[redacted]

回归模型(1)

引言

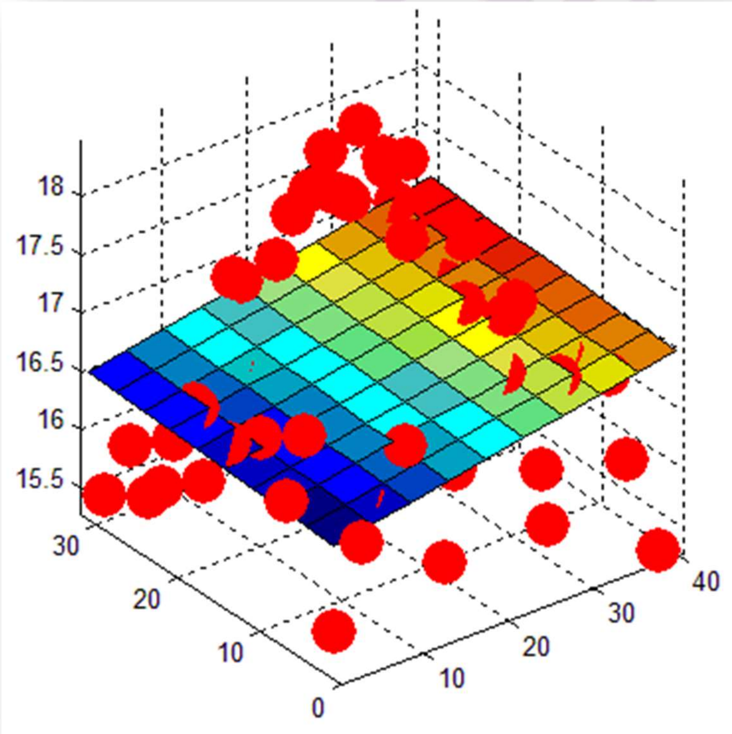
- 术语：回归
  - 趋中效应：如果父代身高高于平均值，则子代身高比他父代还高的概率较小，简单来说就是身高回归平均值。
- 回归问题：数据点沿着一条主轴来回波动的问题
- 回归模型：解决连续值预测问题的数学模型

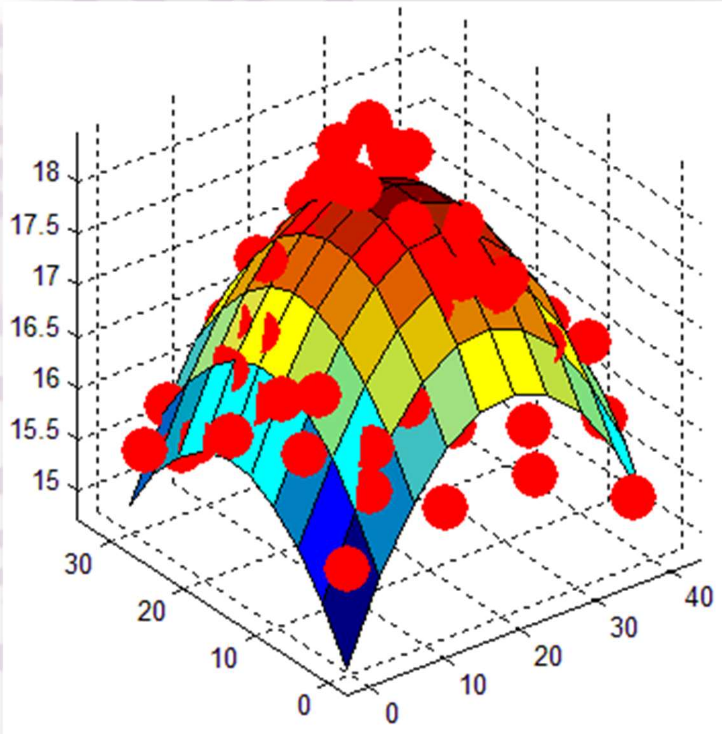
引言

- 回归分析：研究因变量和自变量之间关系的预测性建模技术
- 三个因素
  - 自变量的数目
  - 因变量的类型
  - 回归曲线的形状
- 种类
  - 线性回归(Linear Regression)
  - 逻辑回归(Logistic Regression)
  - 多项式回归(Polynomial Regression)
  - 逐步回归(Stepwise Regression)
  - 岭回归(Ridge Regression)
  - 套索回归(Lasso Regression)
  - 弹性网络回归(ElasticNet)

