

Lab 3

12/02/2016

- 1) For converting the House votes dataset into a Weka readable format, I added nominal attributes, which were viewed from alphabet a through p for 16 headings with values “y”, “n”, and “?” for all of the votes and a nominal attribute for the class determination which was either democrat or republican. I then added all of the data to the data section of the arff file. This file is labeled “house-votes-84.arff”. I opened this file in Weka and ran a Naïve Bayes rule on it with 5 fold cross validation and obtained the following results:

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes
Relation: house-votes-84
Instances: 435
Attributes: 17

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	392	90.1149 %
Incorrectly Classified Instances	43	9.8851 %
Kappa statistic	0.7949	
Mean absolute error	0.0994	
Root mean squared error	0.2975	
Relative absolute error	20.963 %	
Root relative squared error	61.1088 %	
Total Number of Instances	435	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.917	0.109	0.842	0.917	0.877	0.972	republican
	0.891	0.083	0.944	0.891	0.917	0.972	democrat
Weighted Avg.	0.901	0.093	0.905	0.901	0.902	0.972	

=== Confusion Matrix ===

a b <-- classified as
154 14 | a = republican
29 238 | b = democrat

The data accuracy was 90.1149%.

2nd Task

Ran a Naïve Bayes rule on Breast Cancer file with 5 fold cross validation. The result obtained are as follows:

=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes

Relation: breast_cancer_92

Instances: 699

Attributes: 10

clump_thickness
uniformity_of_cell_size
uniformity_of_cell_shape
marginal_adhesion
single_epithelial
bare_nucleoli
bland_chromatin
normal_nucleoli
mitoses
class

Test mode:5-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Attribute	Class
	2 4
	(0.65) (0.35)

=====

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	672	96.1373 %
Incorrectly Classified Instances	27	3.8627 %
Kappa statistic	0.9159	
Mean absolute error	0.0398	
Root mean squared error	0.1961	
Relative absolute error	8.8157 %	
Root relative squared error	41.2581 %	
Total Number of Instances	699	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
---------	---------	-----------	--------	-----------	----------	-------

	0.952	0.021	0.989	0.952	0.97	0.987	2
	0.979	0.048	0.915	0.979	0.946	0.983	4
Weighted Avg.	0.961	0.03	0.963	0.961	0.962		0.986

==== Confusion Matrix ====

```

a  b  <-- classified as
436 22 | a = 2
5 236 | b = 4

```

The accuracy in this case was 96.1373%.

For next step, I used the built in Unsupervised discretization Weka Filter. In addition to that, I ran a Naïve Bayes Rule on it with 5 fold cross validation and obtained the following results.

Task 3

==== Run information ====

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: breast_cancer_92-weka.filters.supervised.attribute.Discretize-Rfirst-last

Instances: 699

Attributes: 10

```

clump_thickness
uniformity_of_cell_size
uniformity_of_cell_shape
marginal_adhesion
single_epithelial
bare_nucleoli
bland_chromatin
normal_nucleoli
mitoses
class

```

Test mode: 5-fold cross-validation

==== Classifier model (full training set) ====

Naive Bayes Classifier

	Class	
Attribute	2	4
	(0.65)	(0.35)

=====

Time taken to build model: 0 seconds

==== Stratified cross-validation ====

=== Summary ===

Correctly Classified Instances	680	97.2818 %
Incorrectly Classified Instances	19	2.7182 %
Kappa statistic	0.9406	
Mean absolute error	0.0284	
Root mean squared error	0.1608	
Relative absolute error	6.2734 %	
Root relative squared error	33.8275 %	
Total Number of Instances	699	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.965	0.012	0.993	0.965	0.979	0.992	2
	0.988	0.035	0.937	0.988	0.962	0.992	4
Weighted Avg.	0.973	0.02	0.974	0.973	0.973	0.992	

=== Confusion Matrix ===

```
a  b  <-- classified as
442 16 | a = 2
3 238 | b = 4
```

The accuracy in this case was 97.2818%.

The results after the use of a discretization filter made the results more accurate. It is also notable that it didn't change the classification of class 4 instances, but it did lower the rate of false negatives for class 2 instances.