线性回归

李 波

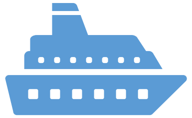
目录

线性回归

- 学习成绩y与学生投入学习时间x的关系

  - ✓ 一般投入时间越长，学习成绩越好。
  - ✓ 投入时间为0，也会有一个成绩。
  - ✓ 关系为 $y = ax + b$，其中a，b为参数。

- 机器寿命y与机器的m个零件质量$x_1, x_2, \cdots, x_m$之间关系



  - ✓ 一般零件质量越好，机器寿命越长。
  - ✓ 关系为$y = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_m x_m + b$。

线性回归

训练数据包含n个数据或样本，每个样本包含输入特征和标签。所有训练数据可以表示为

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

简单线性回归

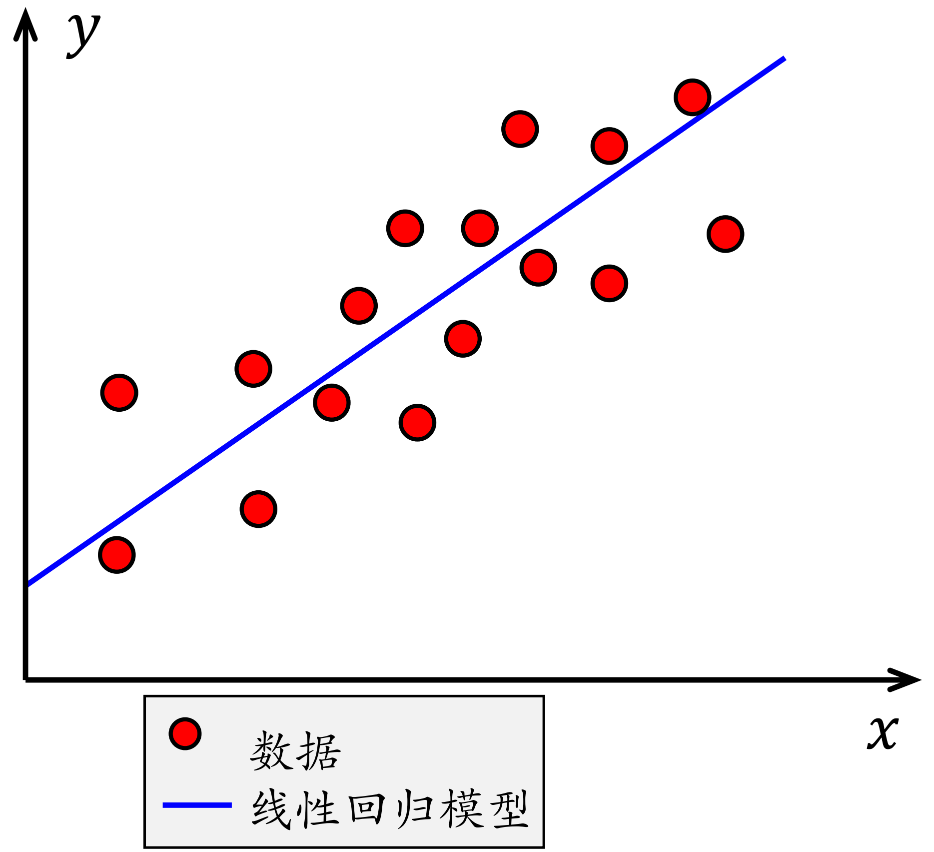
- 每个样本的输入只有一个特征
- 模型为

$$y = \omega x + b$$

ω为权重系数

b为偏置参数

ω和b为模型参数

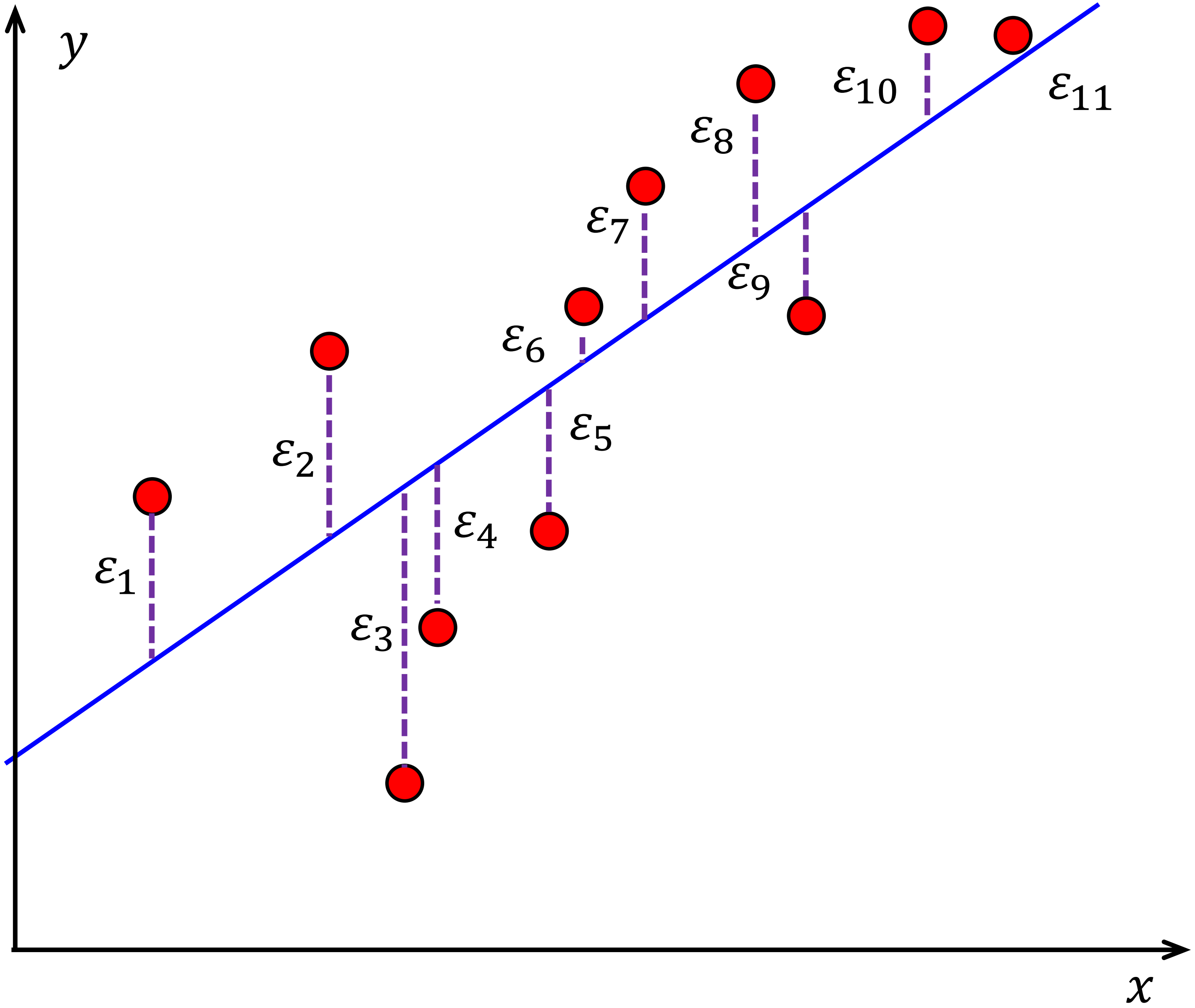


简单线性回归

- 训练数据为 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$
- 模型为 $y=\omega x+b$
- 每个样本不严格落在直线上，纵向误差为：

$$\varepsilon_1 = y_1-(\omega x_1+b)$$
$$\varepsilon_2 = y_2-(\omega x_2+b)$$
$$\cdots$$
$$\varepsilon_n = y_n-(\omega x_n+b)$$

其中ε_i为第i个样本的模型误差。



简单线性回归

- 什么是“好”的拟合直线？ 误差最小
- 需要解决如下优化问题：

$$\min_{\omega,b} \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2$$

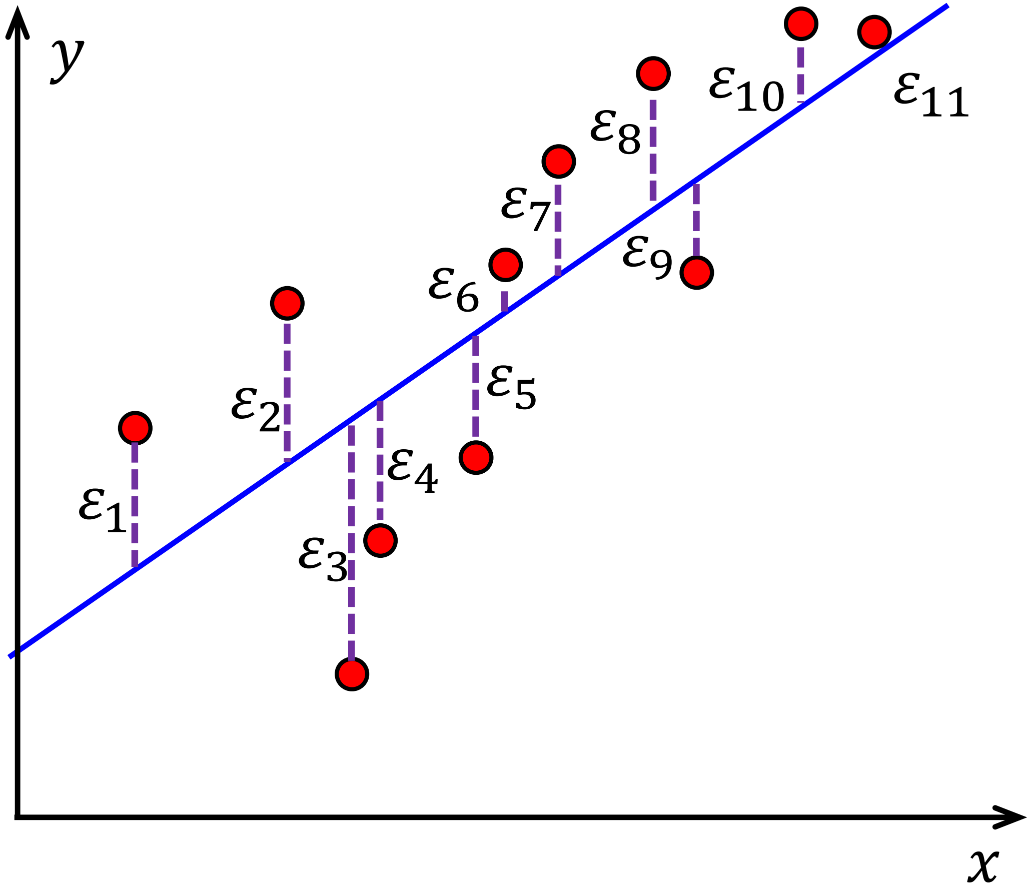
或 $$\min_{\omega,b} \sum_{i=1}^{n} \varepsilon_i^2$$

- 因为$\varepsilon_i = y_i - \omega x_i - b$，所以优化问题可以写为

$$\min_{\omega,b} \sum_{i=1}^{n} (y_i - \omega x_i - b)^2$$

- 一般写为

$$\min_{\omega,b} \frac{1}{n} \sum_{i=1}^{n} (y_i - \omega x_i - b)^2$$



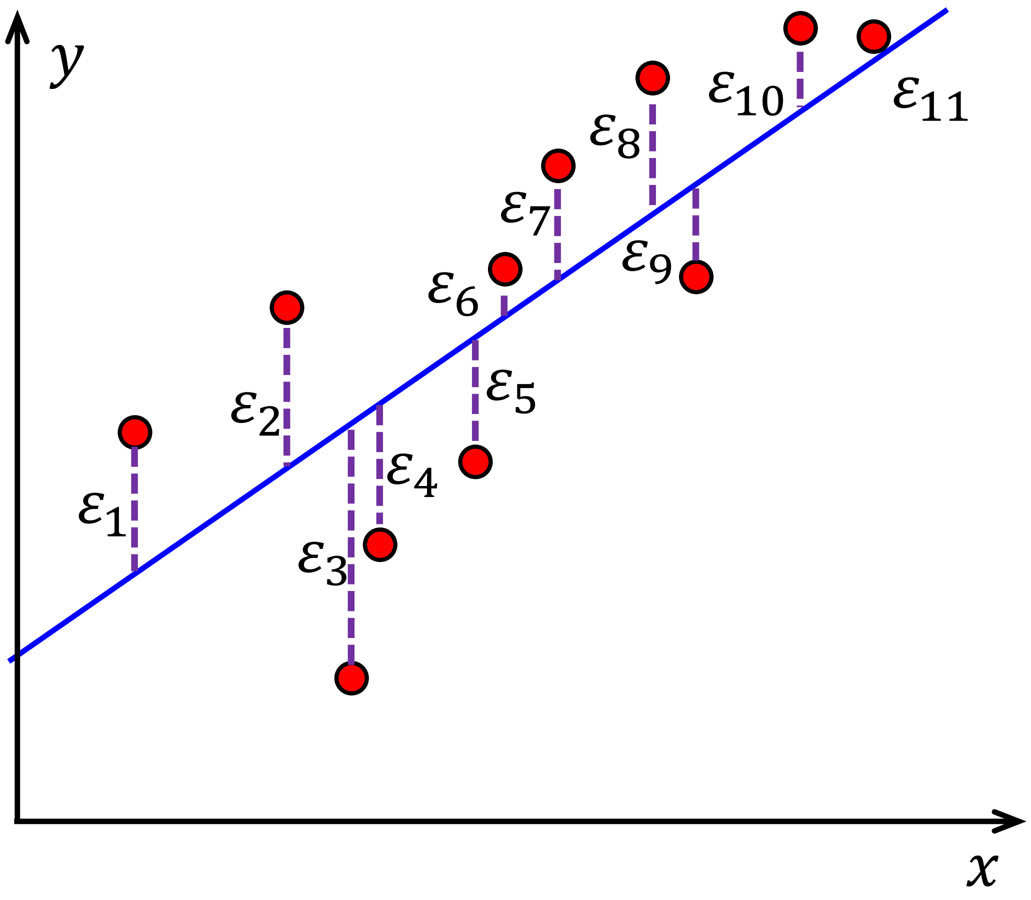
简单线性回归

- $\frac{1}{n}\sum_{i=1}^{n}(y_i-\omega x_i-b)^2$为**均方误差**（mean squared error）
- $\frac{1}{n}\sum_{i=1}^{n}(y_i-\omega x_i-b)^2$被称为**经验损失函数**（empirical loss function）
- 对应的**损失函数**为$(y-\omega x-b)^2$
- 经验损失函数为所有样本损失函数的平均值
- 如下最优化问题即为如何寻找参数ω和b，最小化均方误差经验损失函数

$$\min_{\omega,b}\frac{1}{n}\sum_{i=1}^{n}(y_i-\omega x_i-b)^2$$

简单线性回归

- 经验损失函数$L(\omega,b)=\sum_{i=1}^{n}(y_i-\omega x_i-b)^2$。
- $(y_i-\omega x_i-b)^2$为样本与拟合直线纵向距离平方。
- 函数$L(\omega,b)$为样本到拟合直线纵向距离的平方和。
- 纵向距离大的样本，距离平方和会更大，在经验损失函数占比大。
- 纵向距离小的样本，距离平方和会更小，在经验损失函数占比小。
- 最小化经验损失函数，重点在于调整参数ω和b，使距离大的样本距离变小。



