7.5.2 LOGISTIC REGRESSION FOR $K > 2$ CLASSES

☛ Following the case where $K = 2$, in the logistic case, it may seem that we could assume that, for $k = 1, \ldots, K$,

$$m_k(\mathbf{x}) = E(Y_k|\mathbf{X} = \mathbf{x}) = P(G = k|\mathbf{X} = \mathbf{x}) = \frac{\exp(\beta_{k0} + \beta_k^T \mathbf{x})}{1 + \exp(\beta_{k0} + \beta_k^T \mathbf{x})}.$$

☛ However as seen earlier, we must have $m_1(\mathbf{x}) + \ldots + m_K(\mathbf{x}) = 1$.

☛ This can be satisfied if we use the following slightly different model: for $k = 1, \ldots, K - 1$:

$$m_k(\mathbf{x}) = \frac{\exp(\beta_{k0} + \beta_k^T \mathbf{x})}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T \mathbf{x})}.$$

☛ Then we take

$$\begin{aligned} m_K(\mathbf{x}) =& 1 - \sum_{k=1}^{K-1} m_k(\mathbf{x}) = 1 - \frac{\sum_{k=1}^{K-1} \exp(\beta_{k0} + \beta_k^T \mathbf{x})}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T \mathbf{x})} \\ =& \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_\ell^T \mathbf{x})}. \end{aligned}$$

☛ For $k = 1, \ldots, K-1$, estimate the β_{k0}'s and β_k's by their maximum likelihood estimators $\hat{\beta}_{k0}$, $\hat{\beta}_k$ (again can be applied to PLS components instead) and deduce estimators $\hat{m}_k$ of m_k. Then estimate $m_K(\mathbf{x})$ by

$$\hat{m}_K(\mathbf{x}) = 1 - \sum_{k=1}^{K-1} \hat{m}_k(\mathbf{x}).$$

☛ Classify the new individual in group $\hat{G}$ which gives the max value among $\hat{m}_1(\mathbf{x}), \ldots, \hat{m}_K(\mathbf{x})$:

$$\widehat{G} = \arg \max_{k=1,\ldots,K} \widehat{m}_k(\mathbf{x})$$

☛ Note that this is again a linear method. To see why, note that to find the max it suffices to compute, for $k = 1, \ldots, K-1$

$$\log \frac{\hat{P}(G = k|\mathbf{X} = \mathbf{x})}{\hat{P}(G = K|\mathbf{X} = \mathbf{x})} = \log \frac{\hat{m}_k(\mathbf{x})}{\hat{m}_K(\mathbf{x})} = \hat{\beta}_{k0} + \hat{\beta}_1^T \mathbf{x}\,,$$

and compare the groups 2 by 2 based on those log ratios.

☛ Ex: suppose $K > 5$ and $\hat{m}_5(\mathbf{x})$ is the max. Then

$$\hat{m}_5(\mathbf{x}) > \hat{m}_K(\mathbf{x}) \iff \log\{\hat{m}_5(\mathbf{x})/\hat{m}_K(\mathbf{x})\} > 0$$

and for all $k \neq 5$ and $\neq K$,

$$\hat{m}_k(\mathbf{x}) < \hat{m}_5(\mathbf{x}) \iff \frac{\hat{m}_k(\mathbf{x})}{\hat{m}_K(\mathbf{x})} < \frac{\hat{m}_5(\mathbf{x})}{\hat{m}_K(\mathbf{x})} \iff \log\frac{\hat{m}_k(\mathbf{x})}{\hat{m}_K(\mathbf{x})} < \log\frac{\hat{m}_5(\mathbf{x})}{\hat{m}_K(\mathbf{x})}$$

so we are able to make our decision based only on these log ratios which are linear in $\mathbf{x}$.

LD, QD based on normality assumption.

7.5.3 LD AND QD METHODS FOR $K > 2$ GROUPS (K>2时LD与QD的实施)

☛ Using the same ideas, LD and QD methods can be generalised to the case where data come from $K > 2$ groups.

☛ Let $\pi_1 = P(G = 1), \ldots, \pi_K = P(G = K)$.

