

CSC581 Midterm Prep

due Tuesday 3/3 in class

version 1.1

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. Good places to start are the UCI machine learning repository and the Statlib library at CMU (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 such that for attribute values a_1 and a_2 the relations $a_1 \leq a_2$ or $a_2 \leq a_1$ hold for a given attribute A . Notice, that this does not permit codings such as 1.0 and 2.0 where $1.0 \equiv \textit{blue}$ and $2.0 \equiv \textit{red}$ since $\textit{blue} \leq \textit{red}$ or $\textit{red} \leq \textit{blue}$ is not defined and you are using the reals just as place holders for the names *blue* and *red*. In more technical jargon you can only use data sets where the independent variables/attributes are defined as continuous values **not** as categorical values.
 - Your target attribute needs to be defined in terms of a binary classification problem. The actual labels used are not important since it is trivial to rename them to $\{+1, -1\}$ if you need to. Again, in technical jargon, your target attribute should be a categorical variable with two levels.
 - 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 you computed in the previous point.