Project 1 – Classification

Instructions and Experiments

Note: Please read the entire project description before you begin.

The goal of this project is to analyze the performance of classification algorithms on several synthetic and real-world data sets. This will be done in the following steps:

1. First, you will explore the data sets.
2. Next, you will perform a series of experiments on which you will be asked to answer a series of questions. For these experiments, you will be using the Weka data mining software.
3. Compile your answers in the form of a report.

Weka

Download an appropriate version from:

http://www.cs.waikato.ac.nz/ml/weka/downloading.html

The manual called WekaManual.pdf is provided. Sections 4.1-4.3 would be helpful for this project.
The following are names used by Weka for the classification algorithms that we will be using:

Algorithm	**Weka Name**
Decision Trees (DT)	J48
Naive Bayes	NaiveBayes
K-nearest neighbors (KNN)	IBk (set k appropriately)

Note that Weka assumes by default that the class attribute is the last column.

Before you begin

1. Visually explore the data sets in all experiments, and consider the following:
   - types of attributes
   - class distribution
   - which attributes appear to be good predictors, if any
   - possible correlation between attributes
   - any special structure that you might observe

   Note: The discussion of this exploration is not required in the report, but this step will help you get ready to answer the questions that follow.
2. Use precision and recall to measure performance.
3. Use the default parameters for all the classifiers in Weka, unless specified otherwise.

Report and Submission

- Collect output from your experiments. Submit Weka output electronically as a single zipped file using the Project 1 Moodle submit tool.
- Write a report addressing the experiment questions. Page limit: 6 pages, no smaller than 11pt font, 1-column format. **Your project will be evaluated based only on what you write on the first 6 pages of this report.**
- If you are a **UNITE** student, you should upload your Weka output on moodle like other students and submit your project report via UNITE as homework.
- Do not simply copy and paste the output from Weka into your report. Conserve space. Output should be submitted electronically -- we will look at your output if something is ambiguous in your report.

Experiment 1

Datasets to be used for this experiment:

Dataset	Filename	Description
Dataset 1	Exp1a.arff	Figure 1a.
Dataset 2	Exp1b.arff	Figure 1b.
Dataset 3	Exp1c.arff	Figure 1c.

The classification algorithms to be compared here are: Decision tree, Naive Bayes and k-NN classifier for k = 1, k = 10

- What are the test options provided in Weka?
- What is the default test option? Briefly explain the default test option.

Run the four algorithms for all the three datasets corresponding to Experiment 1 in the table above. For decision tree, change the following settings in classifier property editor: numFolds = 3, reducedErrorPruning=True.

Use the default settings for the other classifiers. For each classifier, report the precision and recall obtained from each of the three datasets for the positive class (+1) and answer the following questions:

1. For k-NN classifier, what is the distance weighting method used in the default option?
2. For k-NN classifier when k=1, how does the performance vary among different datasets?
3. For k-NN classifier when k=10, how does the performance vary among different datasets?
4. For decision tree, how does the performance vary among different datasets?
5. For Naive Bayes classifier, how does the performance vary among different datasets?
6. Depending on your conclusion above, compare the performance of the four classifiers. Explain the reasons for the observed differences in performance.

