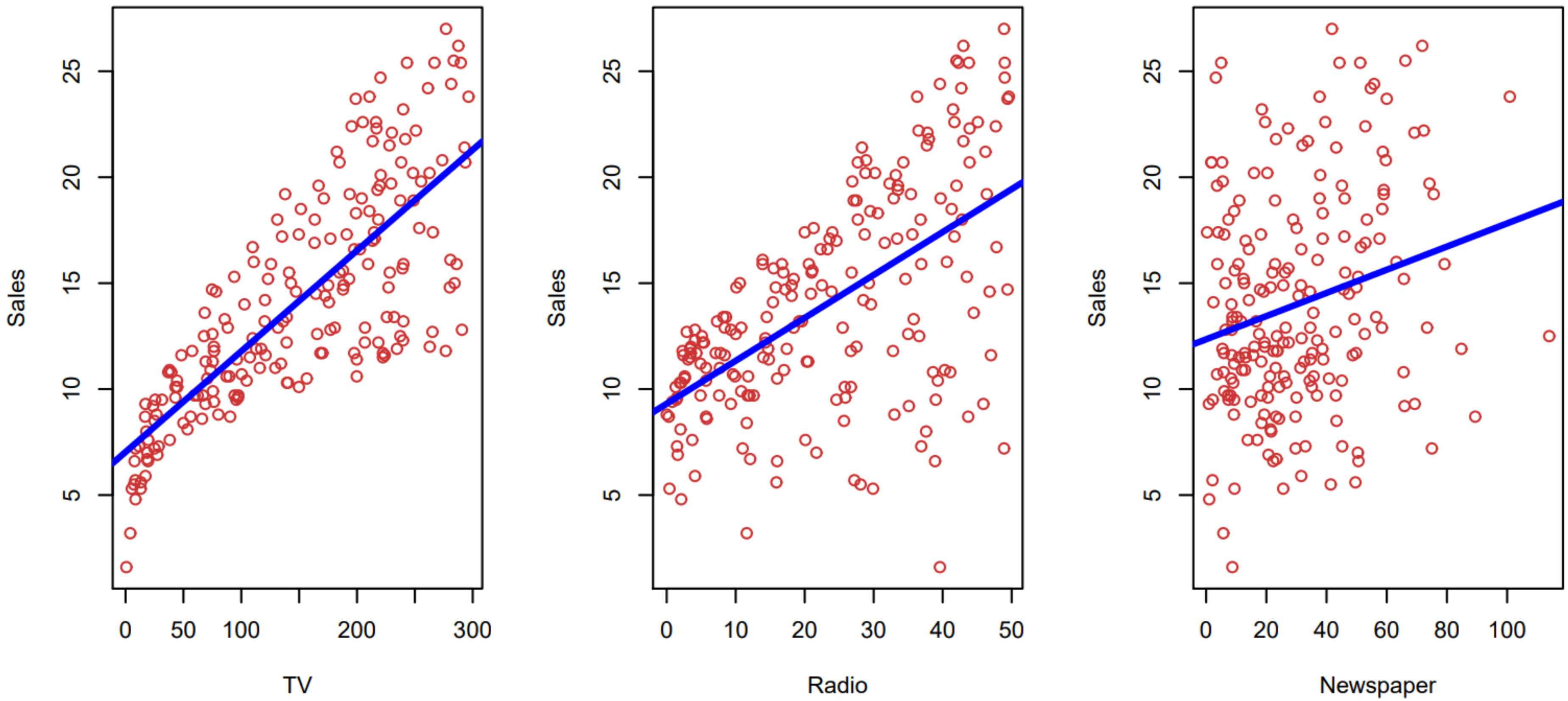
第二章 统计学习

1.什么是统计学习？

统计学习

问题：是否可以用这三种投入来预测销售的结果？



散点图：TV, Radio and Newspaper的投入 vs 销售额

蓝线：线性回归拟合三种结果（linear-regression）

$$\texttt{Sales} \approx f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

标记方法

X：输入变量，特征变量，属性变量，预测变量，自变量，变量。如：电视X1，收音机X2，使用输入向量集合

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

标记方法

Y：输出变量，响应变量, 目标变量，因变量，输出结果等。

如：希望能够预测的销售额，用Y来指代。

模型：预测模型，分类模型等。

$$Y = f(X) + \varepsilon$$

ε：随机误差项，与X独立，均值为0.

f：函数，固定但是未知，f表达了X提供给Y的系统信息

估计f的主要原因：

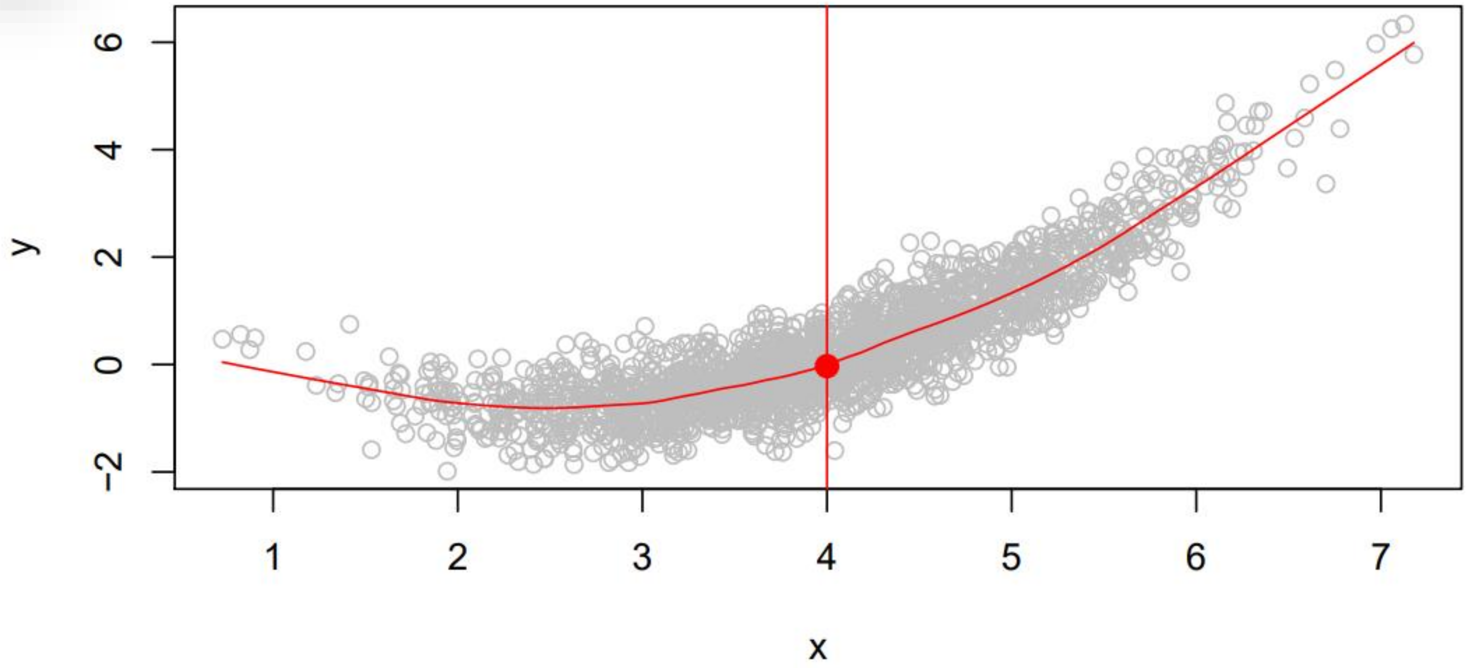
- 有一个好的 f，可以在新点 $X = x$上预测Y。
- 可以理解 $X = (X_1, X_2, \ldots, X_p)$在解释Y时哪些是重要元素，而哪些是不相关的。
- 根据 f 的复杂性，能够理解X的每个分量X_j如何影响Y的。

预测

推断

02

估计*f*



有理想的f (X)吗?

f (X) 在任意选定的X值处的最佳值是多少，比如X = 4？在X = 4处可以有很多Y值。一个好的值是：

$$f(4) = E(Y|X = 4)$$

- $E(Y|X = 4)$表示给定X = 4的Y的期望值(平均值)。
- 理想f (x) = E(Y | X = x)称为回归函数。

02

估计f

对向量X也有定义，如：

$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

对于最优和最理想Y的均方预测误差:

$$f(x) = E(Y \mid X = x)$$

是使函数 g 在所有X= x点上的 $E[(Y - g(X))^2|X = x]$ 最小化的函数。

估计f

对于 $f(x)$ 的任何估计 $\hat{f}(x)$ 有:

$$E[(Y-\hat{f}(X))^2|X=x] = \underbrace{[f(x)-\hat{f}(x)]^2}_{\text{可约误差 reducible error}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{不可约误差 irreducible error}}$$

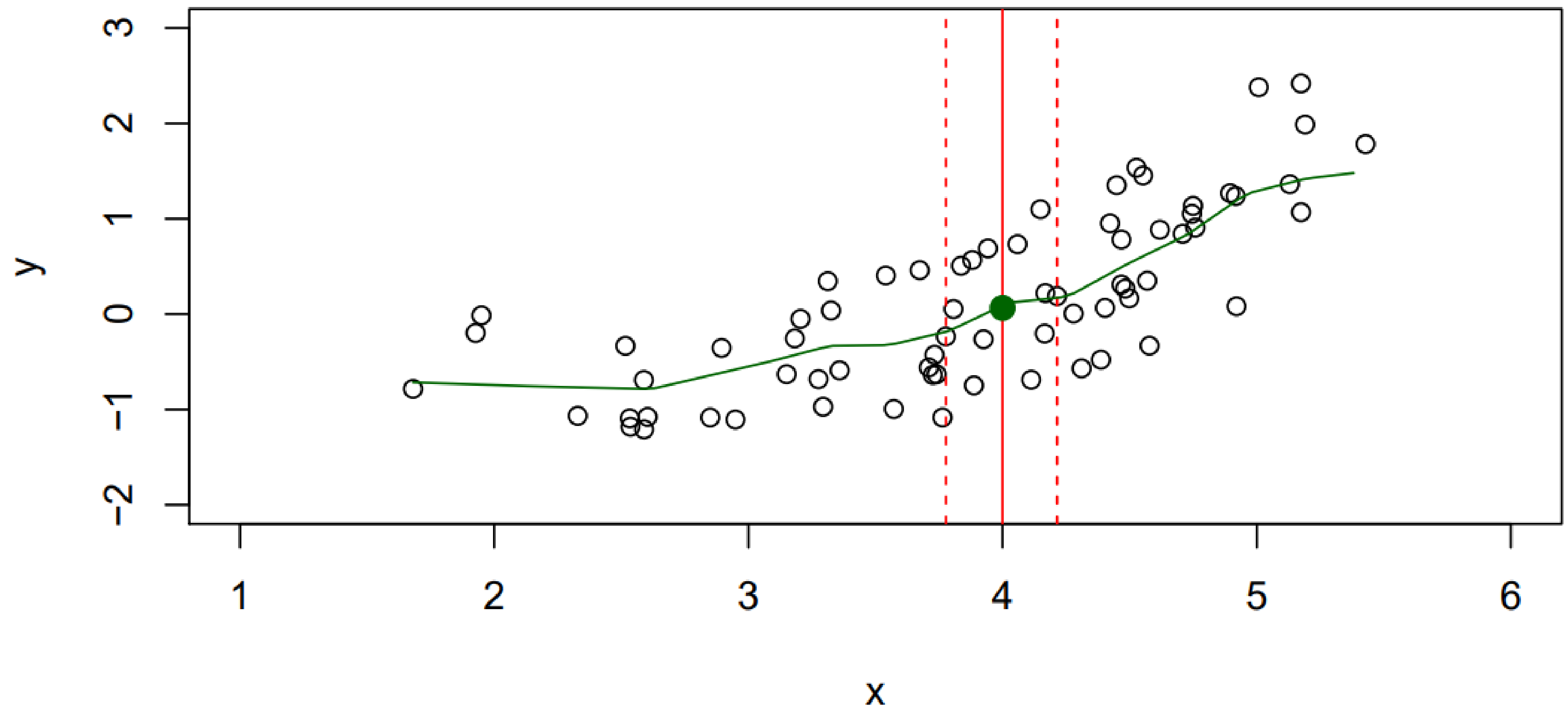
02 估计f

对于 $f(x)$ 的任何估计 $\hat{f}(x)$ 有:

$$E[(Y-\hat{f}(X))^2|X=x] = [f(x)-\hat{f}(x)]^2 + \mathrm{Var}(\epsilon)$$

- $E(Y-\hat{Y})^2$ 代表预测量和实际值Y的均方误差或期望平方误差值，Var(ε)表示误差项ε的方差。
- $\epsilon = Y - f(x)$是不可减少的误差——即使知道$f(x)$，在预测时仍然会出错，因为在每个X= x处，可能的Y值通常都有一个分布。

