FEATURE SELECTION

We have developed a prototype ensemble-based classifier for disease diagnosis. In this final phase of the project we will focus on *feature selection*, the identification of the mutations that provide the most significant insights for cancer research. In part 1 of this activity, you will exploit the power of a fully developed random forest to identify a collection of relevant mutations (this is called feature selection) in the data set that you’ve been analyzing the semester. In part 2 of this activity, you will explore feature associations to discover groups of genetic mutations which may work collectively to cause cancer (this is called feature association selection); feature association analysis will contribute additional features to the feature selection process. These discoveries will provide a starting point for further scientific research studies.

DATA MINING ACTIVITY: *Project Summary*

Part 1: *Deep classification analysis & Feature selection*

* Due date: no later than Thursday November 4, 11:59 pm

Part 2: *Deep association rule mining & Feature association selection*

* Due date: TBD

**PART 1**: *Deep classification analysis & Feature selection*

You have learned how to build and use random forests to solve the classification problem. In this activity you will utilize deep random forests to help medical scientists to identify a set of mutations that appear to have important relationships to cancer. This will involve two major tasks:

* *Deep classification analysis*
* *Feature selection*

*Deep classification analysis* will involve using the full power of random forests. You should build random forests that consist of deep decision trees (previously you built shallow, depth 2 decision trees). Furthermore, you should build a random forest that contains large numbers of decision trees (previously, your random forests consisted of only 10-25 trees).

*Feature selection* will involve interpreting the findings of your random forest analysis. You should select a set of genetic mutations that appear to be related to cancer and, therefore, should be the subject of further scientific study. You should justify your selections based on the importance of the features in the random forest models.

Once you have selected a set of mutations, determine if any of the selected mutations are known to be associated with cancer (e.g., perform a Google search on the names of the mutated genes that you selected). Report your findings.

Try various optimizations that may help to improve the effectiveness of the random forest, such as the following:

* use more (or fewer) than sqrt(n) features for node splitting decisions
* remove features that cover small numbers of cancer samples and large numbers of non-cancer samples *before* you begin to build your random forests;
* after removing such features, consider removing samples that contain none (or only a few) of the selected mutations (any samples that are removed from classification analysis should be reported to the scientists, so that the tumors associated with those samples can be studied in other ways than mutational analyses).

In previous projects, you were not permitted to use pre-existing programs for building decision trees or random forests, or for performing cross-validation. In PART 1 of this activity, you are permitted to use pre-existing code libraries for building random forests and for performing cross-validation (beware that some of the libraries have a steep learning curve, so they may take more time to learn to use than you might anticipate). Alternatively, you may wish to extend the program that you have developed already, constructing deep trees and large forests, and filtering uninformative mutations and samples.

Due date: no later than Thursday November 4, 11:59 pm

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

* The results from deep classification analysis and feature selection.
* Your report should include the following sections:
  1. Research objectives (2-3 bullet points) – what are we trying to discover about cancer?
  2. Key results (2-3 bullet points) – what have we discovered about cancer?
  3. Summary of methods (2-4 bullet points) – describe your software design
  4. Key results (tables, figures, stats, lists, etc.)
  5. Discussion (2-4 bullet points) – how do your findings apply to cancer research?

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

**PART 2**: *Deep association rule mining, Feature association selection, Feature selection update*

THE DETAILS FOR PART 2 WILL BE POSTED ON BLACKBOARD AT A LATER DATE.