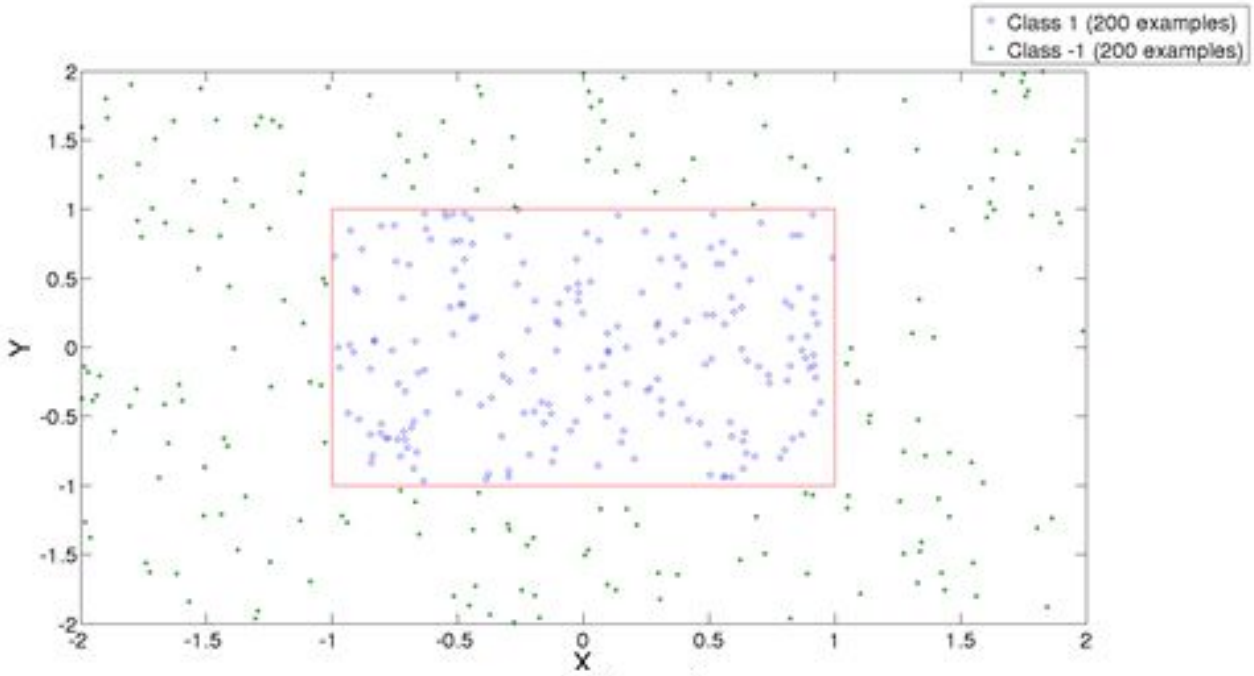

Figure 1a.

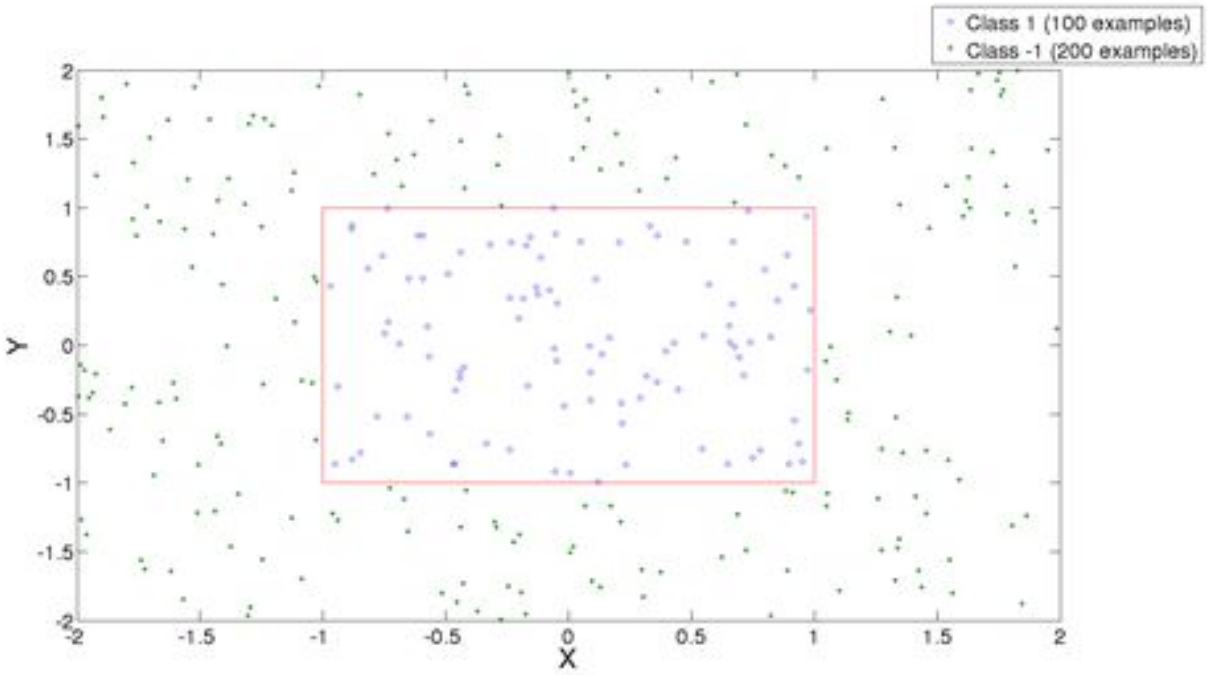

Figure 1b.