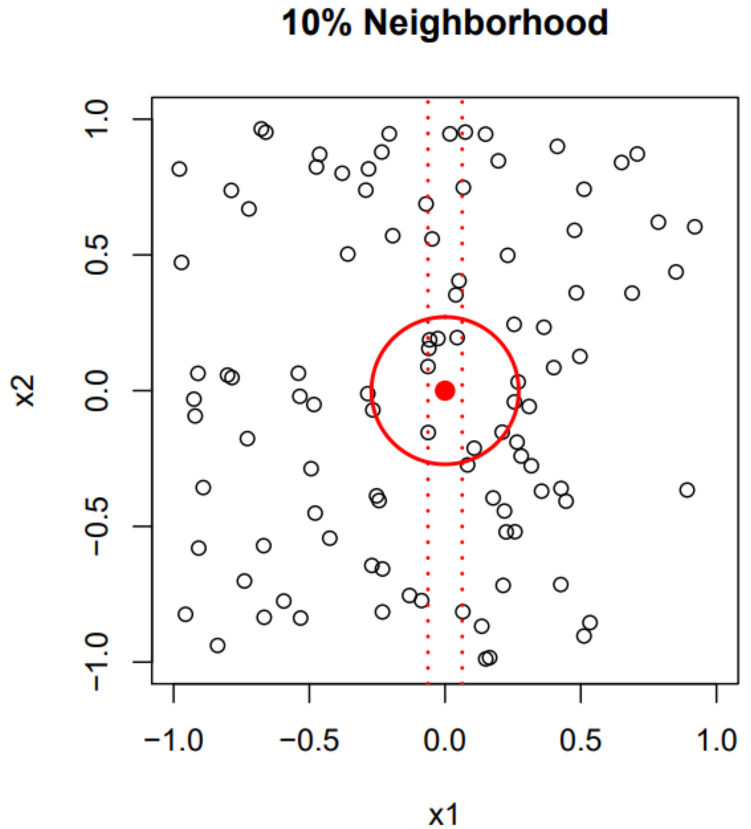
估计f

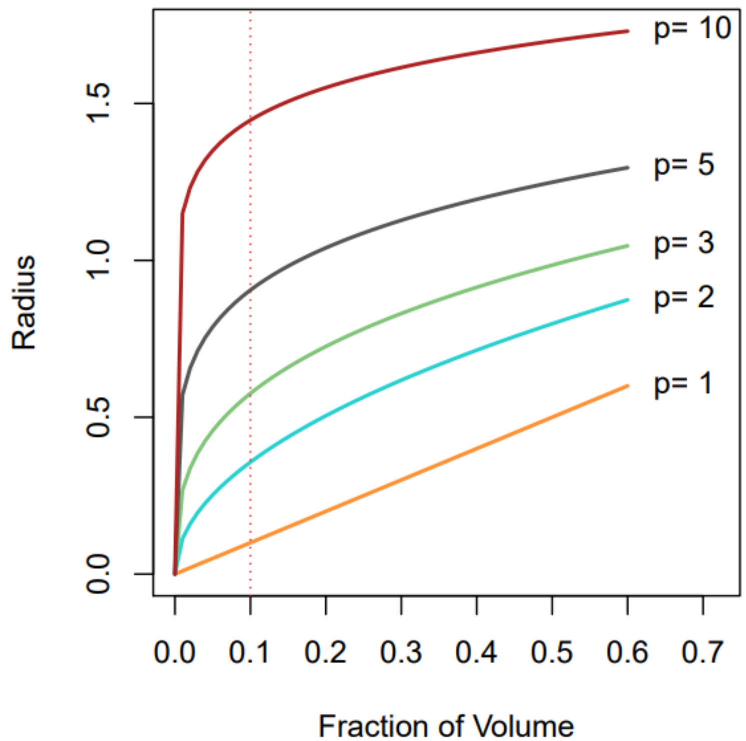


若X = 4的数据点很少，怎么办？

- 放宽定义，设： $\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$
- 其中N (x)是x的邻域。

估计*f*

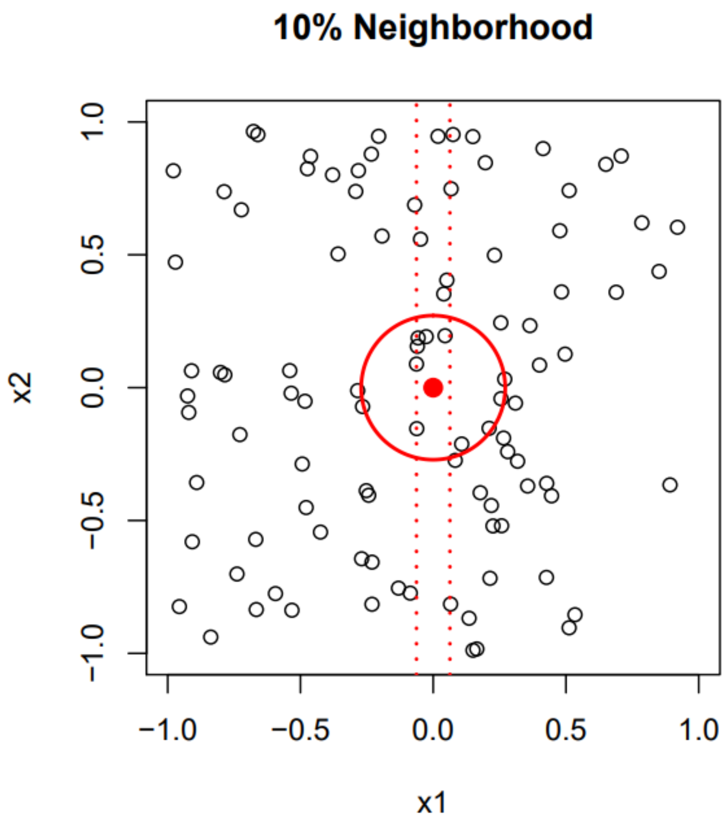


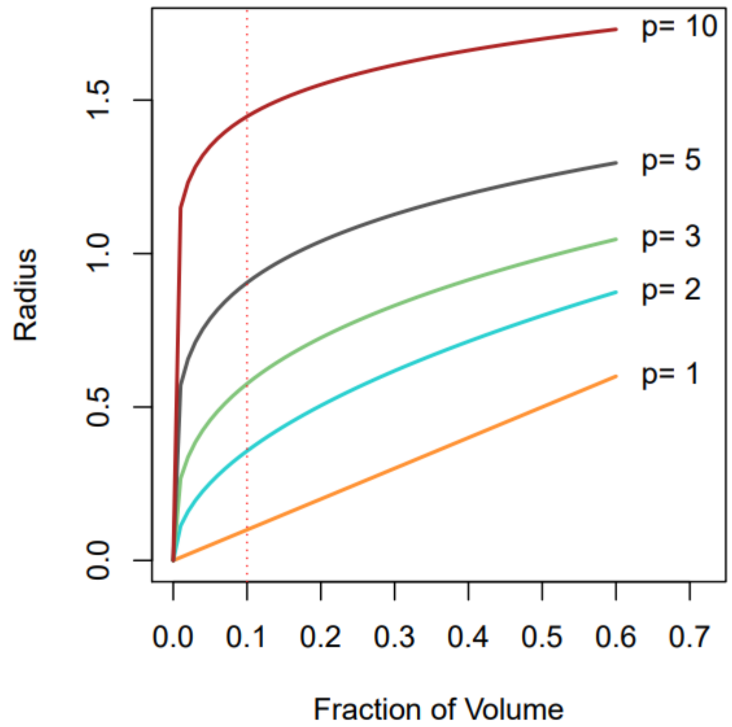


- 对于小的 p，最近邻平均是非常好的，即p≤4和大的N。
- 当 p 很大时，最近邻方法可能很糟糕。

  原因:维数灾难。

估计*f*





在高维空间中，最近的邻居往往离我们很远。

- 需要得到yi 的N值的一个合理的部分来取平均，以降低方差。
- 高维中10%的邻域不再是局部的，因此不能通过局部平均来估计E(Y |X = X)。

03

如何估计?

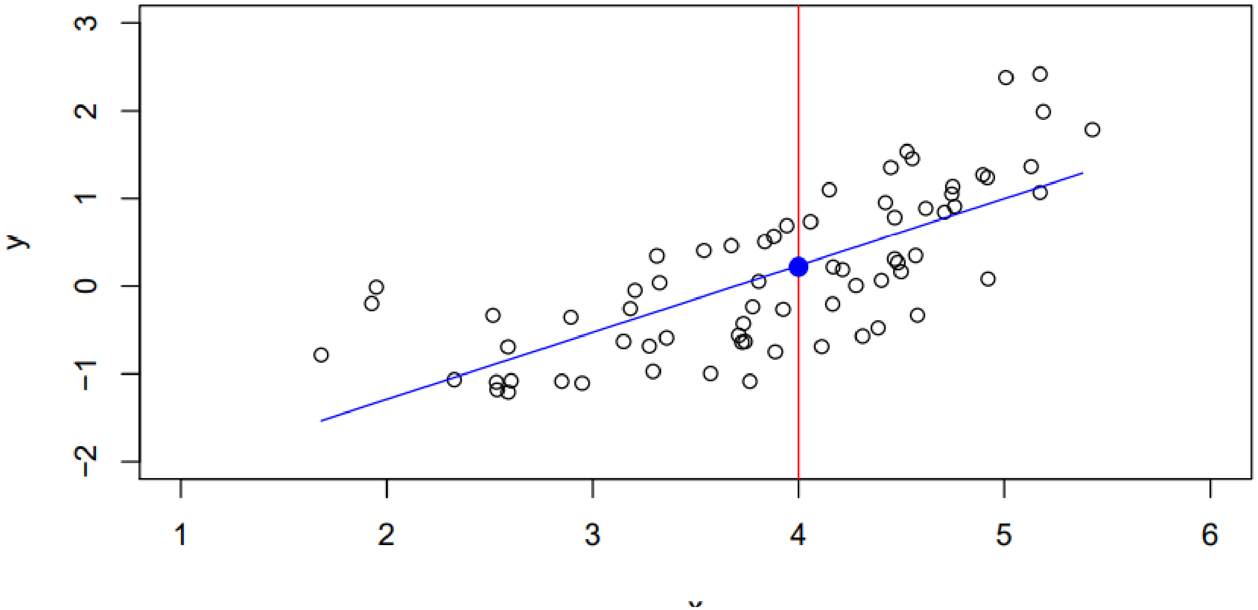
线性模型是参数模型的一个重要例子:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_p X_p$$

线性模型用p + 1参数表示 $\beta_0, \beta_1, \dots, \beta_p$

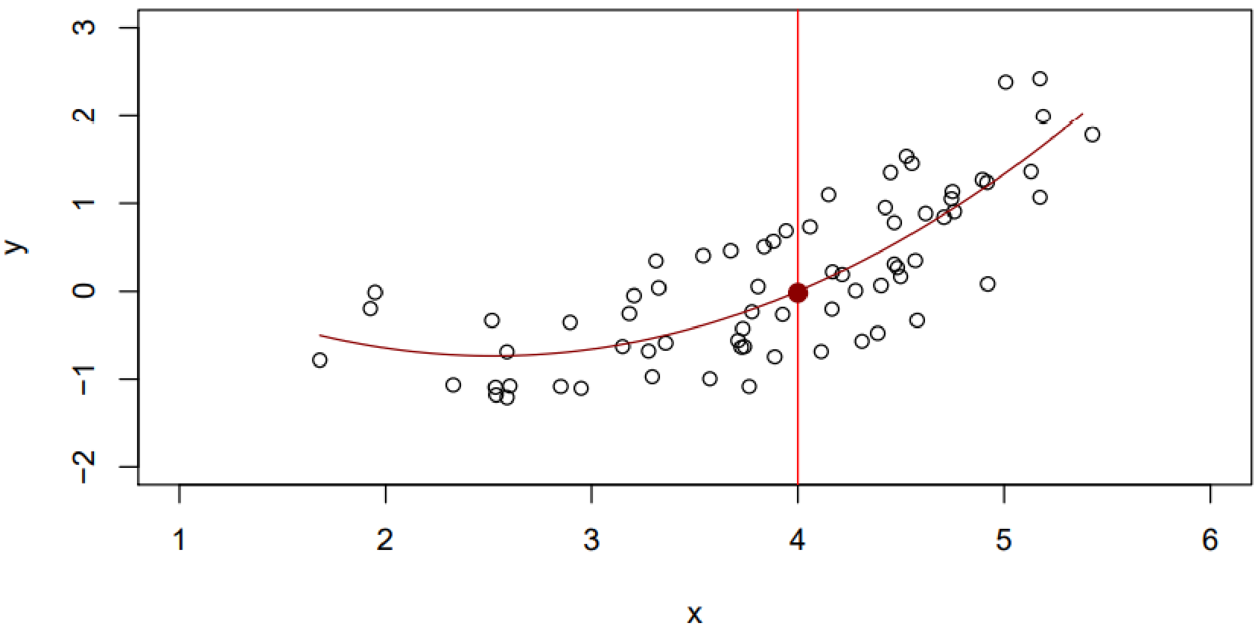
- 参数方法：估计参数，采用训练数据来拟合模型。
- 虽然线性模型常常都不是正确的，但它通常是未知真函数 *f*(X) 的一个良好的、可解释的逼近。

03 如何估计?



