STAT 206 Lab 8

Due Monday, November 27, 5:00 PM

General instructions for labs: You are encouraged to work in pairs to complete the lab. Labs must be completed as an R Markdown file. Be sure to include your lab partner (if you have one) and your own name in the file. Give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used.

Agenda: Fit polynomial regression models to the electricity usage data, use K-fold cross-validation to automatically select degree of the polynomial

Polynomial regression

The polynomial regression model posits that a response variable Y and explanatory variable X are related by the equation.

$$Y = \sum_{j=0}^{d} \beta_j X^j + \epsilon \, .$$

The number d is called the degree of the polynomial. Polynomial regression reduces to linear regression when $d = 1$. Its flexibility and complexity increase as d increases. The cases $d = 2$ and $d = 3$ are usually referred to as quadratic and cubic. The polynomial regression model can be expressed as a $d + 1$ parameter linear model by considering $(X_0, X_1, X_2, \ldots, X_d)$ as explanatory variables. This is done by `poly()` and can be combined with `lm()` to fit a polynomial regression model. In the following example, we fit a degree-3 polynomial, or cubic, regression model using variables y and x in the dataframe `df`.

```
degree <- 3
obj <- lm(y ~ poly(x, degree), data = df)
```

'electemp' dataset

The 'electemp' dataset has 55 observations on monthly electricity usage and average temperature for a house in Westchester County, New York

```
url <- 'http://www.faculty.ucr.edu/~jflegal/electemp.txt'
electemp <- read.table(url)
```

1. Create a scatterplot of `temp` and `usage` with `ggplot2` that includes the least squares fits of a linear and quadratic regression models. You should also include a legend on the plot.
2. Does the linear or quadratic model fit the data better?
3. Write a function `cv_poly()` that performs K-fold cross-validation to estimate the mean squared prediction error (MSPE) of polynomial regression. It takes vectors x and y containing observations of the explanatory and response variables, a vector degree of the degrees of polynomial models to fit, and a number K indicating the number of folds for cross-validation. It returns a $K \times D$ matrix, where K is the number of folds and D is the number of different degree models that are being fit. The entries of the matrix are the MSPE for each fold and degree polynomial model being fit.
4. Use `cv_poly()` to estimate the MSPE of polynomial regression on the electricity usage data by $K = 10$-fold cross-validation for $d = 1, 2, \ldots, 8$. Note that `cv_poly()` should return a matrix, call it `cv_error` with K rows corresponding to the K different validation sets.

5. Plot the estimated MSPE (by averaging across the K folds) versus degree of the polynomial. What degree polynomial would you select according to cross-validation?
6. Repeat the preceding problem for $K = 5$ and leave-one-out cross-validation ($K = n$). What do you notice about the time it takes to compute the cross-validation? How do the results change with K?
7. Plot the estimated MSPE versus degree of the polynomial. What degree polynomial would you select according to cross-validation? Are there differences between $K = 5$, $K = 10$, and leave-one-out estimates of MSPE?
8. Reproduce your first plot and add a layer showing the polynomial regression model selected by cross-validation by modifying the following code.