简单线性回归

$$\min_{\omega,b}\sum_{i=1}^{n}(y_i-\omega x_i-b)^2$$

- 目标函数$L(\omega,b)=\sum_{i=1}^{n}(y_i-\omega x_i-b)^2$
- 目标函数的导数为

$$\frac{\partial L(\omega,b)}{\partial \omega}=-2\sum_{i=1}^{n}x_i(y_i-\omega x_i-b)$$

$$\frac{\partial L(\omega,b)}{\partial b}=-2\sum_{i=1}^{n}(y_i-\omega x_i-b)$$

- 令导数等于零，可得

$$\widehat{\omega}=\left(\sum_{i=1}^{n}x_i^2\right)^{-1}\left(\sum_{i=1}^{n}x_iy_i\right)$$

$$\hat{b}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\omega x_i)=\bar{y}-\widehat{\omega}\,\bar{x}$$

，其中$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$，$\bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}$

线性回归

- 一个具有m个特征的样本为

$$\boldsymbol{x} = [x_1, x_2, \cdots x_m]^T$$

- 模型为 $y = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_m x_m + b$
- 每个样本的模型误差为

$$\varepsilon_1 = y_1 - (\omega_1 x_{11} + \cdots + \omega_m x_{1m} + b)$$

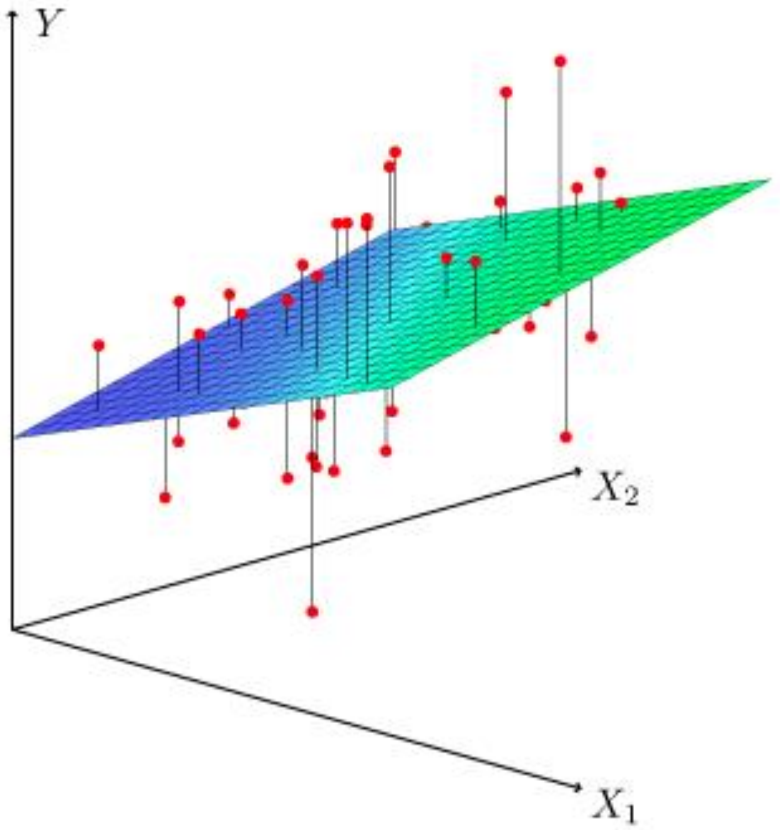
$$\varepsilon_2 = y_2 - (\omega_1 x_{21} + \cdots + \omega_m x_{2m} + b)$$

$$\cdots$$

$$\varepsilon_n = y_n - (\omega_1 x_{n1} + \cdots + \omega_m x_{nm} + b)$$

- 优化问题目标函数或经验损失函数为

$$L(\omega, b) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \omega_1 x_{i1} - \cdots - \omega_m x_{im} - b)^2$$



线性回归示例，每个样本有两个特征

Pic is from http://www.rnfc.org/courses/isl/Lesson%203/Summary/.

简单线性回归

- 优化问题为

$$\min_{\omega_1,\cdots,\omega_m,b} \sum_{i=1}^{n}(y_i-\omega_1 x_{i1}-\cdots-\omega_m x_{im}-b)^2$$

- 目标函数$L(\omega_1,\cdots,\omega_m,b)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\omega_1 x_{i1}-\cdots-\omega_m x_{im}-b)^2$
- 目标函数的导数为

$$\frac{\partial L}{\partial \omega_j}=-2\sum_{i=1}^{n}x_{ij}(y_i-\omega_1 x_{i1}-\cdots-\omega_m x_{im}-b),j=1\cdots m$$

$$\frac{\partial L}{\partial b}=-2\sum_{i=1}^{n}(y_i-\omega x_i-b)$$

- 令导数等于零，可以得到$m+1$个线性方程，求解此线性方程可以得到参数$\omega_1,\cdots,\omega_m,b$的值

线性回归

$$\varepsilon_1 = y_1 - (\omega_1 x_{11} + \cdots + \omega_m x_{1m} + b)$$
$$\varepsilon_2 = y_2 - (\omega_1 x_{21} + \cdots + \omega_m x_{2m} + b)$$
$$\cdots$$
$$\varepsilon_n = y_n - (\omega_1 x_{n1} + \cdots + \omega_m x_{nm} + b)$$

$$\Rightarrow \quad \boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X\omega} - \boldsymbol{1}b$$

其中

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \boldsymbol{\omega} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{bmatrix}, \boldsymbol{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

线性回归

- 经验损失函数为

$$L(\boldsymbol{\omega}, b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \omega_1 x_{i1} - \cdots - \omega_m x_{im} - b)^2 = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega} - \mathbf{1}b\|_2^2$$

- 展开经验损失函数

$$\begin{aligned}
L(\boldsymbol{\omega}, b) &= \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega} - \mathbf{1}b\|_2^2 \\
&= \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega} - \mathbf{1}b)^T\,(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega} - \mathbf{1}b) \\
&= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}^T\mathbf{1}b - \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\omega} + \boldsymbol{\omega}^T\boldsymbol{X}^T\mathbf{1}b - b\mathbf{1}^T\boldsymbol{y} \\
&\quad +b\mathbf{1}^T\boldsymbol{X}\boldsymbol{\omega} + b^2\mathbf{1}^T\mathbf{1}
\end{aligned}$$

线性回归

$$L = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{y}^T\boldsymbol{1}b - \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\omega} + \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{1}b - b\boldsymbol{1}^T\boldsymbol{y} + b\boldsymbol{1}^T\boldsymbol{X}\boldsymbol{\omega} + b^2\boldsymbol{1}^T\boldsymbol{1}$$

- 经验损失函数的导数为

$$\frac{\partial L}{\partial \boldsymbol{\omega}} = -\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\omega} + \boldsymbol{X}^T\boldsymbol{1}b + b\boldsymbol{X}^T\boldsymbol{1}$$

$$\frac{\partial L}{\partial b} = -\boldsymbol{y}^T\boldsymbol{1} + \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{1} - \boldsymbol{1}^T\boldsymbol{y} + \boldsymbol{1}^T\boldsymbol{X}\boldsymbol{\omega} + 2b\boldsymbol{1}^T\boldsymbol{1}$$

- 令导数为零并求解方程，可得

$$\widehat{\boldsymbol{\omega}} = \left(n\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{X}\right)^{-1}\left(n\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{1}\boldsymbol{X}^T\boldsymbol{1}\right)$$

$$\hat{b} = \frac{\boldsymbol{y}^T\boldsymbol{1} - \widehat{\boldsymbol{\omega}}^T\boldsymbol{X}^T\boldsymbol{1}}{n}$$

线性回归

$$\widehat{\boldsymbol{\omega}} = (n\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{X}^T\mathbf{1}\mathbf{1}^T\boldsymbol{X})^{-1}\,(n\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{y}^T\mathbf{1}\boldsymbol{X}^T\mathbf{1})$$

$$\hat{b} = \frac{\boldsymbol{y}^T\mathbf{1} - \widehat{\boldsymbol{\omega}}^T\boldsymbol{X}^T\mathbf{1}}{n}$$

- 已知

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\boldsymbol{y}^T\mathbf{1}}{n}, \qquad \overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \frac{\boldsymbol{X}^T\mathbf{1}}{n}$$

- 可得

$$\widehat{\boldsymbol{\omega}} = (\boldsymbol{X}^T\boldsymbol{X} - n\overline{\boldsymbol{x}} \cdot \overline{\boldsymbol{x}}^T)^{-1}\,(\boldsymbol{X}^T\boldsymbol{y} - n\bar{y}\overline{\boldsymbol{x}})$$

$$\hat{b} = \bar{y} - \widehat{\boldsymbol{\omega}}^T\overline{\boldsymbol{x}}$$

线性回归

- 一个具有m个特征的样本为

$$\boldsymbol{x} = [x_1, x_2, \cdots x_m]^T$$

- 模型为

$$y = \sum_{i=1}^{m} \omega_i x_i + b$$

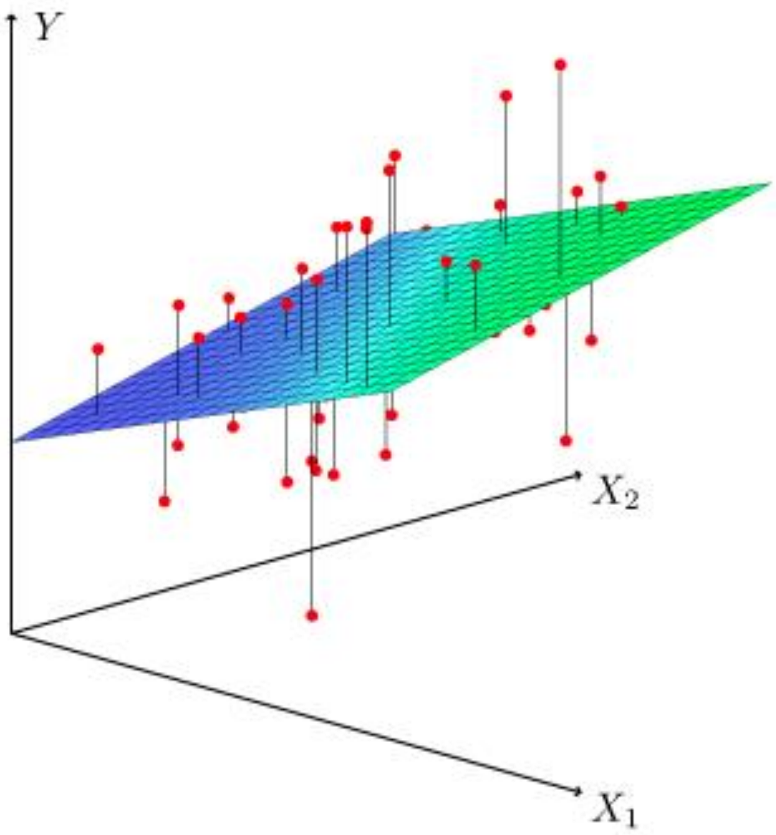
$$\boldsymbol{x} \leftarrow [x_1, x_2, \cdots x_m, 1]^T$$

增加一个值为1的特征

$$\boldsymbol{\omega} \leftarrow [\omega_1, \omega_2, \cdots \omega_m, b]^T$$

b 被吸收到$\boldsymbol{\omega}$

$$y = \boldsymbol{x}^T \boldsymbol{\omega}$$



线性回归示例，每个样本有两个特征

Pic is from http://www.rnfc.org/courses/isl/Lesson%203/Summary/.

线性回归

$$\varepsilon_1 = y_1 - (\omega_1 x_{11} + \cdots + \omega_m x_{1m} + b)$$

$$\varepsilon_2 = y_2 - (\omega_1 x_{21} + \cdots + \omega_m x_{2m} + b)$$

$$\cdots$$

$$\varepsilon_n = y_n - (\omega_1 x_{n1} + \cdots + \omega_m x_{nm} + b)$$

$$\Rightarrow \quad \boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X\omega}$$

其中

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} & 1 \\ x_{21} & x_{22} & \cdots & x_{2m} & 1 \\ \cdots & \cdots & \cdots & \cdots & 1 \\ x_{n1} & x_{n2} & \cdots & x_{nm} & 1 \end{bmatrix}, \boldsymbol{\omega} = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \\ b \end{bmatrix}$$

线性回归

- 经验损失函数为

$$L(\boldsymbol{\omega}) = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\omega}\right)^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}\|_2^2$$

- 损失函数可以展开写为

$$L(\boldsymbol{\omega}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\omega} - \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\omega}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\omega}$$

- 损失函数的导数

$$\frac{\partial F}{\partial \boldsymbol{\omega}} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\omega}$$

- 令导数为零，可解得

$$\hat{\boldsymbol{\omega}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

使用向量的好处：
① 数学表达式和代码简洁
② 向量运算的计算机代码速度更快（软硬件支持）。

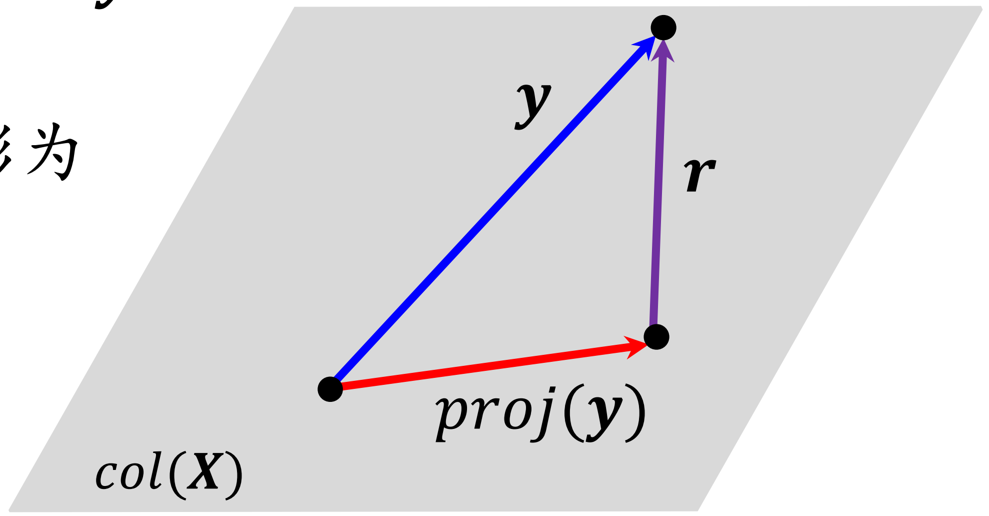
线性回归

$$\min_{\boldsymbol{\omega}} \|\boldsymbol{y} - \boldsymbol{X\omega}\|_2^2 \rightarrow \underbrace{\widehat{\boldsymbol{\omega}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}}_{\boldsymbol{X}^T\boldsymbol{X}\text{ 必须满秩，如何做到？}}$$

$$\min_{\boldsymbol{\omega}} \|\boldsymbol{y} - \boldsymbol{X\omega}\|_2^2$$

- 代表找到一个$\boldsymbol{\omega}$使$\boldsymbol{X}$的列向量的线性组合$\boldsymbol{X\omega}$与向量$\boldsymbol{y}$二范数距离最近
- 向量$\boldsymbol{y}$在$\boldsymbol{X}$的列空间内投影与$\boldsymbol{y}$的距离最近，投影为$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$，投影系数为$\widehat{\boldsymbol{\omega}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- 投影残差为$\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X\omega} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- 投影残差与$\boldsymbol{X}$的列空间正交

$$\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{0}$$



线性回归

线性回归模型训练伪代码

输入： 训练数据$\boldsymbol{X}, \boldsymbol{y}$.
输出： 线性回归参数.