线性模型

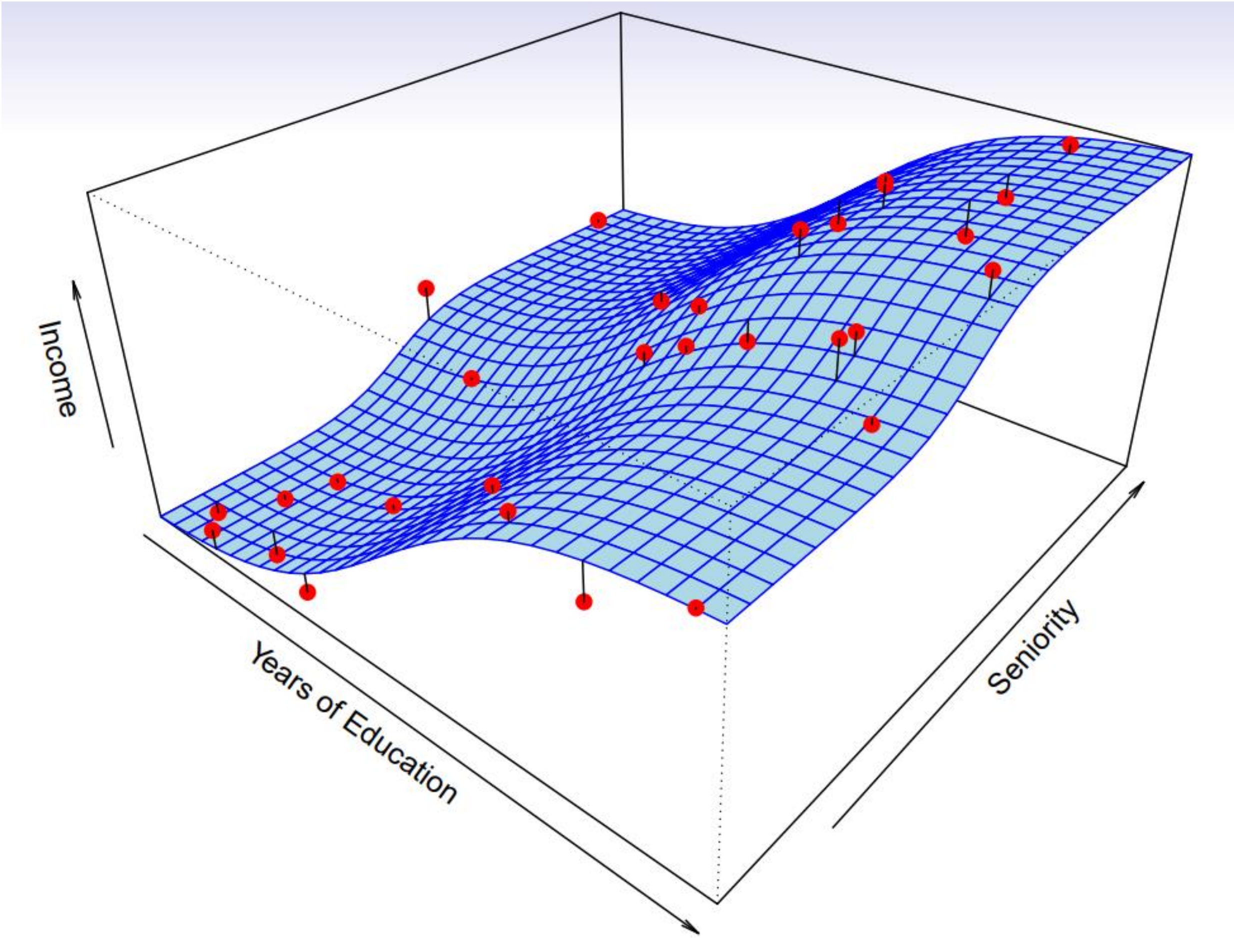
$$\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$



二次方程

$$\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$

如何估计?

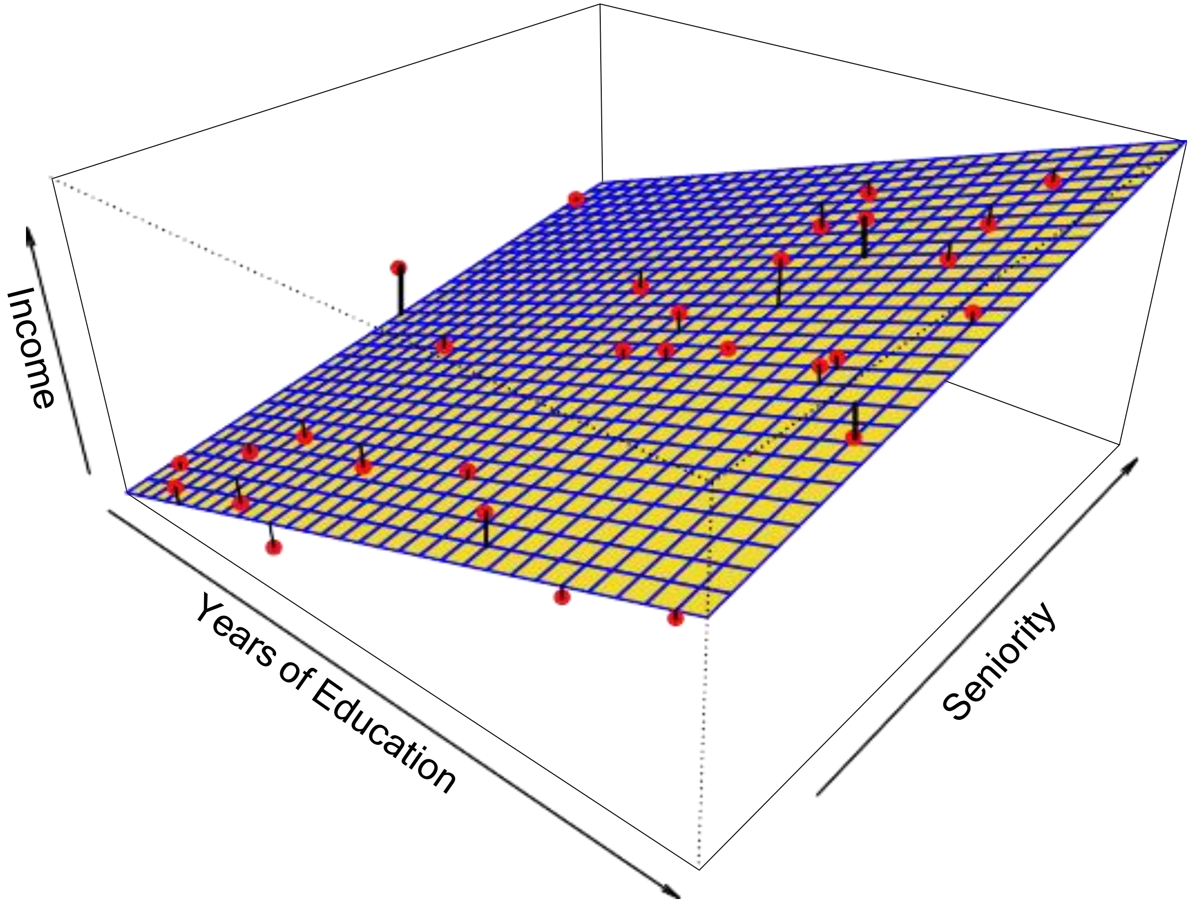


红点是模型所得的模拟值，f 是蓝色的曲面

$$\mathtt{income} = f(\mathtt{education}, \mathtt{seniority}) + \epsilon$$

03

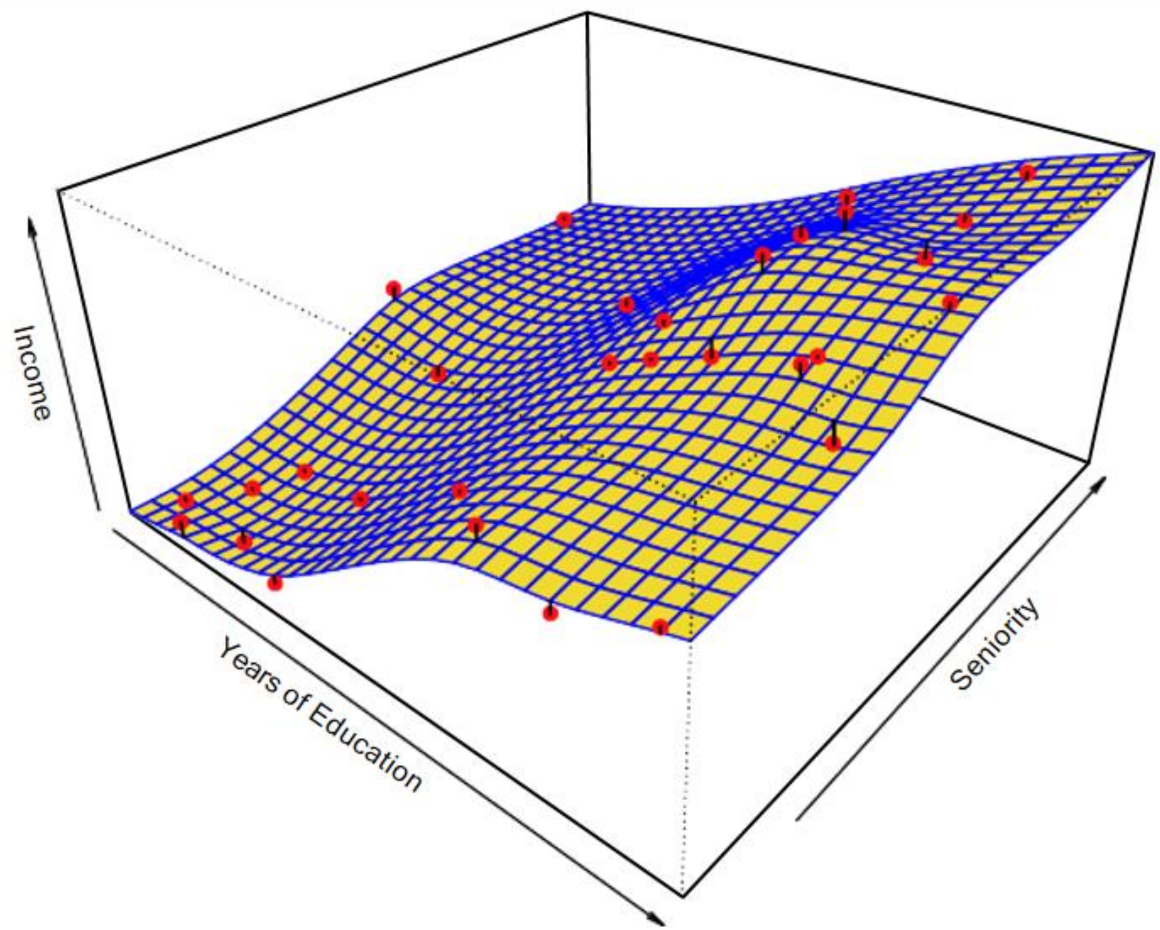
如何估计?