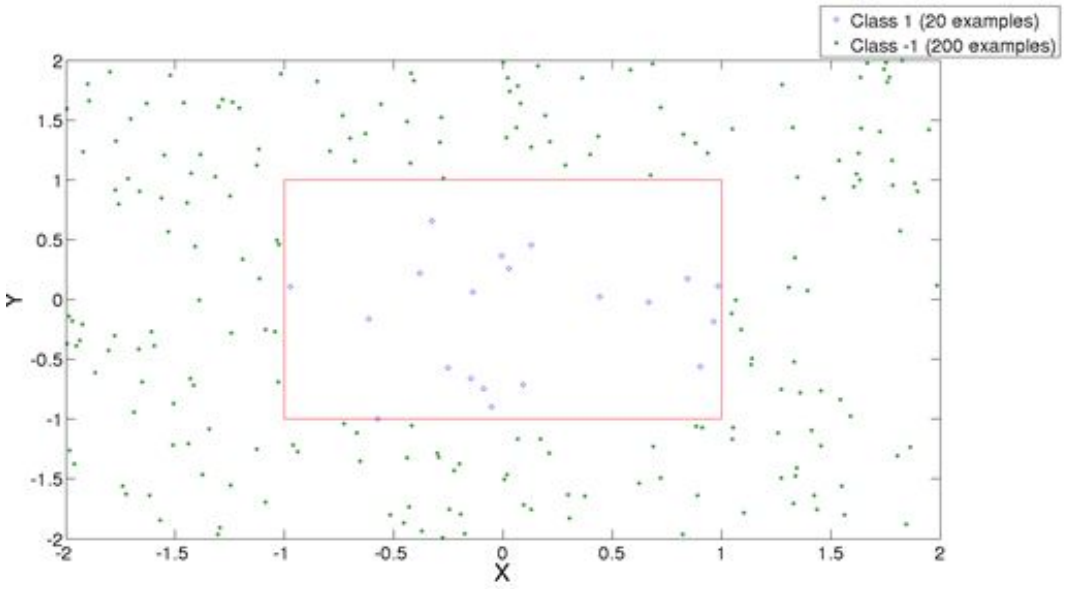

Figure 1c.

Experiment 2

Datasets to be used for this experiment:

Dataset	Filename	Description
Dataset 1 – Training set	Exp2a-train.arff	Figure 2a.
Dataset 1 – Test set	Exp2a-test.arff	This is the test set for Dataset 1
Dataset 2 – Training set	Exp2b-train.arff	Figure 2b.
Dataset 2 – Test set	Exp2b-test.arff	This is the test set for Dataset 2

The above Datasets (1, 2) use different training size (12% and 25% respectively) of the original dataset.

In this Experiment, we will only the Decision tree with the following settings: collapseTree=False, ReduceErrorPruning=False, subtreeRaising=False, unpruned=True, useMDLcorrection=False.

1. Run the algorithm for Dataset 1, first for the training set to train your model and then for the test set by using the appropriate test options. Provide the plot for Training and Test error rates with respect to the number of leaf nodes. Also, fill in the values in Table 1. Briefly explain the results.
   In order to get your results, use the values for setting “minNumObj” given in

Table 1. For these ten different cases, Weka automatically chooses the number of leaves. You should get the values of the second column when you enter in Weka the corresponding values of the first column.

2. Follow the same procedure as in (1) to run the algorithm for Dataset 2 (training and test sets). Provide the plot for Training and Test error rates with respect to the number of leaf nodes, and fill in the values in Table 2. Briefly explain the results.
3. Compare the two plots and explain the reasons for the observed differences if any.

You can use the tool of your choice to create the requested plots.

Minimum Number of Instances per leaf (**minNumObj**)	Number of leaves (leaf nodes)	Training error (%)	Test error (%)
1	147		
2	116		
4	90		
8	70		
16	37		
32	24		
64	17		
128	9		
256	4		
512	2		

Table 1.

Minimum Number of Instances per leaf (**minNumObj**)	Number of leaves (leaf nodes)	Training error (%)	Test error (%)
8	141		
16	91		
32	71		
64	32		
128	18		
256	7		
512	4		
1080	2		

Table 2.

