Discipline Project 2

Ricky Cai (440242475)

Executive Summary

1. Provide a clear statement of the question you intend to address or the topic that you intend to focus on your multi-media discipline report.

Topic

As technology increases, gene expression data is more high resolution and cheaper and more available More than 50k for example When conducting ML techniques, large variables overfitting, high computation time

Question to address is to visualise and understand how the the number of PCA affects the accuracy Also to isolate genes which may be important in predicting acute kidney rejection and visualising how individuals compare to the mean

Approach

2. What is your approach to addressing the question stated in (1) and what is the key technique in your approach (e.g. random forest, lasso, Bayesian network etc.)? Select ONE method and provide a concise technical description.

Using random forest and PCA to reduce the number of dimensions PCA is an example of dimension reduction to reduce number of variables while reatining information Need to optimise the number of PCA to include in the model

Shortcomings

3. Identify potential shortcomings or issues associated with the data analytics that you have performed and discuss a possible approach to address the issue. Here, a strategy doesn't necessarily refer to a model, but it must address the issue.

While researching gene expression data sets, each have their own set of genes and it is unlikely that two data sets have the same exact genes. Therefore, this approach is likely specific to this gene expression data set and optimal numnber of PCA would differ on other gene expression data sets and different genes would be identified as potentially important for each data set.

->> how to address the issue??? - combining multiple data sets with common genes and a wide range of samples.

4. Create and describe the interactive graphics (or shiny app) that illustrate one aspect of your report, and please provide the link to the shiny app (this can either be a web page or a GitHub link).