进行最小二乘线性回归模型拟合

$$\hat{f}_L(\texttt{education}, \texttt{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{education} + \hat{\beta}_2 \times \texttt{seniority}$$

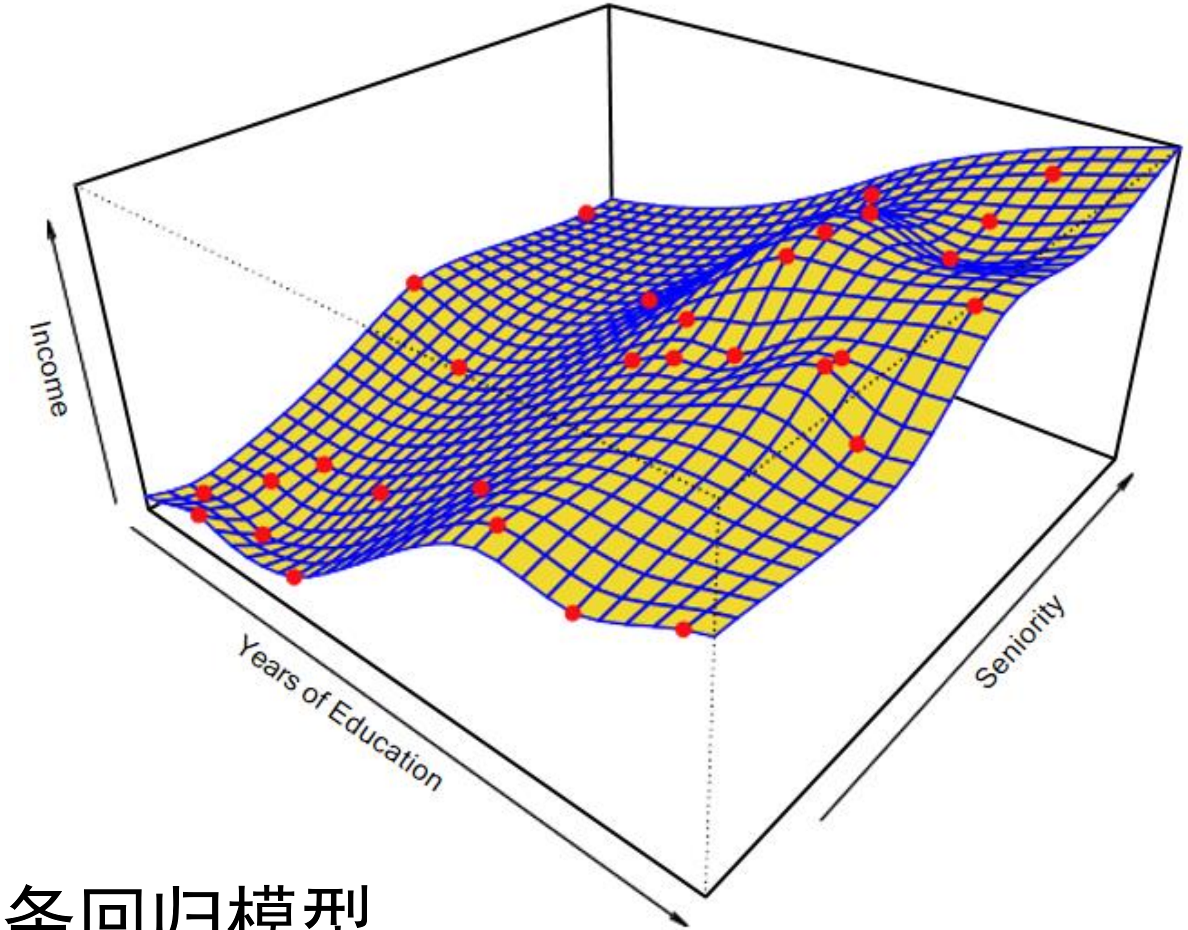
如何估计?



使用光滑薄板样条的技术拟合

$$\hat{f}_S(\texttt{education}, \texttt{seniority})$$

如何估计?



更灵活的样条回归模型

$$\hat{f}_S(\texttt{education}, \texttt{seniority})$$

这里拟合模型在训练数据上没有误差!也被称为过拟合。

- 预测精度VS可解释性。

　　线性模型易于解释;薄板样条则不然。

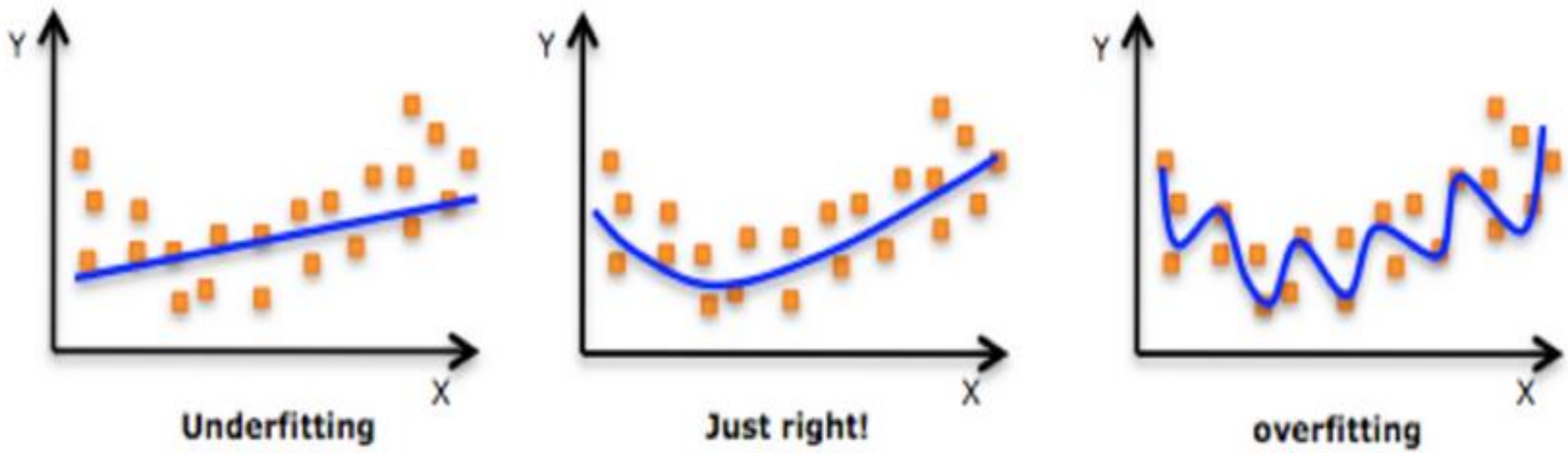
- 好的拟合VS欠拟合与过拟合。

　　怎么知道什么时候合适?

- 简单VS黑盒（black-box）。

　　通常更喜欢包含更少变量的简单模型，而不是包含所有变量的黑箱预测器。

预测精度与模型可解释性的权衡

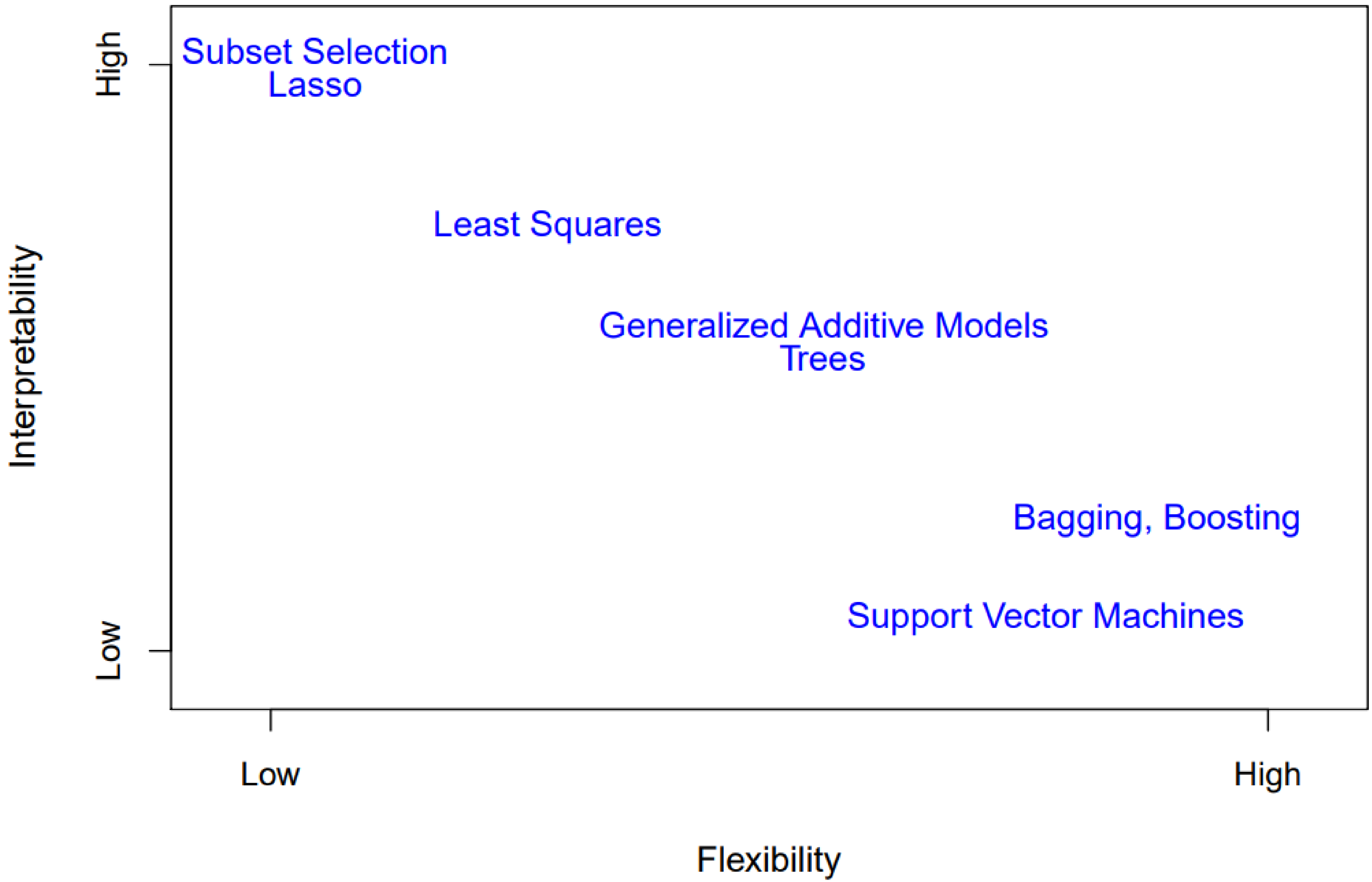


$$\theta_0 + \theta_1 x$$

$$\theta_0 + \theta_1 x + \theta_2 x^2$$

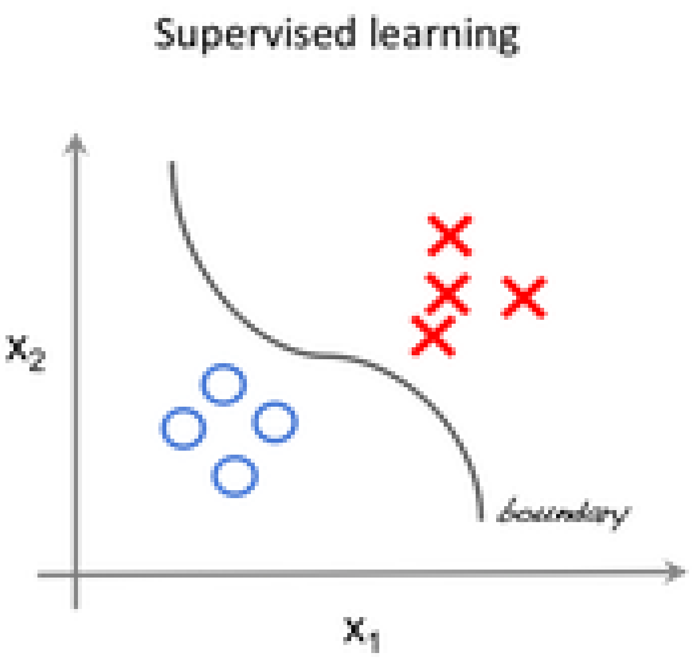
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