引言

- 线性回归(linear regression)是统计学习和机器学习的“御用工具”
- 使用核函数或基函数展开对线性回归进行扩展后，也可以对非线性关系建模





数学基础

- 期望和样本均值

$$\mathbb{E}_{X\sim P}\left\{X\right\}=\mathbb{E}\left\{X\right\}\approx\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

其中：每个 x_i 是来自分布 P 的样本。

- 方差和样本方差

$$Var\left\{X\right\}=\mathbb{E}\left\{\left(X-\mathbb{E}\left\{X\right\}\right)^2\right\}\approx\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$$

数学基础

- 协方差和样本协方差

$$Cov\{X_1, X_2\} = \mathbb{E}\{(X_1 - \mathbb{E}\{X_1\})(X_2 - \mathbb{E}\{X_2\})\}$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)(x_{2i} - \overline{x}_2)$$

其中：(x_{1i}, x_{2i}) 是来自分布 P 的第 i 个联合样本

- 条件期望

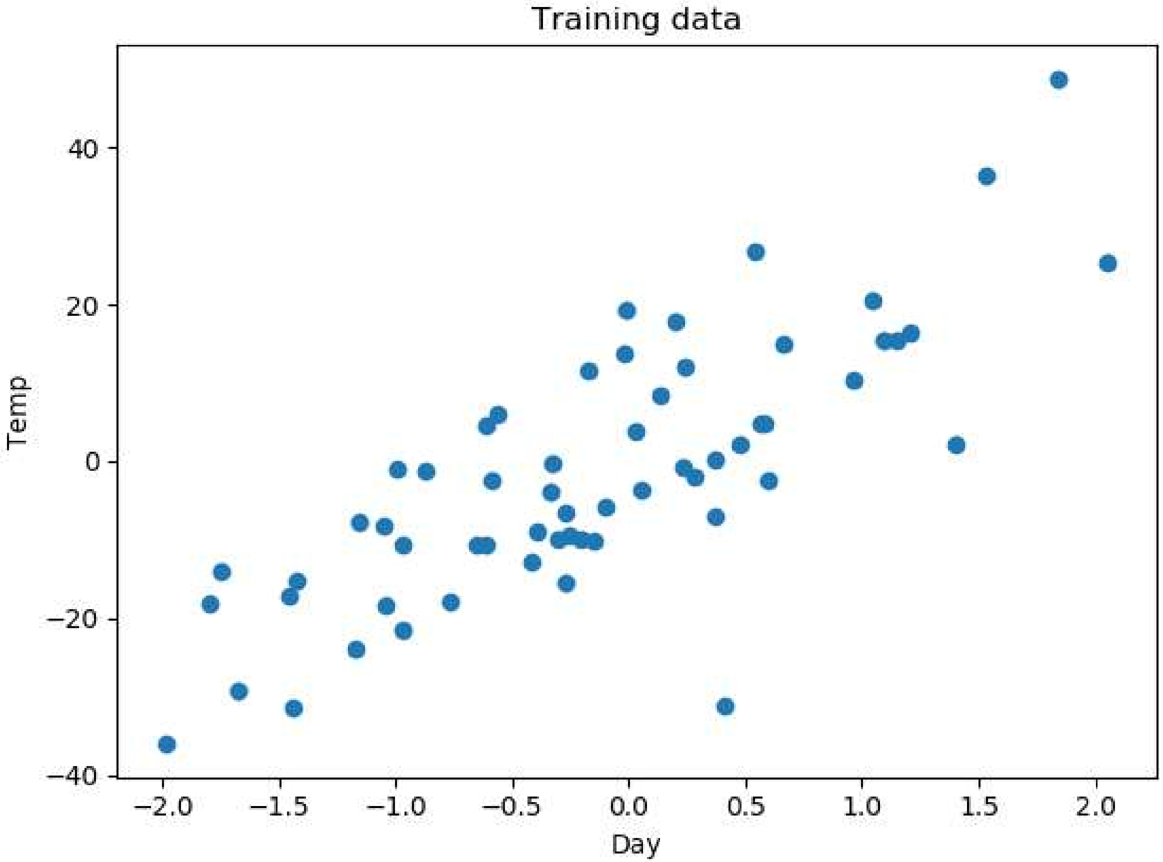
$$\mathbb{E}\{Y|x\} = \int_y yp(y|x)dy$$

