Class Notes

Statistical Computing & Machine Learning

Class 5

Review

Graphics basics

1. API for graphics: `plot()`, `points()`, `lines()`, `polygon()`, `text()`, ...
2. Create a plotting frame: `plot()`
   - Write a function that makes this more convenient to use. What features would you like.

```
blank_frame <- function(xlim, ylim) {

}
```

3. Write a function to draw a circle.
   - What do you want the interface to look like? What arguments are essential? What options are nice to have?

Regression and Interpretability

Regression models are generally constructed for the sake of interpretability:

- Global linearity
- Coefficients are indication of effect size. The coefficients have physical units.
- Term by term indication of statistical significance

Measuring Accuracy of the Model

- R^2 - Var(fitted)/Var(response)
- Adjusted R^2 - takes into account estimate of average increase in R^2 per junk degree of freedom
- Residual Standard Error - Sqrt of Average square error per residual degree of freedom. The sqrt of the mean square for residuals in ANOVA

Bias of the model

- Perhaps effect of TV goes as sqrt(money) as media get saturated?
- Perhaps there is a synergy that wasn't included in the model?

- Whole model ANOVA.
- ANOVA on model parts
- Adjusted R^2

Run an example on `College` data from `ISLR` package

```
data(College, package = "ISLR")
College$Yield <- with(College, Enroll/Accept)
mod1 <- lm(Yield ~ Outstate + Grad.Rate + Top25perc,
    data = College)
```

- What variables matter?
- How good are the predictions?
- How strong are the effects?

Forward, backward and mixed selection

Use the `College` model to demonstrate each of the approaches by hand. Start with `pairs()` or write an `lapply()` for the correlation with `Yield`?

Create a whole bunch of model terms

- "main" effects
- "interaction" effects
- nonlinear transformations: powers, logs, sqrt, steps, ...
- categorical variables

Result: a set of k vectors that we're interested to use in our model.
Considerations:

- not all of the k vectors may pull their weight
- two or more vectors may overlap in how they eat up variance

Algorithmic approaches:

- Try all combinations, pick the best one.
  - computationally expensive/impossible 2^k possibilities
  - what's the sensitivity of the process to the choice of training data?
- "Greedy" approaches

In-class programming activity

Day 5 activity

Drawing a histogram.