预测精度与模型可解释性的权衡



指导学习（supervised learning）

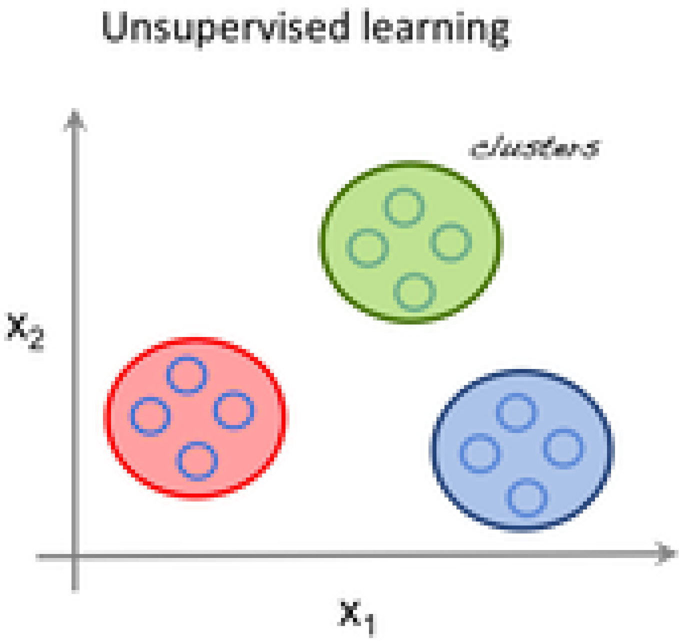
对每一个预测变量观测值$x_i(i=1,\cdots,n)$都有相应的响应变量的观测y_i。



- 线性回归
- 逻辑斯谛回归
- 广义可加模型（GAM）
- 支持向量机（SVM）
- ……

无指导学习（unsupervised learning）

只有预测变量的观测值$x_i(i=1,\cdots,n)$，这些向量没有相应的响应变量y_i与之对应。



典型统计学习工具：
聚类分析
（cluster analysis）

半指导学习（semi supervised learning）

- 假设有n个观测，其中m（m<n）个观测点可同时观测到预测变量与响应变量，而其余n-m个观测点，只能观测到预测变量但无法观测到响应变量

- 这类统计方法就希望能够既用到m个观测点的预测变量和响应变量的信息，同时又包含了n-m个不能获取响应变量观测值的信息。

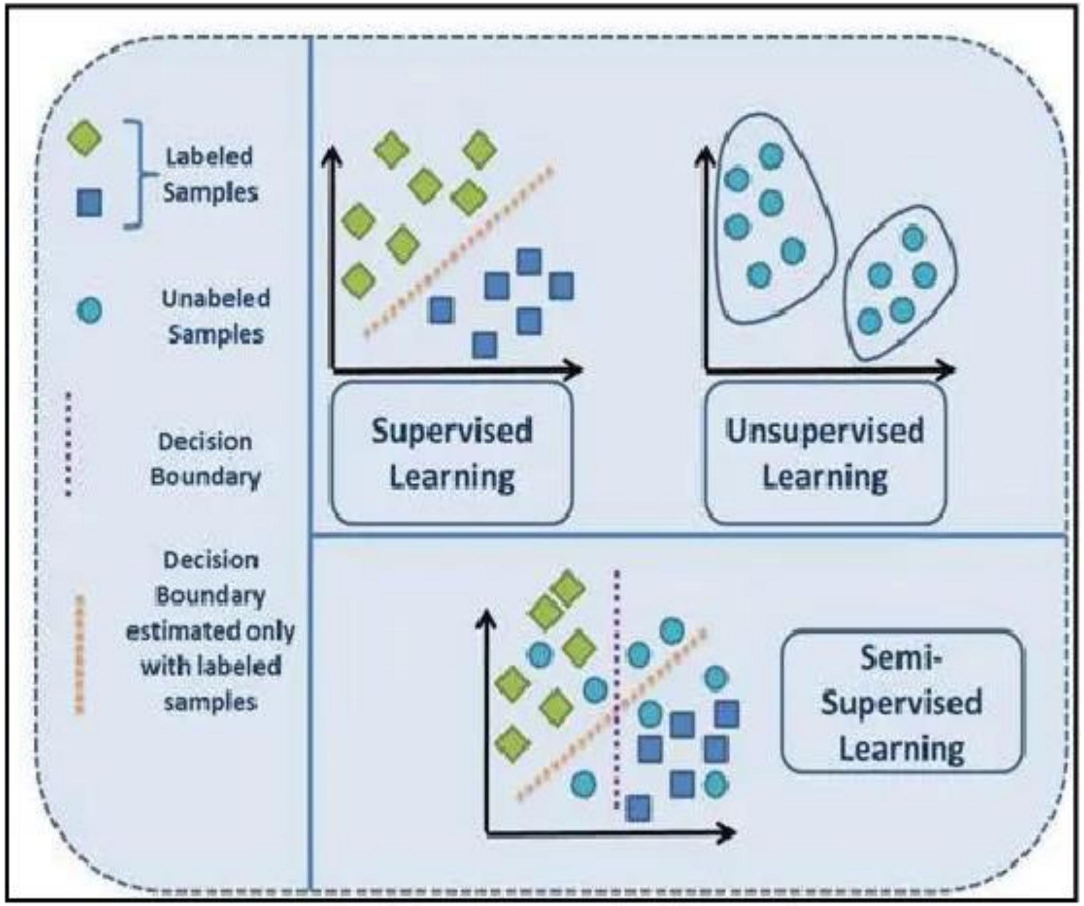


Fig. 1Semi-Supervised Learning

回归与分类问题

变量

定量

呈数值型，如：年龄身高收入房屋的价值以及股票的价格

定性（分类变量）

取K个不同类的其中一个值，如：性别、品牌等

回归问题分析问题：将响应变量为定量的问题。

分类问题：具有定性响应变量的问题

回归与分类问题并不绝对！

2. 评价模型精度

在回归中最常用的评价准则：

均方误差（mean squared error，MSE）

假设我们对训练数据 $\mathsf{Tr} = \{x_i, y_i\}_1^N$ 拟合一个模型 $\hat{f}(x)$ ：

我们可以计算Tr上的训练均方误差:

$$\mathrm{MSE}_{\mathsf{Tr}} = \mathrm{Ave}_{i \in \mathsf{Tr}}[y_i - \hat{f}(x_i)]^2$$

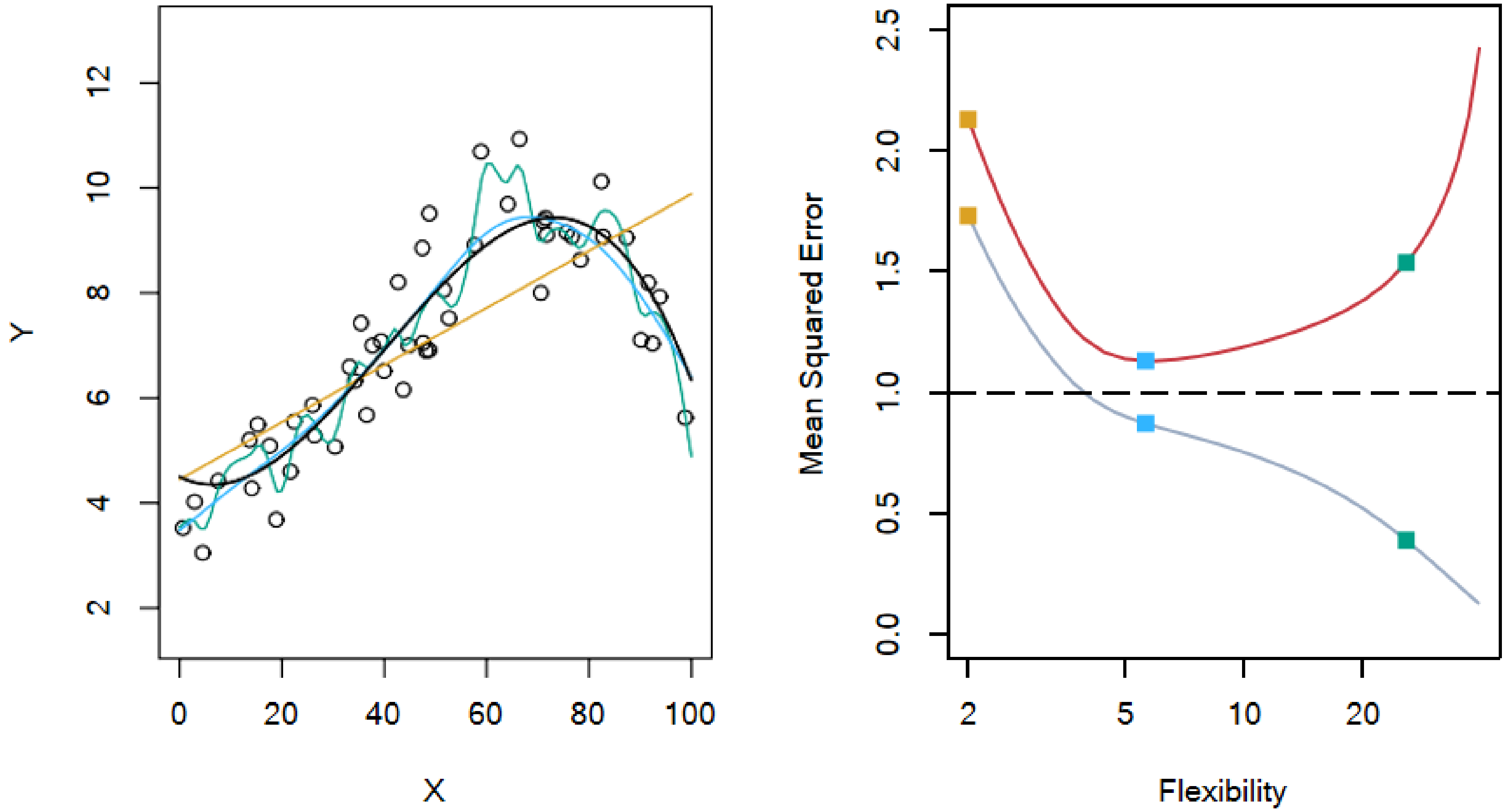
这可能会偏向于更过拟合的模型。

使用新的测试数据 $\mathsf{Te} = \{x_i, y_i\}_1^M$ 计算：

$$\mathrm{MSE}_{\mathsf{Te}} = \mathrm{Ave}_{i \in \mathsf{Te}}[y_i - \hat{f}(x_i)]^2$$