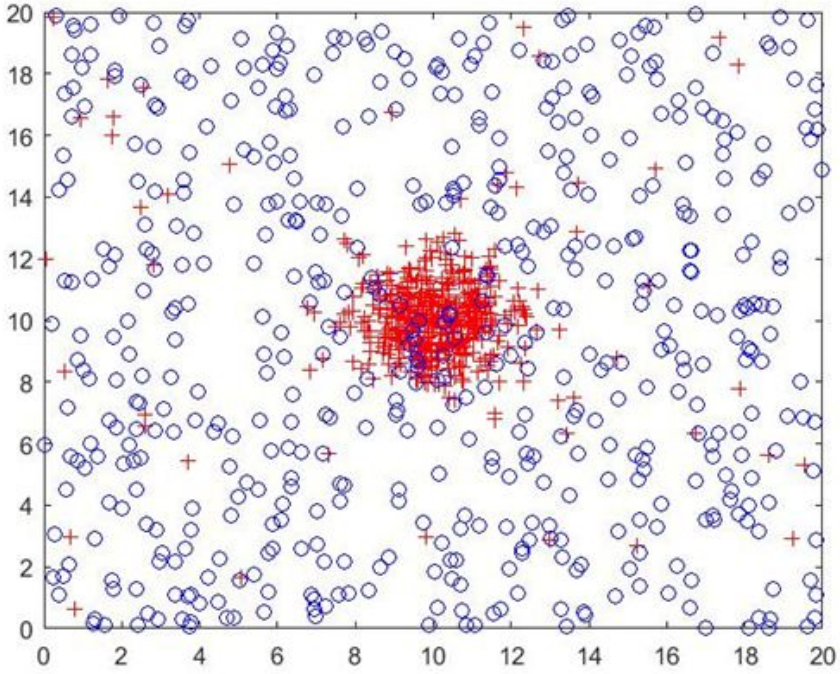

Figure 2a.

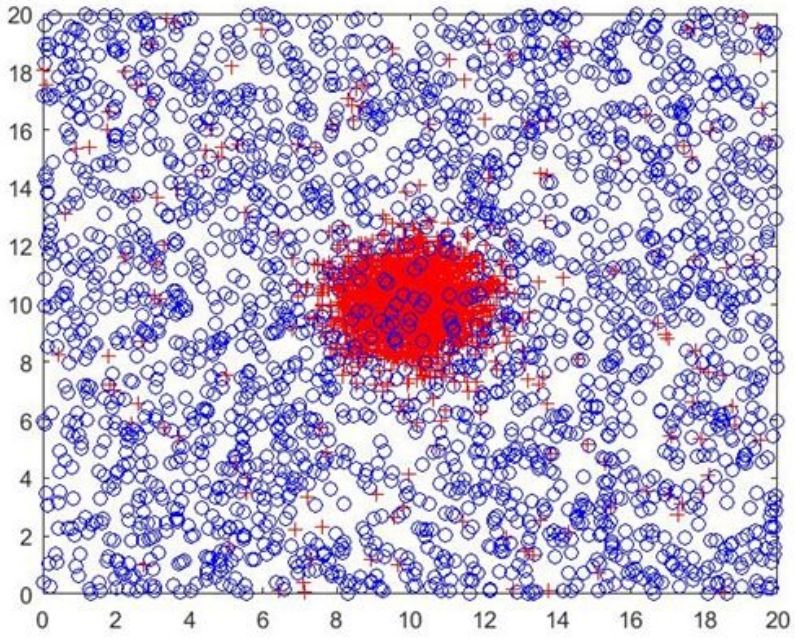

Figure 2b.

Experiment 3

Datasets to be used for this experiment:

Dataset	Filename	Description
Dataset 1 – Training set	Exp3a_train.arff	Figure 3a.
Dataset 1 – Test set	Exp3a_test.arff	This is the test set for Dataset 1.
Dataset 2 – Training set	Exp3b_train.arff	This dataset includes the 2 attributes in "Dataset 1" and 95 additional noisy attributes for the same set of examples. The values for the noisy attributes are randomly assigned from the uniform distribution on [0,1].
Dataset 2 – Test set	Exp3b_test.arff	This is the test set for Dataset 2.

The classification algorithms to be compared here are: Decision tree, Naive Bayes and kNN with k=3. For decision tree, change the following settings in classifier property editor: numFolds=4, reducedErrorPruning=True, useMDLcorrection=False. Use the default settings for the other classifiers.

1. Run the three algorithms on Dataset 1 for Experiment 3 (see table above). First run each algorithm for the training set to train your model and then for the corresponding test set, by using the appropriate test options. Report the precision and recall obtained from the test set of each Dataset for the positive class (+1) and answer the following questions:

   a) Which algorithm(s) are the best performers?
   b) Comment on the collective characteristics of the attributes that lead to such performance from these classifiers.
   c) The Bayes classifier is basically guessing. Explain why.

2. Now, use only the Decision Tree with the following settings: collapseTree=False, ReduceErrorPruning=False, subtreeRaising=False, unpruned=True, useMDLcorrection=False.

   Run the algorithm for Dataset 1, first for the training set to train your model and then for the test set by using the appropriate test options. Provide the plot for Training and

Test error rates with respect to the number of leaf nodes. Also, fill in the values in Table 3. Briefly explain the results.