☛ LD: Assume that for $k = 1, \ldots, K$, $\mathbf{X}|G = k \sim N_p(\mu_k, \Sigma)$ and classify new obs x in group k that maximises $\hat{m}_k$, or equivalently,

$$\delta_k(\mathbf{x}) = \log \hat{\pi}_k + \mathbf{x}^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k.$$

☛ QD: Assume that for $k = 1, \ldots, K$, $\mathbf{X}|G = k \sim N_p(\mu_k, \Sigma_k)$ and classify new obs x in group k that maximises

$$\delta_k(\mathbf{x}) = \log(\hat{\pi}_k) - \frac{1}{2} \log\{\det(\hat{\Sigma}_k)\} - \frac{1}{2} (\mathbf{x} - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (\mathbf{x} - \hat{\mu}_k).$$

☛ As in case $K = 2$ groups, for simplicity, relabel the data $\mathbf{X}_i$ making the distinction of which group they come from: $\mathbf{X}_{i,k}$, $i = 1, \ldots, n_k$ denote all the $\mathbf{X}_i$'s that are from group k;

☛ $\hat{\pi}_k = n_k/n$ is the proportion of training data that are from group k or $\hat{\pi}_k = 1/K$ depending on our beliefs,

$$\hat{\mu}_k = \overline{\mathbf{X}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{X}_{i,k}$$

$$\widehat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{X}_{i,k} - \widehat{\mu}_k)(\mathbf{X}_{i,k} - \widehat{\mu}_k)^T$$

and

$$\widehat{\Sigma} = \frac{1}{n_1 + \ldots + n_K - K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{X}_{i,k} - \widehat{\mu}_k)(\mathbf{X}_{i,k} - \widehat{\mu}_k)^T.$$

then ?

Logistic、LDQD、linear regression 都是参数模型，要估计参数，还有很多的假设(正态，独立)

8 Classification and regression trees and related methods

8.1 Classification and Regression trees (CART)

不确实 我们只能
是一种approximation.

Hastie et al. (2017), section 9.2 (second edition, 12th printing).

8.1.1 Introduction

- ☛ Methods from previous chapter rely on strong parametric assumptions (linear model, logistic model or normality assumptions).

- ☛ When these assumptions are too far from the truth, the performance of classifiers can be poor; e.g. recall the example at page 253.

- ☛ We need more flexible models that are less driven by strong parametric assumptions.

- ☛ Instead of using linear or logistic regression assumptions, one possibility is to use regression trees.

based on non-parametric assumption

8.1.2 Regression trees

n个点,

☛ Suppose we observe an i.i.d. sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ coming from the regression model

X是p维,

$$E(Y_i|\mathbf{X}_i = \mathbf{x}) = m(\mathbf{x}),$$

Y是一个值 not vector.

where $Y_i \in \mathbb{R}$ and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$ is continuous.

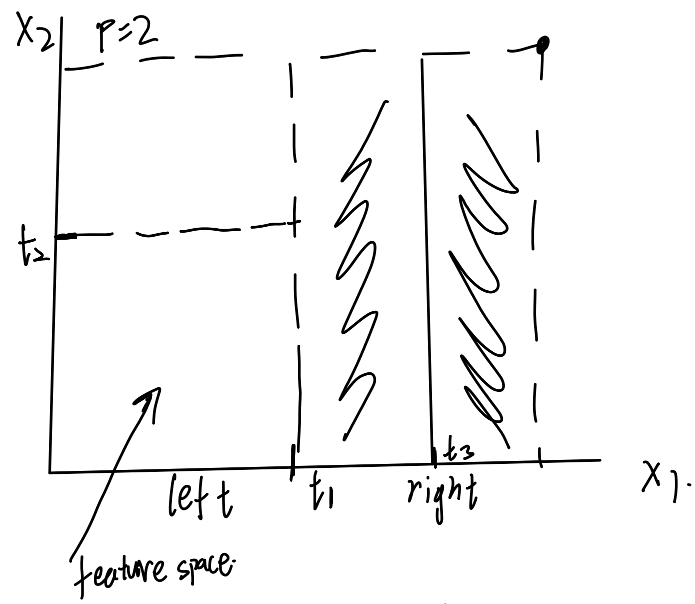
☛ When we don't want to highlight the dependence on the sample, we use the notation $(\mathbf{X}, Y)$, where

not sepecific but generic individual of population.

$$E(Y|\mathbf{X} = \mathbf{x}) = m(\mathbf{x}) \,,$$

$Y \in \mathbb{R}$ and $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ is continuous.