(1) $\boldsymbol{X} = [\boldsymbol{X}, \mathbf{1}_n]$.
(2) $\boldsymbol{\omega} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
(3) 返回$\boldsymbol{\omega}$.

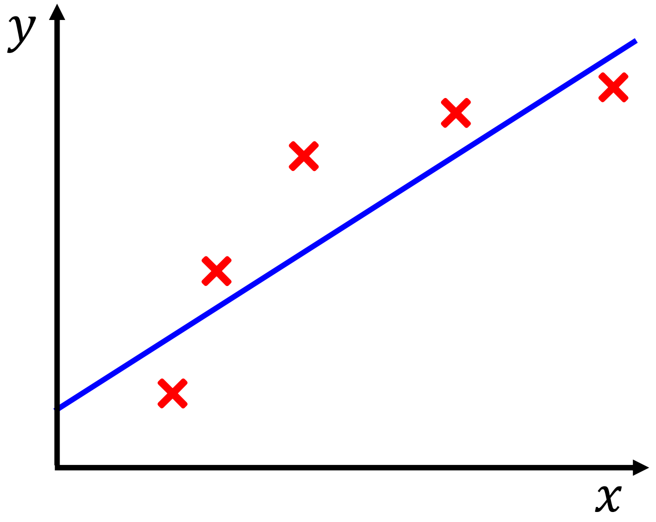
线性回归模型预测伪代码

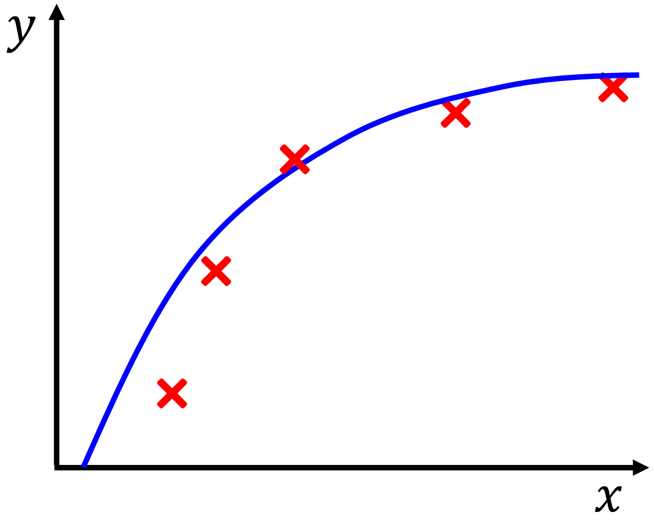
输入: 线性模型参数$\boldsymbol{\omega}$，测试样本$\boldsymbol{x}$。
输出: 测试样本预测标签。

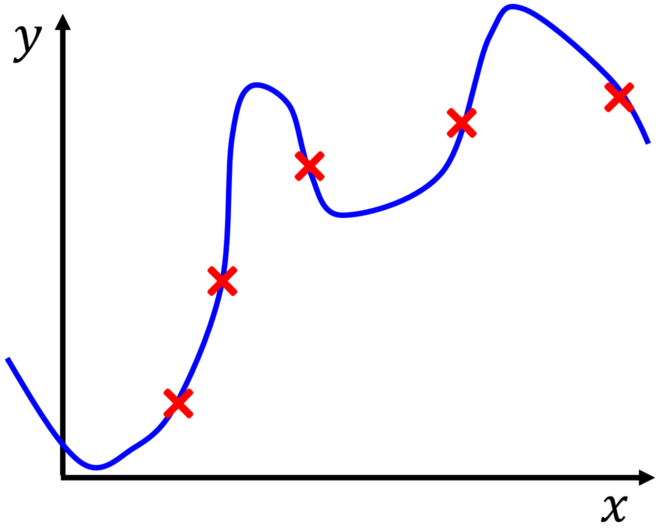
(1) $\boldsymbol{x} = [\boldsymbol{x}^T, 1]^T$.
(2) $\hat{y} = \boldsymbol{\omega}^T\boldsymbol{x}$.
(3) 返回$\hat{y}$.

偏差与方差

哪个模型最好?





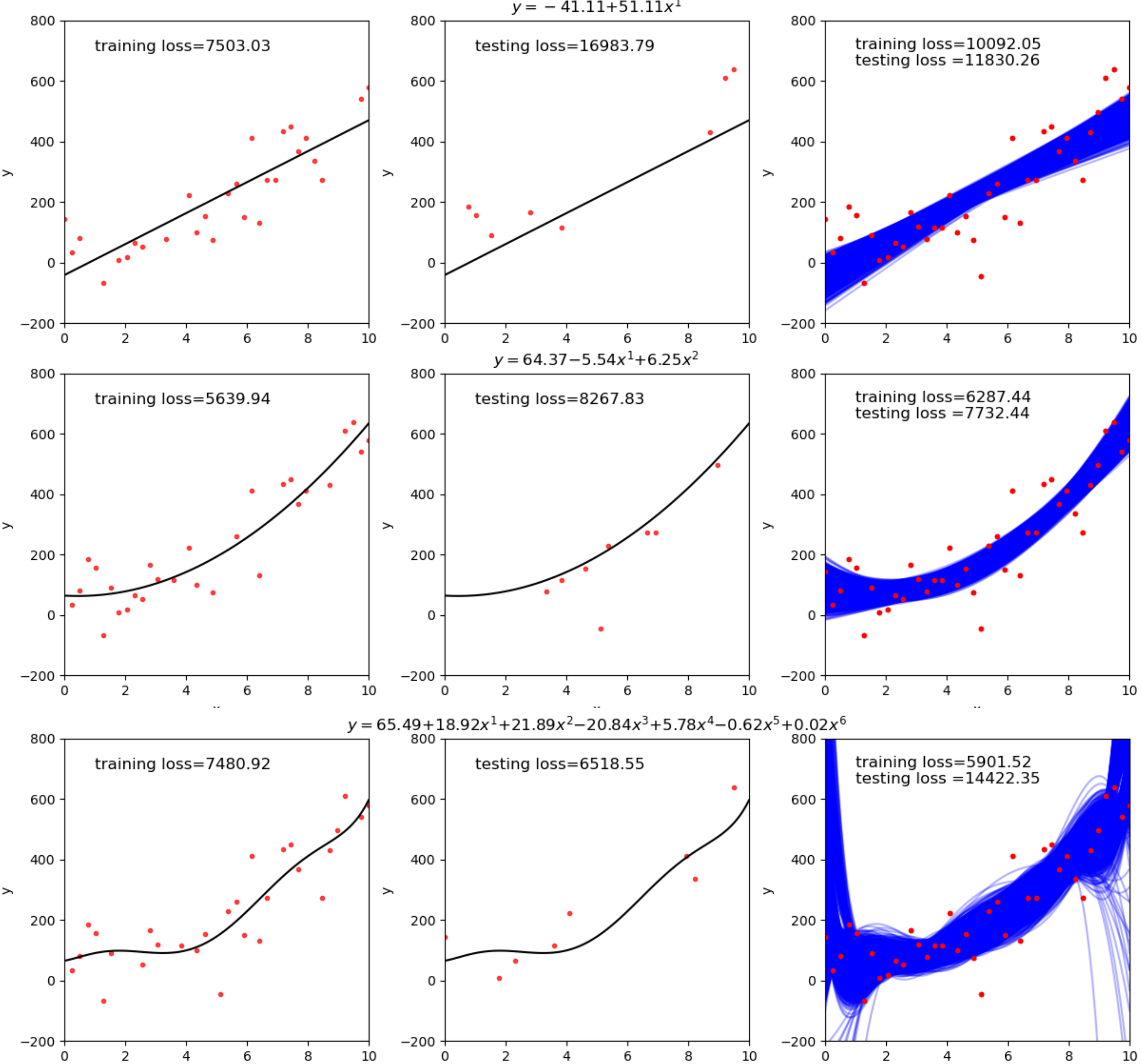


$$y = \omega_1 x + \omega_0$$

$$y = \omega_2 x^2 + \omega_1 x + \omega_0$$

$$y = \omega_p x^p + \cdots + \omega_1 x + \omega_0$$
$$p > 2$$

除了利用给定的特征x，x^r也可以用做特征。



$$y = \omega_0 + \omega_1 x$$

训练损失函数值 = 10092

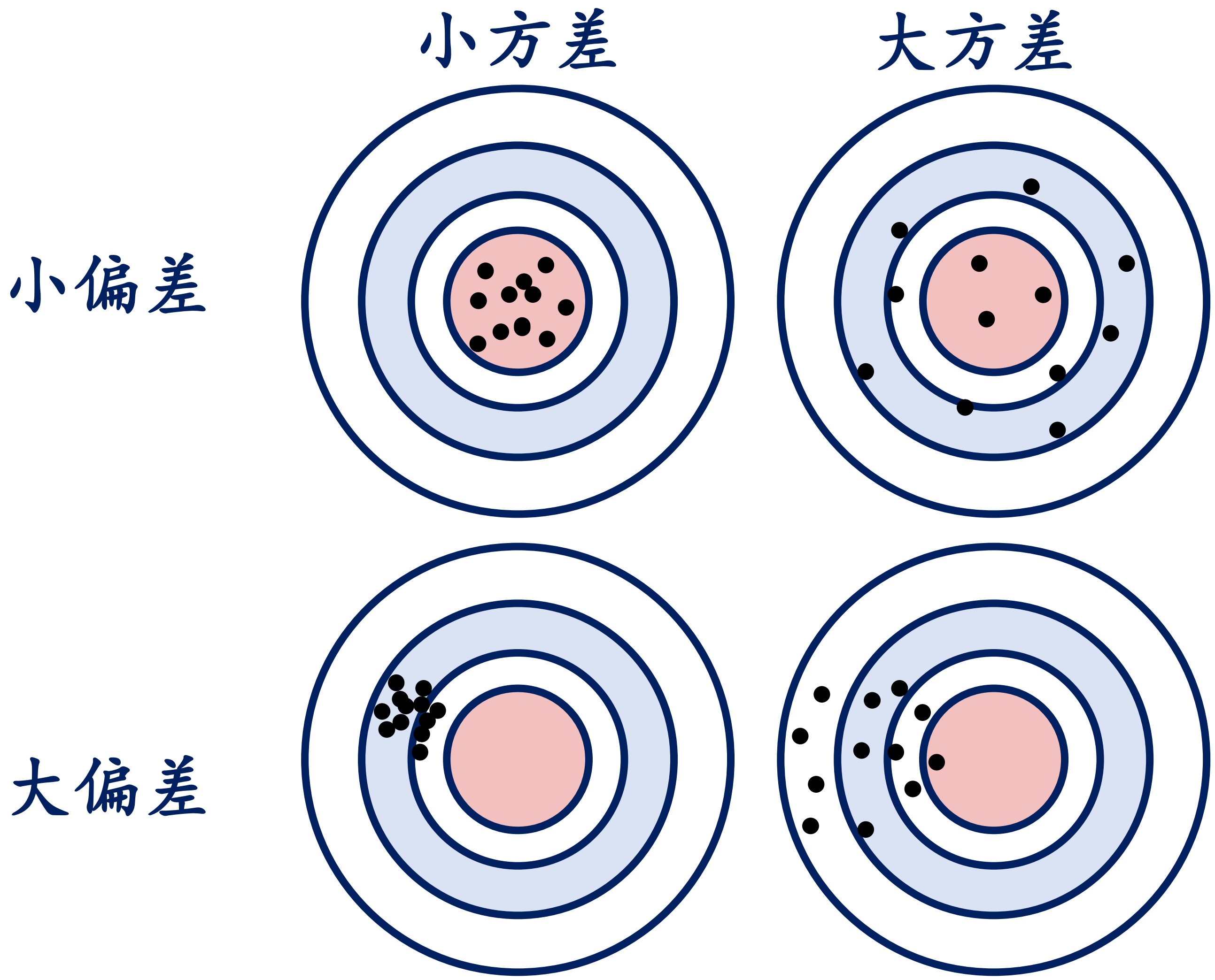
$$y = \omega_0 + \omega_1 x + \omega_2 x^2$$

