Statistical Learning I - Homework # 5, Due November 28

1. Recall that a normal spline on K knots can be constructed using K basis functions. Specifically, for $K = 3$ and knots $\xi_1 < \xi_2 < \xi_3$, these functions are $N_1(X) = 1$, $N_2(X) = X$, and

$$N_3(X) = \frac{(X-\xi_1)_+^3 - (X-\xi_3)_+^3}{\xi_3 - \xi_1} - \frac{(X-\xi_2)_+^3 - (X-\xi_3)_+^3}{\xi_3 - \xi_2}.$$

Show that for $f(X) = \sum_{j=1}^{3} \beta_j N_j(X)$ that $f''(X) = 0$ for $X < \xi_1$ and for $X > \xi_3$, that is, the function f is linear in these regions.

2. A cubic spline with two knots $\xi_1 < \xi_2$ can be constructed using the six basis functions $h_1(X) = 1$, $h_2(X) = X$, $h_3(X) = X^2$, $h_4(X) = X^3$, $h_5(X) = (X-\xi_1)_+^3$, and $h_6(X) = (X-\xi_2)_+^3$. Show that for $f(X) = \sum_{j=1}^{6} \beta_j h_j(X)$, the first and second derivatives at the knots are continuous.

3. Using the *cars04.csv* data set provided, try to predict the Suggested Retail Price from the other variables, using the following suggestions as a guide. Note that there are a few Hybrid cars in the data set.

a) Make a pairs plot of the quantitative variables and use "Hybrid" as the color on the plot so you can see how the few hybrids affect the data.

b) Without making any transformations or adding second order terms, fit the best linear model that you can to predict the SRS from the other variables.

c) Using the *gam* library, investigate improvements on your model by replacing variables with splines or local regression basis functions. Fit the best model possible.

d) Discuss your two models. Which do you prefer and why?

4. Let V be a vector space over the reals with and inner product $\langle \cdot, \cdot \rangle$. Let $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. Show that $\| \cdot \|$ satisfies the properties of a norm on V. You may use the Cauchy-Schwartz inequality where needed.