Chapter 2. Semi-parametric Models (I)
Part 6

February 28, 2007

1 Selecting models based on CV

Suppose we have a number of models to fit data $(X_i, Y_i), i = 1, ..., n$. The question is to select one of the models. For each model, we calculate its CV value. The model with the smallest CV value is the model we choose.

Example 1.1 (Simulation) *50 samples are generated from model*

$$y = (\mathbf{x}_1 - \mathbf{x}_2 + 0.5\mathbf{x}_3)^2 + 0.2 * \varepsilon$$

Suppose we don't know the true model and need to select a model between linear regression model

$$LM: \quad y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \xi$$

and PPR model (with 1 component, i.e. single-index model)

$$PPR1: \quad y = \phi(\beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3) + \xi$$

*(**(code)**). The calculation shows that most of the time,*

$$\text{CV of LM} > \text{CV of PPR1}$$

the CV criterion suggests that we need to choose a PPR model (with 1 component)

If the data are generated from model

$$y = \mathbf{x}_1 - \mathbf{x}_2 + 0.5\mathbf{x}_3 + 0.2 * \varepsilon$$

The calculation shows that most of the time,

$$\textit{CV of LM} < \textit{CV of PPR1}$$

we choose linear regression model.

Remark 1.2 *We always prefer simplest model if all models are correct*

Example 1.3 (ozone data) *If we select a model between linear regression model and PPR models with 1, 2, ... components. We have the CV values are*

$$\begin{aligned}
\textit{CV of linear regression model}: &\quad 468.4915 \\
\textit{CV of PPR with 1 component}: &\quad 346.0969 \\
\textit{CV of PPR with 2 component}: &\quad 340.3468 \\
\textit{CV of PPR with 3 component}: &\quad 334.6160 \\
\textit{CV of PPR with 4 component}: &\quad 328.4993 \\
\textit{CV of PPR with 5 component}: &\quad 330.4823
\end{aligned}$$

*(**(code)**) Thus, a PPR model with 5 components is suggested.*

2 Classification and Regression Tree (CART)

Suppose we have $(\mathbf{x}_{i1}, y_i), i = 1, ..., n.$ the plot of y against $\mathbf{x}$ is shown in Figure 1. We can fit the relation between y and x by

$$y = \begin{cases} 1, & \text{if } x \le 0.2 \\ 0, & \text{if } x > 0.2 \end{cases}$$

If it looks like figure 2, we fit the relation by

$$y = \begin{cases} \stackrel{x \le 0.2}{=} \begin{cases} \stackrel{x \le -0.5}{=} 1.2 \\ \stackrel{x > -0.5}{=} 0.8 \end{cases} \\ \stackrel{x > 0.2}{=} 0. \end{cases}$$

Please note the connection between this idea and the NW kernel estimation.

More generally, if we have $(\mathbf{x}_1, ..., \mathbf{x}_p, Y)$ and samples $N = \{(\mathbf{x}_{i1}, ..., \mathbf{x}_{ip}, y_i) : i = 1, ..., n\}$. Here, $X = (\mathbf{x}_1, ..., \mathbf{x}_p)$ is in a p dimensional space. Our interest is still the regression surface

