STAT 530 Final Paper - Outline

Andrew Mehrmann

May 2, 2016

Biological problem

- What is the biological problem
- Why is it important

Data

- Where do the data come from (e.g. give source if public data)
- What are the details of the experiment (e.g. how many mice, how many conditions)
- How were the data collected (e.g. what array or sequencer, what version, from what company)

Statistical methods

- How were the data preprocessed

The training and test labels were given and The data were normalized by [N1]-[N4] separately for training and test in the paper.

- What were the analysis steps
- What statistical methods were used and why
- Must use at least 3 different statistical methods

Results – report the results of your analyses

- Their model

I reproduced their results using the equation…

- Their Data

As mentioned above, they normalized their training and testing sets separately, so I did the same originally. When I began fitting my own models, I noticed they were predicting very well on the training set but very poorly on the test set. I believe there are two possible reasons for that:

1) I was overfitting the training data. But I used two methods (LASSO regression and Random Forest) that are designed to combat overfitting. LASSO in particular doesn't usually overfit, at least not this badly.
2) The data between training and testing were different. This could be because they actually were very different when they were given to me. If the data were from two different statistical populations, then a model fit on one population would certainly not generalize to another population. Another option was that the normalization process created differences in the data that the models could not overcome. To test this, I went back and normalized the data together and still had bad (albeit better) results. Specifically…

- My analyses

Abandoning the methodology of the paper, I simply used their raw data, unnormalized. I thought methods that implement Bootstrapping would be best for this dataset with so few observations, so I did not proceed further with Naive Bayes.

Conclusions – interpret your results (e.g. what do you results reveal about the biological mechanisms of your experiment)