Class 9 overview: 21 March

Announcements & Carry-forwards

1. Plan for the upcoming Project assignments
2. Next two classes cover regularized regression, bootstrap, significance testing, ROC curves

First hour

- Presentation of first (of the three) class projects

Second hour

- Stat unit 3: Resampling and bootstrap interval estimation

Third hour

- HW review (part of the hour)
- Stat unit 4: “Ridge” regression and shrinkage methods

Fourth hour

- Continuation of Stat unit 4
- Description of next project assignments.

Second project assignment - building on previous assignments

Programming: Improving regression predictions

Regression when there are many features - explanatory variables — suffer from high variance and correspondingly degraded accuracy. The premise is that many of the variables are not particularly relevant, or their relevance is captured in other variables.

Using your results from last week and the techniques for ridge regression for this week you now have three methods to improve accuracy

1. Select variables that are more informative for the dependent variable. This may be seen from the value of the regression coefficients, or by measures such as covariance or mutual information.
2. Generate a small set of components that capture most of the variance of the “raw features” using PCA
3. Vary the lambda parameter in ridge regression.

These three methods can be used in combination.

Your next project is to come up with a regression method that demonstrably improves on linear regression using the entire set of variables. Continue using the same dataset from the previous project.

You should have some ideas from the previous project on how to evaluate your regression in terms of accuracy. Please look at the notebook for this week "ridge_regression.ipynb" on how to create a learning curve to demonstrate accuracy. It might be helpful to apply the sampling methods explained in the notebook "embedding_intervals.ipynb" to place error bounds on your evaluation. That is optional — more about that in the next project task.

1. Decide on your evaluation criterion. How will you decide to measure performance of your model? The accuracy measure used in the ridge regression notebook is one possibility, but depending on what the goal you set is, you may propose other measures.
2. Given the dataset and information you've discovered about the feature variables experiment with a combination of the techniques provided to find a modified set of feature variables that perform better according to your evaluation criterion.
3. ***Visualize your findings.***
4. (Thinking ahead for next week:) How might bootstrapping be applied to this problem?

Files

- embedding_intervals.ipynb
- ridge_regression.ipynb
- Statistics unit 3
- Statistics unit 4