Class Notes

Statistical Computing & Machine Learning

Class 14

Model & Theory building

Occam's Razor: A heuristic

Non sunt multiplicanda entia sine necessitate.

- Entities must not be multiplied beyond necessity.

  But what's "beyond necessity?"

Einstein's proverb

Man muß die Dinge so einfach wie möglich machen. Aber nicht einfacher.

- Make things as simple as possible, but not one bit simpler.

  But what's "too simple?"

Statistical parallel

Bias versus variance

Bias refers to the ways that $\hat{f}(\mathbf{X})$ is systematically different from the idealized $f(\mathbf{X})$.

Variance refers to the ways that $\hat{f}_i(\mathbf{X})$ varies with the particular sample i used for training.

Goal for model building

Make model error small

Sub-Goals:

- Make bias as small as possible
  - use many parameters, flexible function architectures.
- Make variance as small as possible.
  - use large training sample size n
  - use few parameters/stiff function architectures

Basic modeling-building question:

- Does adding this [variable, model term, potential flexibility] help?
  - For reducing bias: yes
  - For reducing variance: maybe (if it eats up residual variance)
  - For reducing error: maybe

Operationalizing model choice

Think in terms of comparing two models to select which one is better.

"Better" needs to be transitive: if A is better than B, and B is better than C, then A is better than C.

Then, any number of models can be compared to find the one (or more) that is the best.

Some definitions of "better"

- Larger likelihood (non-iid Gaussian error models)
- Smaller mean square prediction error (same as larger likelihood with iid Gaussian error model)
- Classification error rate is smaller
- Loss function is smaller

Training and Testing

Evaluation of performance using training data is biased to give larger likelihood (smaller MSE or classification error or loss error).

Unbiased evaluation is done on separate testing data.

Trade-off

- Need large testing dataset for good estimate of performance

- Need large training dataset for reducing variance in fit.

How to get both:

1. Collect a huge amount of data. When this works, go for it!
2. K-fold cross validation
   - Pull out 1/K part of the data for performance testing.
   - Fit to the other (K-1)/K part of the data.
   - Repeat K times and average the prediction results over the K trials.
3. Once you've found the best *form* of model, fit it to the whole data set. That's your model.

Programming Basics: Loops/Iteration

Loops are the programming control structure that allows you to repeat the same commands many times.

A definition of insanity: Doing something over and over again and expecting a different result.

Parts of a loop

1. Preparation — creating a place to hold the results

   This is called the "accumulator."

2. Identify a set to loop over.

3. Inside the loop, modify the accumulator

4. When the loop is done, package up the results.

Trivial examples

- Find the sum of squares of a vector
- Find the biggest element of a vector
- Find the kth Fibonacci number

More realistic example

Leave-one-out cross-validation.

```
# preparation
my_data <- mosaicData::KidsFeet
error <- numeric(nrow(my_data))

# The looping set: each row in my_data
for (k in 1:nrow(my_data)) {
  # the body of the loop
  mod <- lm(width ~ length * sex, data = my_data[ -k, ])
  mod_value <- predict(mod, newdata = my_data[k, ])
  error[k] <- my_data$width[k] - mod_value
}

# packaging up the results
result <- sum(error^2)
```

Look at the result:

```
result
```

```
## [1] 6.304381
```

```
regular_model <- lm(width ~ length * sex, data = my_data)
sum(resid(regular_model)^2)
```

```
## [1] 5.329327
```

```
anova(regular_model)
```

```
## Analysis of Variance Table
##
## Response: width
##            Df Sum Sq Mean Sq F value
## length      1 4.0557  4.0557 26.6353
## sex         1 0.4790  0.4790  3.1456
## length:sex  1 0.0037  0.0037  0.0246
## Residuals  35 5.3293  0.1523
##              Pr(>F)
## length     9.86e-06 ***
## sex         0.08484 .
## length:sex  0.87638
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
```

In ANOVA, we use a degrees of freedom to adjust for the under-estimate of residuals.

```r
sum(resid(regular_model)^2) / 35
```

```
## [1] 0.1522665
```

In leave-one-out, we can simply average the errors:

```r
result / 38
```

```
## [1] 0.1659048
```

Jubilee Week

Work on completing the first 12 in-class activities.