训练损失函数值 = 6287

$$y = \omega_0 + \omega_1 x \cdots + \omega_6 x^6$$

训练损失函数值 = 5901

偏差与方差



偏差与方差

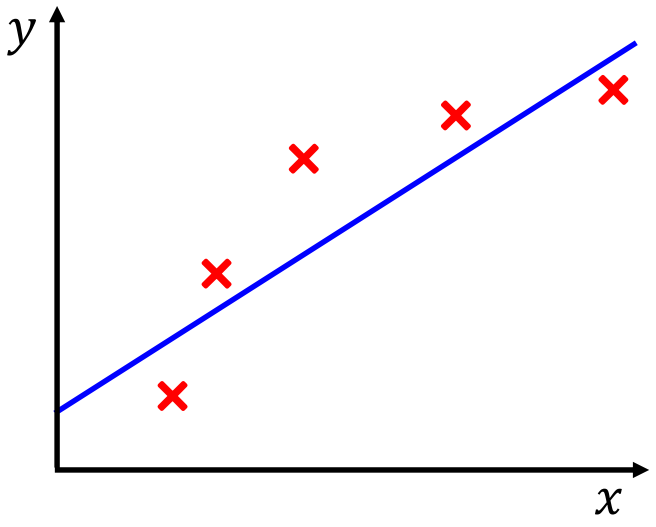
- 方差

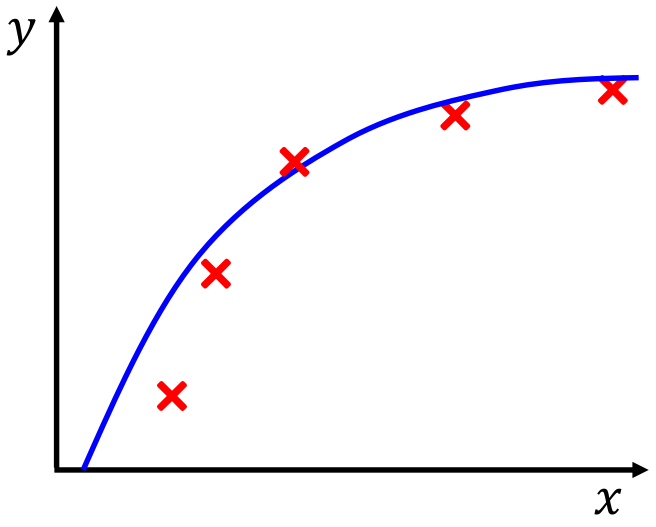
随机变量的方差描述的是它的离散程度，也就是随机变量在其期望值附近的波动程度。

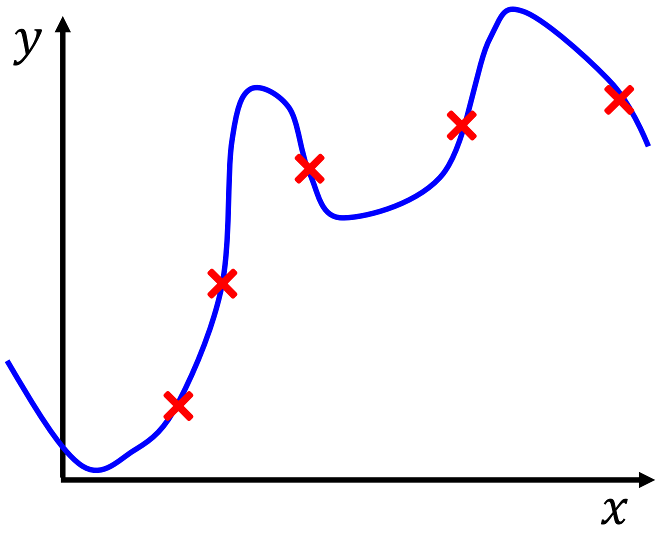
- 偏差

随机变量均值或期望值与真实值的差值或偏离程度。

偏差与方差







$$y = \omega_1 x + b$$

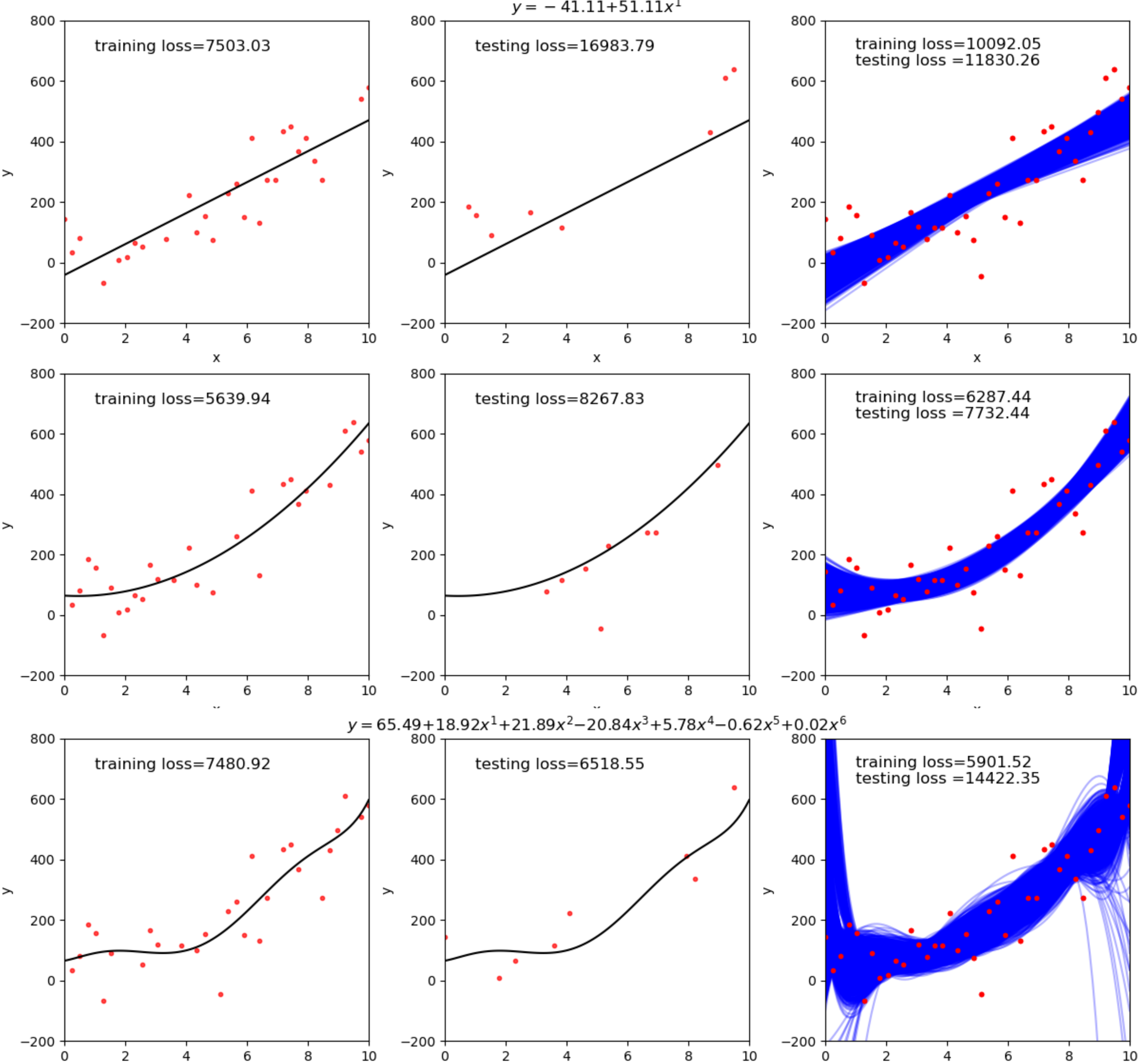
$$y = \omega_2 x^2 + \omega_1 x + b$$

$$y = \sum_{i=1}^{n} \omega_i x^i + b, n > 4$$

大偏差

恰好

大方差



$y = b + \omega_1 x$

训练损失函数值= 10092

测试损失函数值= 11830

$y = b + \omega_1 x + \omega_2 x^2$

训练损失函数值= 6287

测试损失函数值= 7732

$y = b + \omega_1 x \cdots + \omega_6 x^6$

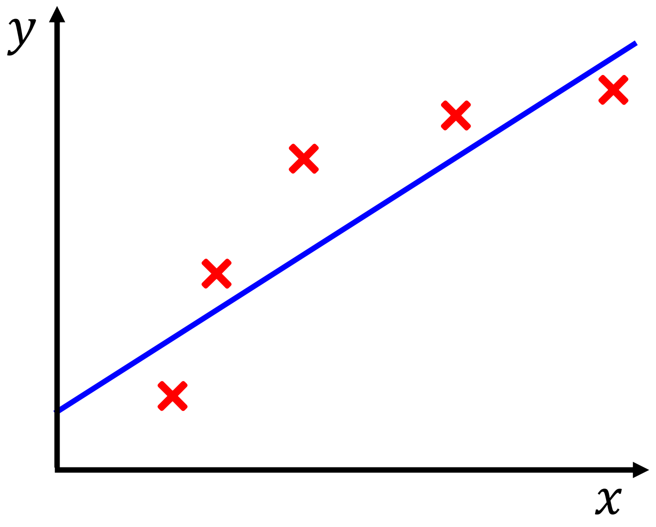
训练损失函数值= 5901

测试损失函数值= 14422

过拟合

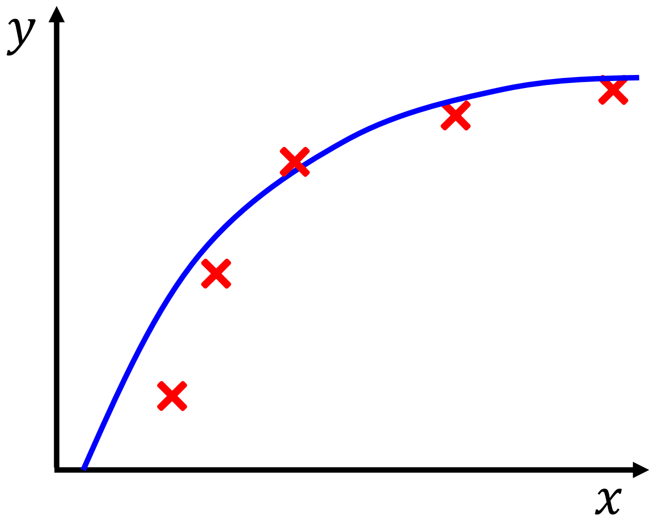
- 模型在**<u>训练数据</u>**上表现良好，但在**<u>测试数据</u>**上表现极差。
- 模型不仅从训练数据中学习到了数据的规律，还学习了数据中的噪声。
- 模型具有小的偏差，但有大的方差。

偏差与方差



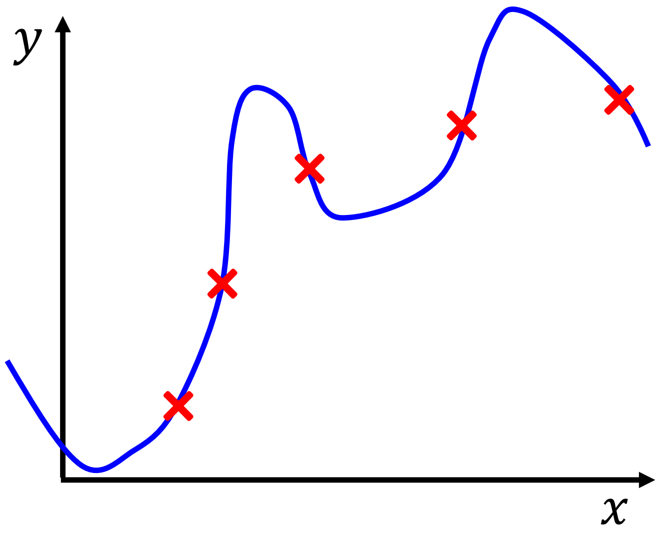
$$y = \omega_1 x + b$$

大偏差
欠拟合



$$y = \omega_2 x^2 + \omega_1 x + b$$

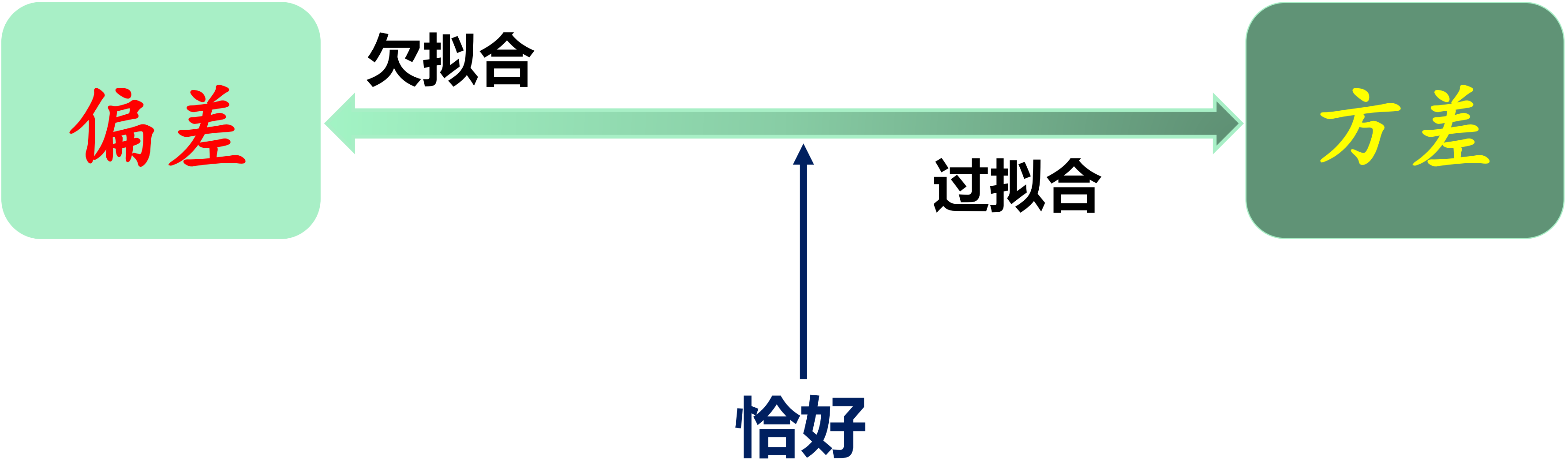
恰好



$$y = \sum_{i=1}^{n} \omega_i x^i + b, n > 4$$

大方差
过拟合

偏差与方差



偏差与方差

有监督学习建立一个函数，将样本转换为标签

$$\hat{y} = g(\boldsymbol{x})$$

给定一个样本 $\boldsymbol{x}$, 标签的分布函数为$f_{Y|X}(y|\boldsymbol{x})$。可得到

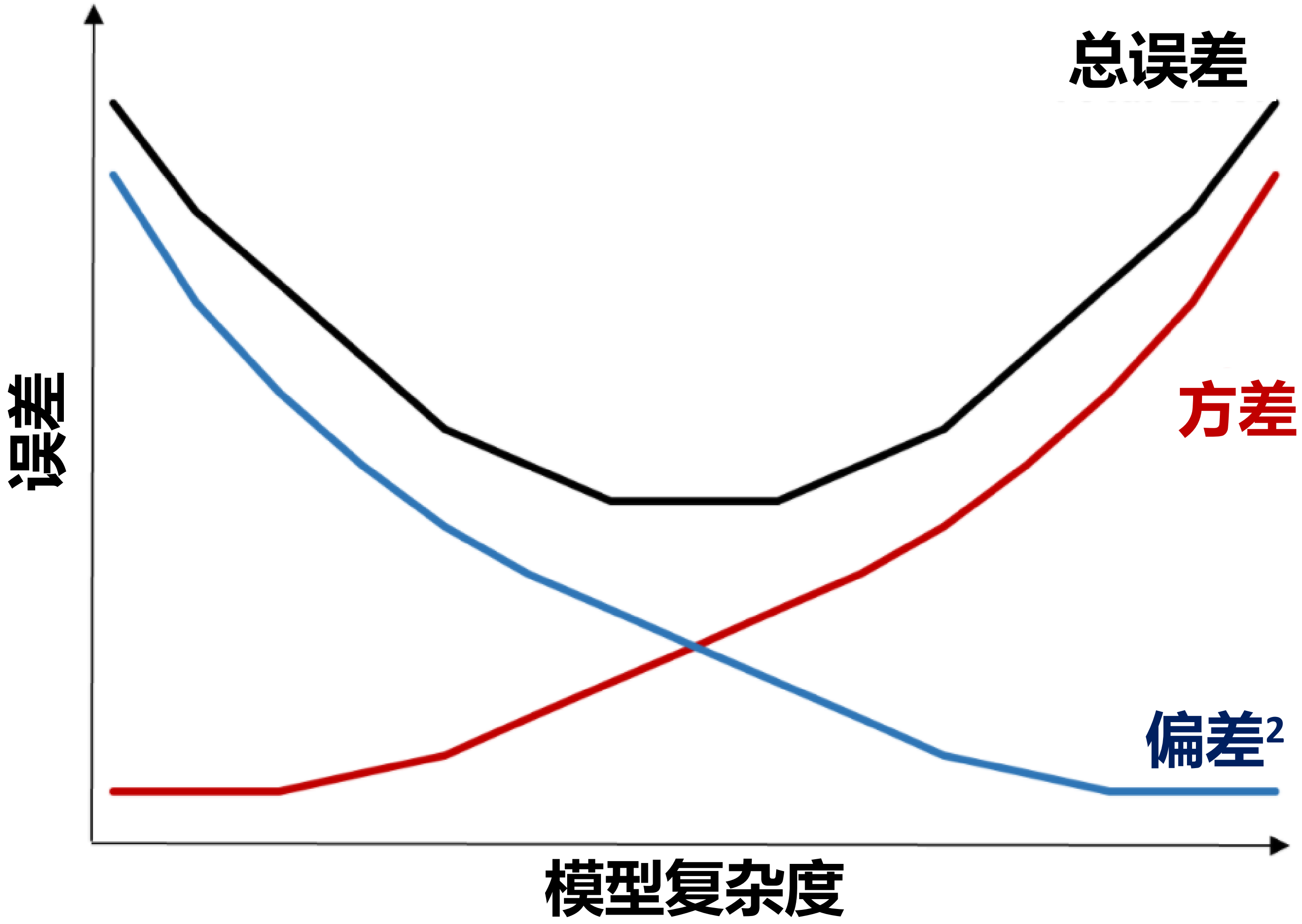
$$\bar{y} = \mathrm{E}[y] = \int y f_{Y|X}(y|\boldsymbol{x})dy$$

