# Homework 4

Advanced Methods for Data Analysis (36-402/36-608)

Due Tues February 25, 2014 at 11:59pm

Instructions: each homework will be submitted in two parts.

1. *Writeup.* An important part of the learning the data analysis trade is learning how to communicate. Prepare a writeup to the questions below and work hard to make your submission as readable as you possibly can—this means no raw R code as part of the writeup, unless you find it necessary for clarity; no raw R output with 12 significant digits, etc., unless again you deem it necessary; making sure figures are of proper (large) size with labeled, readable axes; and so forth. Your submitted writeup must be in PDF format. You can produce this PDF using LaTeX, Word, or whatever you want, but at the end of the day, it must be a PDF. Any other type of file will not be accepted.

2. *Code.* Also submit your R code for the homework as a plain text file. You should demarcate sections of the code that correspond to the different homework problems using clearly visible comments (e.g., as in `##### Problem 1 #####`).

## 1 Oh training error, so young and naive, part deux

This problem will check our claim in lecture about degrees of freedom and optimism, namely, that

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n (y_i' - \hat{y}_i)^2\Big] = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2\Big] + \frac{2\sigma^2}{n}\mathrm{df}(\hat{y}). \tag{1}$$

The setup is

$$\begin{aligned} y_i &= r(x_i) + \epsilon_i, \quad i = 1, \ldots n, \\ y_i' &= r(x_i) + \epsilon_i', \quad i = 1, \ldots n. \end{aligned}$$

The points $x_i$, $i = 1, \ldots n$ are considered fixed. The errors $\epsilon_i$, $i = 1, \ldots n$ are uncorrelated, with mean zero and variance $\sigma^2$. The same is true for $\epsilon_i'$, $i = 1, \ldots n$, and the two sets of errors are independent. I.e., here $(x_i, y_i)$, $i = 1, \ldots n$ is the training set, on which we fit $\hat{y}_i = \hat{r}(x_i)$, $i = 1, \ldots n$, and $(x_i, y_i')$, $i = 1, \ldots n$ is the independent test set.

**(a)** Start off with

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n (y_i' - \hat{y}_i)^2\Big] = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n \big(y_i' - r(x_i) + r(x_i) - \hat{y}_i\big)^2\Big],$$

and then expand the right-hand side, to prove that

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n (y_i' - \hat{y}_i)^2\Big] = \sigma^2 + \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n \big(\hat{y}_i - r(x_i)\big)^2\Big].$$

**(b)** Now following the result in part (a), write

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(y_i' - \hat{y}_i)^2\Big] = \sigma^2 + \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\big(\hat{y}_i - r(x_i)\big)^2\Big] = \sigma^2 + \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\big(\hat{y}_i - y_i + y_i - r(x_i)\big)^2\Big],$$

and expand the right-hand side to prove the desired result in (1).

**(c)** Think about the result (1): this shows that the optimism—i.e., the difference in expected test error and expected training error—is equal to $2\sigma^2/n$ times the degrees of freedom. When $\mathrm{df}(\hat{y})$ is positive, this means that the expected training error is always smaller than the expected test error. When $\mathrm{df}(\hat{y})$ is negative, it is the other way around. So the question is: can $\mathrm{df}(\hat{y})$ ever be negative? If you think it can, give an example. If not, explain why not.

# 2 Kernels and splines: brothers, or distant cousins?

In this problem you'll working again with the abalone data set from Homework 3. Refer back to Problem 2 of Homework 3 for a description of the data set. This time we'll be looking at predicting the "whole weight" of male abalones from the diameter measurements (columns of 5 and 3 of the abalone data frame, respectively).

**(a)** Plot the weight versus diameter for male abalones.

**(b)** We're going to consider using both smoothing splines and kernel regresion (with a Gaussian kernel) to model this trend. For smoothing splines, we'll use the `smooth.spline` function in R, as in

```
ssmod = smooth.spline(x,y,df=10),
```

which fits a smoothing spline of $y$ on $x$ with 10 degrees of freedom. Smoothing splines are usually parametrized in terms of the smoothing parameter $\lambda$, but conveniently, this function allows us to specify a target degrees of freedom value directly. Use (5 or 10 fold) cross-validation to estimate the test error of the smoothing spline fit of weight on diameter for male abalones, at each of the degrees of freedom values `df = 2:60`.