$$\mathbb{E}\{\mathbb{E}\{Y|X\}\} = \mathbb{E}\{Y\}$$

其中： X 和 Y 是遵循分布 p 的随机变量； x 是 X 的可能取值。

回归

- **目标**：预测输入的响应/输出
- 考虑
  - 回归曲线的形状：考虑预测类型
  - 拟合标准（损失）：衡量对数据的拟合程度



线性回归

- 单变量线性函数（两个参数）

$$f(x;\mathbf{w}) = w_0 + w_1 x,\ \ \mathbf{w} = \begin{bmatrix} w_0 & w_1 \end{bmatrix}^{\mathrm{T}}$$

平方损失：

$$\mathrm{Loss}\left(y, f(x;\mathbf{w})\right) = \frac{1}{2}\left(y - f(x;\mathbf{w})\right)^2$$

$$= \frac{1}{2}\left(y - w_0 - w_1 x\right)^2$$

- 损失函数

$$J_n(\mathbf{w}) = \sum_{i=1}^{n} \mathrm{Loss}\left(y_i, f(x_i;\mathbf{w})\right)$$

线性回归

- 最小化平方损失

$$J_n(\mathbf{w})=\sum_{i=1}^{n}\mathrm{Loss}\big(y_i,f(x_i;\mathbf{w})\big)=\sum_{i=1}^{n}\frac{1}{2}(y_i-w_0-w_1x_i)^2$$

- 令对 w_0 和 w_1 的偏导数为零，获取最优参数值的必要条件

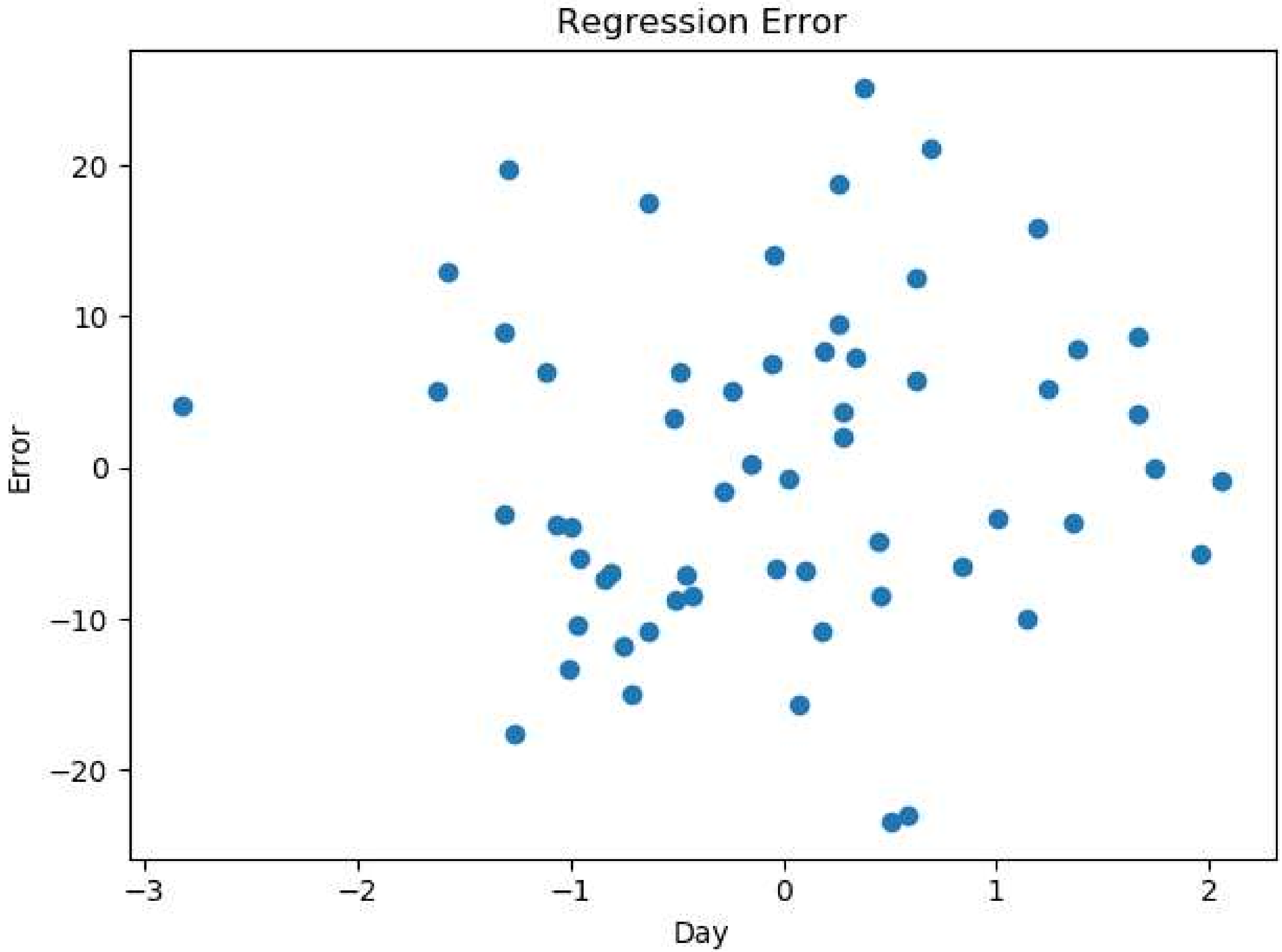
$$\frac{\partial}{\partial w_0}J_n(\mathbf{w})=-\sum_{i=1}^{n}(y_i-w_0-w_1x_i)=0$$

