Midterm Proposal

due Monday 3/5 in Sakai (version 1.0)

Problems

1. Select a data set for machine learning/data mining purposes. This data set will be with you for a while, it is the data set you will use for your midterm work. A good place to start is the UCI machine learning repository (see course homepage for URLs). But if you have a data set you are interested in investigating that would be fine too.

   You need to keep a couple of things in mind, in other words, your data set needs to fulfill the following criteria:

   - The independent variables/attributes in your data set need to be over the reals or integers, that is, they should be continuous or numerical attributes. If an attribute is categorical (i.e. consists of labels) then you will have to turn it into a factor (R's version of a categorical variable) as follows,

     ```
     > x <- c("red", "red", "blue", "green", "blue")
     > x
     [1] "red"   "red"   "blue"  "green" "blue"
     > y <- factor(x, levels = c("red", "green", "blue"))
     > y
     [1] red   red   blue  green blue
     Levels: red green blue
     ```

     Turns out that the svm() function we will use to construct SVM models knows how to deal with factors.

   - Your target attribute needs to be defined in terms of a binary classification problem. The actual labels used are not important

since theoretically it is trivial to rename them to $\{+1, -1\}$. Again, in technical jargon, your target attribute should be a categorical variable with two levels.[1]

- Your data set should be non-trivial, by that I mean it should have at least 50 rows and not less than 5 independent attributes.

2. Format your data so you can import it into R.

3. Perform an exploratory data analysis on the data (at minimum): basic statistical summary for each attribute (including the dependent attribute), graphs of the distributions for each independent variable, a histogram for the dependent variable.

4. Write a 1-2 page proposal why you picked this data set incorporating the basic statistics you computed in the previous point.

NOTE: You can use any language you like that supports graphics and a support vector machine library for your work. However, if you use a language other than R or Python I will not be able to answer any questions regarding libraries, graphics, implementation, etc.

Submit your proposal in Sakai.

[1] You probably don't want to use the labels +1 and -1 because R will get confused and interpret the target as numerical. Labels such as POS and NEG or PLUS and MINUS *etc.* should be used.