Practical 4 Week 5

**Classification: Basic Practice**

# What are we doing?

Using what you learned from lectures (Lecture 4) and relevant reading materials, you will answer some review questions. These questions are for your self-review on topics covered: Supervised learning (classification). You will need to review lecture and reading materials or seek for other resources (e.g. Googling), in order to answer questions. You will also use WEKA to run two construct simple classifiers using a given training data set, based on two basic classification methods: OneR and Decision Tree.

**Submission:**

You are required to submit one document containing your answers via the weekly-practical submission box (available on CP1407 LearnJCU)

For Practical tasks, screen capture your computer screen after the completion of each task and include the captured image in your document to submit.

# Self-Review Questions

1. Define the following terms (compare and contrast concepts) :
2. Supervised learning and Unsupervised learning
3. Classification and Prediction
4. Confusion matrix (Contingency table), Accuracy, and Error rate
5. Cross-validation and Bootstrap
6. Outline major steps required by the classification process.
7. With examples, explain the difference between accuracy and precision.
8. Why is a confusion-matrix-based evaluation method better than a simple accuracy based measure?
9. Compare the advantages and disadvantages of eager classification versus lazy classification.
10. Briefly explain overfitting and underfitting. Which of these are important issues in classification?

# Practical Tasks: Classification through WEKA

This session demonstrates the general process of using two basic classification methods (1R and Decision Tree) in WEKA.

Before starting the classification practice provided in this document, quickly go through two slides provided in LearnJCU: “Prac2 (Week3) Weka Introduction.ppt” and “Weka\_Classifier.ppt”.

The first one was already provided and covered in a previous prac session (Prac 02) of this subject but recommended to be reviewed again by you if needed.

The second slide sets presents the basic/common practice in using WEKA for classification tasks.

In this session, we are going to use a machine learning software (WEKA) to learn some examples of classification tasks.

**Task 1: Using OneR (1R) Classification**

OneR classification generates a one-level decision tree, which is expressed in the form of a set of rules that all test one particular attribute. OneR is a simple, cheap method that often comes up with quite good rules for characterizing the structure in data. It turns out that simple rules frequently achieve surprisingly high accuracy. Perhaps this is because the structure underlying many real-world datasets is quite rudimentary, and just one attribute is sufficient to determine the class of an instance quite accurately. In any event, it is always a good plan to try the simplest things first.

The idea is this: We make rules that test a single attribute and branch accordingly. Each branch corresponds to a different value of the attribute. It is obvious which is the best classification to give each branch: use the class that occurs most often in the training data. Then the error rate of the rules can easily be determined: just count the errors that occur on the training data, that is, the number of instances that do not have the majority class. Each attribute generates a different set of rules, one rule for every value of the attribute. Evaluate the error rate for each attribute’s rule set and choose the best.

Figure 1 shows statistics about the weather data. “Outlook” and “Humidity” exhibit the smallest error rate (4/14), thus either can be the best attribute that classifies all the examples.

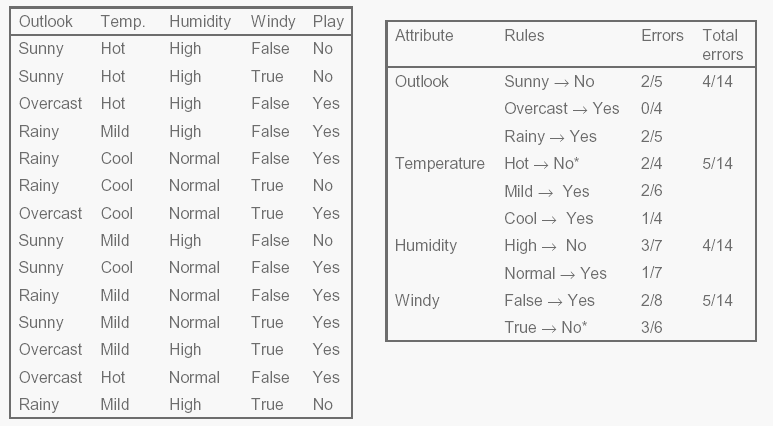
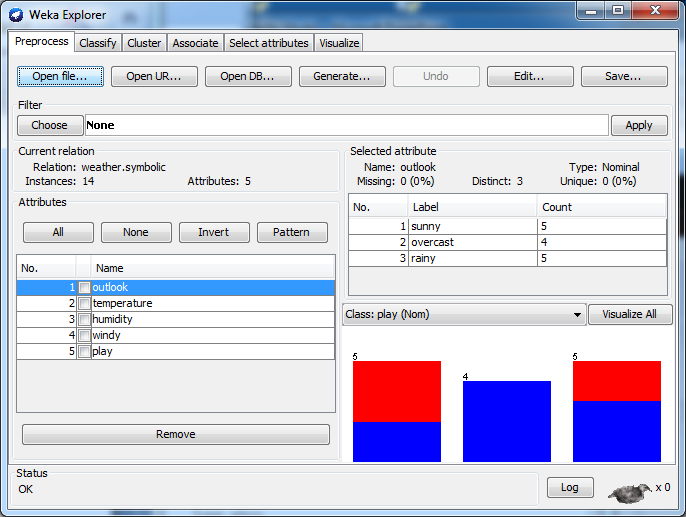


