CECS 551 Programming Assignment 1

Dr. Todd Ebert

Due Date: 3:30pm on February 8th

Preparatory Reading

Read Appendix B, and the parts of Chapter 2 that pertain to the R language and programming system.

Presenting Your Work.

Submit a hard copy of a text or word document that, for each exercise, displays i) the sequence of R commands that were used to obtain the answer, and ii) the answer itself. Preface each R command with a comment that describes what the command is doing. Also, for each block of code that pertains to some exercise, say Exercise 1, preface the code with the comment (all in caps) The easiest way to accomplish this is by copying and pasting from the R console to your document. Also, for each block of code that pertains to some exercise, say Exercise 1a, preface the code with the comment (all in caps)

```
#EXERCISE 1a
```

Plagiarism

Plagiarism is defined for this assignment as the practice of taking someone else's code and passing it off as one's own. Although it is OK to discuss this assignment with your peers, in the end each student must implement his or her own source code. a student suspected of plagiarism will receive an automatic course grade of "F". Therefore, it is important for you to keep your code private.

Mammographic Mass Data Set

Review and download the mammographic mass data set at

`https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass`

In examining the data, notice that some datapoints have missing attribute values. In these cases "?" is substituted for each missing value. Add a header line to the csv file that labels the attributes as `Birads,Age,Shape,Margin,Density`, and `Severity`.

Exercises

1. In R the more appropriate indicator for missing data is "NA" (not available). Therefore, replace each occurrence of "?" with "NA".

   a. For this exercise, create an R data frame for the mammographic data using only datapoints that have no missing values. This can be done using the `complete.cases` function which inputs a data frame and returns a Boolean vector v, where $v[i]$ equals `TRUE` iff the i th data-frame sample is complete (meaning it does not possess an NA). For example, if the data-frame is stored in `mammogram.frame`, then

      ```
      mammogram2.frame = mammogram.frame[complete.cases(mammogram.frame),]
      ```

      creates a new data frame called `mammogram2.frame` that has all the complete mammogram data samples.

   b. Use R's `summary` function to provide a statistical summary of each of attribute of the altered data frame.

   c. Use the `e1071 svm` function to construct a linear classifier for the data set. Report on the percentage of datapoints that are correctly classified by the svm model.

   d. Repeat part b) using a degree-2 polynomial classifier. This particular type of svm can be constructed using the input options

      ```
      kernel = ‘‘polynomial’’, degree = 2, type = ‘‘C-classification’’
      ```

2. Repeat each part of the previous exercise, but, for Part a, instead of removing datapoints with missing attribute values, replace each NA with the nominal value -1 and use the entire data set.