☛ Note the difference between the notation $\mathbf{X}_i$ for the ith training vector, and X_i for the ith component of the vector $\mathbf{X}$. We will use both notations a lot in this chapter.



1st split: devide $(X_1 < t_1 \quad X_1 \geq t_1)$ into 2 regions.

2st split $(X_2 < t_2 \quad X_2 \geq t_2)$

$t_1 < X_1 < t_3 \qquad X_1 > t_3$ . . . . ，不断地划分区域来找X的范围来创造特征空间，

gradually => divide big rectangles into some smaller rectangles each time. and each time work on one rectangle. (either vertically or horizontally).

$(X_1/X_2$ less/greater to $t_2)$

☛ The main idea of regression trees is to partition the "feature space", i.e. the domain of $\mathbf{X}$, i.e. the subset of $\mathbb{R}^p$ of all possible values of $\mathbf{X}$.

☛ On each partition we approximate $m(\mathbf{x})$ by a constant.

☛ To understand regression trees, consider the case where $p = 2$. Here we need to estimate the regression curve

$$m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$$

where $Y \in \mathbb{R}$ and $\mathbf{X} = (X_1, X_2)^T \in \mathbb{R}^2$ is continuous.

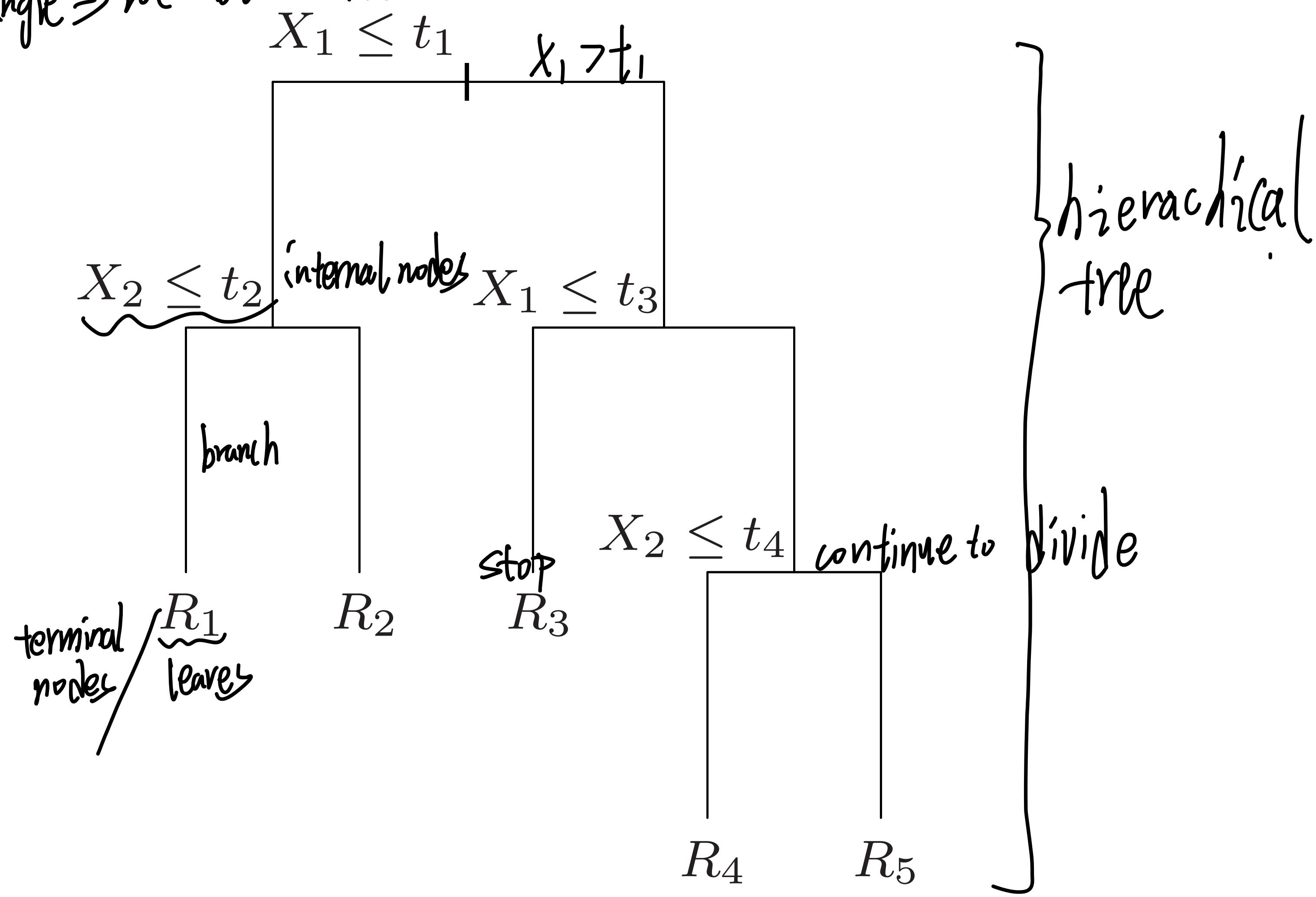
☛ Then we construct a sequence of partitions of the type:

