Modeling Mathematics

Project 1 - Statistical Learning Methods

1. Pick (and document) your data set(s). The official suggestion is to use kaggle.com, but you may be able to find data sets more interesting to you if you look at the documentation for a paper (studies often are required to post their results). Whatever data set you choose, make sure it comes from a publicly available, reputable source and be wary of data from think tanks or studies from corporations who have incentives to make the results look a particular way.

2. In an ideal world you would begin with a question you'd like to answer, but the availability of data will be the biggest restriction we work under. *Based on your data*, think of the kinds of predictions and inferences you might be able to make and formulate as many questions as you can.

3. Do some light research attempting to answer your questions. Some may be impossible to answer. Some may have answers already, though it does not guarantee you will find them. It could be that an obscure paper in a language that is not English has already answered and you won't know better. This step is so that you can make hypotheses as to what you will find and then compare.

4. As an optional, step, you may want to pre-process your data. This can mean filtering to narrow the scope of your project, combining data entries from different sets into vectors, changing the format you found the data in to something easier to work with, etc. Document this step (what you did exactly and why).

5. Use at least three of the methods on classification, resampling, deep learning or clustering that are mentioned in *An Introduction to Statistical Learning* to analyze your data sets. If it will help, you may use regressions (but they will not count as one of the three methods). Make sure you can justify the use of the modeling approach based on the type of data you are working with. Keep tidy records of all scripts used and make sure they are well documented so that anyone reading them can see what they're meant to do.

6. Draw conclusions based on the results you obtained and discuss their validity. If possible and appropriate, run analyses on the error of your predictions, perform hypothesis tests, or plot results obtained with varying parameters to explain your choices.

7. Discuss what you would change if you had more time and resources to answer your question. How could your work be better?

8. Create a project repository on GitHub and upload all your scripts to it. Each group member can make an account and be added as administrator of the repository (so that you can also split up the writing for the readme). This will be your final deliverable. The repository should include a readme file with:

    - instructions for others to run your code on the data
    - an introduction to the problem you are hoping to solve, including references to the work that already exists on the topic (from your light research).
    - documentation for where you got your data,
    - analysis of your results and discussion,
    - credits for who in your project did what.

    Make sure to set the visibility to public so that your project can be graded.

    Strictly speaking, GitHub repositories are not meant to be as wordy as this one will come out to be. But getting used to writing code others can read and learning the markdown necessary to make a nice readme file is as good a start as any. You can delete the repository after it is graded if you are not particularly proud of it.

    One of your group members should email a link to your repository by **Sunday, Feb. 23rd.**