误差
$$
\begin{aligned}
&= \mathrm{E}[(\hat{y} - y)^2|\boldsymbol{x}] = \mathrm{E}[(\hat{y} - \bar{y} + \bar{y} - y)^2|\boldsymbol{x}] \\
&= \mathrm{E}[(\hat{y} - \bar{y})^2|\boldsymbol{x}] - 2\mathrm{E}\big[(\hat{y} - \bar{y})(\bar{y} - y) \ \ |\boldsymbol{x}\big] + \mathrm{E}[(\bar{y} - y)^2|\boldsymbol{x}] \\
&= \mathrm{E}[(\hat{y} - \bar{y})^2|\boldsymbol{x}] - 2\mathrm{E}[(\hat{y} - \bar{y})|\boldsymbol{x}]\mathrm{E}[(\bar{y} - y)|\boldsymbol{x}] + \mathrm{E}[(\bar{y} - y)^2|\boldsymbol{x}] \\
&= \mathrm{E}[(\hat{y} - \bar{y})^2|\boldsymbol{x}] - 2\mathrm{E}[(\hat{y} - \bar{y})|\boldsymbol{x}](\bar{y} - \mathrm{E}[y|\boldsymbol{x}]) + \mathrm{E}[(\bar{y} - y)^2|\boldsymbol{x}] \\
&= \mathrm{E}[(\hat{y} - \bar{y})^2|\boldsymbol{x}] + \mathrm{E}[(\bar{y} - y)^2|\boldsymbol{x}]
\end{aligned}
$$

方差　　**偏差²**

偏差与方差

误差 = 方差 + 偏差^2



过拟合

- **过拟合发生原因是什么?**

过拟合

- **过拟合发生原因是什么?**

模型过于复杂，训练数据不够大

过拟合

- **过拟合发生原因是什么?**

模型过于复杂，训练数据不够大

- **如何避免过拟合?**

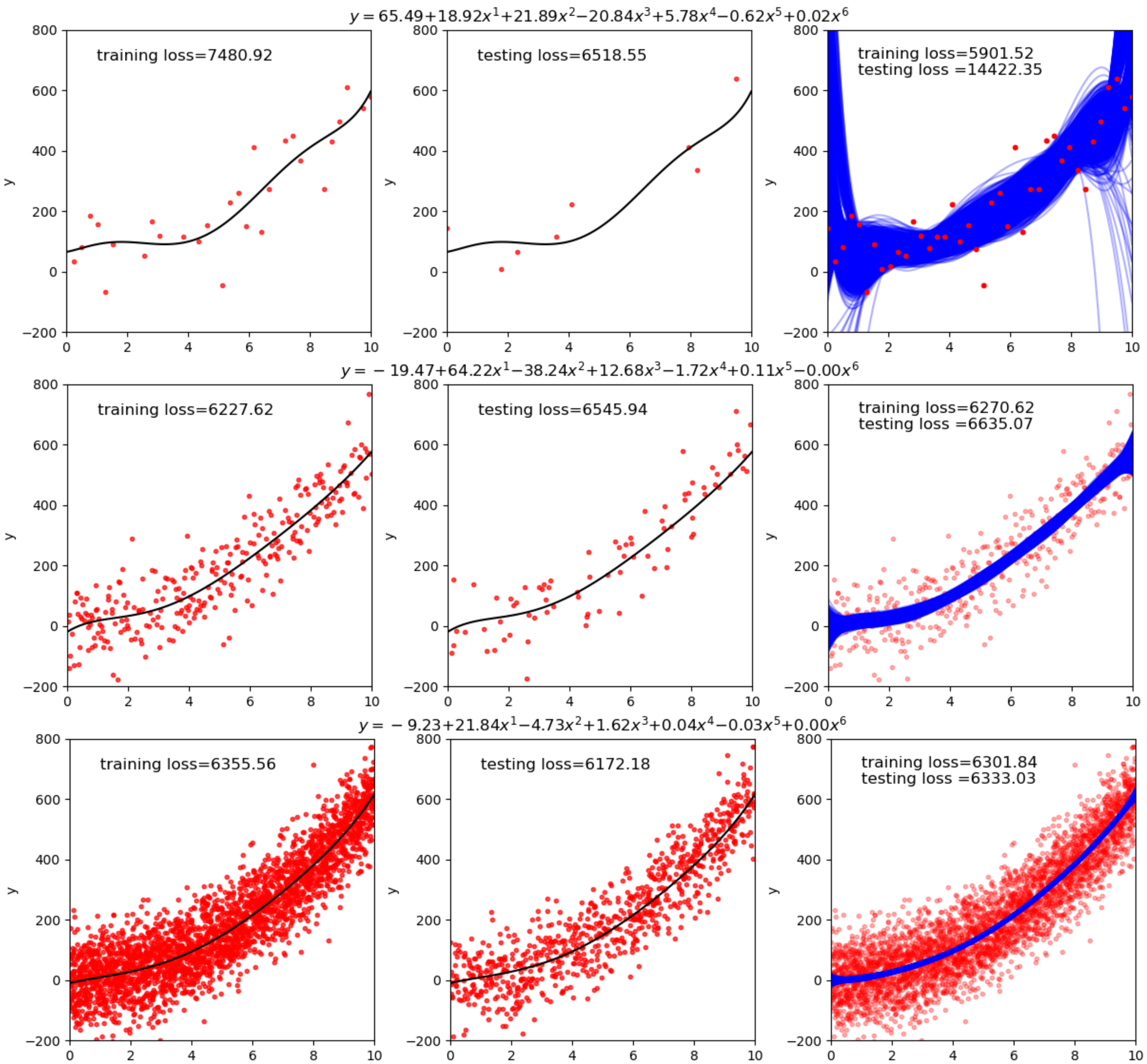
过拟合

- **过拟合发生原因是什么?**

模型过于复杂，训练数据不够大

- **如何避免过拟合?**

✓ 增大训练数据
✓ 降低模型复杂度，比如经验损失函数加正则项



模型 $y = b + \omega_1 x \cdots + \omega_6 x^6$

训练样本数 = 40
训练损失函数值 = 5901
测试损失函数值 = 14422

训练样本数 = 400
训练损失函数值= 6270
测试损失函数值= 6635

训练样本数 = 4000
训练损失函数值= 6301
测试损失函数值= 6333

正则化

- **损失函数**

λ 为以参数，平衡两个式子

$$\min_{\boldsymbol{\omega}} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X\omega}\|_2^2+\lambda\psi(\boldsymbol{\omega})$$

鼓励模型拟合数据

正则项，鼓励模型不拟合数据

正则项 $\psi(\boldsymbol{\omega})$ 可以为

- $\|\boldsymbol{\omega}\|_2^2$, 问题被称为**岭回归**.
- $\|\boldsymbol{\omega}\|_1$, 问题被称为**LASSO** (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator).
- $\frac{\alpha}{2}\|\boldsymbol{\omega}\|_2^2+(1-\alpha)\|\boldsymbol{\omega}\|_1$, 问题被称为**Elastic net.**
- $\|\boldsymbol{\omega}\|_p^p$, $0<p\le 1$.

正则化

- **岭回归**

$$\min_{\boldsymbol{\omega}} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_2^2 \rightarrow \boldsymbol{\omega}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

意外收获：$\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}$ 满秩

正则化

- **岭回归**

$$\min_{\boldsymbol{\omega}} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_2^2 \quad \longrightarrow \quad \boldsymbol{\omega}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- 如果 $\lambda=0$，解为什么样子?

正则化

- **岭回归**

$$\min_{\boldsymbol{\omega}} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_2^2 \rightarrow \boldsymbol{\omega}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- 如果 $\lambda=0$，解为什么样子?

  如果$\lambda=0$，目标函数为 $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2$ ，问题退化为线性回归，解为 $\boldsymbol{\omega}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

- 如果$\lambda=+\infty$，解为什么样子?

正则化

- **岭回归**

$$\min_{\boldsymbol{\omega}} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_2^2 \rightarrow \boldsymbol{\omega}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- 如果 $\lambda=0$，解为什么样子?

  如果$\lambda=0$，目标函数为 $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2$ ，问题退化为线性回归，解为 $\boldsymbol{\omega}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

- 如果$\lambda=+\infty$，解为什么样子?

  如果$\lambda=+\infty$，目标函数等价于 $\|\boldsymbol{\omega}\|_2^2$ ，解为$\boldsymbol{\omega}=\boldsymbol{0}$.

- 如何理解$\lambda=+\infty$的解?

正则化

- **岭回归**

$$\min_{\boldsymbol{\omega}} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_2^2 \rightarrow \boldsymbol{\omega}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- 如果 $\lambda=0$，解为什么样子?

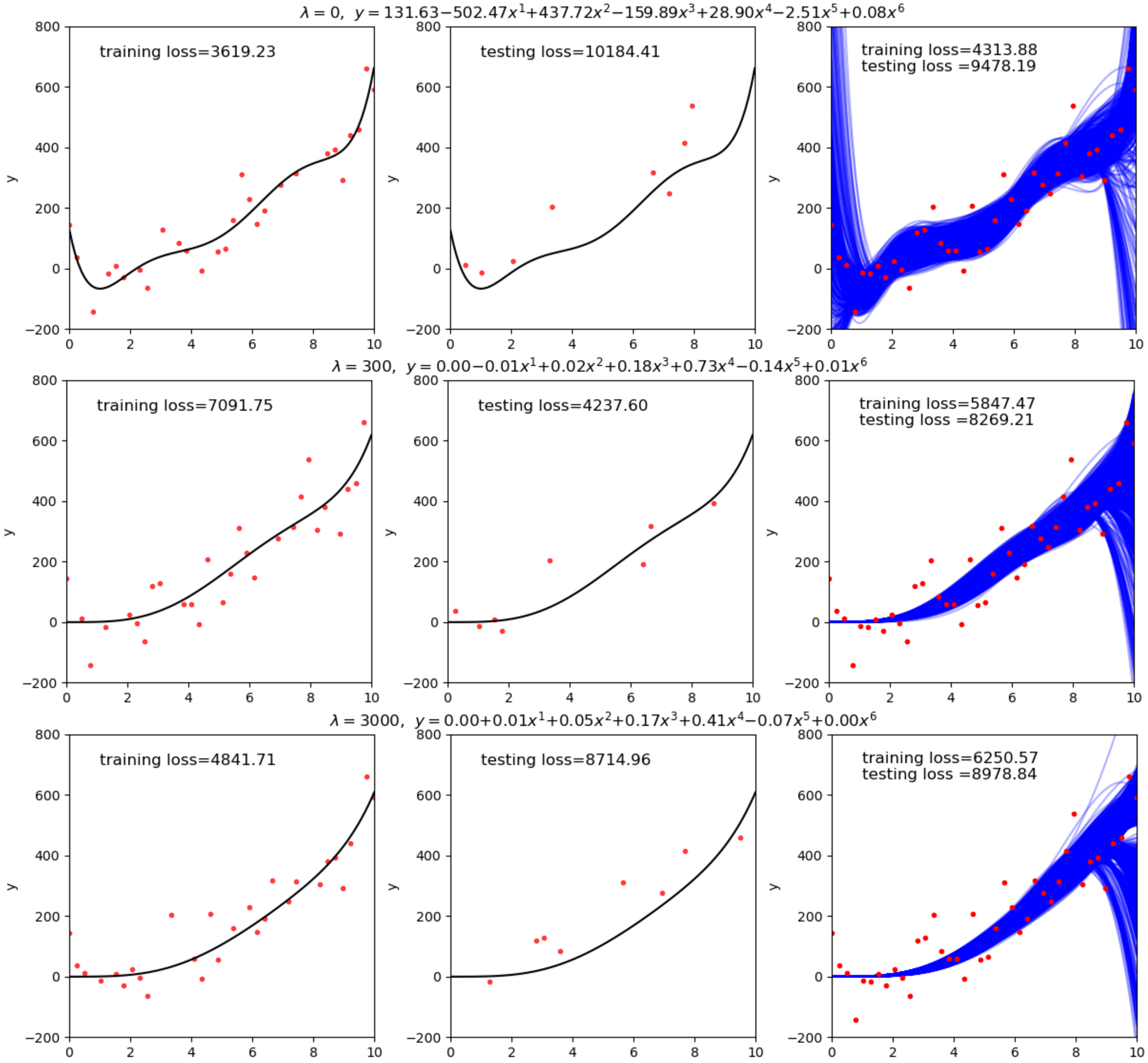
  如果$\lambda=0$，目标函数为 $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2$，问题退化为线性回归，解为 $\boldsymbol{\omega}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$。

- 如果$\lambda=+\infty$，解为什么样子?

  如果$\lambda=+\infty$，目标函数等价于 $\|\boldsymbol{\omega}\|_2^2$，解为$\boldsymbol{\omega}=\boldsymbol{0}$。

- 如何理解$\lambda=+\infty$的解?

  目标函数与训练数据无关。模型偏差极大。



岭回归 $\min\limits_{\boldsymbol{\omega}} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_2^2$

