Introduction to Kernel Regression

Xiang Ao

March 24, 2009

1 Nonparametric Density Estimation

Recall the definition of pdf of a random variable X is

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h) \tag{1}$$

We can approximate this for a given value of h as

$$\hat{f(x)} = \frac{1}{N} \frac{1}{2h} [\text{number of } X_1, \cdots, X_N \text{ falling in the interval}(x - h, x + h)] \tag{2}$$

The "naive" estimator is

$$\hat{f(x)} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} w(\frac{x - X_i}{h}) \tag{3}$$

where w is a weighting function (rectangular kernel) defined as

$$w(z) = \begin{cases} \frac{1}{2} & \text{if } |z| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

A broad class of density functions, can be defined as

$$\hat{f(x)} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} K(\frac{x - X_i}{h}) \tag{5}$$

where K refers to the kernel and h is the bandwidth.

2 Nonparametric Regression

The fitted values from a regression model are estimates of the expectation of the dependent variable conditional on the values of the explanatory variables for each observation. Linear regression models assumes that the expectation of the dependent variable conditional on independent variables is an affine function of independent variables. An alternative approach is to use a nonparametric

regression, which estimates $\mathrm{E}(\mathrm{y_t}|\mathrm{X_t})$ directly, without an assumptions about functional form.

The simplest approach is kernel regression. We suppose that two random variables Y and X are jointly distributed, and we wish to estimate the conditional expectation $\mu(x) \equiv \mathrm{E}(Y|x)$ as a function of x, using a sample of paired observation (y_t, x_t) for $t = 1, \ldots, n$.

Consider the function $G(x)$ defined as:

$$G(x) = \mathrm{E}(Y \cdot I(X \leq x)) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} y f(y, z) dy dz \tag{6}$$

Let $g(x) \equiv G'(x)$ denotes the first derivative of $G(x)$. Then

$$g(x) = \int_{-\infty}^{\infty} y f(y, x) dy = f(x) \int_{-\infty}^{\infty} y f(y|x) dy = f(x)\mathrm{E}(Y|x) \tag{7}$$

We use the biased but smooth estimator of $G(x)$

$$\hat{G}_h(x) = \frac{1}{n} \sum_{t=1}^{n} y_t K(\frac{x - x_i}{h}) \tag{8}$$

Defining $\hat{g}_h(x)$ as the derivative of $\hat{G}_h(x)$ and using the kernel estimator to estimate the marginal density of X leads to the following estimator of $\mu(x)$:

$$\hat{\mu}_h(x) = \frac{\hat{g}_h(x)}{\hat{f}_h(x)}. \tag{9}$$

This is called Nadaraya-Watson estimator. It simplifies to

$$\hat{\mu}_h(x) = \frac{\sum_{t=1}^{n} y_t k_t}{\sum_{t=1}^{n} k_t}, \; k_t \equiv k(\frac{x - x_t}{h}), \; k_t \equiv K_t'. \tag{10}$$

We define the cross-validation function by the formula

$$\mathrm{CV}(h) = \frac{1}{n} \sum_{t=1}^{n} w(x_t)(y_t - \hat{y}_h^{(t)})^2 \tag{11}$$

where $\hat{y}_h^{(t)}$ is the leave-one-out estimator, $w(x_t)$ is a weight. When we use cross validation, we evaluate $\mathrm{CV}(h)$ for a number of values of h and pick the value that minimizes it.

3 Intuition

The idea of kernel regression is to use a non-parametric method to estimate the relationship between Y and X. Say we have m pairs of x_i and y_i observed, in the interval of a and b. The idea is to put a kernel function at every point of x_i observed. Then, between , divide the interval into n equal intervals, x_1 , x_2,...,

x_t, ..., x_n. At each of these points, some of them do not have observed Y and X. We need to predict y_t for each x_t. The way to do it is to use weighted average of all the y_i's observed as an estimate of $\hat{y}_t$. The weight is the pdf of each kernel density at the point of x_i. Therefore, the larger the distance between x_i and x_t, the smaller the weight. The bandwidth determines the variance of the density. The larger the bandwidth, the flater the density, the larger the weight a far point can get.