$$\frac{\partial}{\partial w_1}J_n(\mathbf{w})=-\sum_{i=1}^{n}(y_i-w_0-w_1x_i)x_i=0$$

注意：这些条件意味着预测误差 $(y_i-w_0-w_1x_i)$ 是零均值，而且与输入 x_i 不相关。

线性回归

❖ 预测误差与输入不相关



线性回归

- 通过逆矩阵求解

$$\frac{\partial}{\partial w_0} J_n(\mathbf{w}) = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) = 0$$

$$\frac{\partial}{\partial w_1} J_n(\mathbf{w}) = -\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) x_i = 0$$

$$w_0 \left(\sum_{i=1}^{n} 1 \right) + w_1 \left(\sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} y_i$$

$$w_0 \left(\sum_{i=1}^{n} x_i \right) + w_1 \left(\sum_{i=1}^{n} x_i^2 \right) = \sum_{i=1}^{n} y_i x_i$$

线性回归

- 用矩阵表示，最小化

$$\frac{1}{2}\left\|\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}\right\|^2 = \frac{1}{2}\left\|\mathbf{y} - \mathbf{X}\mathbf{w}\right\|^2$$

  - 令导数为零，求解正规方程（最小二乘法的矩阵形式）

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{0} \Rightarrow \boxed{\hat{\mathbf{w}} = \underbrace{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}}_{\Phi} \underbrace{\mathbf{X}^{\mathrm{T}}\mathbf{y}}_{\mathbf{b}}}$$

  - 注意：解 $\hat{\mathbf{w}}$ 是输出 y 的线性函数

线性回归

- 表示成矩阵形式 $\Phi\mathbf{w}=\mathbf{b}$，其中

$$\Phi=\begin{bmatrix}\sum_{i=1}^{n}1 & \sum_{i=1}^{n}x_i\\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2\end{bmatrix},\ \mathbf{b}=\begin{bmatrix}\sum_{i=1}^{n}y_i\\ \sum_{i=1}^{n}y_ix_i\end{bmatrix}$$

- 如果矩阵 Φ 是可逆的，则可获得参数估计的解 $\hat{\mathbf{w}}=\Phi^{-1}\mathbf{b}$

几何解释

- 假设：样本数目大于特征维数，即：$N > D$
- 样本矩阵

$$\mathbf{X}_{N\times(D+1)} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,D} \\ 1 & x_{2,1} & \cdots & x_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,D} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