训练样本数为40

$y = b + \omega_1 x \cdots + \omega_6 x^6$

$\lambda = 0$

训练损失函数值= 4313

测试损失函数值= 9478

$\lambda = 300$

训练损失函数值= 5847

测试损失函数值= 8269

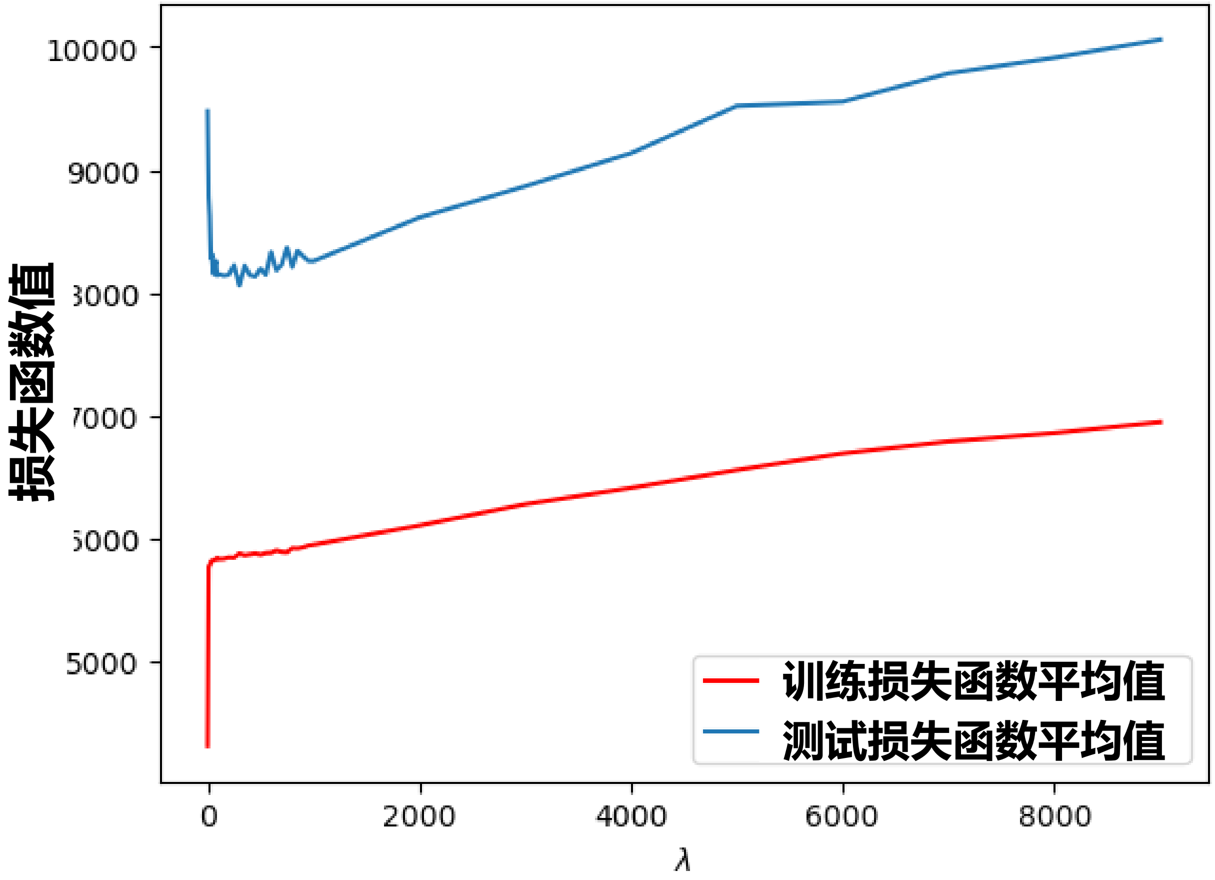
$\lambda = 3000$

训练损失函数值= 6250

测试损失函数值= 8978

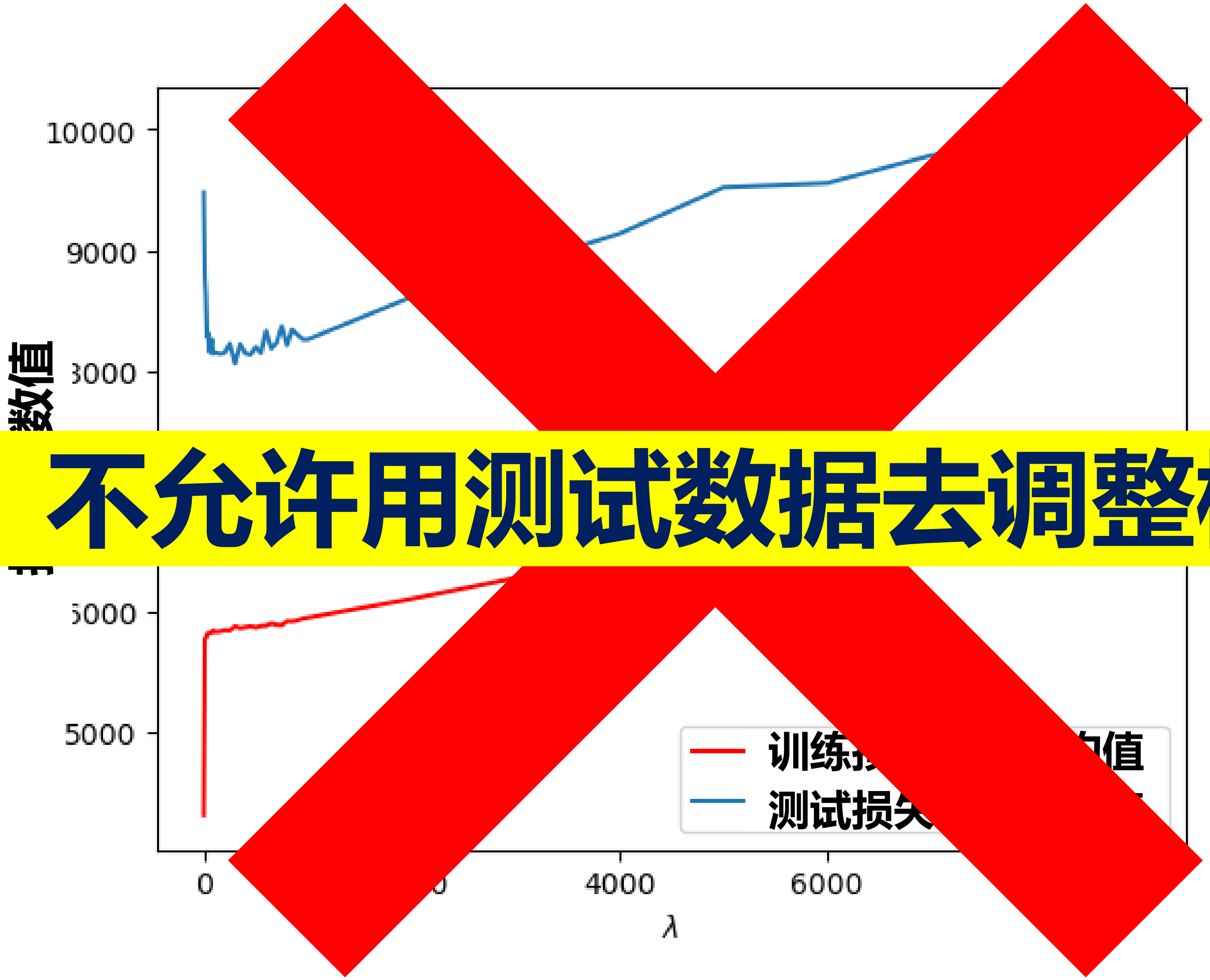
正则化

- **岭回归**


10000
9000
3000
7000
5000
5000
0
2000
4000
6000
8000
λ
损失函数数值
训练损失函数平均值
测试损失函数平均值

正则化

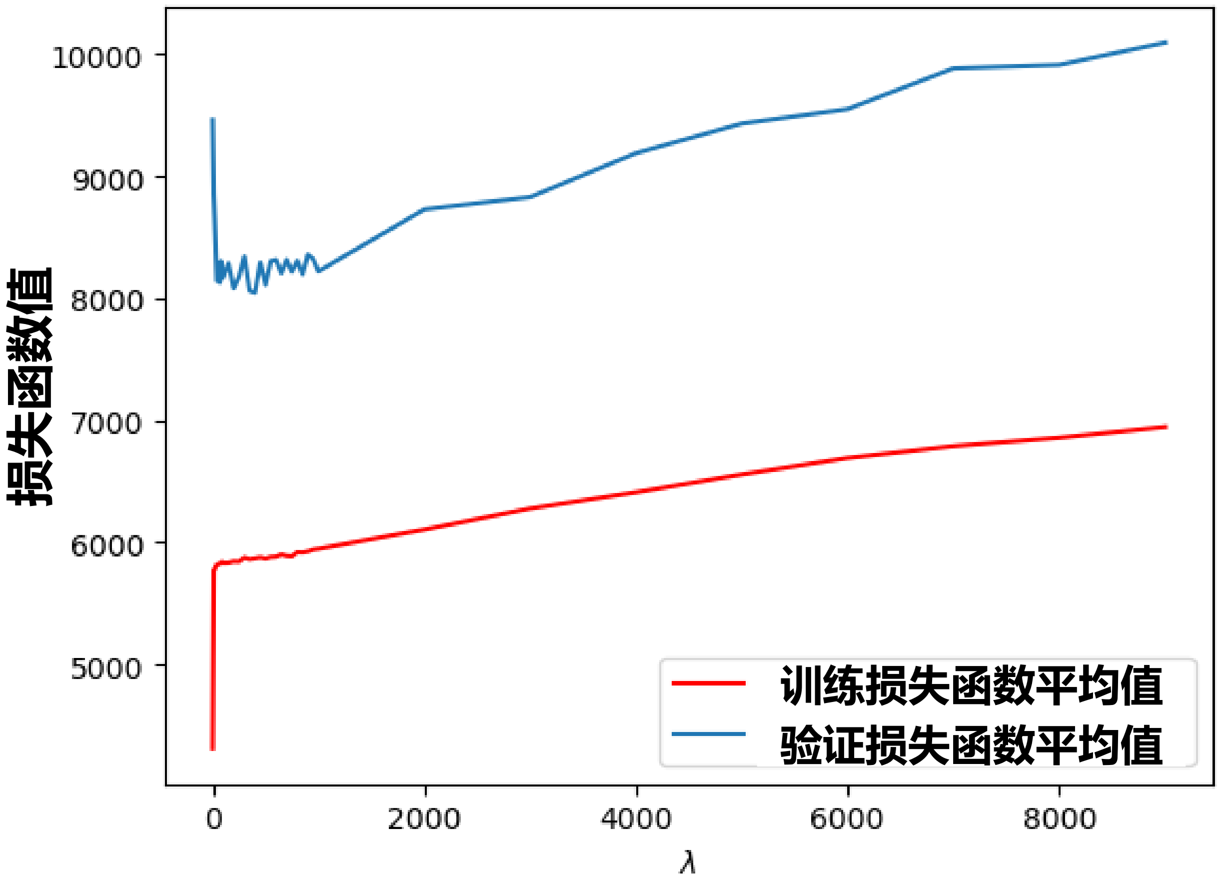
- **岭回归**



不允许用测试数据去调整模型

Regularization（正则化）

- **岭回归**



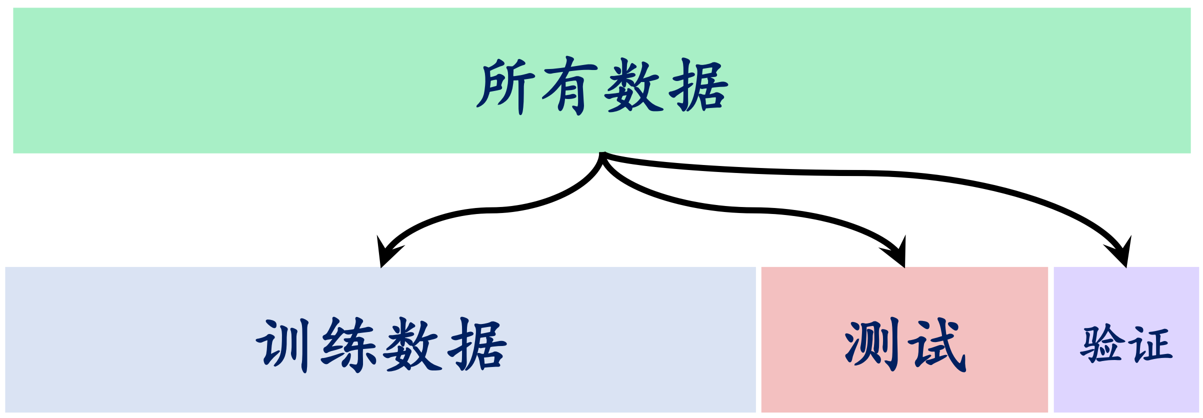
$$\min_{\omega} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_2^2$$

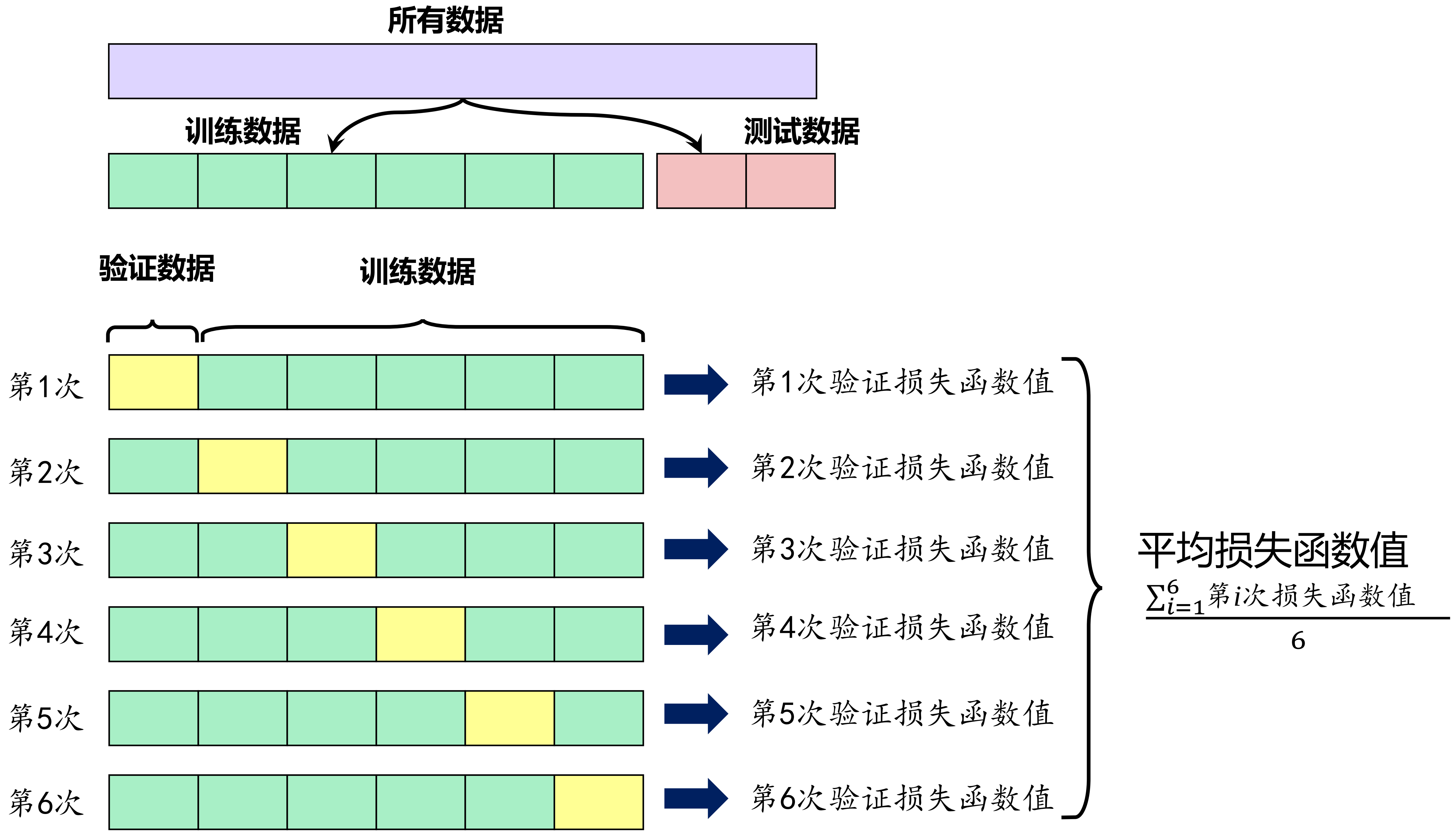
λ 为超参数

验证数据是用来选择最优的超参数

交叉验证

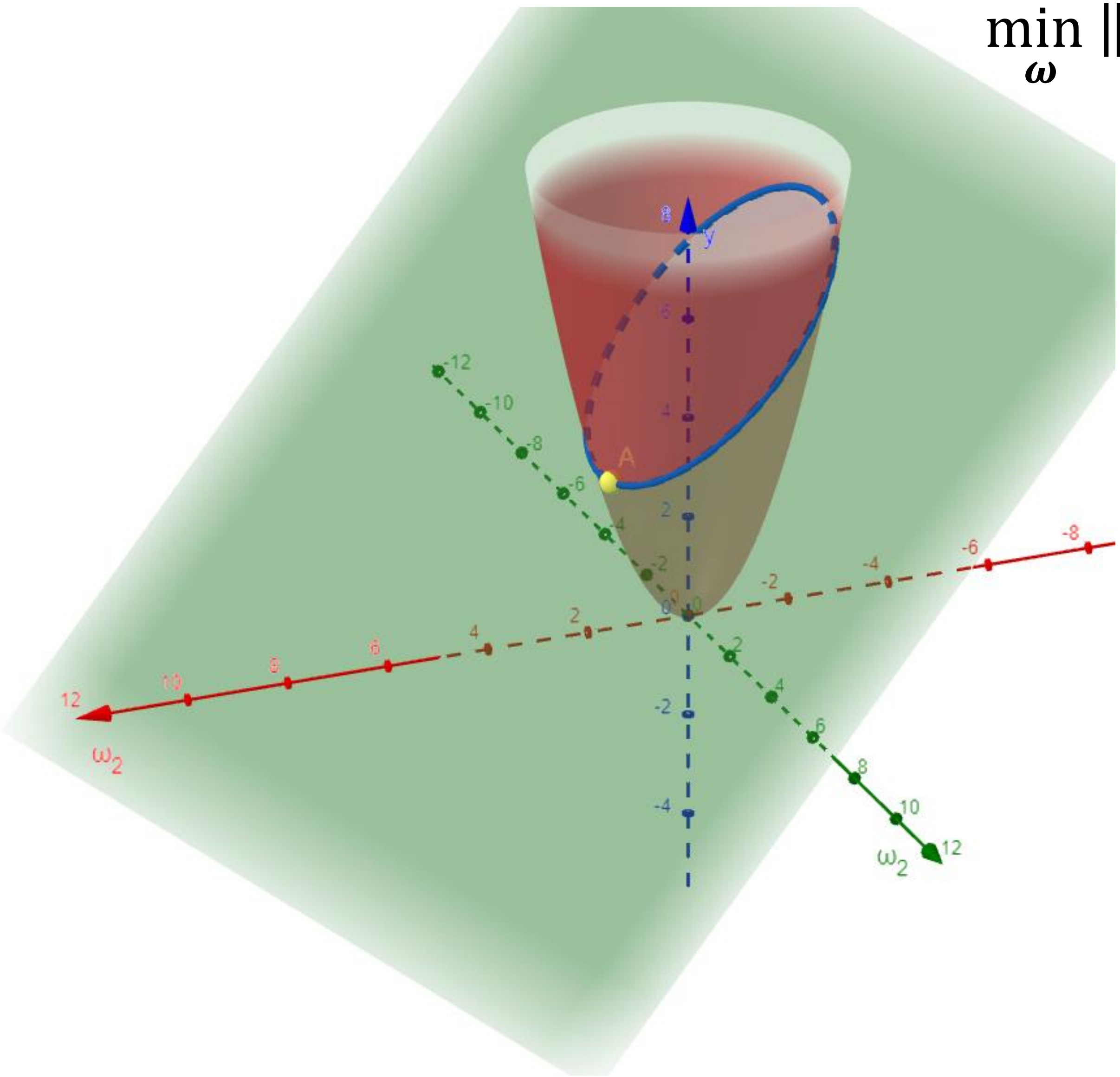
- 验证数据用来选择超参数。
- 但是验证数据集一般很小。
- 模型在验证数据上的指标不准确，因为验证数据集太小。

