=== Confusion Matrix ===
   a   b   <-- classified as
```