- $\{X_1 \leq t_1\}$, $\{X_1 > t_1\}$
- $\{X_2 \leq t_2\}$, $\{X_2 > t_2\}$
- $\{X_1 \leq t_3\}$, $\{X_1 > t_3\}$

- etc. See handwritten construction on video.

page 306 of Hastie et al. (2017): Constructing a tree by a sequence of binary partitions of the type $\{X_j \leq t\}$, $\{X_j > t\}$. After a number of partitions, stop the splitting process and obtain regions $R_1, \ldots, R_L$.

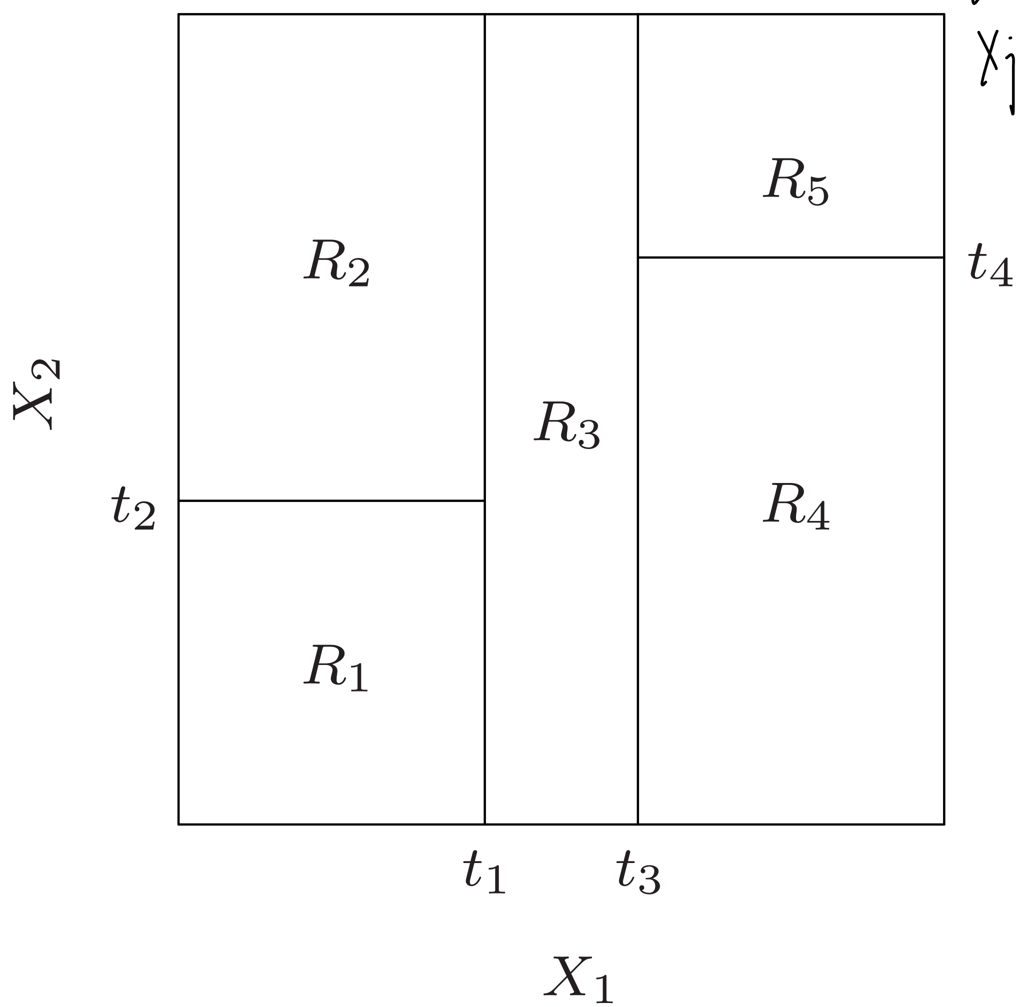
by dividing rectangle => we obtain a tree.