  - 列矢量定义一个 D+1 维线性子空间，第 j 列 $\tilde{\mathbf{x}}_j$ 是 N 维实空间中的矢量
  - $\mathbf{y}$ 也是 N 维实空间中的一个矢量
- 在该线性子空间中，寻找与 $\mathbf{y}$ 尽可能近的矢量 $\hat{\mathbf{y}}$ ，即

$$\hat{\mathbf{y}} = \underset{\tilde{\mathbf{y}}\in\text{span}(\{\tilde{\mathbf{x}}_0,\ldots,\tilde{\mathbf{x}}_D\})}{\arg\min} \|\mathbf{y}-\tilde{\mathbf{y}}\|_2$$

几何解释

- 因为 $\tilde{\mathbf{y}} \in \mathrm{span}(\mathbf{X})$，故存在着一个矢量，满足

$$\tilde{\mathbf{y}} = w_0 \tilde{\mathbf{x}}_0 + w_1 \tilde{\mathbf{x}}_1 + \cdots + w_D \tilde{\mathbf{x}}_D = \mathbf{X}\mathbf{w}$$

- 为了最小化残差矢量的模 $\|\mathbf{y} - \tilde{\mathbf{y}}\|$，令残差矢量与 $\mathbf{X}$ 的列矢量均是正交的，有

$$\tilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{y} - \tilde{\mathbf{y}}) = 0 \quad \text{for } j = 0 : D$$

因此

$$\tilde{\mathbf{x}}_j^{\mathrm{T}} (\mathbf{y} - \hat{\mathbf{y}}) = 0 \Rightarrow \mathbf{X}^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\mathbf{w}) = 0 \Rightarrow \mathbf{w} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- $\mathbf{y}$ 到 $\mathbf{X}$ 列矢量张开空间的正交投影

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

其中，投影矩阵 $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}$ 称为帽子矩阵(hat matrix)

