Tutorial 7: Connectionist and Statistical Language Processing

Introduction to Machine Learning

Date: 19.12.2001
Student's name:

1 Setting up the Weka Machine Learning Environment

First, add the following line to your `.bashrc` file:

```
export PATH=/proj/corpora/tools/bin:$PATH
```

Then log on to `gnome` using:

```
ssh gnome
```

Now you should be able to start the Weka Machine Learning Environment using the command:

```
weka &
```

Click on `Explorer` to launch the Weka Explorer. The files for this tutorial reside in:

```
/courses/connectionism.winter.01/tutorial7/
```

2 The Weka Input Format

We will first look at a simple data set to illustrate the input qformat that Weka is using. Have a look at the files `weather.nominal.arff` and `weather.arff` in the tutorial directory (using `more`, for instance).

Can you make sense of the format? What do you think the declarations @attribute and @data mean? Describe the syntax of these declarations.

3 Visualizing the Data

In the Weka Explorer choose the field `Preprocess` and click on the button `Open file` and open the file `weather.nominal.arff`. Check that the attributes of this data file are listed correctly.

Now choose the field `Visualize` that allows you to visualize the data set. Is there a single attribute that predicts the classification of `play` into `yes` and `no` reasonably accurately? Hint: To

explore this question, plot `instance_number` on the x axis, `play` on the color axis, and use the y axis to vary the attribute you are looking at.

Now use the visualizer to look at the file `weather.arff`. What changes?

4 Learning a Decision Tree

Load the file `weather.nominal.arff`. In the Weka Explorer choose the field `Classify` and click on the field `Classifier`. In the window that pops up click on `weka.classifiers.ZeroR` and choose `weka.classifiers.j48.J48` instead. This is a decision tree induction algorithm called C4.5 (that we will discuss in detail in the next lecture).

Under `test options` click on `use training set` to tell Weka to test the resulting decision tree on the training set. Then click on `start` to run the classifier.

Inspect the output of the classifier. Draw the decision tree that it came up with and give the performance on the test set (precision and recall). Is this an example of realistic learning behavior?

5 Testing on Unseen Data

Now test the classifier on an unseen test data set. Under `test options` click on `supplied test set` and select the file `weather.test.arff` (7 instances).

How does the decision tree perform on the test test? Compare precision and recall. Weka also outputs a confusion matrix. Can you guess what this is? How could it be used in evaluating the performance of the model?