☛ page 306 of Hastie et al. (2017): At the end of the sequence of binary partitions, we have partitioned the feature space in rectangles, say $R_1, R_2, \ldots, R_L$. Example when $p = 2$ and $L = 5$:

divided

X_j greater/less than



☛ The regions $R_1, \ldots, R_L$ obtained at the end of the process are called terminal nodes or leaves of the tree.

终点 叶子

☛ The splits such as $\{X_1 \leq t_1\}$, inside the tree, are called internal nodes.

☛ The segments of the tree that connect the nodes are called branches of the tree.

☛ Once we have partitioned the space into regions $R_1, \ldots, R_L$, on each region we approximate the regression curve m by a constant:

sum of all of my L regions 这个结好 (SST区)

$$\text{For all } \mathbf{x} \text{ in feature space, } m(\mathbf{x}) \approx \sum_{\ell=1}^{L} c_\ell \underbrace{I\{\mathbf{x} \in R_\ell\}}_{\text{Indication Fuction.}}$$

where

$$c_\ell I\{\mathbf{x} \in R_\ell\} = \begin{cases} c_\ell & \text{if } \mathbf{x} \in R_\ell \\ 0 & \text{otherwise.} \end{cases}$$

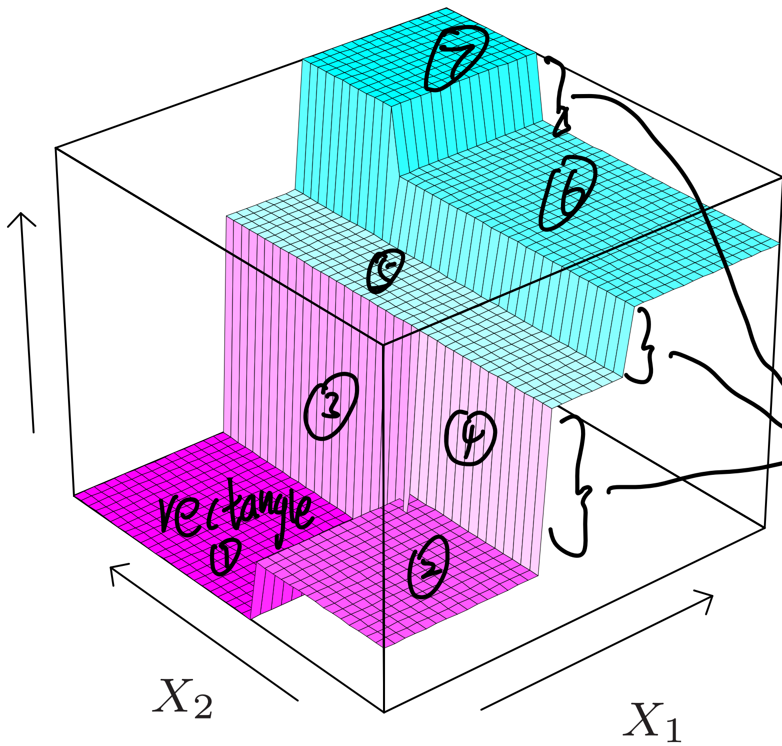
c_j constant used in that particular region x indicator of X in the particular region.

系

page 306 of Hastie et al., 2017: piecewise constant approximations on regions $R_1, R_2, \ldots, R_5$ of $\mathbb{R}^2$ constructed earlier:

by each of region we approximate the regression curve by a constant

不同的 rectangle 跟不同的 constant



说到和上面的一个意思

height is difference on each rectangle

☛ Why is this flexible? As long as we partition a feature space in small enough pieces, we can always approximate well a regression curve by constants on each piece.

just use constant in each region instead of using sophiscated functions.