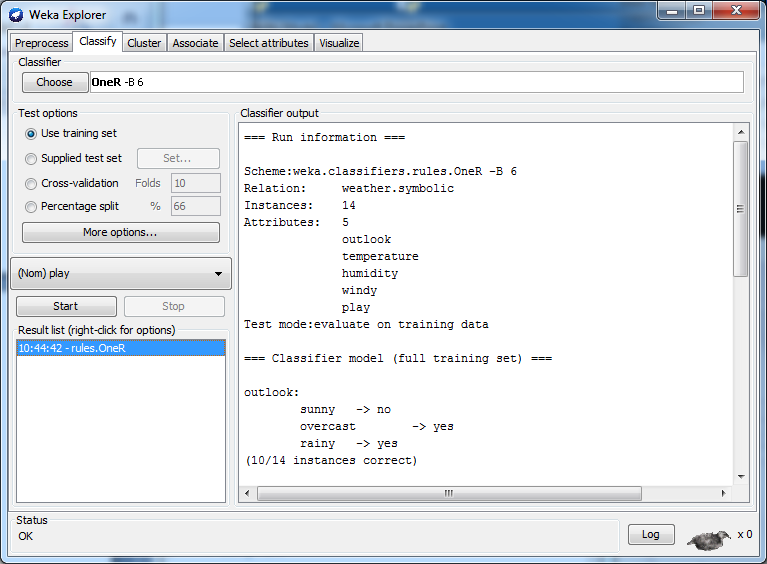
Figure 1: Evaluating the attributes in the weather data.

**Running 1R in Weka with the weather data**

1. Run WEKA program
2. Click “Explorer” from the initial Weka GUI Chooser to activate “Explorer” window.
3. Click “Open file..” under “Preprocess” tab to open “weather.nominal” file (weather.nominal.arff) under **data** directory (under Weka program folder).



1. Click “Classify” button and choose “OneR” (under “rules” classifier folder) as your classifier.
2. Choose “Use training set” as Test options.
3. Click “Start” to get OneR classification.



1. You will get the result as follows:

=== Run information ===

Scheme: weka.classifiers.rules.OneR -B 6

Relation: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

Test mode:evaluate on training data

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

outlook:

sunny -> no

overcast -> yes

rainy -> yes

(10/14 instances correct)

Time taken to build model: 0 seconds

=== Evaluation on training set ===

=== Summary ===

…

…

Correctly Classified Instances 10 71.4286 %

Incorrectly Classified Instances 4 28.5714 %

Kappa statistic 0.3778

Mean absolute error 0.2857

Root mean squared error 0.5345

Relative absolute error 61.5385 %

Root relative squared error 111.4773 %

Total Number of Instances 14

=== Detailed Accuracy By Class ===

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

0.778 0.4 0.778 0.778 0.778 0.689 yes

0.6 0.222 0.6 0.6 0.6 0.689 no

Weighted Avg. 0.714 0.337 0.714 0.714 0.714 0.689

=== Confusion Matrix ===

a b <-- classified as

7 2 | a = yes

2 3 | b = no

**Task 2: Constructing a decision tree with the weather data**

1. Load “weather.nominal” file under **data** directory.
2. Click “Classify” button and choose “J48” (under “trees” classifier folder).
3. Choose “Use training set” as Test options.
4. Click “Start” to get a decision tree classification.
5. You will get the results as follows:

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: weather.symbolic

Instances: 14

Attributes: 5

outlook

temperature

humidity

windy

play

Test mode:evaluate on training data

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

J48 pruned tree

------------------

outlook = sunny

| humidity = high: no (3.0)

| humidity = normal: yes (2.0)

outlook = overcast: yes (4.0)

outlook = rainy

| windy = TRUE: no (2.0)

| windy = FALSE: yes (3.0)

