Nonparametric regression

Khanh Nguyen

March 12, 2015

We are given n pairs of observations (x_1, y_1), ..., (x_n, y_n), presumably drawn from the distribution $P(X, Y)$. We want to use those observations to estimate the regression function $r(x) = \mathbb{E}(Y \mid X = x)$ using weak assumptions on $r(x)$. The estimator of $r(x)$ is denoted by $\hat{r}_n(x)$.

1 Linear Smoothers

Definition. An estimator $\hat{r}_n$ is a **linear smoother** if, for each x, there exists a vector $l(x) = [l_1(x), ..., l_n(x)]$ such that:

$$\hat{r}_n(x) = \sum_{i=1}^{n} l_i(x) y_i = l(x)^\top y$$

We gather all the vector $l(x_i)$ into a matrix L. Entry (i, j) of L is $l_j(x_i)$.

Example. Suppose that $a \leq x_i \leq b$, **regressogram** divides the interval (a, b) into m equally spaced bins denoted by $B_1, ..., B_m$. Then for $x \in B_j$:

$$\hat{r}_n(x) = \frac{1}{|B_j|} \sum_{i: x_i \in B_j} Y_i$$

Example. Let $h > 0$ be a positive integer, called the bandwidth. The **Nadaraya-Watson** (NW) **kernel estimator** is defined by:

$$\hat{r}_n(x) = \sum_{i=1}^{n} l_i(x) y_i$$

where

$$l_i(x) = \frac{K(\frac{x - x_i}{h})}{\sum_{j=1}^{n} K(\frac{x - x_i}{h})}$$

with K being a kernel.

For linear smoothers, there is always a hyper-parameter h that controls that the degree of smoothness. To choose the optimal h, we conduct cross-validation, i.e. minimizing the cross-validation score $\hat{R}(h)$:

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{r}_n(x_i)}{1 - L_{ii}} \right)^2$$

2 Confidence interval

The confidence interval for $\bar{r}_n = \mathbb{E}(\hat{r}_n(x))$ is

$$\hat{r}_n(x) \pm c\hat{\sigma}(x) \, \|l(x)\|$$

The constant c is chosen as the $(1-\alpha/2)$ quantile of the standard normal distribution, where α is the desired level of confidence.

To estimate the variance $\hat{\sigma}^2(x)$, we apply the following process recommended by Wasserman (2006):

1. Define $Z_i = \log(y_i - \hat{r}_n(x_i))^2$.
2. Regress the Z_is on the x_is (using nonparametric method) to get an estimate $\hat{q}_n(x)$ of $\log \sigma^2(x)$.
3. Set $\hat{\sigma}^2(x) = e^{\hat{q}_n(x)}$.

3 Local likelihood

When y is a binary variable, using local likelihood seems to be more appropriate. In this case, each y_i is drawn from a Bernoulli distribution with parameter $r(x_i)$. We use the estimate:

$$\hat{r}_n(x) = \frac{e^{\hat{\theta_*}(x)}}{1 + e^{\hat{\theta_*}(x)}}$$

where $\hat{\theta_*}(x)$ maximizes the log-likelihood function:

$$l_x(\theta) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \left(y_i \theta(x)^\top (x_i - x) - \log(1 + e^{\theta(x)^\top (x_i - x)}) \right)$$

References

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.