☛ The finer the partition of the feature space, the better is the approximation of m by constants $c_1, c_2, c_3, \ldots$ on the regions $R_1, R_2, R_3, \ldots$.

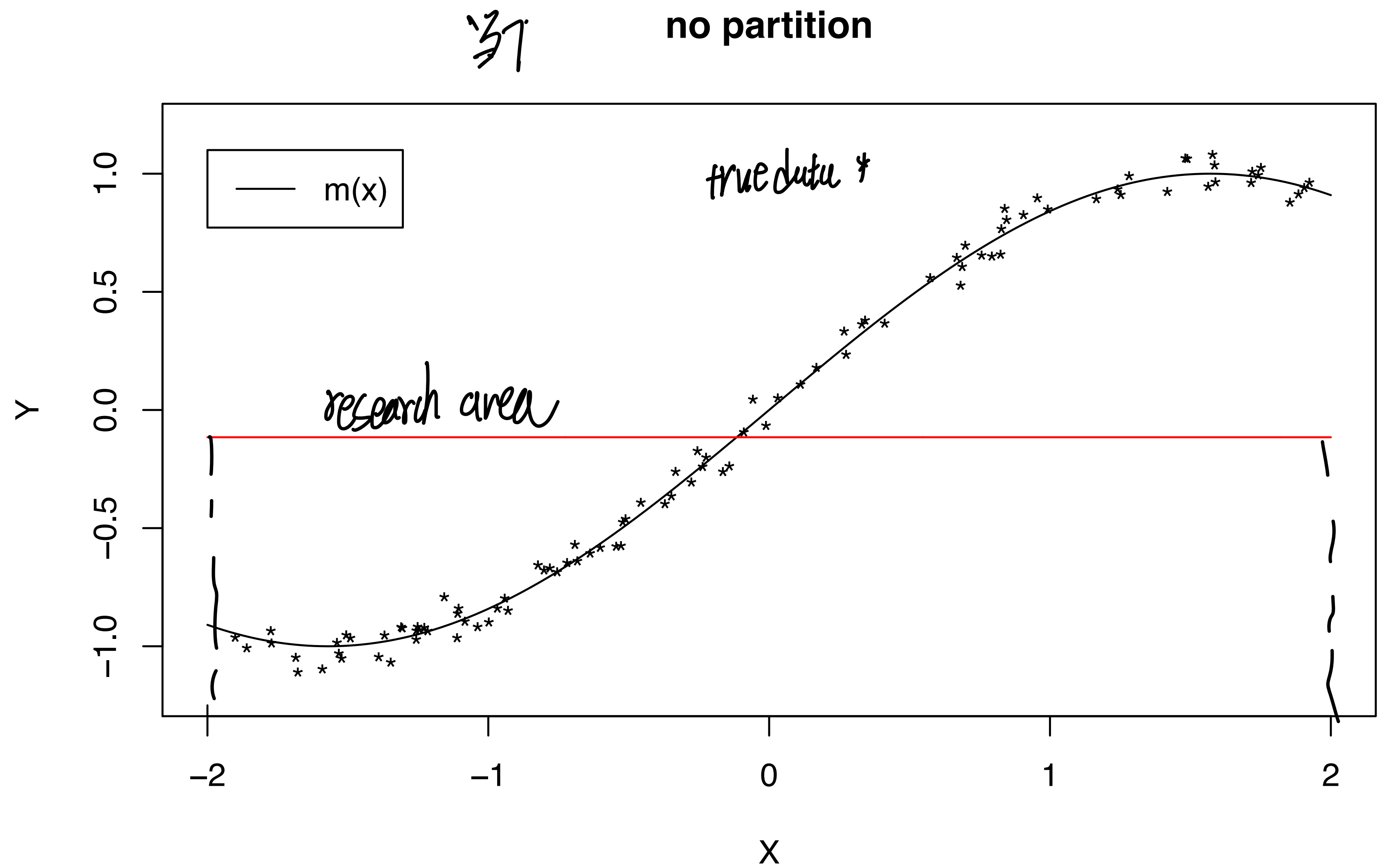
☛ To understand this, here is an example in the case where $p = 1$.

分区分得越细，regression就越精准.