几何解释

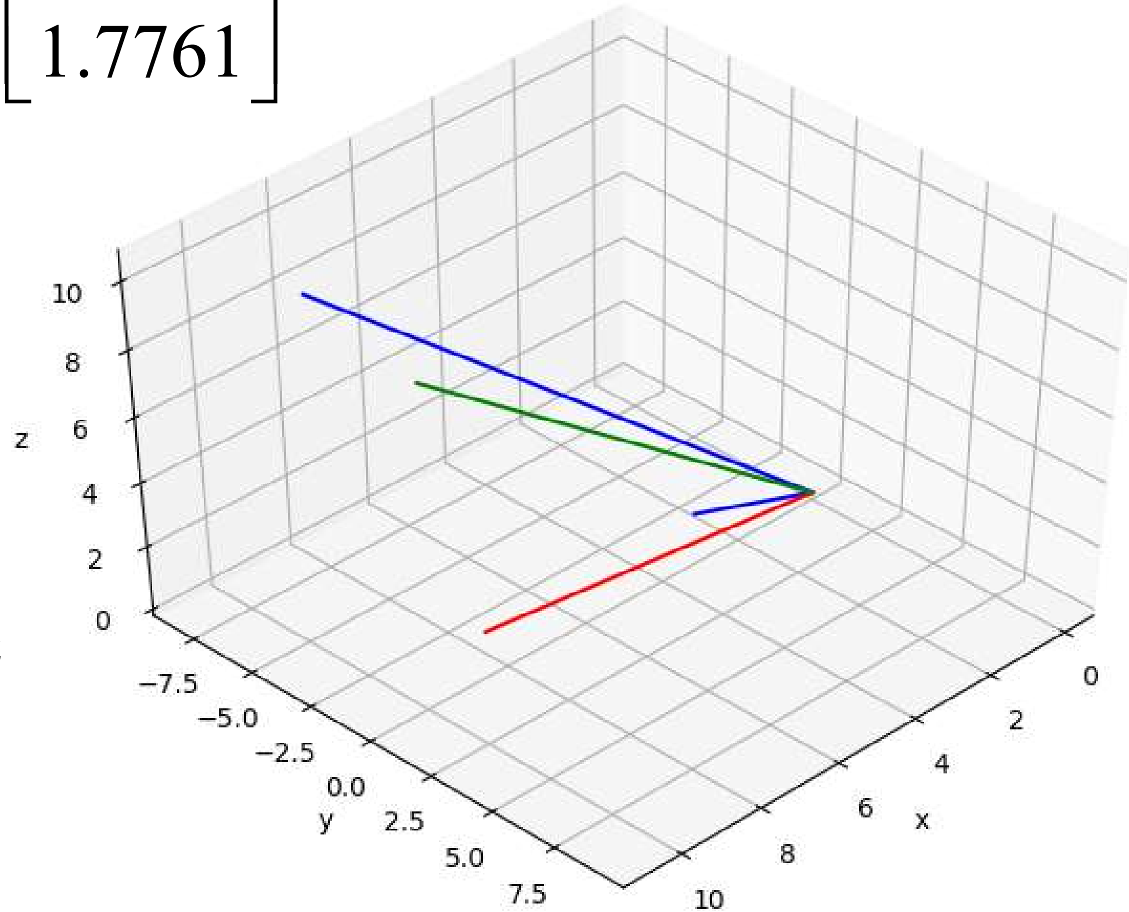
- 例：$N=3, D=1$

$$\mathbf{X}=\begin{bmatrix}1 & 2\\1 & -2\\1 & 2\end{bmatrix},\ \mathbf{y}=\begin{bmatrix}8.8957\\0.6130\\1.7761\end{bmatrix}$$

- 解：

$$\hat{\mathbf{y}}=\begin{bmatrix}5.3359\\0.6130\\5.3359\end{bmatrix}$$

  - 蓝色为 X 表示两条直线
  - 红色为 y 表示的直线
  - 绿色为 y 的正交投影



线性回归

- 矩阵的秩
  - 矩阵的列秩就是矩阵中线性独立列的最大数目，行秩定义相同
  - 方阵的列秩和行秩总是相等，简称矩阵的秩，表示为rank(A)
- 逆矩阵
  - 一个矩阵存在逆矩阵的前提是该矩阵是一个满秩的方阵
  - 如果矩阵 A 是满秩方阵，则有$\det(\mathbf{A}) \neq 0$，且矩阵 A 存在逆矩阵
- 伪逆矩阵
  - 奇异矩阵和非方阵不存在逆矩阵
  - 假设，矩阵 A 的奇异值分解是：

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$$

则，矩阵 A 的伪逆矩阵 $\mathbf{A}^{\dagger}$ 为：

$$\mathbf{A}^{\dagger} = \mathbf{V}\Sigma^{\dagger}\mathbf{U}^{\mathrm{T}}$$

线性回归

- 对于矩阵 A，其伪逆矩阵 $\mathbf{A}^{\dagger}$ 与 $\mathbf{A}^{\mathrm{T}}$ 同型，且有如下四个性质：

$$\begin{cases} \mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A} \\ \mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger} \\ \mathbf{A}\mathbf{A}^{\dagger} = \left(\mathbf{A}\mathbf{A}^{\dagger}\right)^{\mathrm{T}} \\ \mathbf{A}^{\dagger}\mathbf{A} = \left(\mathbf{A}^{\dagger}\mathbf{A}\right)^{\mathrm{T}} \end{cases}$$

- 二维样本空间：$\hat{\mathbf{w}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{\dagger}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

$$y_i \approx f\left(\mathbf{x}_i;\hat{\mathbf{w}}\right) = \hat{w}_0 + \hat{w}_1 x_{1i} + \hat{w}_2 x_{2i} = \hat{\mathbf{w}}^{\mathrm{T}} \begin{bmatrix} 1 \\ x_{1i} \\ x_{2i} \end{bmatrix}$$

- 计算得到的解在样本张开的子空间中（在正交的维上权矢量为零）

线性回归的统计观点

- 当参数矢量 $\mathbf{w}$ 确定后，预测结果 $\mathbf{w}^{\mathrm{T}}\mathbf{x}_i$ 和真实值 y_i 满足：

$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \varepsilon_i$$

  其中： ε_i 是预测误差。

- 如果误差满足均值为零的正态分布，那么 $\mathbf{x}_i$ 和 y_i 的条件概率可以表示为

$$p\left(y_i \middle| \mathbf{x}_i;\mathbf{w}\right) = \mathcal{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i, \sigma^2\right)$$

- 因为样本是独立同分布的，故样本集合的概率密度可以表示为每个样本概率密度之积的形式

$$p\left(\mathcal{D} \middle| \mathbf{w}\right) = \prod_{i=1}^{N} p\left(y_i \middle| \mathbf{x}_i;\mathbf{w}\right)$$

- 最小化模型预测误差的参数就是最大化样本集合概率的参数

最大似然估计

- 最大似然估计：统计模型参数估计的常用方法

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p\left(\mathcal{D} \middle| \mathbf{w}\right)$$

