Lecture 29: Review

Reading: All chapters in ISLR.

STATS 202: Data mining and analysis

Sergio Bacallado
December 6, 2013

Announcements

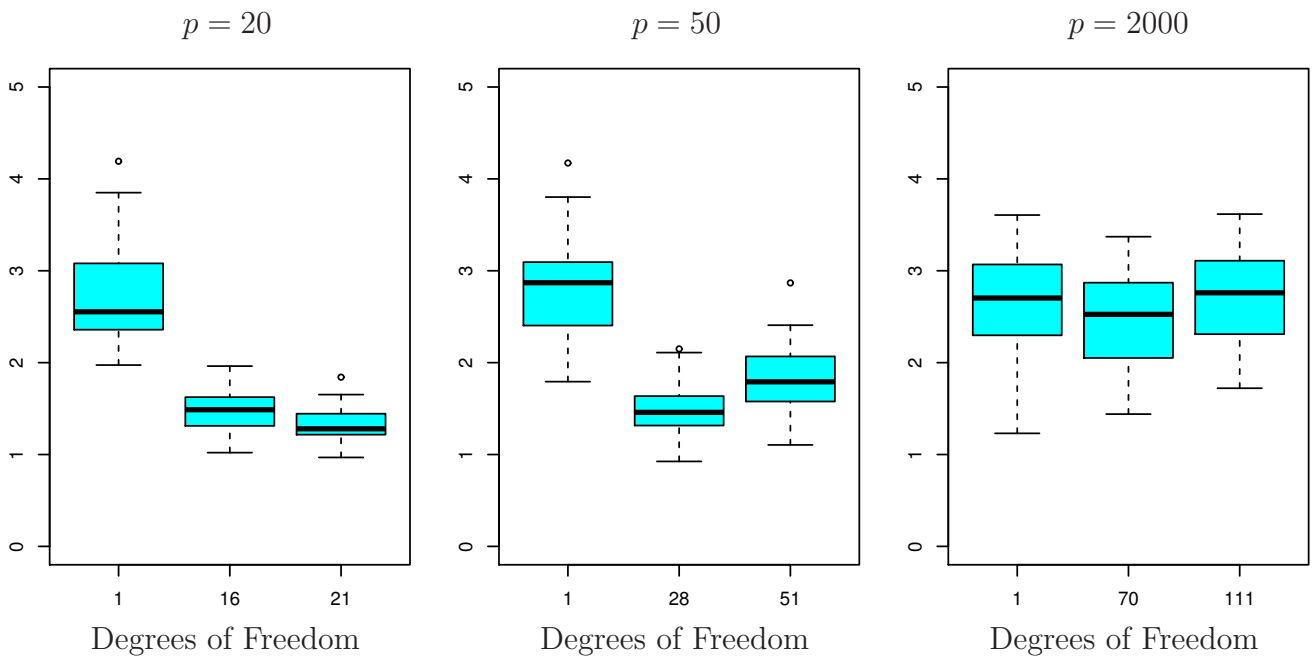
- All homework grades will be posted on Coursework by the end of the day today.
- You may pick up any leftover homework from my office (Sequoia 202) this afternoon or on Monday afternoon.
- Please check Coursework tomorrow and let us know if you are missing any grades.
- Send all regrade requests to the graders mailing list by **Wednesday, December 11**.

Kaggle competition

- In most Kaggle competitions, the top positions in the public and private leaderboards are the same.
- In our case, only the top team, GobbleGobble, was in both leaderboards.
- This was probably due to the fact that the dataset is small, and the response is hard to predict.
- The top 4 teams in the private leaderboard were near the top of the public leader board.
- Surprisingly, the 2nd team in the public leaderboard did poorly in the private leaderboard. This is probably because they tried too many different methods (46 submissions!).
- Since many teams were small (1 or 2 people), we rewarded more teams than we planned.

What worked, what didn't?

In a high-dimensional problem such as this one, variable selection methods do not work so well. Most predictors are noise.



From section 6.4.3 in the book.

What worked, what didn't?

Some of the winning teams used prior information to manually filter out irrelevant variables.

Team GobbleGobble did a manual selection of variables aided by a set of meta-features, describing each predictor:

- Is the predictor significantly different in the training vs. test set?
- Is the predictor significantly different for high vs. low values of the response?
- What is the variance of the predictor?

What worked, what didn't?

Two of the winning teams used Bayesian Additive Regression Trees (BART) after Lester Mackey's Lecture, which gave them their highest public score.

The performance of team DLFK was superior (averaging the public and private leaderboard MSE) because they managed to cross-validate the parameters.

Team Bryan applied an ensemble of Boosting and Random Forests, trained on the 10% of predictors with the highest correlation with the response.

Self testing questions

For each of the regression and classification methods:

1. What are we trying to optimize?
2. What does the fitting algorithm consist of, roughly?
3. What are the tuning parameters, if any?
4. How is the method related to other methods, mathematically and in terms of bias, variance?
5. How does rescaling or transforming the variables affect the method?
6. In what situations does this method work well?

Regression methods

- Nearest neighbors regression
- Multiple linear regression
- Stepwise selection methods
- Ridge regression and the Lasso
- Principal Components Regression
- Partial Least Squares
- Non-linear methods:
  - Polynomial regression
  - Cubic splines
  - Smoothing splines
  - Local regression
  - GAMs: Combining the above methods with multiple predictors
- Decision trees, Bagging, Random Forests, and Boosting

Classification methods

- Nearest neighbors classification
- Logistic regression
- LDA and QDA
- Stepwise selection methods
- Decision trees, Bagging, Random Forests, and Boosting
- Support vector classifier and support vector machines