Data Mining Assignment #1

Predicting Malignancy of Tumors

**Due Date: Midnight, Sunday, June 5**

Researchers at the University of Wisconsin-Madison Medical School examined digitized images of fine needle aspirate of suspicious breast masses. Using the images, a number of characteristics of the cells were measured. For each biopsy and each characteristic, the mean, the standard deviation, and the most extreme observation were recorded along with the diagnosis (malignant/benign).

Using *only* the mean values for each characteristic, examine both a Naïve Bayes classifier and a logistic regression classifier to predict if the cell is malignant or benign. Determine which classifier should be deployed to make classifications when the diagnosis is unknown. The data is found in the file *BreastCancerData.txt*.

Submissions by the deadline via Moodle are to consist of two documents:

1. A report consisting of
   1. An executive summary indicating the type of model to be applied and the success that can be expected with this model.
   2. A **high level** description of the modeling process consisting of no more than a three pages. Identify the primary steps needed to eventually select the final model including:
      1. data issues and resolution methods (if any),
      2. the considered models (for any functional models, provide the function definition and all coefficients), and
      3. model performance.
   3. The report is to be a Word document (not a pdf)
2. An R code file. Your R code should be in a “final form” meaning you have removed unnecessary commands and statements that were necessary for development but no longer serve a purpose when determining the key information for the report. When the R code is run, only key metrics should be returned to the console and key graphs should be displayed. In addition, precede each console output with a command that states a title, for example:

*print*(“The following table is the confusion matrix for the training data”)

IMPORTANT NOTES RELATED TO YOUR R CODE:

* Prior to submitting your R code, edit the code so the data file *BreastCancerData.txt* is read from the directory: *C:\Users\bgubser\OneDrive\Class Files\Data Mining - Summer 2015\Assignments* (the purpose here is I will be able to run your code without having to edit your submissions – I don’t want to edit 25 submissions). **If your code will not read the data file *BreastCancerData.txt* from this location, a deduction of 10 points (one letter grade) will be assessed.**
* The documentation of the R code is essential. Much of the details of the analysis will not appear in your report. To evaluate if the analysis was properly conducted, I will need to walk through the R code line by line. Without thorough documentation I will be unable to quickly understand what the code is attempting to accomplish. Sometimes it is obvious from the code itself what the code is accomplishing; however sometimes it is not. Generally the author finds the code more readable than someone else but it is this other person, not the author, who is the judge if the code is transparent or not. Be liberal with your documentation. **Although documentation specifically contributes 4% to the grade of this assignment, if I cannot determine if the analysis is correct, this will be reflected in the grade for valid analytics whose contribution is 60%.**

**Note on working together**: I think it is beneficial to work together. However this still means that the work you submit should be your own. It is fine to discuss issues with someone else, ask if they have suggestions for resolving a difficulty you are facing, or bounce ideas off each other. But when implementing idea, you should be doing this on your own. Specifically, you should be writing your own report and R code. The language, syntax, and structure should reflect your understanding and the decisions you have made to address this assignment.

**Evaluation**: Submissions will be evaluated by the following rubric:

1. Report (25 points) evaluation criteria:
2. Clarity of explanation (8 points)
3. Inclusion of essential analytical elements (8 points)
4. Appropriate use of analytical terms (5 points)
5. Tone (a professional tone is expected) (4 points)
6. Readability of R code (8 points) evaluation criteria:
7. Organization and Logic of code (little, if any, extraneous code, easy-to-follow logic): 4 points
8. Documentation: 4 points
9. Output of R code (7 points) evaluation criteria:
10. Key assessment measures are displayed with title (4 points)
11. Key graphical elements for the analysis are generated (3 points)
12. Valid Analytics (60 points). Your analytics will be evaluated via both the report and the R code (so documentation is important). The evaluation criteria are:
13. Data preparation (20 points)
14. Model development (20 points)
15. Model Assessment (20 points)