Homework Assignment 5

36-350, Data Mining

Due at the start of class, Friday, 31 October 2008

The general form of a linear model is

$$\mathbf{E}\left[Y|X_1 = x_1, X_2 = x_2, \ldots X_p = x_p\right] = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

where we need to estimate the weights β_i.

An extension is to use an **additive** model,

$$\mathbf{E}\left[Y|X_1 = x_1, X_2 = x_2, \ldots X_p = x_p\right] = \alpha + \sum_{j=1}^{p} f_j(x_j)$$

where we need to estimate the functions f_j. This can be done using any of our non-parametric smoothing methods (local linear regression, kernel regression, k-nearest-neighbors regression, etc.). Additive models are more flexible than linear models, but less flexible than fully non-parametric regressions.

In the linear model, β_0 is chosen to make sure that the regression line goes through the center of the data. Similarly, with additive models, we choose $\alpha = \mathbf{E}\left[Y\right]$, and require that the functions obey $\mathbf{E}\left[f_j(X_j)\right] = 0$.

Define the **partial residuals** of the i^{th} observation on X_k as

$$r_{ik} \equiv y_i - \alpha - \sum_{j \neq k} f_j(x_{ij})$$

where y_i is the value of the response at i, and x_{ij} is the value of the j^{th} predictor variable. This is how much difference the k^{th} feature makes to the predictions at i.

You can read more about additive models on pp. 393–395 of the textbook, though notice that they leave out the constant α and don't require $\mathbf{E}\left[f_j(X_j)\right] = 0$. This is mathematically equivalent but makes fitting harder. Also, there is an R package on CRAN called `gam` which fits generalized additive models, including additive models as a special case.

1. (a) Show that

$$f_k(x_k) = \mathbf{E}\left[Y - \alpha - \sum_{j \neq k} f_j(X_{ij}) \middle| X_k = x_k \right]$$

i.e., the function f_k is the conditional expectation of the partial residuals. *Hint:* Use smoothing, a.k.a. the law of total expectation.

(b) Write a function to implement the following **back-fitting** procedure for estimating additive models. It should take as its inputs a response variable, predictor variables, and a tolerance level δ. (The response and predictors can either be different columns of the same data frame or separate objects, whichever you find easier.)

i Set $\widehat{\alpha}$ to the sample mean of the response

ii Set $\widehat{f_j} = 0$ for all j

iii For each j in $1 : p$, calculate the partial residuals r_{ij} using the current $\widehat{f_j}$. Set $\widehat{g}_j$ to be the function obtained by smoothing the partial residuals with your favorite smoothing procedure. Set $\widehat{f_j}$ equal to $\widehat{g}_j$, minus the sample mean of $\widehat{g}_j$.

iv Check whether the predicted value at any point has changed by more than δ. If so, go back to (iii); if not, return the estimated functions and exit.

Hint: You may find the trick for storing and accessing many variables with similar names used in the lecture handout for 20 October helpful.

(c) Explain how you would modify your code to choose the degree of smoothing by cross-validation. Remember that each function f_j might be more or less smooth than the others, so it needs its own bandwidth.

(d) Why is it helpful to set $\alpha = \mathbf{E}[Y]$ and require that $\mathbf{E}[f(X_j)] = 0$? *Hint:* What happens if you add 19740228 to f_1 and subtract it from f_2?

2. Download the California housing data set used on the exam, `http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=83`

   (a) Linearly regress the *log* of the median house price on all the other variables. Report your regression coefficients, your mean squared error, and the distribution of residuals. Is the latter Gaussian? Do scatter-plots of residuals against predictors show any trends?

   (b) Using a kernel smoother, regress the log of the median house price on the median income. Use cross-validation to pick the bandwidth. What is the mean squared error? Plot the estimated regression function; is it linear? Plot the distribution of residuals; are they Gaussian? Plot the residuals versus the median income; do you see any trends?

   (c) Fit an additive model to the same data. Report the mean squared error and plots of the estimated functions. Are they close to linear? What is the distribution of total (i.e., not partial) residuals? (If you cannot complete the coding in the first question, use the `gam` package on CRAN to do this part.)