- 为了计算方便，通常表示为对数似然的形式

$$\ell\left(\mathbf{w}\right) = \log p\left(\mathcal{D} \middle| \mathbf{w}\right) = \sum_{i=1}^{N} \log p\left(y_i \middle| \mathbf{x}_i; \mathbf{w}\right)$$

- 最大似然估计等价于最小化负对数似然(NLL)

$$\text{NLL}\left(\mathbf{w}\right) = -\sum_{i=1}^{N} \log p\left(y_i \middle| \mathbf{x}_i; \mathbf{w}\right)$$

最大似然估计

- 假设似然函数为高斯函数，则有

$$\mathrm{NLL}(\mathbf{w}) = -\sum_{i=1}^{N} \log\left[\left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)^2\right)\right]$$

$$= \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)^2 + \frac{N}{2}\log\left(2\pi\sigma^2\right)$$

- 定义：残差平方和(residual sum of squares, RSS)，也称为误差平方和(sum of squared errors, SSE)

$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)^2$$

- 最小负对数似然估计等价于最小残差平方和估计

均方误差

- 定义：均方误差(mean squared error, MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)^2$$

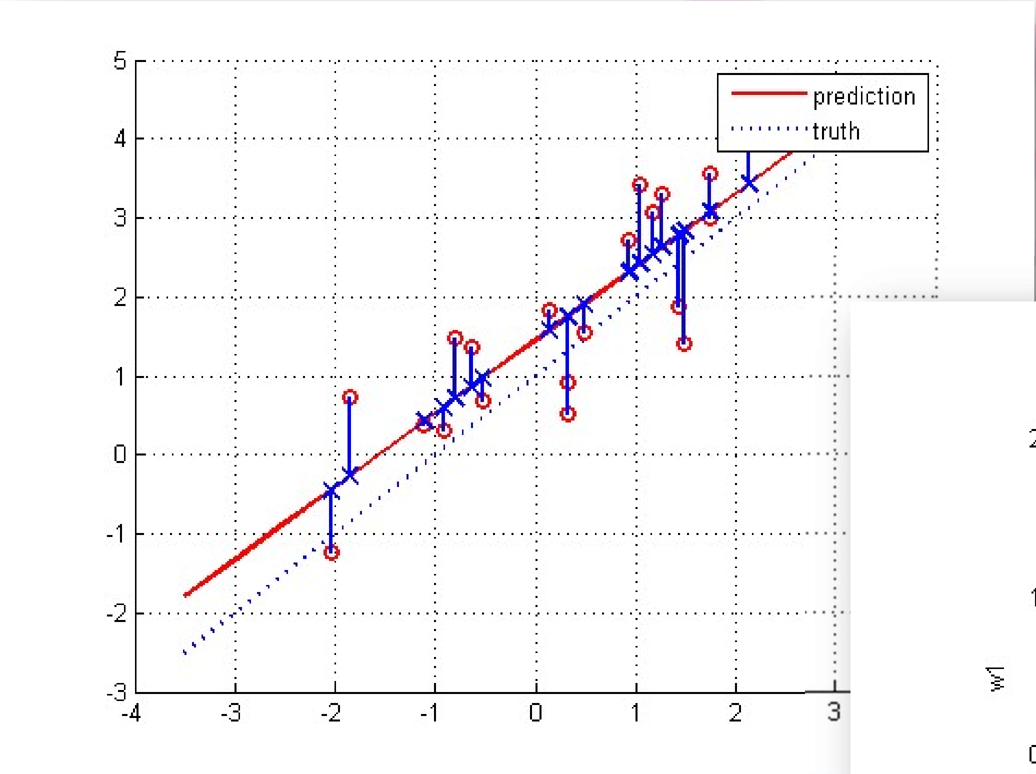
或

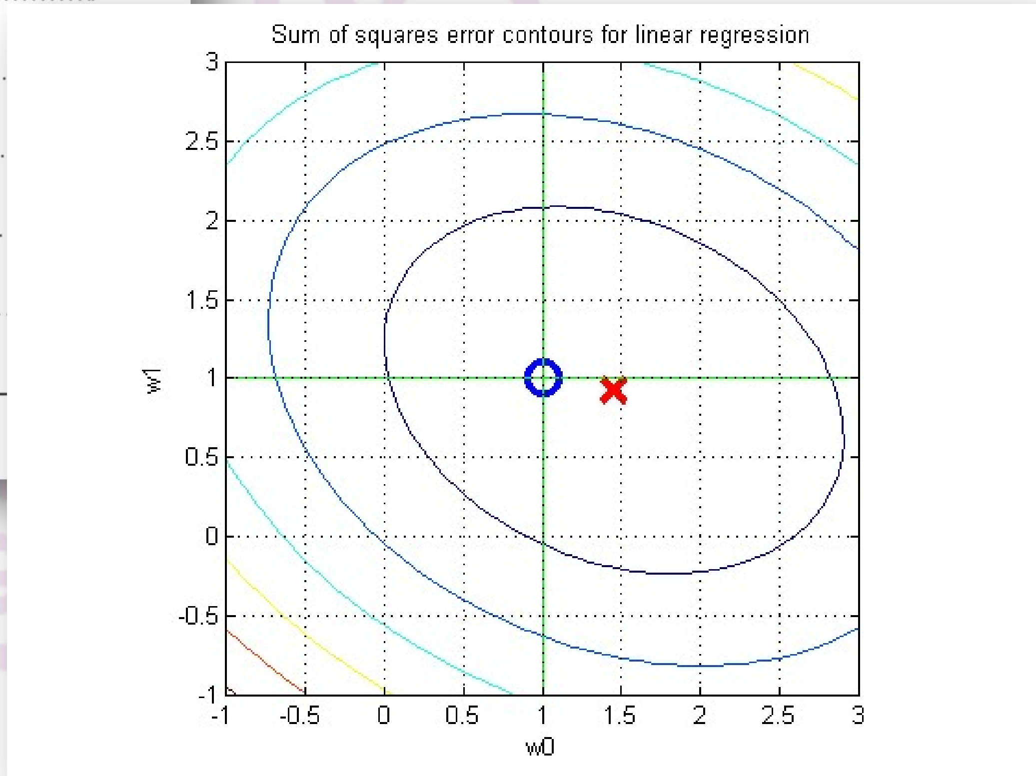
$$MSE = \frac{1}{N}\|\boldsymbol{\varepsilon}\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2$$

其中 $\varepsilon_i = y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i$

最小均方误差法

- 最小均方误差法：通过最小化 MSE 可以得到对参数 w 的最大似然估计





最大似然估计

❖ 目标函数：

$$\mathrm{RSS}(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\mathbf{w})$$

$$= \frac{1}{2}\mathbf{w}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{w} - \mathbf{w}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)$$