01

拟合效果检验



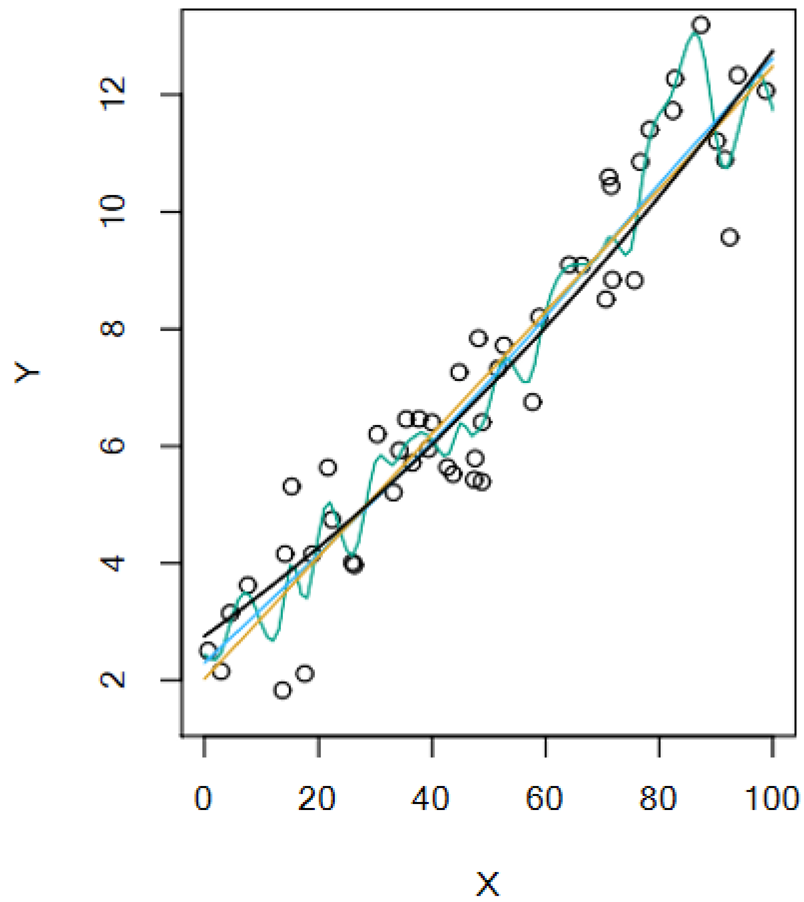
三种 f 的估计：

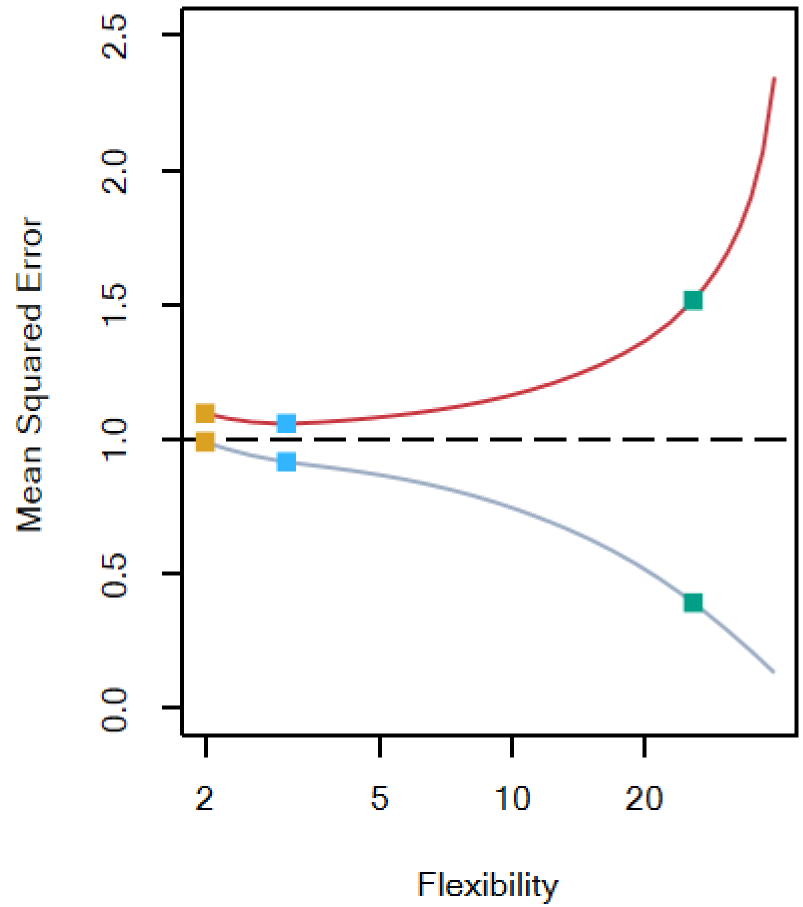
- 线性回归（橙色曲线）
- 两条光滑样条拟合（绿色和蓝色曲线）

左：真实函数 f 模拟产生的数据，黑色曲线表示。

右：训练均方误差（灰色曲线），测试均方误差（红色曲线），所有方法都已使测试均方误差尽可能小。

01 拟合效果检验

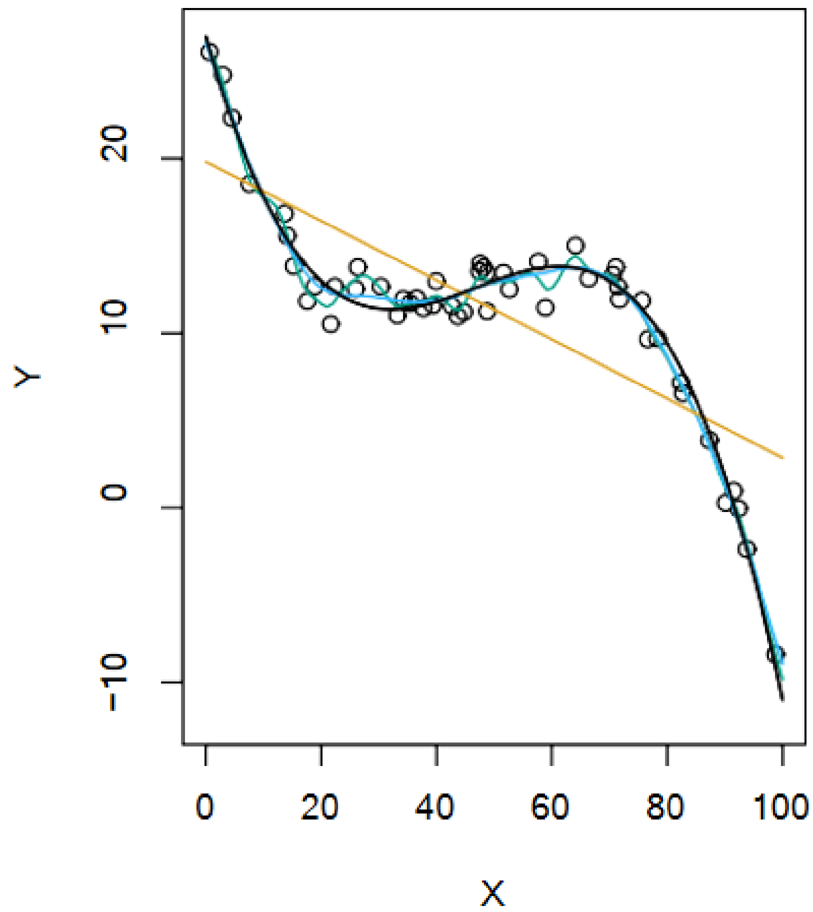


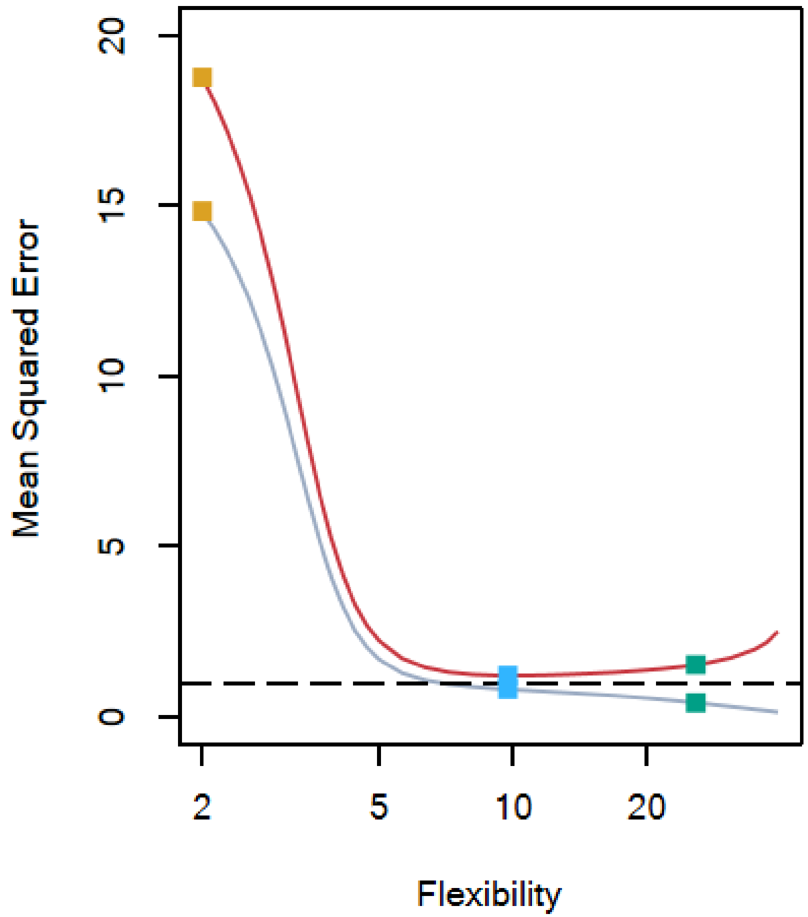


这里真实的 f 平滑，

所以更平滑的拟合和线性模型拟合的较好。

01 拟合效果检验





这里的真实的 f 是摇摆的和低噪声，
所以更灵活的拟合做得更好。

偏差-方差权衡

假设我们对某些训练数据Tr拟合一个模型 $\hat{f}(x)$ ，并让(x_0, y_0)是从总体中提取的一个测试观察值。如果真正的模型是：

$$Y = f(X) + \epsilon \ (\text{with } f(x) = E(Y|X = x)),$$

期望测试均方误差：

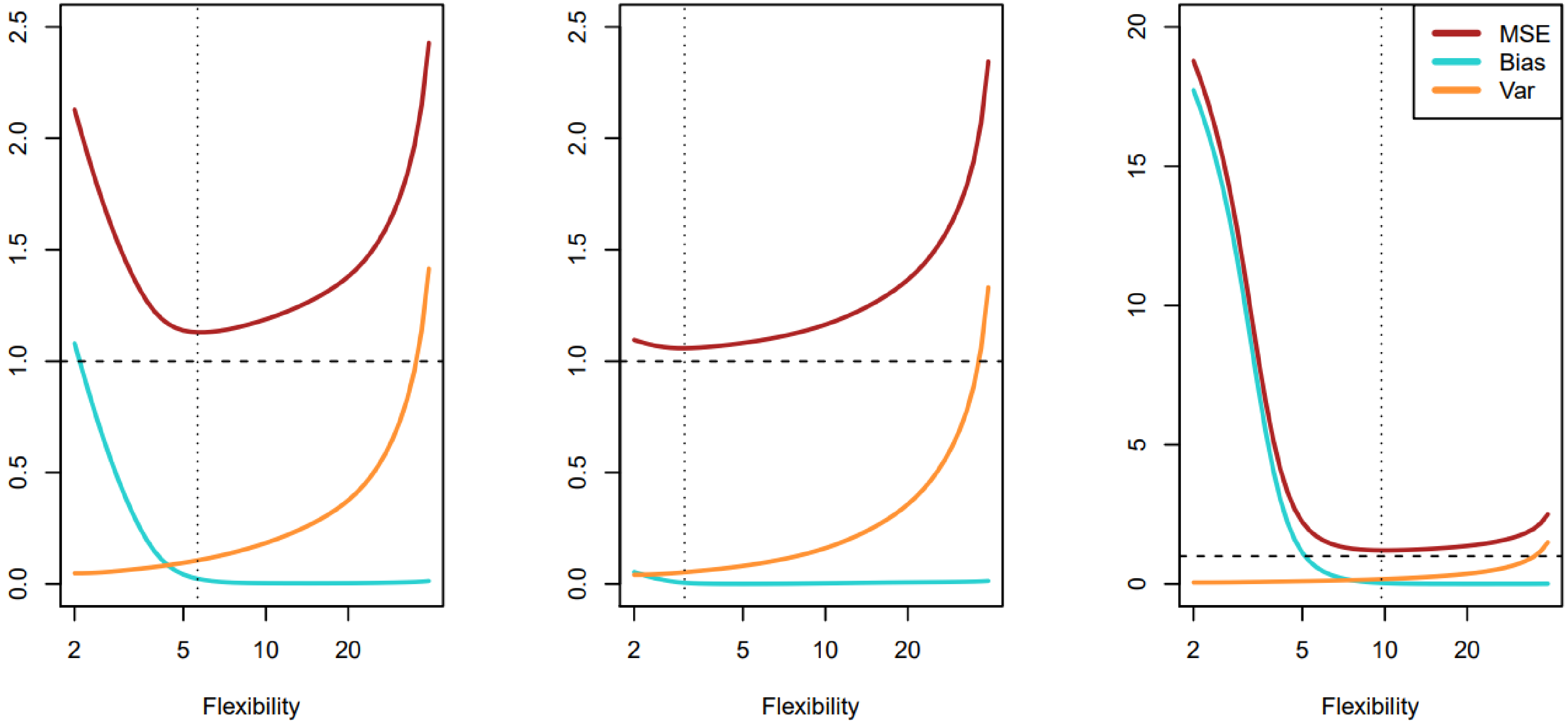
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

