Joshua Anderson

Data Mining

HW2

**Naïve Bayesion Classification of the Gene Relations Data**

* **How to run the program**
  + First, if not already done, extract the zip file containing the R script titled “gene\_analysis.R” and the data files titled “Genes\_relation.data” and “Genes\_relation.test” to the folder of your choice. Copy the filepath to this folder to use for setting the working directory in the RStudio session the program will be run in.
  + Next, open an RStudio session. Press “Ctrl-O” or navigate to File > Open File. Navigate to and open the Rscript “gene\_analysis.R”
  + Execute the script within RStudio. To do this, click inside the window containing the code and press “Ctrl-Alt-R” or alternatively press “Ctrl-A” followed by “Ctrl-Enter”
  + The script will ask you to enter the filepath for the location where the zip file was extracted to. Enter the filepath and press enter.
  + Please note: This script has a very long runtime.
  + When the script has finished executing, RStudio will display the list of Keys providing correct results, the data frame containing the program predictions and the accuracy of the model applied to the test data.

For further examination, a list of intermediate data frames are also returned. The returned list is titled “Results” and may be accessed by typing Results[[i]] into the terminal for values of i between 1 and 10. The objects stored in each index of the list are given below for reference

* + - 1 – Modes: This returns the mode for each attribute as determined from Genes\_relation.data. These modes are used to fill missing values in both the training and test data.
    - 2 – Genes\_relation.data: Present the original training data, without alterations
    - 3 – Genes\_relation.test: Presents the original test data, without alterations
    - 4 – trimmedGeneRelation: Presents the data frame containing the original training data with the “Function” column removed and any resulting duplicate rows removed as well.
    - 5 – trimmedGeneRelationTest: Presents the data frame containing the training data modified as described in the above entry.
    - 6 – filledTrimmedGeneRelation – Presents the data frame containing all values from trimmedGeneRelation but has replaced all missing values with the mode of the respective attribute.
    - 7 – filledTrimmedGeneRelationTest – Presents the test data modified as described in the previous entry.
    - 8 – Results – Presents the data frame containing the results of classification.
    - 9 – Keys – Presents the set of correct classifications by GeneID
    - 10 – Accuracy - Displays the accuracy of the classification model.
* **Handling Missing Data**

Missing data values were replaced using the mode of the respective attribute as taken from the Genes\_relation.data file. The mode for each attribute was calculated and then for every row, each missing value was replaced with the mode for it’s associated attribute. The modes from Genes\_relation.data were used to replace missing values in both the training data and the test data (Genes\_relation.data and Genes\_relation.test).

Note: Interaction data was not included in this model

* **Classifier Method**

The classification method implemented was a Naïve Bayesian Classifier. After filling in missing data values based on modes, each row from Genes\_relation.test was fed into the classifier as a sample vector. The classifier then used the filled in training data to predict the most probable localization for that given tuple. The localization for this tuple was then classified into the most probable class value for localization.

* **Accuracy**

The accuracy of the solution was just over 50%.