CSC4008 2019 Term2

Assignment #9

Deadline: 2020/5/3, 23:59
(This is a strict deadline. Submissions are not acceptable after the deadline)
Percentage: 10%

Purpose: Learn to naïve Bayes classifier and model evaluation; implement naïve Bayes classifier, cross-validation, and model performance measures.

Problem for implementation

Dataset: credit_g.arff (1000 instances, 21 attributes) (Path: Weka-3-8-4/Data/) or credit_g.csv provided.
Algorithms: NaïveBayes
Cross-validation: k-fold cross-validation

(1) Use a programming language that you are familiar with, such as Python, C++, or R to implement an algorithm for naïve Bayes classifier (30%). An implementation of k-fold cross-validation is required (10%). Compute precision, sensitivity (recall), specificity, and F measure (10%). It is not allowed to apply any packages (in a programming language) that are directly relevant to the problem. For numeric attributes in the data set, assume Gaussian distribution (package allowed). Show your code and execution results. Refer to the Weka results (Figure 1) to present your results.

(2) Compare your result with the ones you get using Weka, "NaïveBayes". Write a report to compare the results of models established at different parameter settings, including different k (10%), and selection of attributes (20%). Describe your methods for attribution selection, manually selection is not prohibited.

(3) Implement the "Bagging" strategy to improve the performance of models established in (1) (20%).

Notice: you can choose either (1)(2) or (1)(2)(3). The points you can get are different due to the degree of difficulty.

Submission: You need to submit a **studentID.zip** (replace your ID with studentID) which contains your codes and report.

```
Relation:     german_credit
Instances:    1000
Attributes:   21
              checking_status
              duration
              credit_history
              purpose
              credit_amount
              savings_status
              employment
              installment_commitment
              personal_status
              other_parties
              residence_since
              property_magnitude
              age
              other_payment_plans
              housing
              existing_credits
              job
              num_dependents
              own_telephone
              foreign_worker
              class
Test mode:    10-fold cross-validation

=== Summary ===

Correctly Classified Instances         754               75.4    %
Incorrectly Classified Instances       246               24.6    %
Coverage of cases (0.95 level)          96.4    %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      Class
                 0.864    0.503    0.800      0.864    0.831      0.385    good
                 0.497    0.136    0.611      0.497    0.548      0.385    bad
Weighted Avg.    0.754    0.393    0.743      0.754    0.746      0.385

=== Confusion Matrix ===

   a   b   <-- classified as
 605  95 |   a = good
 151 149 |   b = bad
```

Figure 1. Naïve Bayes Classifier to credit_g data set using WeKa.