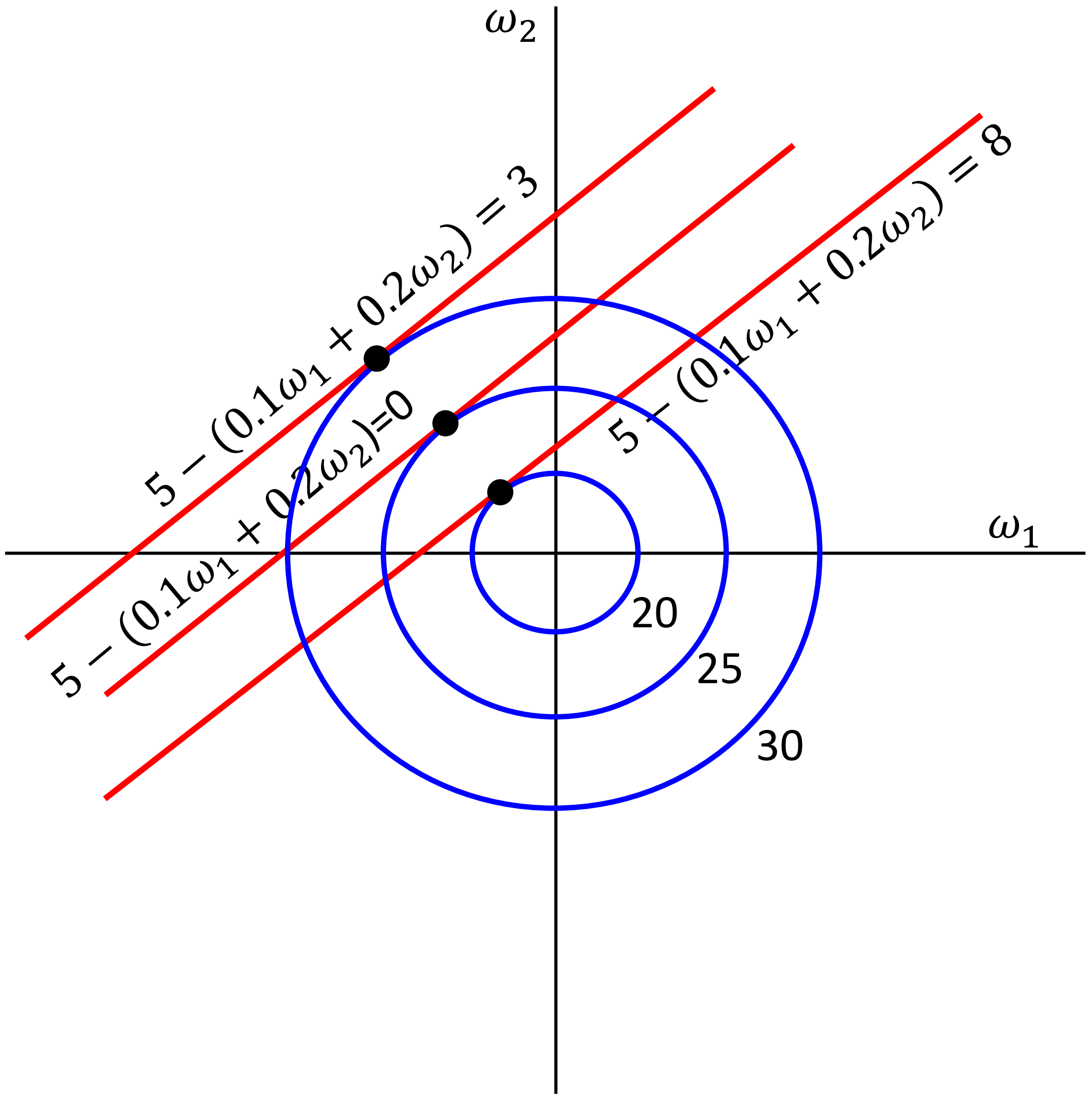



所有数据
训练数据
测试数据
验证数据
训练数据
第1次
第2次
第3次
第4次
第5次
第6次
第1次验证损失函数值
第2次验证损失函数值
第3次验证损失函数值
第4次验证损失函数值
第5次验证损失函数值
第6次验证损失函数值
平均损失函数值
$\frac{\sum_{i=1}^{6} 第i次损失函数值}{6}$

正则化：岭回归

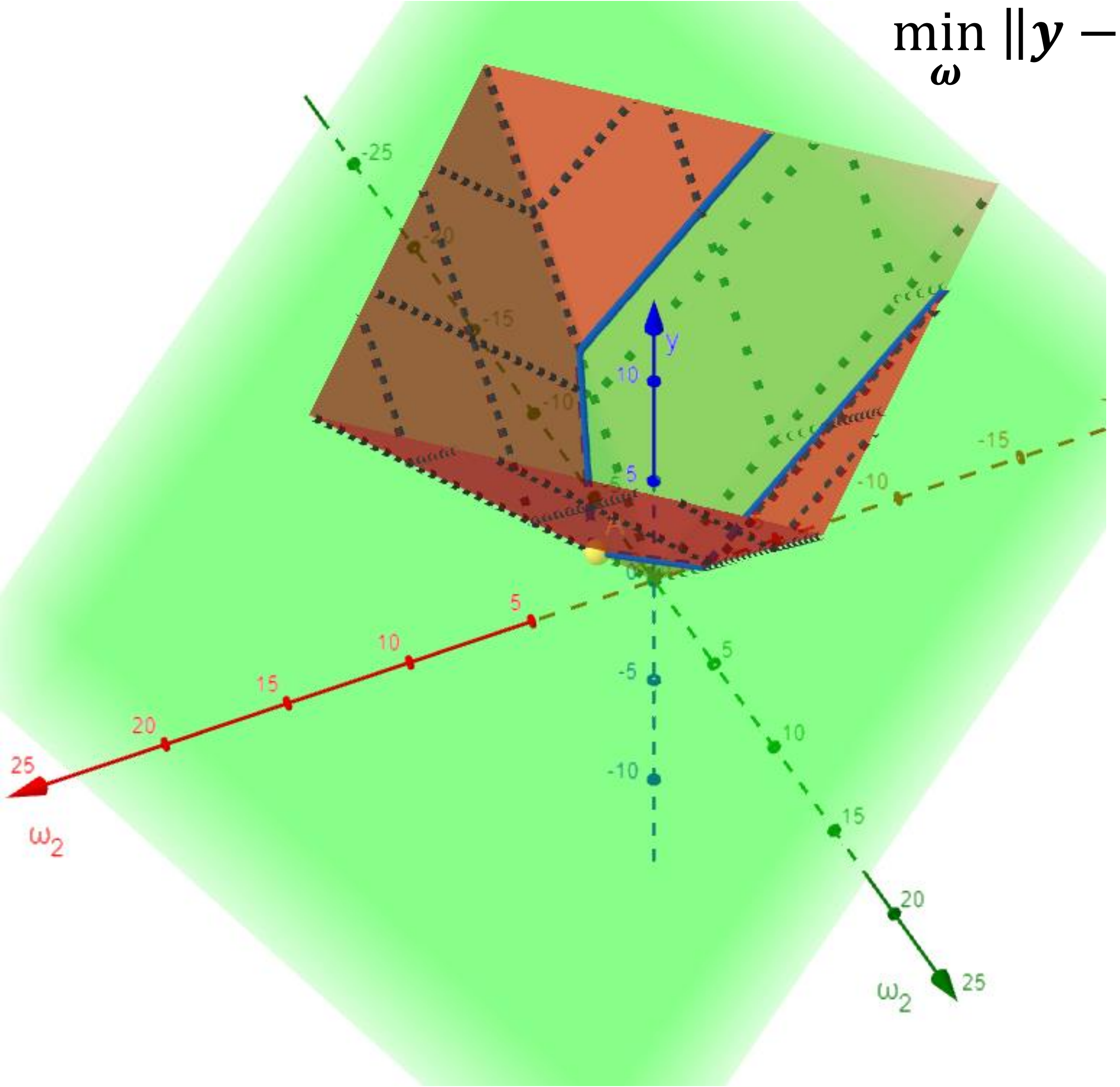
$$\min_{\boldsymbol{\omega}} \|\boldsymbol{y} - \boldsymbol{X\omega}\|_2^2 + \lambda\|\boldsymbol{\omega}\|_2^2$$

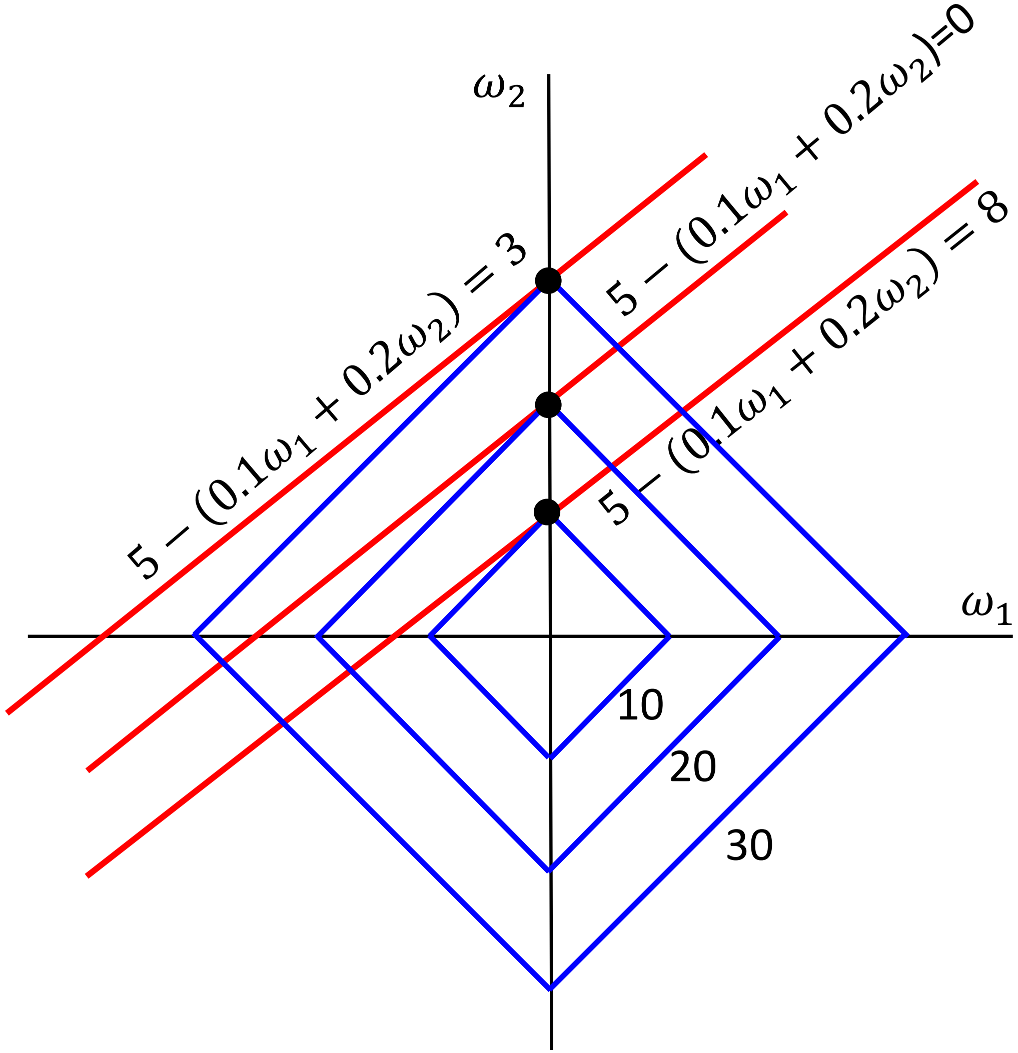




正则化: LASSO

$$\min_{\boldsymbol{\omega}} \|\boldsymbol{y} - \boldsymbol{X\omega}\|_2^2 + \lambda \|\boldsymbol{\omega}\|_1$$ ⇒ 解为稀疏向量