For kernel regression, we'll use the `npreg` function in the `np` package, as in

```
kernmod = npreg(y~x,bw=5),
```

which fits a kernel regression (with Gaussian kernel, by default) of $y$ on $x$ with bandwidth 5. Use (5 or 10 fold) cross-validation to estimate the test error of the kernel regression fit of weight on diameter for male abalones, at each of the bandwidth values `bws = 3:50`.

(Hint: don't perform these separately, just write one cross-validation loop for both tasks. It will keep the coding at a minimum.)

For each of these procedures, plot the cross-validation errors as a function of the underlying tuning parameter; i.e., for smoothing splines, plot the cross-validation error as a function of the degrees of freedom of the fit, and for kernel regression, plot the cross-validation error as a function of the bandwidth.

**(c)** What is the optimal value of degrees of freedom for smoothing splines, in terms of the minimum cross-validation error? What is the optimal bandwidth for kernel regression? Which has a better minimum cross-validation error, smoothing splines or kernels? Do you believe the two to be significantly different in this regard? (Hint: for the last part, look at the cross-validation standard errors.)

**(d)** Suppose that the optimal value of degrees of freedom for smoothing splines that you chose in part (c) was $d$, and the optimal bandwidth for kernel regression was $h$. Produce a plot of estimated regression line from smoothing splines, fit to the *entire* data set, with degrees of freedom $d$, on top of the weight-diameter pairs. Also produce a plot of the estimated regression line from kernel regression, fit again to the *entire* data set, with bandwidth $h$, on top of the weight-diameter pairs. Which looks more complex, the smoothing spline model or the kernel regression model?

**(e)** Prove that the degrees of freedom of a kernel regression fit with bandwidth $h$ is

$$\sum_{i=1}^{n} \frac{K(0)}{\sum_{j=1}^{n} K(\frac{x_i - x_j}{h})}. \tag{2}$$

(Hint: in lecture, we showed that the degrees of freedom of a linear smoother was $\sum_{i=1}^{n} w(x_i, x_i)$. Kernel regression is just a linear smoother with a particular choice of weights ...)

**(f)** Using the result of part (e), for each bandwidth value `bws = 3:50`, calculate the corresponding value of degrees of freedom of the kernel regression fit—remember, in you kernel regression, you used the Gaussian kernel, so that

$$K(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

You should perform this calculation using the entire data set, i.e., the $x_i$ points that you use in (2) are the diameter values from the entire data set of male abalones. (Hint: it may help to write two functions: one to compute the term $1/\sum_{j=1}^{n} K(\frac{x_i - x_j}{h})$ for a particular point $x_i$, and another to sum this over $x_i$, $i = 1, \ldots n$.)

What is the trend you see? I.e., is the degrees of freedom high for large bandwidths, or small for large bandwidths? And does this make sense?

**(g)** Reproduce a plot of the cross-validation errors of kernel regression, but this time, plot the corresponding degrees of freedom values (that you calculated in (f)) on the x-axis, rather than the bandwidth values. This now allows you to make a direct comparison between the cross-validation error curves for kernel regression and smoothing splines. Do so: describe the similarities/differences between the general shapes of the two curves, as functions of degrees of freedom, and explain what this means in the context of modeling. Comment on the value of degrees of freedom at which the kernel regression cross-validation error curve is minimized, and whether this is different from that for smoothing splines.

**Bonus 1:** There is a difference in behavior for the cross-validation errors of kernel regression and smoothing splines, towards the end of the degrees of freedom spectrum (i.e., for low and high degrees of freedom values). Though this may not seem like a drastic difference, it is pretty characteristic. Can you explain why this might be the case?

**Bonus 2:** One might argue that, since the cross-validation error curves for both kernel regression and degrees of freedom are so flat in the middle of the degrees of freedom range, they are both pretty much minimized by any number of a large degrees of freedom values. Brainstorm: supposing we had a preference for lower complexity models, what rule could we use to choose an optimal value for degrees of freedom for either of these two methods, instead of just using the value that minimizes the cross-validation error curve? (Hint: the cross-validation errors may come in handy here.)