只有一个变量

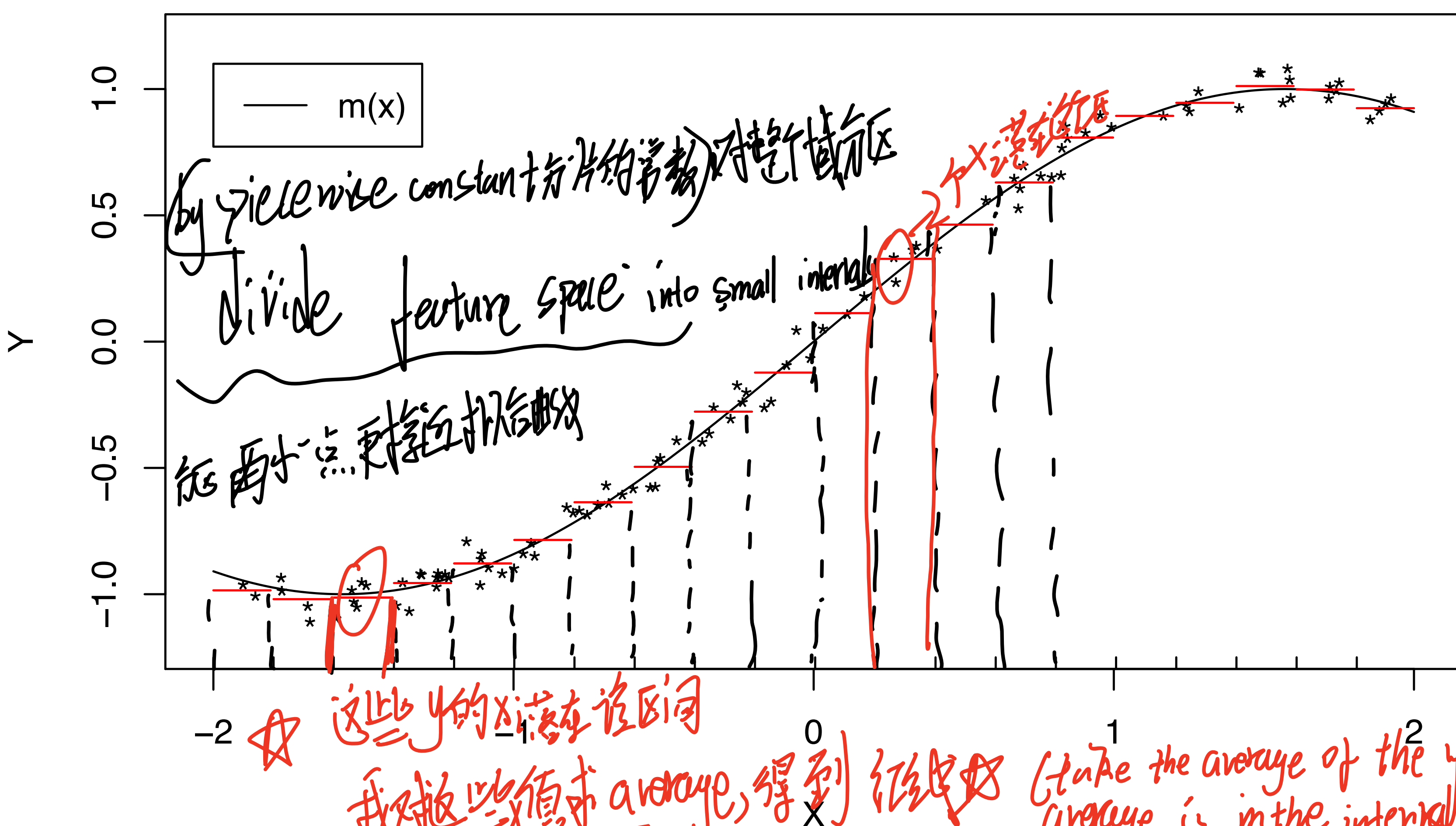
Example with $p = 1$. Here, partitioning the feature space means splitting the range of values of the variable X into intervals.

