RANDOM FOREST CLASSIFIERS

This week we will learn about an advanced classification method called *random forest*. We will construct a random forest classifier that will perform a multifactor genetic test to determine if an individual may develop cancer.

**Materials**:

1. Prof Welch’s intro to random forests:

<https://www.youtube.com/watch?v=u6Ocjv8jiUM>

1. The wisdom of Crowds (length: 3:37)

<https://www.pbs.org/video/wisdom-crowds-ys0jjp/>

1. Building, using and evaluating Random Forests (length: 9:53)

<https://www.youtube.com/watch?v=J4Wdy0Wc_xQ>

1. Bootstrap sampling (length: 2:12)

<https://www.youtube.com/watch?v=tTZybQTE0dw>

1. Bootstrap aggregating

<https://en.wikipedia.org/wiki/Bootstrap_aggregating>

**Concepts to learn from the materials:**

* The 2 major steps of building a random forest
* The decision procedure used by a random forest classifier
* The definitions of the following terms:
  + “bootstrap sample”
  + “bagging”
  + “out-of-bag dataset”

**Quiz**:

After learning the concepts listed above, complete the BlackBoard quiz no later than Monday October 25, 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*)

Due date: no later than Thursday October 28, 11:59 pm

1. Develop a procedure to generate a random decision tree:
   1. Construct a bootstrap dataset from the full set of samples. The bootstrap data set should contain the same number of samples as the original dataset.
   2. Generate a depth 2 decision tree

* for *all* node splitting decisions, use EITHER the F function OR information gain (choose one method and always use it)
* for *each* node splitting decision, randomly select *sqrt*(***n***) features to consider (where, ***n***is the total number of genetic mutations)

Report the size of the out-of-bag dataset

List the samples in the out-of-bag dataset

Report the feature selected to split the root node of the tree

Report the features selected to split the child nodes of the tree

1. Generate a random forest
   1. Generate 10-25 random decision trees (use the procedure from steps 1a and 1b)
   2. Report a summary of the random forest:
      1. List of features selected to split the root nodes of the trees in your random forest (also report the number of times that each feature was selected to split a root node)
      2. List of features selected to split child nodes of the trees in your random forest (also report the number of times that each feature was selected to split a child node)
      3. A list of the sizes of the out-of-bag datasets
      4. Average size of the out-of-bag datasets
2. Build a random forest *classifier*, which can be used to classify any sample, ***S***:
   1. The *Random Forest Classifier* performs the function of classifying a single sample.

The specification for the *Random Forest Classifier* function is as follows:

* + 1. INPUT: a sample *S*
    2. PROCESSING: use the random forest to determine the classification for *S*
    3. OUTPUT: the classification for sample *S*, consisting of a value in the set {C,NC}
  1. Use *each* of the 10-25 random decision trees (from step 2) to classify the sample, ***S***.
  2. Keep track of the number of trees that classified ***S*** ascancer (**C**)and the number of trees that classified ***S*** asnon-cancer (**NC**).
  3. Using the result from step 4b, perform majority voting to make the classification decision for sample ***S***.

1. Use your random forest to classify the following:
   1. Samples *C1, C10, C100, NC5,* and *NC15*. Report the following for each sample:
      1. the classification result from the random forest
      2. the number of trees in the random forest that classified the sample as**C**
      3. the number of trees in the random forest that classified the sample as**NC**
   2. All samples in the out-of-bag dataset in step 1 (above). Report the following:
      1. TP, FP, TN, FN
      2. Accuracy, sensitivity, specificity, precision, miss rate, false discovery rate, false omission rate

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

* The results from steps 1, 2, and 4.
* Your report should include the following sections:
  1. Goals and objectives (2-3 bullet points)
  2. Summary of results (2-3 bullet points)
  3. Summary of methods (2-4 bullet points)
  4. Key results (tables, figures, stats, lists, etc.)
  5. Discussion (2-4 bullet points)

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.