即偏差（bias）为：

$$\mathrm{Bias}(\hat{f}(x_0))] = E[\hat{f}(x_0)] - f(x_0)$$

通常，随着 $\hat{f}(x)$ 灵活性的增加，其方差增加，偏差减少。因此，基于平均测试误差来选择灵活性就相当于一种偏差-方差权衡。

02 偏差-方差权衡



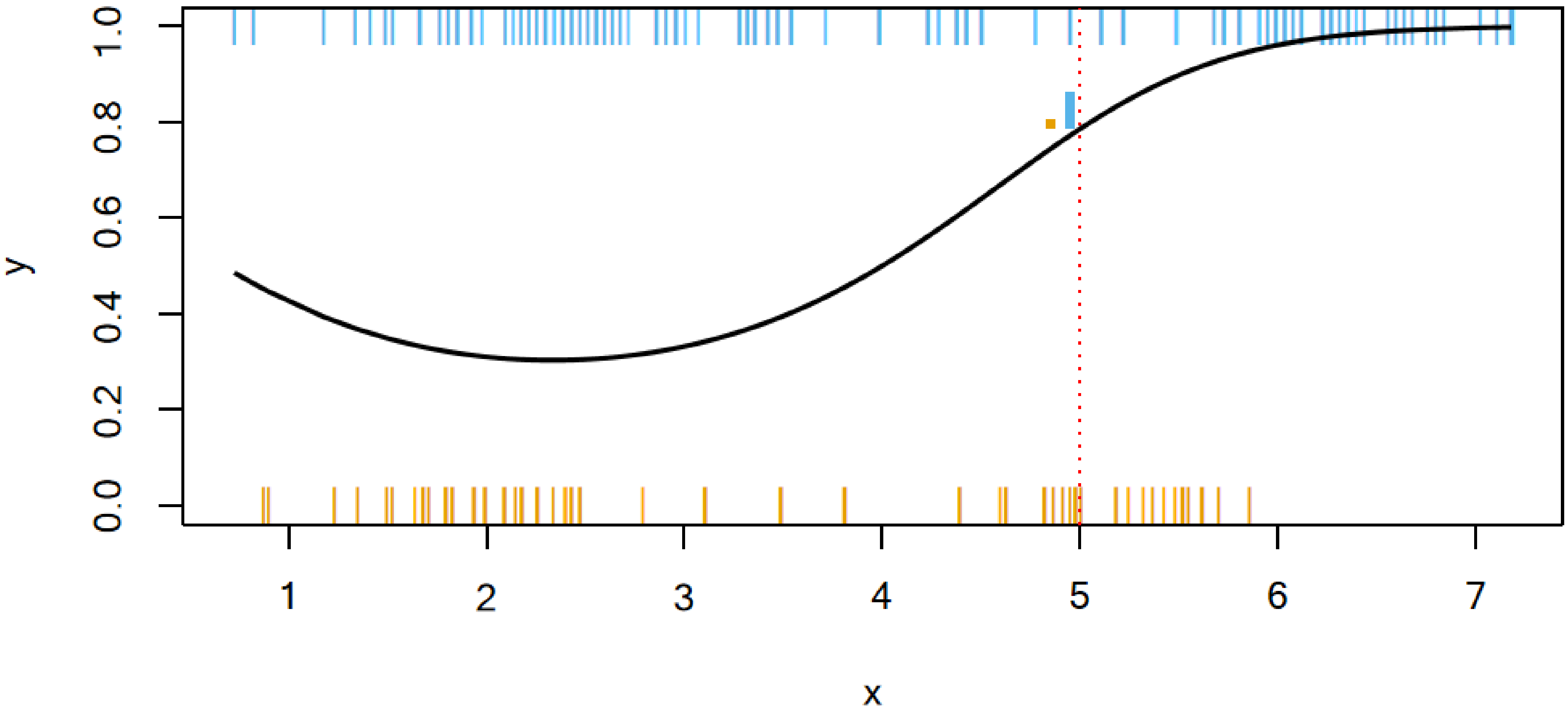
分别表示上图三个数据集的平方偏差（蓝色曲线）、方差（橙色曲线）、不可约误差（虚线）、测试均方误差（红色曲线）。垂直的点线表示最小测试均方误差所对应的光滑度。

分类模型

分类模型中的响应变量Y是定性的，例如，电子邮件是C = (spam, ham) (ham=好的电子邮件)中之一，数字类别是C ={0,1，…9}。目标是:

- 构建一个分类器C(X)，将C的类标签分配给未来未标记的观察X。
- 评估每种分类模型的不确定性。
- 了解不同预测因素之间的作用$X = (X_1, X_2, \ldots X_p)$。

有理想的C(X)吗?



贝叶斯分类器

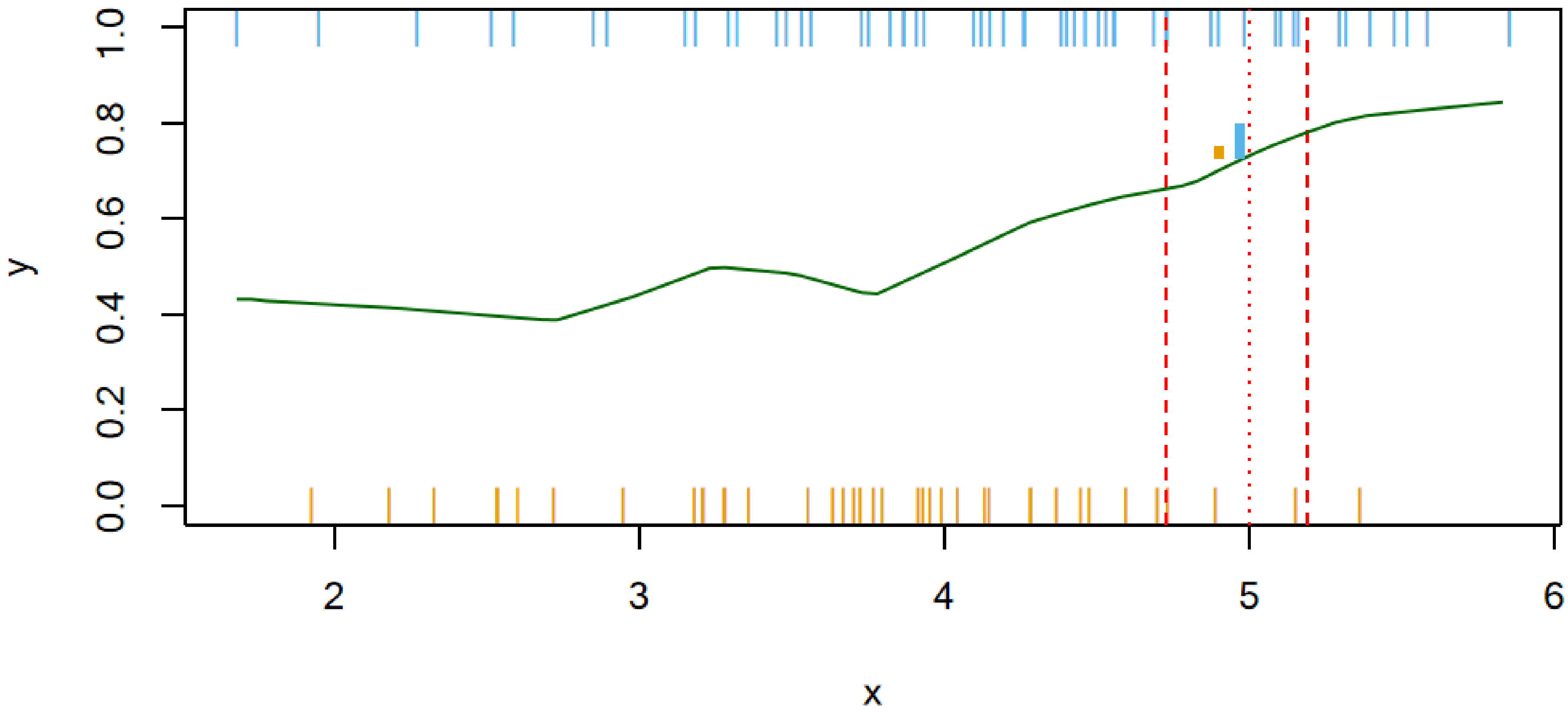
假设C中的K个元素编号为1,2，…, k .让

$$p_k(x) = \Pr(Y = k|X = x),\ k = 1, 2, \ldots, K$$

这些是在x点的条件概率。例如，见x = 5处的条形图。那么x处的贝叶斯分类器为：

$$C(x) = j \text{ if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

k最近邻



最近邻平均法可与之前一样使用，也会随着维度的增长而分解。但对 $\hat{C}(x)$ 的影响小于对 $\hat{p}_k(x),\ k=1,\ldots,K$ 。

03

分类模型

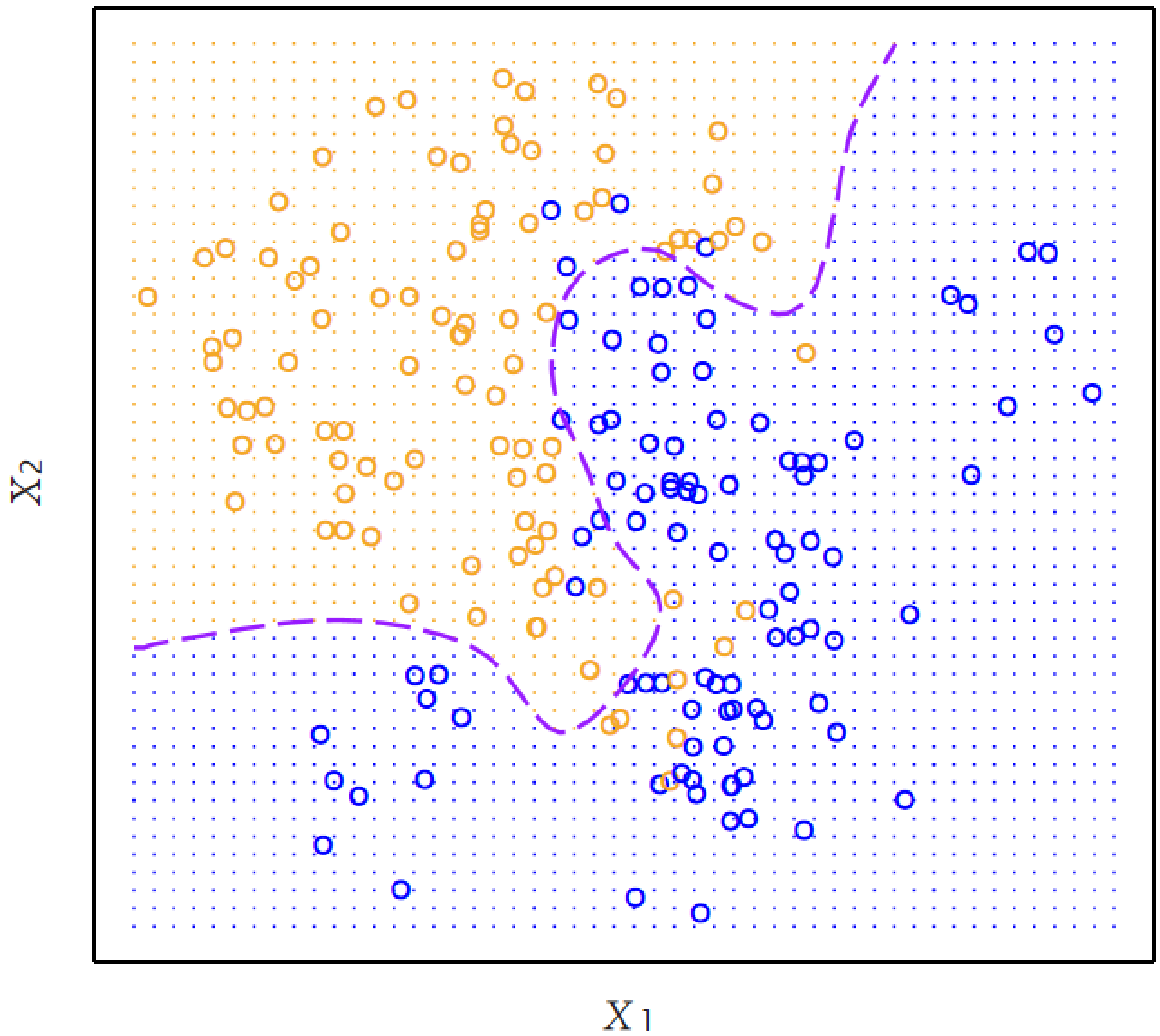
$$C(x) = j \text{ if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$

- 通常我们用误分类错误率来衡量 $\hat{C}(x)$ 的性能:

$$\text{Err}_{\mathsf{Te}} = \text{Ave}_{i \in \mathsf{Te}} I[y_i \neq \hat{C}(x_i)]$$