Data points:$*$. Approximate m by a constant (red line) without partitioning feature space does not give a good approximation of m:



Partition the real line into small intervals and approximate m by a well chosen constant on each interval gives ~~a~~ much better approximation of m

reasonable partition



(by piecewise constant 分片的常数)对整个域分区

Divide feature space into small intervals

比两个点更接近拟合曲线

个x落在这区间

☆ 这些y的x落在该区间

我对这些值求average，得到红线☆ (take the average of the y, whose average is in the interval)

参数值 in each piece is best possible approximation to my curve

FITTING THE TREE (CHOOSING THE CONSTANTS)

实际

- In practice don't know m and can't choose c_j that approximates m "the best" on each R_j.

- Instead, for each j, using only the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, approximate "the best" c_j by an estimator $\hat{c}_j$. How?

把特征空间分成L个

- Suppose we have partitioned the feature space of $\mathbf{X}$ into $R_1, \ldots, R_L$. For $\ell = 1, \ldots, L$, for $\mathbf{x} \in R_\ell$, we approximate $m(\mathbf{x})$ by

(的获取方式: estimator of regression curve (常数) ⇒ 在 R_ℓ 内的 X_i 对应的 Y_i 的均值

$$\hat{m}(\mathbf{x}) = \hat{c}_\ell = \text{average of } Y_i's \text{ whose } \mathbf{X}_i \in R_\ell$$

$$= \sum_{i=1}^{n} Y_i \cdot I(\mathbf{X}_i \in R_\ell) \Big/ \sum_{i=1}^{n} I(\mathbf{X}_i \in R_\ell).$$

sum of indicators

(indicator of whether X_i are in the region or not)

(的选择逻辑

- Same as choosing $\hat{c}_j$'s that minimise residual sum of squares (RSS)

所有区域

$$\sum_{i=1}^{n} \Big\{Y_i - \underbrace{\sum_{\ell=1}^{L} c_\ell\, I(\mathbf{X}_i \in R_\ell)}_{\hat m(X_i)}\Big\}^2 = \sum_{\ell=1}^{L} \sum_{\mathbf{X}_i \in R_\ell} (Y_i - c_\ell)^2.$$

wrt $c_1, \ldots, c_L$

Σ (每个区域的真实值-估计值)

BUILDING THE TREE (CHOOSING CONSECUTIVE SPLITS)

☛ Recall that for $\mathbf{X} = (X_1, \ldots, X_p)^T$, a split is of the type $\{X_j \le t\}, \{X_j > t\}$ for some $1 \le j \le p$ and some value t.

☛ Ideally, choose splits so that final tree minimises RSS but not computationally possible.

☛ Why? Would involve constructing all possible sequences of splits $\{X_j \le t\}, \{X_j > t\}$ for all $j = 1, \ldots, p$ and all values of t.

☛ Note: in fact, for split on X_j, only need to consider $t \in \{X_{1j}, \ldots, X_{nj}\}$.

Why? To find $\hat{c}_\ell$ on R_ℓ, only need to know $I(\mathbf{X}_i \in R_\ell)$. Changing t for split $\{X_j \le t\}, \{X_j > t\}$ used in R_ℓ affects $I(\mathbf{X}_i \in R_\ell)$ through $I(X_{ij} \le t)$, whose value only changes when t changes from $t < X_{ij}$ to $t \ge X_{ij}$ (or vice versa).

☛ Even with this simplification, too time consuming to consider all partitions.

about tree: at the end, the entire feature space was divided into a number of regions.

多数情况下，我们只有树没有图 ⇒ 有很多[illegible] → 一棵树

做 regression tree 时: we approximate a regression curve by piecewise constant.
(一个区域一个常数)

在每个区域 ⇒ we approximate regression curve by a constant (one constant per region).

In every small region

对于每个区域的 constant: (the fit that we use in each region)

= average of y_i (whose x_i belongs to that region ($x_i \in R_i$)) [we need to check which data falls in a particular region. ⇒ we take the average of these y_i of these particular individuals whose x_i was in the region]

选择能最小化残差平方和的 c (RSS)

$$\sum_{i=1} \sum_{x_i \in R_i} (y_i - c_i)^2$$

How do we choose the splitting point.

$$X = (X_1, \cdots, X_p)^T \Rightarrow \overbrace{\{X_j \le t\}. \{X_j > t\}}^{\text{two regions}}$$ (how do we choose j & t).

Split is based on residue sum of squares

choose the tree that can minimize the RSS → 但树太多

所以无法计算太多树 (计算资源)

we use the Indicator of variable to fit the tree (X在该区域就是1 不在是0)