In order to get your results, use the values for setting "minNumObj" given in Table 3. For these five different cases, Weka creates automatically the number of leaves. You should get the values of the second column when you enter in Weka the corresponding values of the first column.

3. Run the three algorithms on Dataset 2 for Experiment 3 (see table above). First, run each algorithm for the training set to train your model and then for the corresponding test set, by using the appropriate test options. Report the precision and recall obtained from the test set of each Dataset for the positive class (+1) and answer the following questions:
   a) Which algorithm(s) are the best performers?
   b) Which algorithm(s) have close to random performance?
   c) Which algorithm(s) show a big drop of performance compared to "Dataset 1", and what are the reasons for such drop?
   d) Which algorithm(s) perform the same as they did for "Dataset 1" and why?

4. Now, only use the Decision Tree with the following settings: collapseTree=False, ReduceErrorPruning=False, subtreeRaising=False, unpruned=True, useMDLcorrection=False.

Run the algorithm for Dataset 2, first for the training set to train your model and then for the test set by using the appropriate test options. Provide the plot for Training and Test error rates with respect to the number of leaf nodes. Also, fill in the values in Table 4. Briefly explain the results.

In order to get your results, use the values for setting "minNumObj" given in Table 4. For these nine different cases, Weka creates automatically the number of leaves. You should get the values of the second column when you enter in Weka the corresponding values of the first column.

You can use the tool of your choice to create the requested plots.

Table 3.

Minimum Number of Instances per leaf (**minNumObj**)	Number of leaves (leaf nodes)	Training error (%)	Test error (%)
1	10		
2	9		
32	7		
50	4		
70	3		

Minimum Number of Instances per leaf (**minNumObj**)	Number of leaves (leaf nodes)	Training error (%)	Test error (%)
1	49		
2	47		
4	37		
8	27		
16	15		
32	7		
50	5		
70	4		

Table 4.

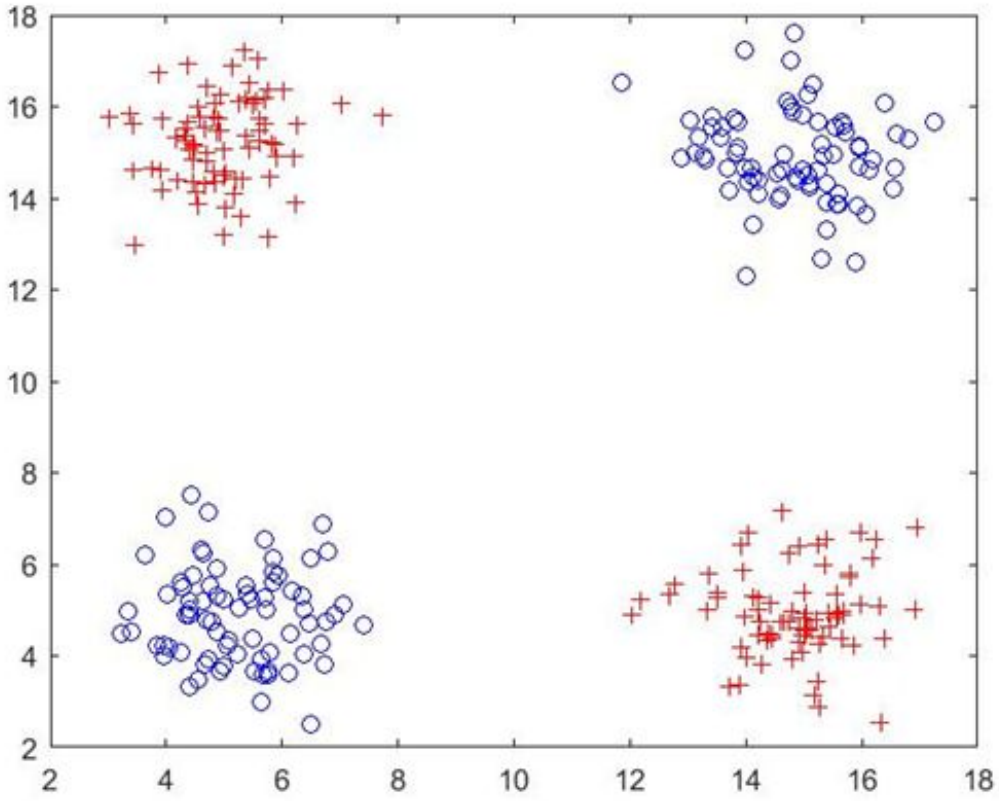

Figure 3a.