## Week 5 Weka Assignment (75 points)

## **Assignment Due Date and Time**

* Sunday at 11:59 p.m. ET.

## **Purpose**

To gain an understanding of how to apply a decision tree model.

## **Description**

Consider the following two sets of data to be analyzed separately:

1. The **mushroom** data set with the "edible/poisonous" attribute as the dependent variable. This dataset is clean and will load into Weka without any editing.
2. The **usnews** data set. This dataset contains college data taken from the U.S. News & World Report’s Guide to America’s Best Colleges. Here the “private/public” attribute is the dependent variable. Note that even though the values of this attribute are 0s and 1s, this is a categorical (not a numeric!) attribute. This dataset is not clean and may contain some minor errors. You may need to clean the data and fix anomalies before it will load into Weka. Some preprocessing will be needed so that columns have the correct data types and values.

You are to construct J48 (C4.5) decision tree models that (a) explain the data as best as possible, and (b) generalize as much as possible. Use both the **hold-out method** (70-30 or 80-20 split) and **10-fold cross-validation** to demonstrate that your model parameters do indeed construct models that generalize well and to illustrate that your model explains your data well.

Both data sets represent raw problem domain data which you will need to first translate into the ARFF format and transform in order to build appropriate models. Once you have the data in the appropriate format you might want to consider the following data preparation questions:

1. Should instances with missing values be deleted?
2. Should attributes with missing values be deleted?
3. Should missing values be coded in a special way and then used in the data mining task?
4. Should numerical values be binned, or should we let J48 figure out the optimal splits in the data?
5. Are some of the attributes categorical even though their levels are expressed as numbers?

You will need to perform a certain amount of experimentation in order to devise a data set that allows you to construct an optimal model. You will also need to experiment with the pruning parameters for J48 to find the optimal model.

## **Your Assignment Report**

Your report should describe three aspects of your data mining effort for each data set:

1. data preparation,
2. the models, their parameters, and their performance,
3. model description – does the model provide any interesting insights?

## **Submission Instructions**

## Upload your written report to the Week 5 Weka Assignment submission area.