Number of Leaves : 5

Size of the tree : 8

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances 14 100 %

Incorrectly Classified Instances 0 0 %

Kappa statistic 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0 %

Root relative squared error 0 %

Total Number of Instances 14

=== Detailed Accuracy By Class ===

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

1 0 1 1 1 1 yes

1 0 1 1 1 1 no

Weighted Avg. 1 0 1 1 1 1

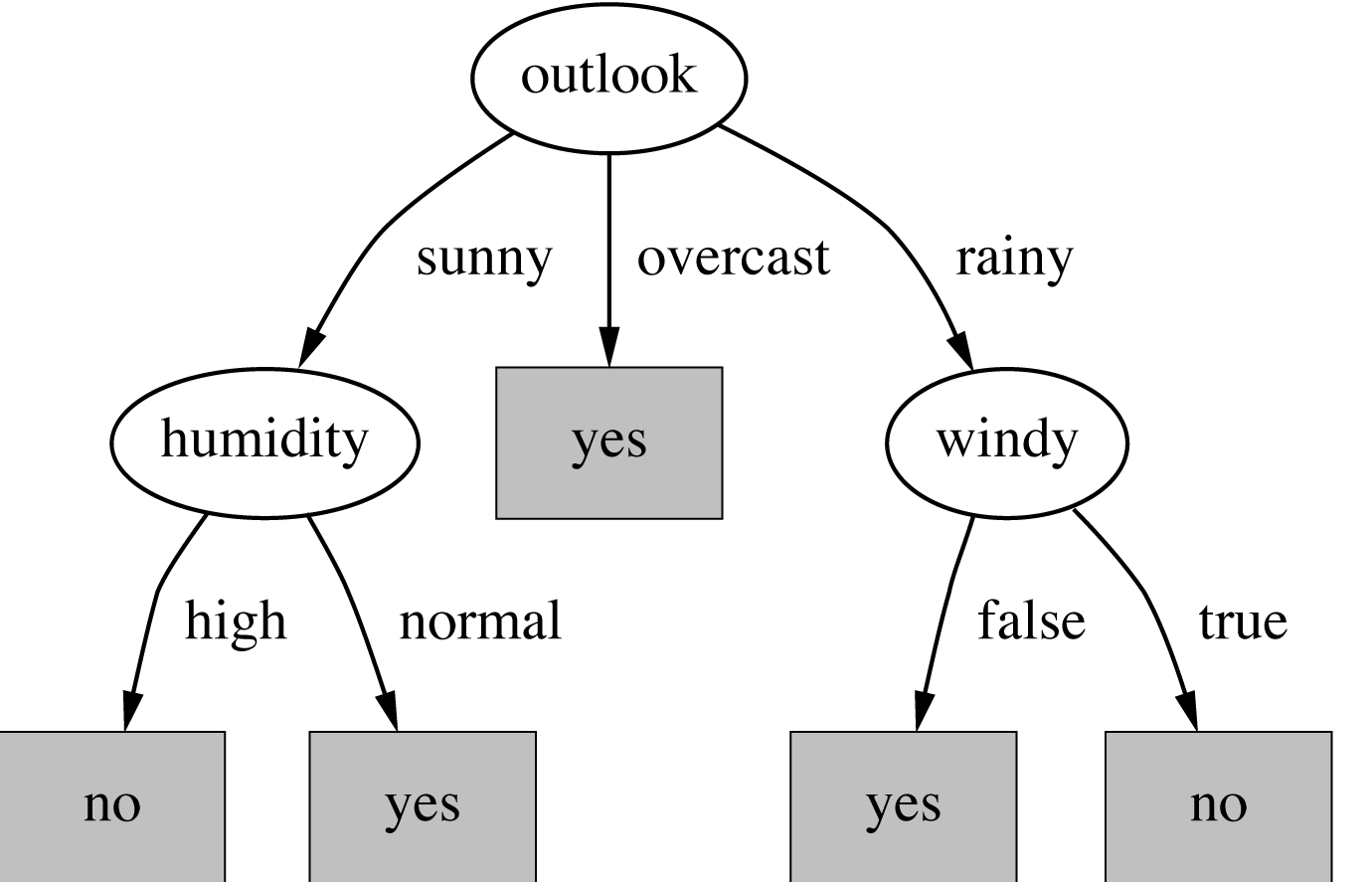
=== Confusion Matrix ===

a b <-- classified as

9 0 | a = yes

0 5 | b = no

1. In the result above, there are 5 pruned rules:
   1. If Outlook = sunny and Humidity = high Then Play = NO
   2. If Outlook = sunny and Humidity = normal Then Play = YES
   3. If Outlook = overcast Then Play = YES
   4. If Outlook = rainy and Windy = true Then Play = NO
   5. If Outlook = rainy and Windy = false Then Play = YES
2. A decision tree from the rules can be constructed as below.



1. Check if the decision tree covers all the examples in the training dataset.

**Task 3: Testing the decision tree classifier constructed**

We want to test if we play tennis in a new day with the values below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A new day | | | | |
| Outlook | Temperature | Humidity | Windy | Play |
| Sunny | Cool | High | True | ? |

1. Create an empty ARFF test file named “weather.test1.nominal.arff” in the data subdirectory using your favourite text editor (e.g. notepad).
2. Add all the necessary tags and add a data entry with “sunny,cool,high,TRUE,no” values. The test file “weather.test1.nominal.arff” will be as follows:

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}

@attribute temperature {hot, mild, cool}

@attribute humidity {high, normal}

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data

sunny,cool,high,TRUE,no

1. Choose “Supplied test set” as test options and it will bring up a “Test instances” window.
2. Open the test file created (weather.test1.nominal.arff).
3. Click “Start” in the “Weka Knowledge Explorer” window to activate classification. (By doing this, the decision tree classifier developed in the previous task will be used to classify this new test data, and will show the classification results).