FEATURE SELECTION

We have developed a prototype ensemble-based classifier for disease diagnosis and we have performed association rule mining to discover relationships between feature pairs and triplets. 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 part 2 of this activity, you will fully explore feature associations to discover groups of genetic mutations which may work collectively to cause cancer (this is called feature association selection). These discoveries will provide a starting point for scientific research studies to investigate the mechanisms of the disease processes involved in cancer.

DATA MINING ACTIVITY: *Project Summary*

Part 1: *Deep classification analysis & Feature selection*

* Due date: no later than Sunday November 15, 11:59 pm

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

* Due date: no later than Sunday November 22, 11:59 pm

**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 components:

* *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 a large number 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 scientific study. You should justify your selections based on the importance of the features in the random forest models.

Consider various optimizations that may help to improve the effectiveness of the random forest. For example, try using more than sqrt(n) features for node splitting decisions. Also, you may find it helpful to remove features that cover small numbers of 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 can be studied in other ways).

In previous projects, you were not permitted to use pre-existing programs for building decision trees or random forests, nor for performing cross-validation. In this final project, you ARE ENCOURAGED to use pre-existing code libraries for building random forests and for performing cross-validation (for example, Python has some very good libraries for performing these tasks). Alternatively, you may wish to simply extend the program that you have develop already, constructing deep trees and large forests.

Due date: no later than Sunday November 15, 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*

You have learned how to perform association rule mining in order to identify pairs and triples of genetic mutations that are strongly related. In this activity you will exploit the full power of association rule mining to help medical scientists to identify full extent of relationships among sets of mutations that appear to work together in the cancer phenotype. This will involve the following major components:

* *Deep association rule mining*
* *Feature association selection*
* *Feature selection update*

*Deep association rule mining* will involve performing additional iterations of the *a priori* algorithm. Previously, you used the *a priori* algorithm to find the sets of frequent features (C1), frequent feature pairs (C2), and frequent feature triplets (C3). In this activity, you should perform additional iterations of the *a priori* algorithm to find C4, C5, … CN. That is, find the set of frequent feature quadruplets, the set of frequent feature quintuplets, …, and the set of frequent feature N-tuplets. Eventually, the *a priori* algorithm will produce an empty set of tuplets, which will be the terminating point of your algorithm.

*Feature association selection* will involve the interpretation of the results of deep association rule mining. You should select sets of related genetic mutations (from C2, C3, …, CN) that appear to function cooperatively in individuals with cancer and, therefore, should be the subject of further scientific study. Justify your selections.

*Feature selection update*:

You should relate the selected feature associations to the features selected in PART 1. Are there any overlaps between selected features and selected feature associations? Are there any additional features that should be selected, based on the selected feature associations?

**NOTE***: you must develop your own computer program to accomplish deep association rule mining and feature association selection. You ARE NOT permitted to use pre-existing programs for association rule mining.*

Due date: no later than Sunday November 22, 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 association rule mining, feature association selection, and feature selection update.
* 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).