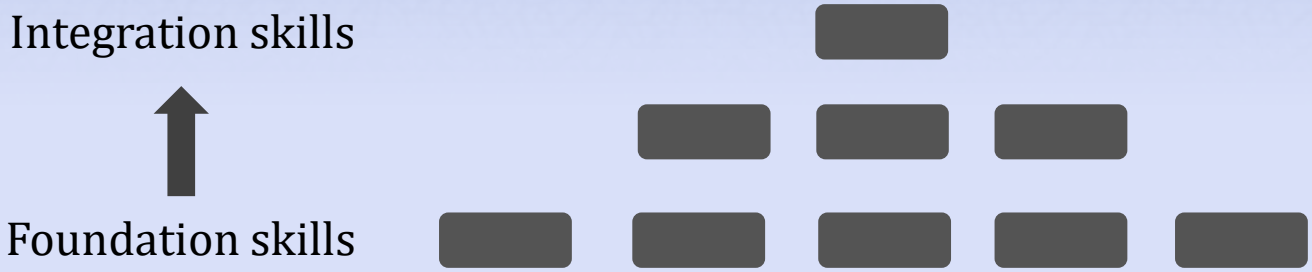
Teaching the Building Blocks of Undergraduate Research

by
John Aleshunas and Alyssa Dalton

Partitioning the Research Process into its Building Blocks



- Viewed as a hierarchy
- Start with foundation skills
- Develop toward integrated skills

A Hierarchy of Research Building Blocks

- Foundation skills
  - Polya's problem solving methodology
  - Manipulating data in text files
  - Visualizing data with descriptive statistics
- Integration skills
  - Organizing an experiment
  - Choosing/designing data to test a hypothesis
  - V-fold training/testing

An Example Assignment

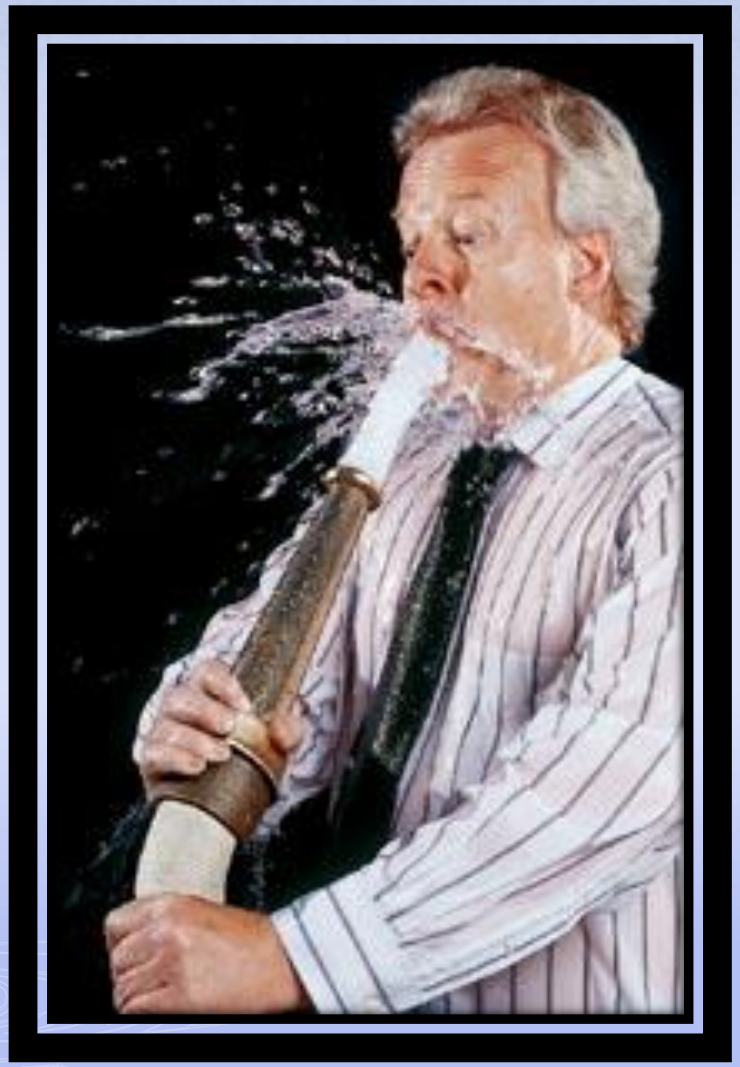
MATH 3210
Assignment #4 (The Sommelier Project Revisited)

Decision Tree Comparison: The dataset Wine.xls contains the values of thirteen attributes for three classes of Italian wines. In this assignment you will construct a classification decision tree for the data using C4.5 and compare that result to the result you developed in Assignment #2. (The Sommelier Project).

This assignment requires the following operational tasks:

- Download and verify the C4.5 executable
- Download the Wine.xls dataset
- Create a training dataset and a test dataset
- Format the datasets to conform to the application input file structure for the C4.5 application
- Ensure that you save your execution output for analysis

Data Mining: A Student Perspective



Alyssa Dalton

Executive Summary

This report explains how a rule set was constructed using the C4.5 application to classify three types of wine based on the attributes of each wine. There are 153 instances of three classes of wine. All of these instances are listed in the wine data set along with values for 13 attributes of each of the wines. Using these values, a rule set was constructed, which when applied can classify wines within the three classes. This was done by running the data through the C4.5 application, which gave a classification tree (based on entropy) Parenthesis not needed here – remove them for the training data, which the application also expresses as rule set #2. The test set was used to measure accuracy of the rule set. The accuracy was then compared to the accuracy of the previously developed rule set #1 from Assignment #2. The training set gave a 0.82% error for both rule sets, and the test sets gave a 9.38% error for rule set #1 and a 12.90% error for rule set #2. Therefore, rule set #1 gives a higher accuracy to classify new wines within the three classes and should be used rather than rule set #2.

Explain the wine data set using phrasing like, "The wine data set used in this analysis consists of 153 instances…(etc.)". Address how the rule set was developed in Assignment #2. State and briefly explain the technique you used. This is an important point of this contrast study. While I understand what you are trying to do by referring to the two rule sets as rule set #1 and rule set #2 this is an awkward phrasing. It is much clearer to just refer to them as the Assignment #2 rule set and the C4.5 rule set. Briefly explain what entropy is and how it is used to develop a decision tree rule set. Good comparison of the error rates for both rule sets but is there anything you can state regarding the comparison of these two methodologies? [see my comment in your analysis section below for some other potential points of comparison]

Problem Description

This analysis should compare the rule set developed in Assignment #2 to the decision tree (which is also given in a written rule set- rule set #2) constructed using C4.5 for the wine data set.

Address how the rule set was developed in Assignment #2. State the technique you used. This is an important point of this contrast study.

Refer to the C4.5 output as a decision tree rule set. That way you can eliminate the parenthetical content. While I understand what you are trying to do by referring to the two rule sets as rule set #1 and rule set #2 this is an awkward phrasing. It is much clearer to just refer to them as the Assignment #2 rule set and the C4.5 rule set.

Analysis Technique

The entire wine data set contains 153 wines. Each of the wines is categorized into class 1, 2, or 3. Class 1 contains 47 wines, Class 2 contains 61 wines, and Class 3 contains 45 wines. Along

```
Running the C4.5 decision tree induction process

C4.5 [release 5] decision tree generator        Fri F
----------------------------------------

    Options:
            File stem <clusterfuzzy3>

ERROR:  cannot open file clusterfuzzy3.names
```

The Result