Stat 435 Intro to Statistical Machine Learning

Week 5: Additional exercises

Richard Li

May 2, 2017

Before we start, questions about this?

Inference the old way (pre-1980?) :

1. Devise a model
2. Collect data
3. Test hypotheses

Classical inference

Inference the new way:

1. Collect data
2. Select a model
3. Test hypotheses

Post-selection inference

Classical tools cannot be used post-selection, because they do not yield valid inferences (generally, too optimistic)

The reason: classical inference considers a fixed hypothesis to be tested, not a random one (adaptively specified)

(slide credit to Robert Tibshirani (NIPS,2015))

Linear or cubic model?

Problem 1(d) in midterm

(d) Consider a test observation (x_0, y_0). Answer (c) using the expected test error, $\mathrm{E}\left(y_0 - \hat{f}(x_0)\right)^2$, rather than the training RSS.

There's not enough information to tell: cubic model will have lower bias but higher variance. So we can't tell what bias^2 + variance will be.

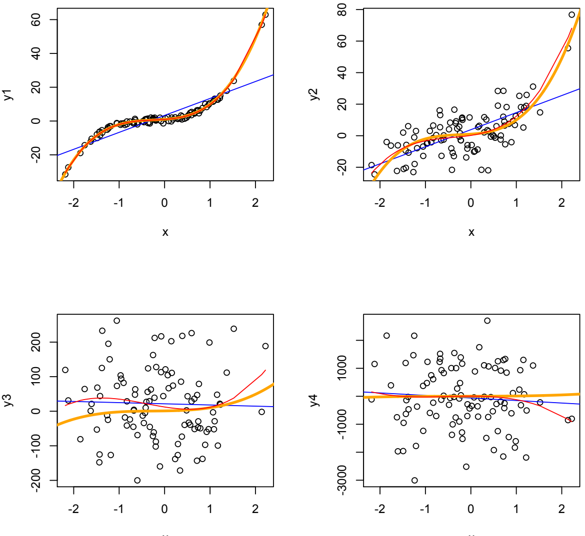
Linear or cubic model?

Problem 1(d) in midterm

```
set.seed(10)
x <- rnorm(100)
y1 <- 1 + 2 * x + 3 * x^2 + 4 * x^3 + rnorm(100, sd = 1)
y2 <- 1 + 2 * x + 3 * x^2 + 4 * x^3 + rnorm(100, sd = 10)
y3 <- 1 + 2 * x + 3 * x^2 + 4 * x^3 + rnorm(100, sd = 100)
y4 <- 1 + 2 * x + 3 * x^2 + 4 * x^3 + rnorm(100, sd = 1000)
```

Linear or cubic model?

Blue: linear fit. Red: Cubic fit. Orange: true relationship



Linear or cubic model?

Correct model is not always the best model in terms of testing MSE!

Bias-variance tradeoff revisited

Bias-variance tradeoff revisited

Bias-variance tradeoff revisited

QDA decision boundary

Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Prove that in this case, the Bayes classifier is not linear. Argue that it is in fact quadratic.

QDA decision boundary

QDA v.s. LDA

- If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?
- If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?
- In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

QDA v.s. LDA

QDA v.s. LDA

Curse of dimensionality

Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly (evenly) distributed on $[0, 1]$. Associated with each observation is a response value.
Suppose that we wish to predict a test observations response using only observations that are within 10% of the range of X closest to that test observation.
For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction ?

Curse of dimensionality

Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observations response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. On average, what fraction of the available observations will we use to make the prediction ?

Curse of dimensionality

What if $p = 100$?

Curse of dimensionality

Comparing resampling methods

Consider a very simple model,

$$Y = \beta + \epsilon$$

where Y is a scalar response variable, β is an unknown parameter, and ϵ is a noise term with $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$. Assume that we have n observations with uncorrelated errors. Show that

1. Validation set approach over-estimate the expected test error.
2. LOOCV does not substantially over-estimate the expected test error, provided that n is large.
3. K-fold CV provides an over-estimate of the expected test error that is somewhere between the other two approaches

Comparing resampling methods

Comparing resampling methods

Comparing resampling methods

Comparing resampling methods

Additional question: How about variance of test error?