$$m(x_1, ..., x_p) = E(Y|\mathbf{x}_1 = x_1, ...\mathbf{x}_p = x_p).$$

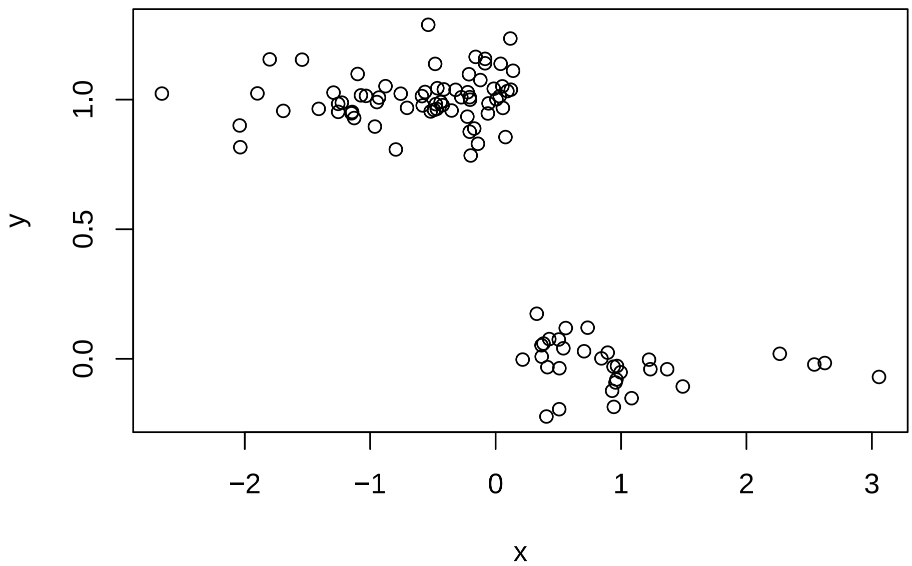

Figure 1:

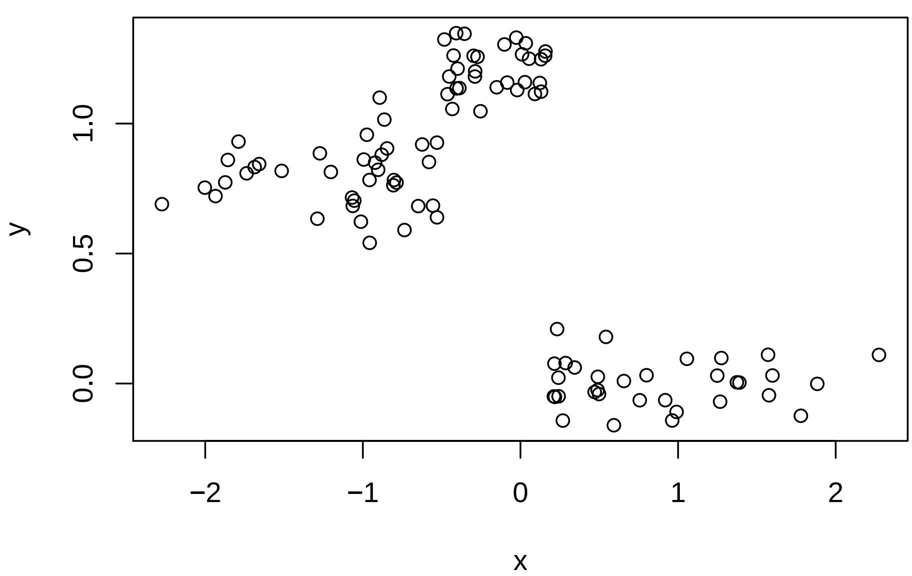

Figure 2:

We try to find a partition $N_1, ..., N_H$ of the space and approximate the function by

$$m(x_1, ..., x_p) \approx \sum_{h=1}^{H} c_h I(x \in N_h)$$

where $I(.)$ is the indicator function.

In practice, if the partition is given, then c_h is estimated as

$$\hat{c}_h = \frac{\sum_{i=1}^{n} y_i I(X_i \in N_h)}{\sum_{i=1}^{n} I(X_i \in N_h)}.$$

The main difficulty is in partitioning the space. The **regression tree** partitions the space by binary recursive method. For each variable $\mathbf{x}_k$ and a node x'_k, we consider a model

$$Y = \begin{cases} \stackrel{\mathbf{x}_k \le x'}{=} c_1 \\ \stackrel{\mathbf{x}_k > x'}{=} c_2 \end{cases}$$

Calculate its CV value (how?). Compare all the CV values (could be very many), the one with the smallest CV is first partition. Suppose it is $\mathbf{x}_1$ with x'_1. Denote the corresponding CV values by CV'. Calculate also the CV for model

$$Y = c + \xi$$

denote it by CV_0, i.e.

$$CV_0 = n^{-1} \sum_{i=1}^{n} (y_i - \bar{Y}_i)^2$$

where $\bar{Y}_i = (Y_1 + ... + Y_{i-1} + Y_{i+1} + ... + Y_n)/(n-1)$. If $CV_0 < CV'$, stop and the final model is

$$Y = c + \xi$$

If $CV_0 > Cv'$. consider $N_1 = \{(X_i, Y_i) : \mathbf{x}_{i1} > x'_1\}$ and $N_2 = \{(X_i, Y_i) : \mathbf{x}_{i1} \le x'_1\}$. Applied the same procedure (as to N) to N_1 and N_2. continue the procedure until no more partitioning is needed.

