EVALUATING DECISION TREE CLASSIFIERS

This week we will learn how to perform cross-validation to measure the effectiveness of a decision tree classifier.

**Materials**:

* Dr. Welch’s Intro: <https://www.youtube.com/watch?v=XppToCDjF8o>
* Chapter 6 – Preparing to model the data

Chapter 14 –Model evaluation techniques (sections 14.3-14.7,14.9,14.10)

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

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

* <https://en.wikipedia.org/wiki/Cross-validation_(statistics)>

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

* overfitting
* supervised learning
* cross-validation
* confluence of results
* measures of performance used in model evaluation for the classification task

**Quiz**:

After learning the concepts listed above, complete the BlackBoard quiz no later than Monday October 4, 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.

**Supplementary Materials**:

* <https://en.wikipedia.org/wiki/Overfitting>
* <https://en.wikipedia.org/wiki/Supervised_learning>

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

* **Due date**: no later than Thursday Oct. 7, 11:59 pm
* Submit results, report and program by email to [welch@ohio.edu](mailto:welch@ohio.edu)

In this exercise you will use the method of cross-validation to assess the effectiveness of the decision tree classifier that you constructed previously.

**STEP 1**: *training*

Construct a decision tree as you did previously, but *construct the tree by considering only the* *training set, consisting of 2/3 of the samples, selected at random.*

**STEP 2**: *testing*

Use the decision rules derived from your decision tree (constructed in step1 by considering the training set) to *classify* *a test set of* *samples* (the remaining 1/3 of samples, which are not contained in the training set). Calculate the following metrics to characterize the resulting classification of all samples in the *test set*:

* + TP, FP, TN, FN
  + Accuracy
  + Sensitivity
  + Specificity
  + Precision
  + Miss rate
  + False discovery rate
  + False omission rate

**STEP 3**: *3-fold cross-validation*

Repeat the training and testing steps (see steps 1 & 2 above) three times, using different training and testing sets each time. Report the evaluation metrics for each test (see step 2 above). Also report the average evaluation metrics (across all 3 tests).

Submit an email to [welch@ohio.edu](mailto:welch@ohio.edu) that contains

1. a brief report that includes
   1. the genetic mutations selected by *each* of the 3 decision trees
   2. the metrics computed by 3-fold cross-validation
   3. a discussion and interpretation of the 3-fold cross-validation results
2. the computer program that you developed for this activity
3. 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, performing cross-validation, 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.