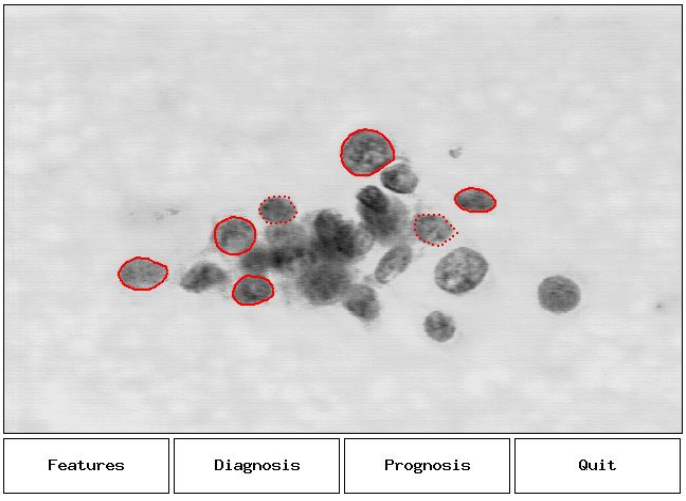
IDS Course Project

Overview:
In this project, you will analyze a set of data using three of the methods you have learned for classification in the course. You will then create a video presentation that explains the practical side of using machine learning techniques for classification.

Background:
A group of physicians have become interested in using machine learning techniques to assist in identifying malignant tumors. They have asked your group to demonstrate machine learning techniques.

The audience:
Your presentation should be directed to a group of physicians. These individuals are generally well versed in basic science but have limited background in statistics and generally no exposure to machine learning techniques.

The data:
The physicians have identified a data set that consists of over 500 measurements from Fine Needle Aspiration (FNA) of breast tissue masses. In an FNA, a small needle is used to extract a sample of cells from a tissue mass. The cells are then photographed under a microscope. The resulting photographs are entered into graphical imaging software. A trained technician uses a mouse pointer to draw the boundary of the nuclei. The software then calculates each of ten characteristics for the nuclei. An example image is given below. This process is repeated for most or all of the nuclei in the sample.



The data consists of measurements of the cell nuclei for the following characteristics:

1. radius
2. texture
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension ("coastline approximation" - 1)

Measurements of these ten characteristics are summarized for all cells in the sample. The dataset consists of the mean, standard error of the mean, and maximum of the 10 characteristics, for a total of 30 observations for each. Additionally, the data set includes an identification number and a variable that indicates if the tissue mass is malignant (M) or benign (B).

The task:
You have been asked by the physicians to conduct an analysis of the data using three of the classification methods we have seen in this class, and provide a video presentation that describes those results.

For your analysis, you should:

1. Download the data from NeXus: FNA_cancer.csv
2. Perform basic exploratory data analysis.
3. Split the data into test and training data.
4. Build a classification algorithm using decision trees. Prune your tree appropriately.
5. Build a classification algorithm using random forests/bagging. Adjust the parameters of the forest appropriately.
6. Build a classification algorithm using Kth Nearest Neighbors. Tune the value of K appropriately.

For your presentation, you should describe the results of your analysis. In your presentation, you should describe each algorithm at a level that is appropriate to your audience. You should also discuss the issue of misclassification. You may assume misclassifying a malignant as benign is worse than misclassifying a benign tumor as malignant.

Submissions:
Analysis code: You should submit your R code used to conduct the analysis by 11:59 pm on the Friday of Week 10. You will submit this file to Box in NeXus. This code should be appropriately commented to make it easily used by other programmers. This code should be submitted as a .pdf file (not .r or .rmd).

Slide Deck:
You should prepare a set of powerpoint slides to be used in your presentation. These slides should be prepared in a professional manner. They should accurately convey the information from your analysis and provide an overview of the methods you used. This should be presented at a technical level appropriate to the audience. This should be submitted to Box on NeXus by 11:59 pm on the Friday of Week 13, and should be in .ppt format.

Video presentation:
You should submit a video of your group presenting the slides. Your video must be in mp4 format. The video should not exceed 15 minutes in length in total and should be submitted by 11:59 pm on the Friday of Week 14 to Box in NeXus.

Some notes:

- You do not need to do detailed research into the medical aspects of this project.
- You do not need to use techniques that were not discussed in the IDS course.
- This is not a project that expects you to do “fancy” programming in R. You will not get extra points for doing something not covered in the course. You may also lose points if you do something wrong.
- Your presentation should include at least one member of your group speaking; however, not everyone needs to speak. You may NOT use an outside narrator.