Example 2.1 *Examples from R*

For cpus data, a new X

$$X = (203, 2867, 11796, 25, 5, 18)$$

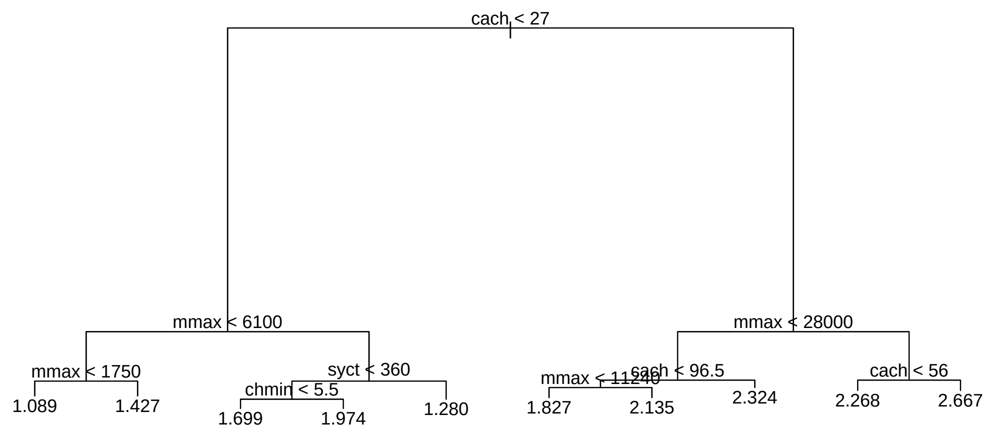

Figure 3: The estimated regression tree for cpus data. **(c2f3.R)**

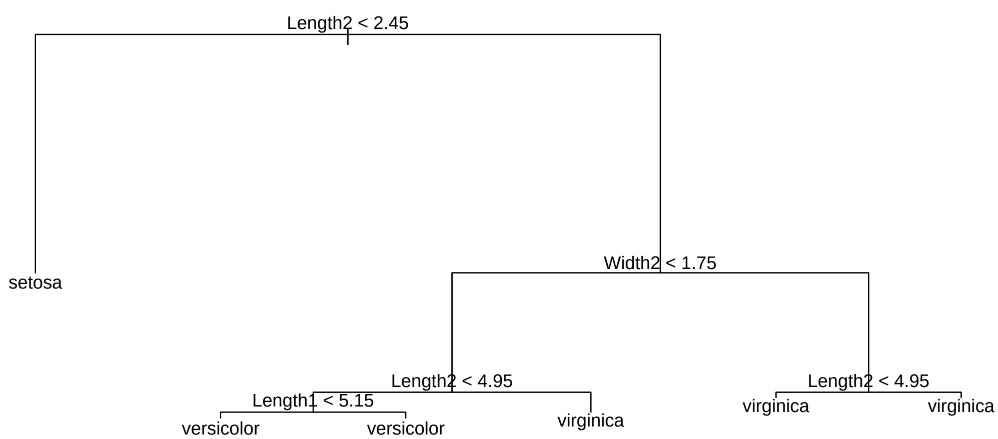

Figure 4: the estimated regression tree for data iris **(c2f4.R)**

The predicted $Y = log10(perf)$ *is 1.698613 (or* $10^{1.698613}$ *for perf).*

For the iris data, a new X *is*

$$Length1 = 5.8,\ Width1 = 3,\ Length2 = 3.7,\ Width2 = 1.2$$

The predicted $Y = Species$ *is: "versicolor"*

References

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.