DECISION TREES and DECISION RULES

During the next two weeks you will develop a method that will produce a *multifactor* genetic test to determine if an individual may develop cancer.

**Materials**:

* Professor Welch’s intro to this week: <https://youtu.be/BRLLUAgOT50>
* <https://www.youtube.com/watch?v=PbzjDIKWKSw>
* <https://en.wikipedia.org/wiki/Decision_tree_learning>
* <https://www.youtube.com/watch?v=7VeUPuFGJHk&t=9s>
* Chapter 8 – Decision Trees and Decision Rules (*sections 8.1, 8.2, 8.5*)

*Discovering Knowledge in Data*. D.R. Larose and C.D. Larose. Wiley. 2014.

<https://alice.library.ohio.edu/record=b5187242?>

(click on the link “Connect to resource OhioLink”)

* <https://en.wikipedia.org/wiki/Decision_tree>
* <https://en.wikipedia.org/wiki/Classification_rule>

C**oncepts to learn from the materials:**

* Decision tree classifiers (classification trees)
* Decision rules (classification rules)

**Quiz**:

After learning the concepts listed above, complete the BlackBoard quiz no later than Monday September 13, 9:39 am. The quiz will cover your understanding of the *concepts to learn from the materials* (see above). The quiz may include multiple choice, true-false, fill-in-the-blank, and/or matching questions.

Data Mining Activity: (*to be started after you complete the quiz*)

**PART 1:**

* **Due date**: no later than Thursday Sept. 16, 11:59 pm
* Complete steps 1-4 (see below)
* Submit results, report and program by email to welch@ohio.edu

**PART 2:**

* **Due date**: no later than Thursday Sept. 23, 11:59 pm
* Complete all steps (see below)
* Submit results, report and program by email to welch@ohio.edu

--------------------------------------PART 1 BEGINS HERE--------------------------------------

In this exercise you will construct a two-level decision tree to classify an individual sample as a member of either the **C** class or the **NC** class. To construct the decision tree, you will consider the quantity “**TP** – **FP**”in orderto select the best features for classifying samples. Complete the following steps to construct and test your tree for classifying samples.

1. Compute the quantity “**TP** – **FP**” for each feature (genetic mutation). Which feature yields the maximum value of “**TP** – **FP**”? Show the table of the top 10 features (ranked by their corresponding values of “**TP** – **FP**”) and their corresponding “**TP** – **FP**” values.
2. Select the most useful feature, **F**, for correctly classifying samples (i.e., the genetic mutation **F** that yields the maximal value for the quantity “**TP** - **FP**”).
3. Divide the samples by using the feature **F** to classify each feature as either
   1. group-A: has the mutation denoted by feature **F**

or

* 1. group-B: does not have the mutation denoted by feature **F**

1. Make a confusion matrix to represent the classification of all samples with genetic mutation **F**.

**Part I**:

Submit an email to [welch@ohio.edu](mailto:welch@ohio.edu) that contains a brief report, including the following:

* table of the top 10 features (see item 1 above)
* a list of samples in group-A and a list of samples in group-B (see item 3 above), and
* a confusion matrix (see item 4 above),
* a brief discussion of the results.

Additionally, attach the computer program that you developed for this activity and the output of your program (either a screenshot(s) or a file).

**NOTE***: you must develop your own computer program to accomplish this assignment. You ARE NOT permitted to use pre-existing programs for building decision trees or any other component of this project.*

**NOTE**: I may respond to your email submissions with questions about your methods, results, and/or interpretation. Please respond promptly to my questions.

--------------------------------------PART 2 BEGINS HERE--------------------------------------

1. Repeat steps 1 & 2 for group-A, resulting in selection of a feature **A** *that* *best classifies the samples* *in group-A*. Show the table of the *top 10 features* *for classifying samples in group-A* (ranked by their corresponding values of “**TP** – **FP**” *for group-A*) and their corresponding “**TP** – **FP**” values *for group-A*. Make a confusion matrix to represent the *classification of samples in group-A* with genetic mutation **A**.
2. Repeat steps 1 & 2 for group-B, resulting in selection of a feature **B** *that best classifies the samples in group-B*. Show the table of the *top 10 features* *for classifying samples in group-B* (ranked by their corresponding values of “**TP** – **FP**” *for group-B*) and their corresponding “**TP** – **FP**” values *for group-B*. Make a confusion matrix to represent the *classification of samples in group-B* with genetic mutation **B**.
3. The result is a ‘decision tree’ that contains the following:
4. a root node **R**
5. two children of **R**:

* node **A**
* node **B**

1. two children of **A**:

* node **A1**
* node **A2**

1. two children of **B**:
   * node **B1**

* node **B2**

1. Draw your resulting decision tree (*you may manually draw the resulting decision tree; you are not required to develop a computer program that displays your tree*). Label each internal (non-leaf) node (i.e., nodes **R**, **A**,and **B**) with the genetic mutation used to divide samples at that node. Label the arcs coming from a node with each of the possible values (‘0’ or ‘1’) of the genetic mutation that is represented by the node. Label each leaf node (i.e., nodes **A1**, **A2**,**B1**, and**B2**) with the appropriate class label, signifying the specific class that would be assigned by the tree to all samples at that node (see step 9 for a description of the classification procedure).
2. Your ‘decision tree’ can be used to classify each sample **S** by using the following classification rules:

If **S** has mutation **F** then

if **S** has mutation **A**

then classify **S** as **C**

else classify **S** as **NC**

else

if **S** has mutation **B**

then classify **S** as **C**

else classify **S** as **NC**

1. Use your decision tree to classify the following samples: *C1, C10, C50, NC5, and NC15*. For each of these samples, report which classification your decision tree would assign to the sample. How many times did your decision tree make the correct classification?

**Part II**:

Submit an email to [welch@ohio.edu](mailto:welch@ohio.edu) that contains a brief report, including the following:

* an annotated drawing of your decision tree (see item 8 above),
* tables of the top 10 features for the internal nodes of the decision tree, ranked by their corresponding values of “**TP** – **FP**” (see items 5 & 6 above),
* confusion matrices for the internal nodes of the decision tree (see items 5 & 6 above),
* your *specific* classification rules (see item 9 above),
* the results of classifying specific samples (see item 10 above), and
* a discussion and interpretation of your results.

Additionally, attach the computer program that you developed for this activity and the output of your program (either a screenshot(s) or a file).

**NOTE***: you must develop your own computer program to accomplish this assignment. You ARE NOT permitted to use pre-existing programs for building decision trees or any other component of this project.*

**NOTE**: I may respond to your email submissions with questions about your methods, results, and/or interpretation. Please respond promptly to my questions.