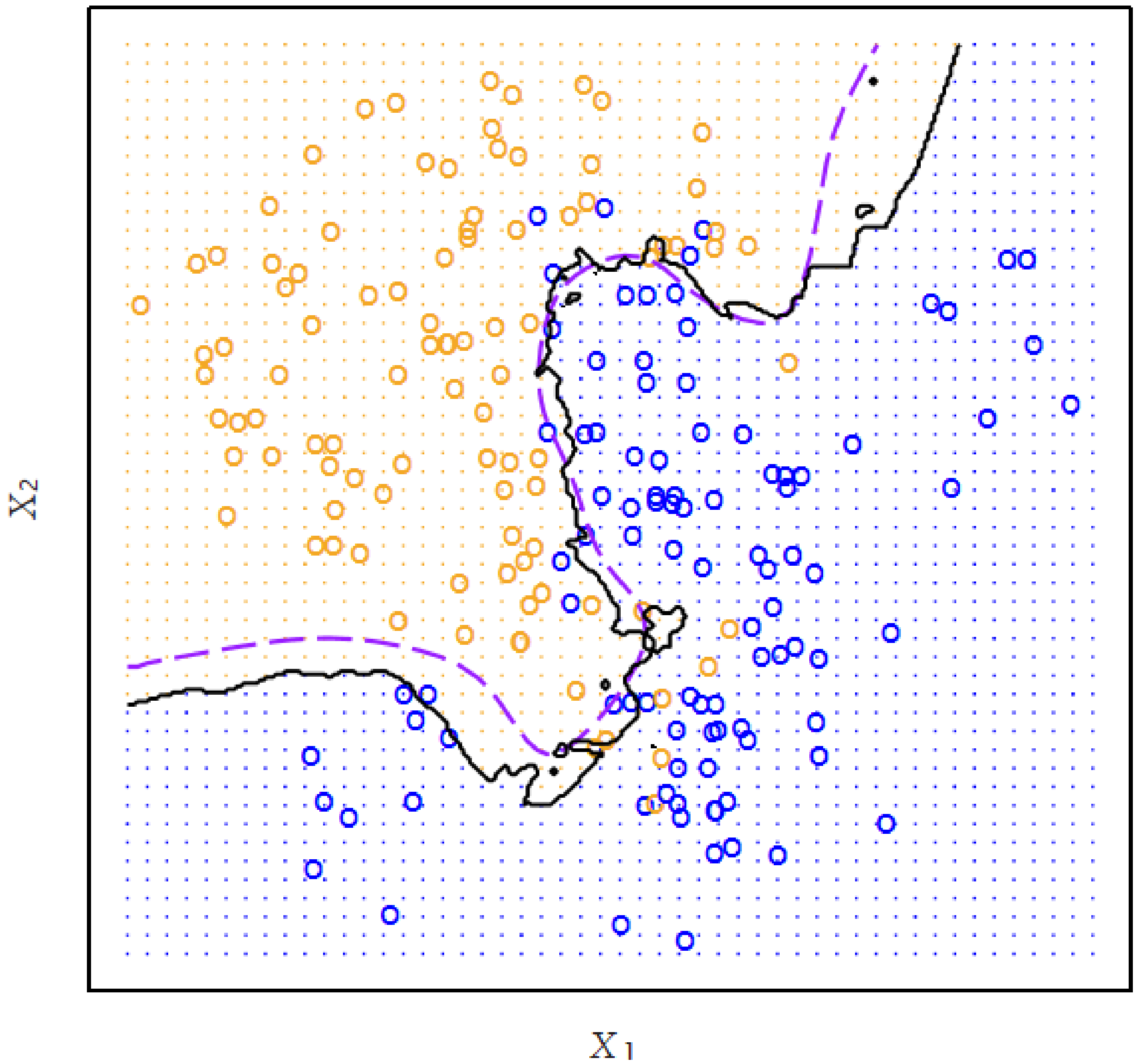
- 贝叶斯分类器(使用真实的$p_k(x)$)的误差最小(在总体中)。
- 支持向量机为C(x)构建结构化模型。
- 我们还将构建结构化模型来表示$p_k(x)$。例如逻辑回归，广义可加性模型。

分类模型

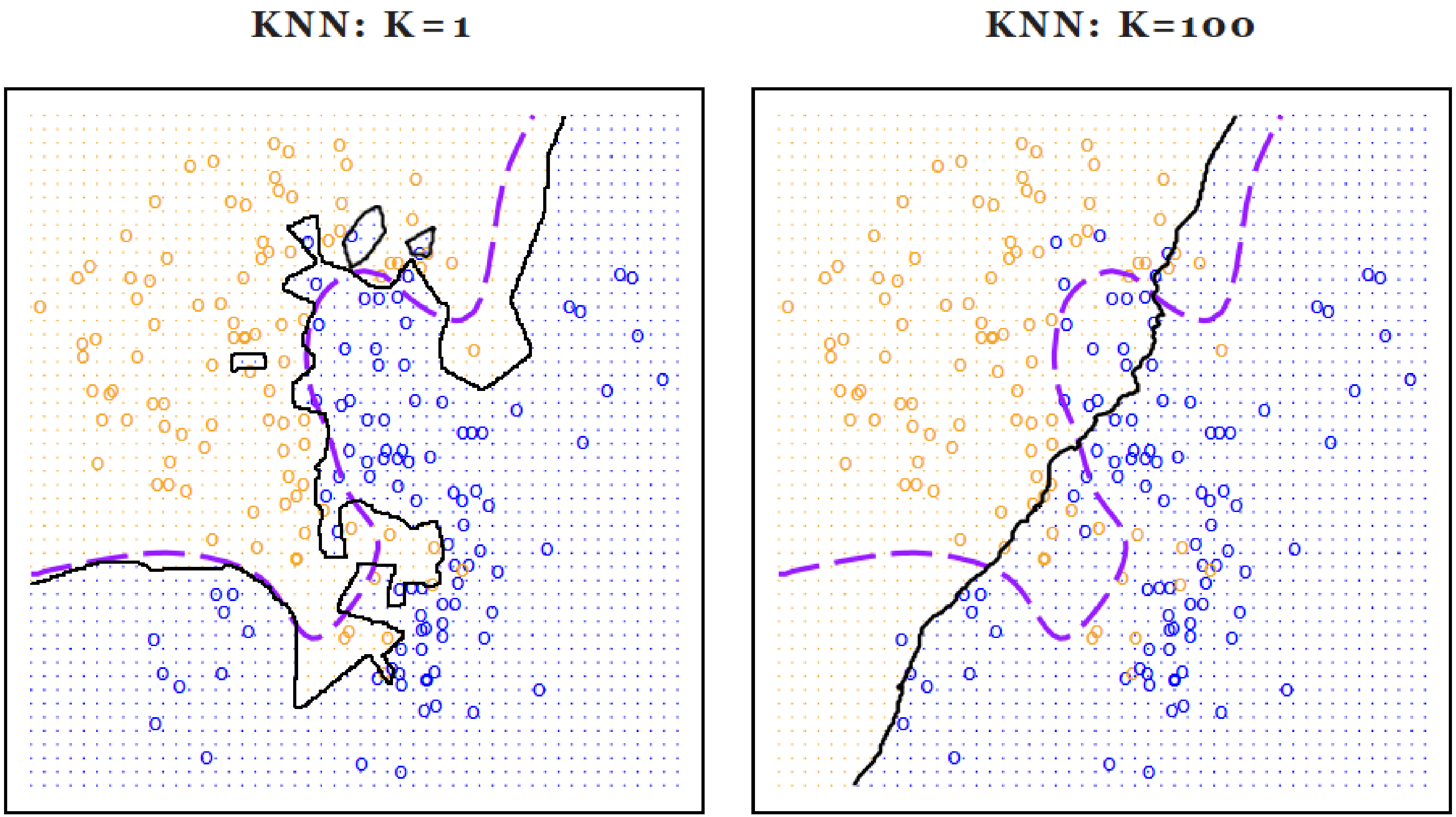


- 从二分类数据中的每一类中抽取100个观测值组成一个模拟数据集。
- 观测点由蓝色和橙色表示。紫色的虚线表示贝叶斯决赛边界。

KNN: K=10

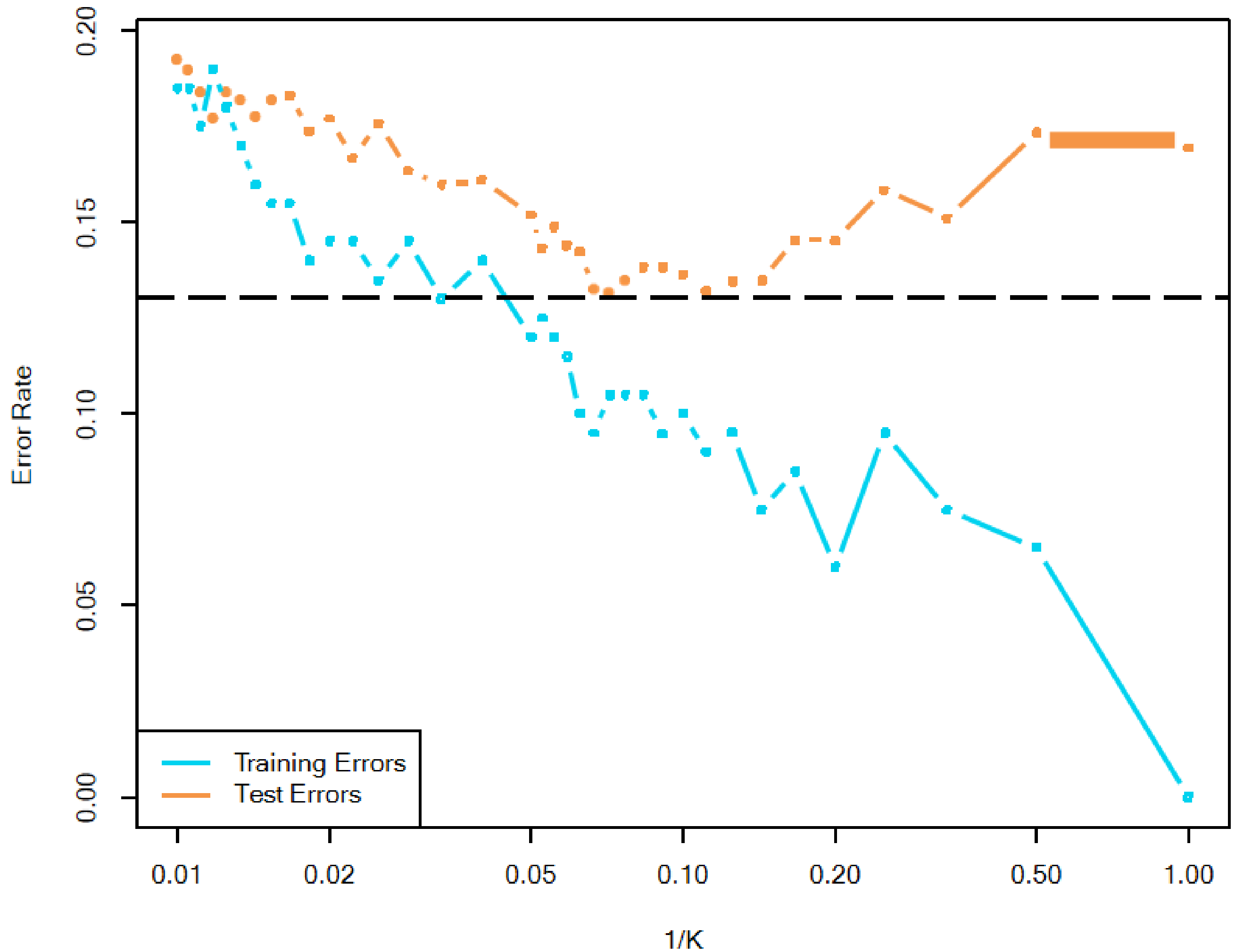


黑色曲线表示KNN方法用于上图数据生成的决策边界，这里k=10.



在K=2和K=100两种设置下KNN决策边界的比较：

- 当K=1时，决策边界相当不规则；
- 当K=100时，模型光滑度下降。紫色虚线为贝叶斯决策边界。



用上述数据所生成的KNN分类器训练错误率（蓝色，200观测点），和测试错误率（橙色，5000观测点）对模型光滑度的变化曲线。