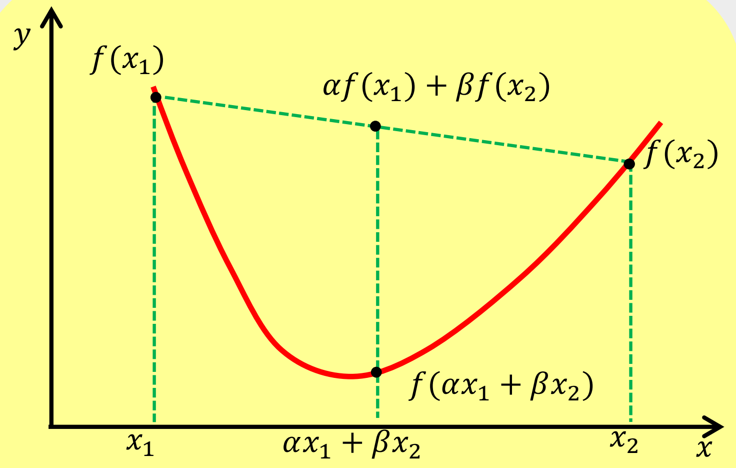


正则化: LASSO

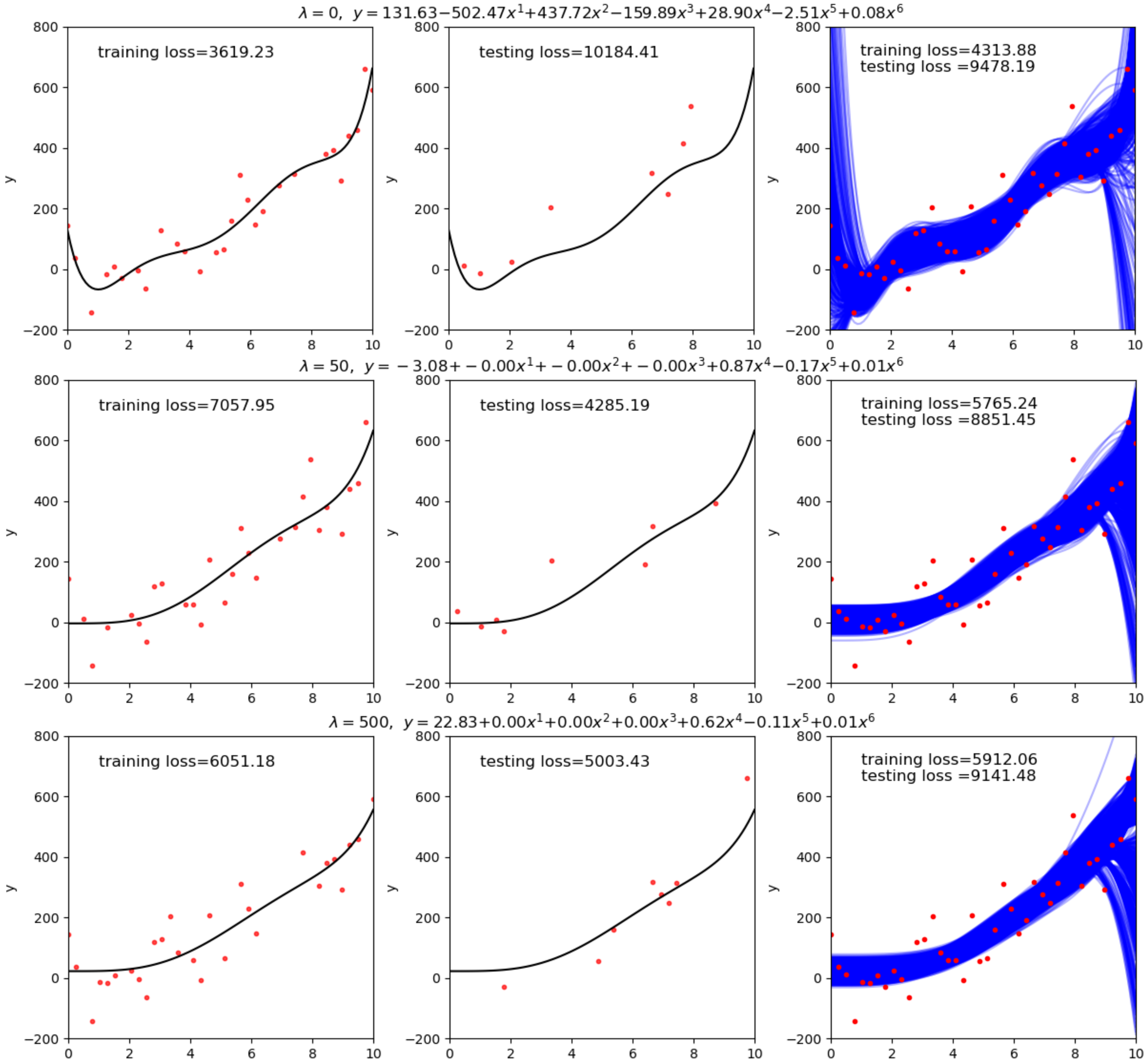
LASSO 的优化问题为

$$\min_{\boldsymbol{\omega}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega}\|_2^2 + \lambda\|\boldsymbol{\omega}\|_1$$

- 这个问题也被称作 **Basis Pursuit DeNoising (BPDN)**。
- 目标函数为**凸函数**，所以只有一个最优解。
- 最优解为一个稀疏向量，绝大多数元素为零。
- 无封闭解.
- 现有多种方法解决LASSO问题，包括
  - 二次规划算法(Quadratic programming)
  - Homotopy算法
  - ADMM (Alternating Direction Method of Multiplier)
  - ……



$$f(\alpha x_1 + \beta x_2) \leq \alpha f(x_1) + \beta f(x_2)$$
$$\alpha + \beta = 1, \alpha > 0, \beta > 0$$



LASSO $\min_{\omega} \frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_1$

训练样本数为 40

$y=b+\omega_1 x\cdots+\omega_6 x^6$

$\lambda=0$

训练损失函数值 = 4313

测试损失函数值= 9478

$\lambda=50$

训练损失函数值= 5765

测试损失函数值= 8851

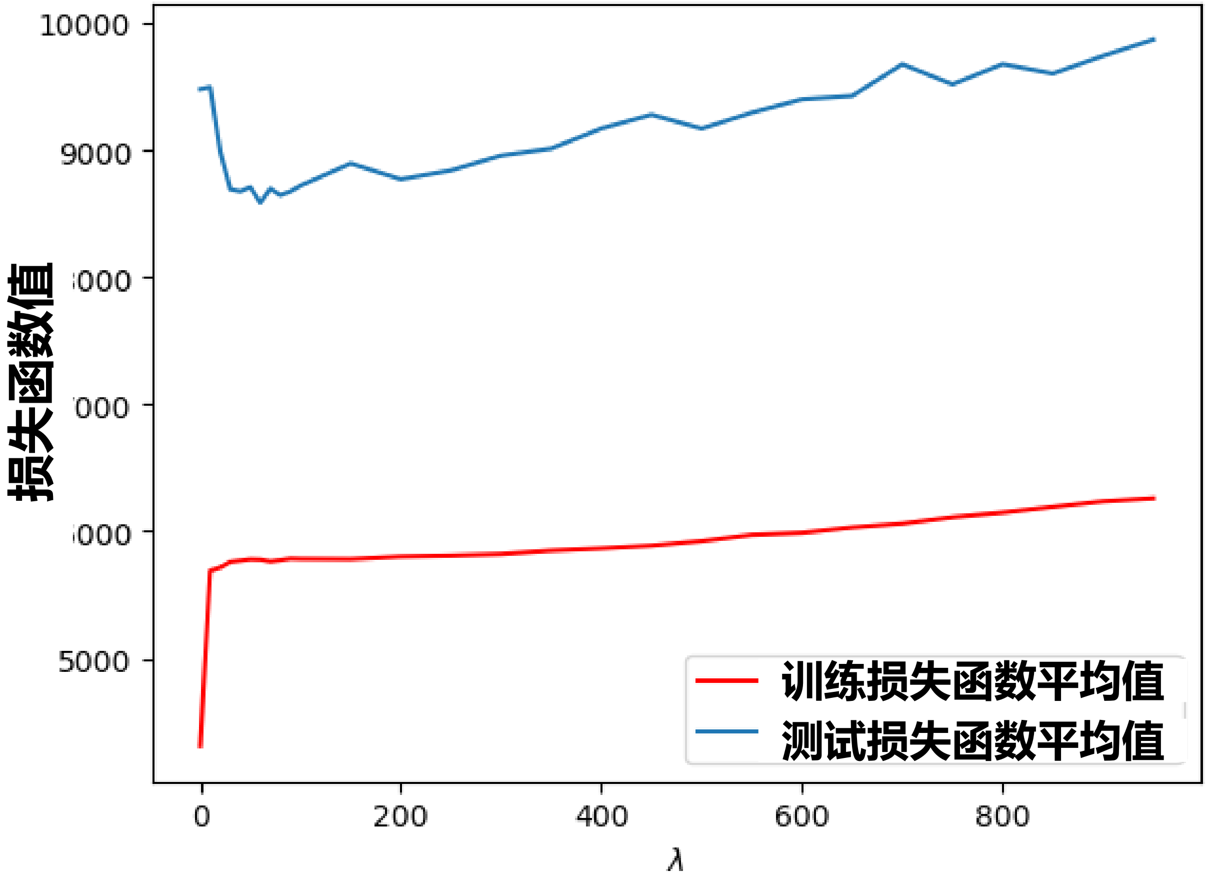
$\lambda=500$

训练损失函数值= 5912

测试损失函数值= 9141

正则化

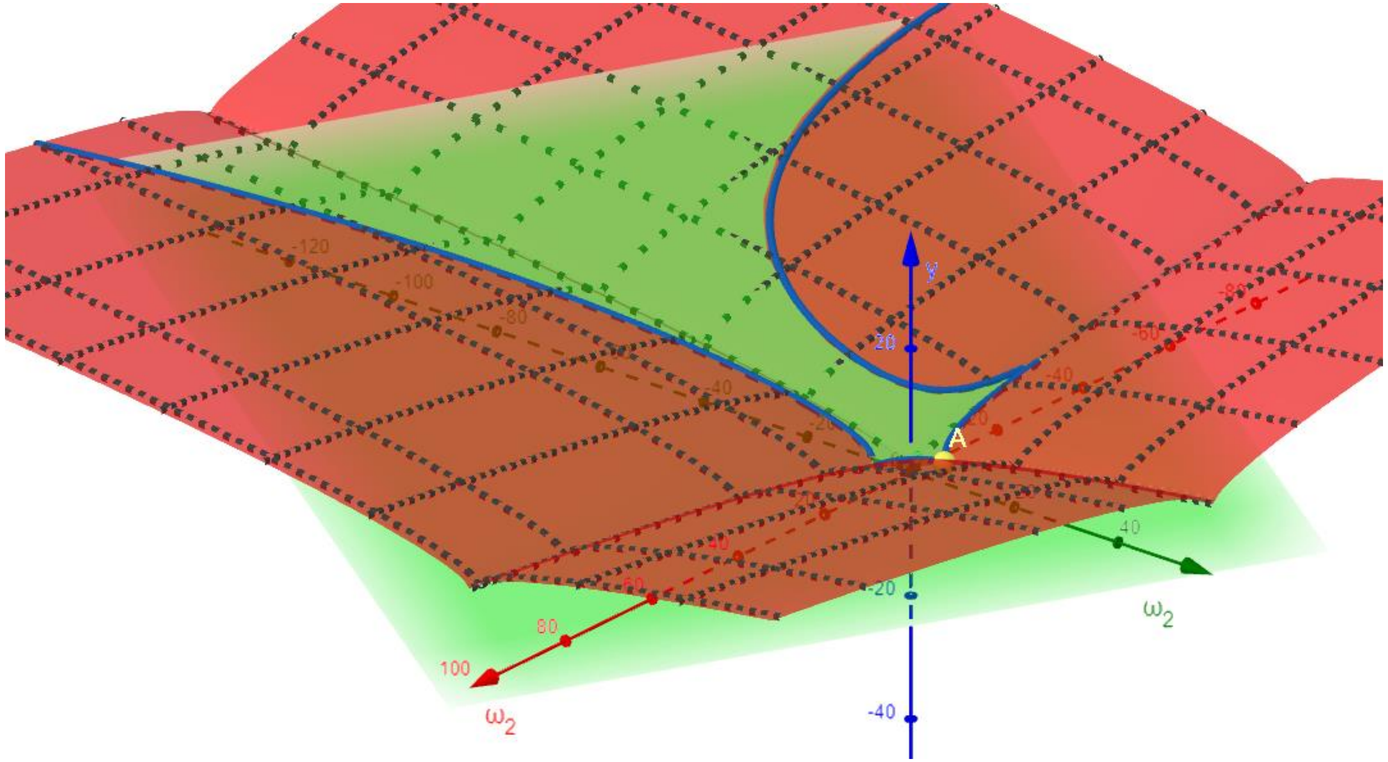
- **LASSO**

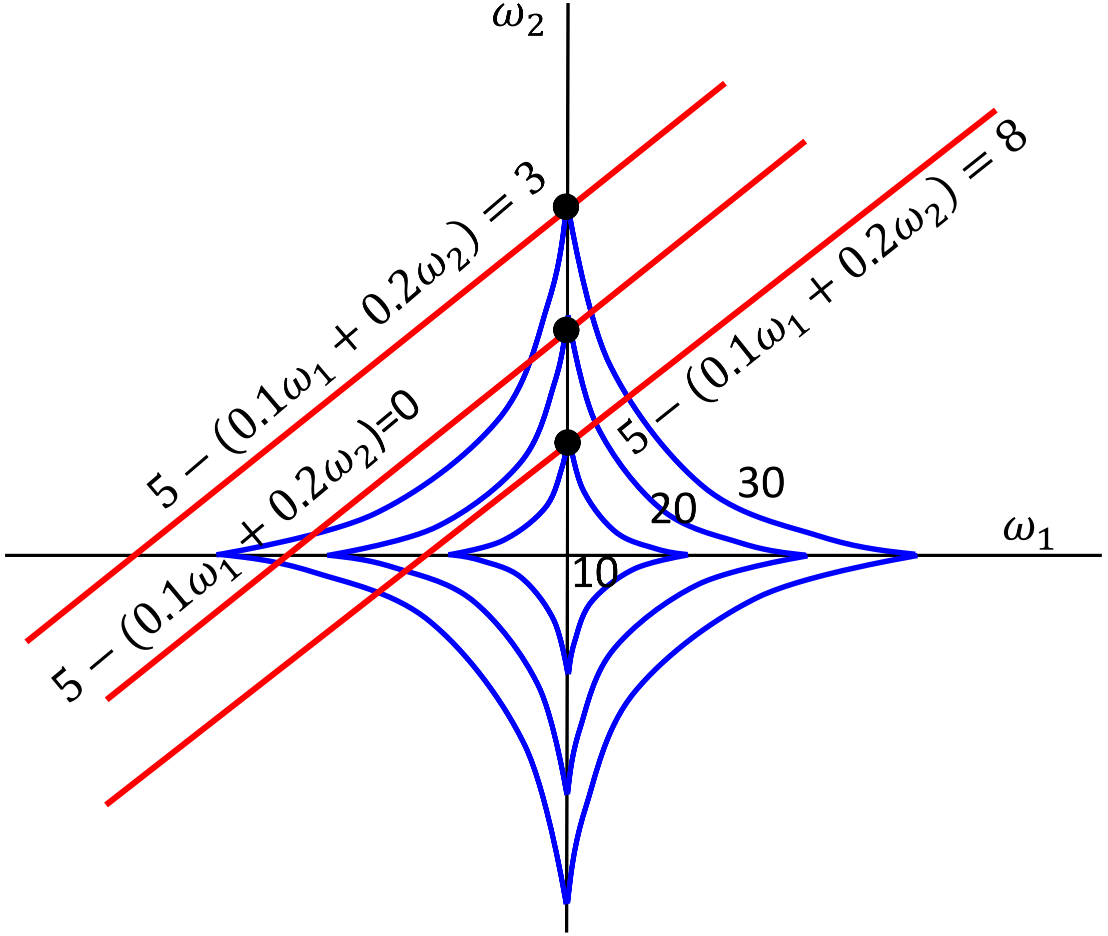


10000
9000
000
000
000
5000
0
200
400
600
800
λ
损失函数值
训练损失函数平均值
测试损失函数平均值

正则化: Lp 范数正则化项

$$\min_{\boldsymbol{\omega}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_p^p \longrightarrow \text{解为稀疏向量}$$





正则化: Lp 范数正则化项

带有Lp 正则化项的线性回归优化问题为

$$\min_{\boldsymbol{\omega}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\omega}\|_2^2+\lambda\|\boldsymbol{\omega}\|_p^p$$

- $0 < p \leq 1.$
- 目标函数为非凸函数.
- 解为稀疏向量，大部分元素为零.
- 无封闭解.
- 解决优化问题的方法为迭代类算法.