Homework 1
Due Friday [04/19/13] in APM 5151 3pm
MATH 287D – Statistical Learning

Be concise. Always comment on your results/findings. Send your R code, cleaned, polished and commented. Make it so that running the R code is straightforward. Send it to jbradic@ucsd.edu with the exact subject line 287D Homework (number).

Problem 1. Fit the simple linear model $y = \beta_1 x + \beta_0$ to the following (true) model:

(a) $E[y|x] = \alpha_1 x + \alpha_0$

(b) $E[y|x] = \alpha_2 x^2 + \alpha_1 x + \alpha_0$

(c) $E[y|x] = \alpha_1 \sin(x) + \alpha_2 \cos(x)$

(d) another model of your choice (make it interesting!)

(a) Each time, x and ε are independent of each other, with x non constant and $E(\varepsilon) = 0$. For models (a) and (b), express the least squares coefficients in terms of the α coefficients and the moments of x, denoted $\mu_j = E(x^j)$. For model (c), assume that x is uniform in $(0, 2\pi)$.

(b) In R, simulate each model in Problem 1. (Try different distributions for x and ε.) Varying the sample size (e.g., $n \in \{50, 200, 1,000\}$), compute the least squares coefficients (see the function `lm()`) and show that they converge to the true values. (Bonus) What is the rate of convergence?

Problem 2. Describe and propose K-NN method for data with a categorical response. Implement your own function for it. Test it for a dataset of size $n = 100$ in dimension $p = 2$. See what would change if p gets larger and larger ?

Problem 3.
Understanding bias and variance of k-NN method.

(a) Compute the variance and bias trade-off for local averaging with in k-NN for a model of the form $y_i = f(x_i) + \varepsilon_i$, $i = 1, \dots, n$, where f is Lipschitz with constant A, and the ε_i are independent of the x_i's and uncorrelated among themselves, with zero mean and same variance σ^2. For simplicity, assume that the x_i span the grid $\{1/m, 2/m, \dots, (m-1)/m, 1\}^d$ where $m = n^{1/d}$ (assumed to be an integer).

(b) Can you illustrate that with simulations?

Problem 4. Let $\{(x_i, y_i), i = 1, \cdots, n\}$ be a dataset and let $w_i(x) = K((x - x_i)/h)$ be a weight function. One way to justify NW estimators by the following optimization. Consider finding the c that minimizes the weighted sum of squared errors

$$\sum_{i=1}^{n} w_i(x)(y_i - c)^2.$$

(a) Show that the minimum is achieved for NW estimator.

(b) Consider trying to improve the local constant estimator to a local polynomial kernel estimator. To do this, write the Taylor expansion on a neighborhood of x,

$$g_x(u; c) = c_0 + c_1(x - u) + \frac{c_2}{2!}(x - u)^2$$

In this form the goal is to find $\hat{c} = (\hat{c}_0, \hat{c}_1, \hat{c}_2)^T$ a sum of squares

$$\sum_{i=1}^{n} w_i(x)(y_i - g_x(x_i; c))^2.$$

(a) Why does $\hat{c}$ depend on x now ?

(b) Show that $\hat{g}_x = g_x(x; \hat{c}) = \hat{c}_0$. Why is this different from local constant estimator NW ?

(c) Show the estimator is linear estimator ? [Hint: consider writing the model in matrix form and use matrix notation to deduce linearity]

Problem 5. (Bonus) Illustrate the curse of dimensionality with a couple of graphs is such a way that any person with a minimum background in math would understand the basic principle.