其中：

$$\mathbf{X}_{N\times(D+1)} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,D} \\ 1 & x_{2,1} & \cdots & x_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,D} \end{bmatrix}, \quad \mathbf{y}_{N\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

最大似然估计

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}=\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}=\sum_{i=1}^{N}\begin{bmatrix}1 & x_{i,1} & \cdots & x_{i,D}\\ x_{i,1} & x_{i,1}^2 & \cdots & x_{i,1}x_{i,D}\\ \vdots & \vdots & \ddots & \vdots\\ x_{i,D} & x_{i,D}x_{i,1} & \cdots & x_{i,D}^2\end{bmatrix}$$

是平方和矩阵，且

$$\mathbf{X}^{\mathrm{T}}\mathbf{y}=\sum_{i=1}^{N}\mathbf{x}_i y_i$$

最大似然估计

- RSS(w)的梯度

$$\mathbf{g}(\mathbf{w}) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - \mathbf{X}^{\mathrm{T}}\mathbf{y}\right) = \sum_{i=1}^{N} \mathbf{x}_i \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i\right)$$

- 令梯度为零，得到正规方程(normal equation)

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- 求解正规方程，得到一般最小二乘解

$$\boxed{\hat{\mathbf{w}}_{\mathrm{OLS}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}}$$

- 局限性：噪声不总是服从高斯分布的！

简单推导

- 考虑权值为$\hat{\mathbf{w}}$ 的对数似然函数

$$\mathrm{L}\left(\mathcal{D};\hat{\mathbf{w}},\sigma^2\right)=-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)^2-\frac{N}{2}\log\left(2\pi\sigma^2\right)$$

  - 前期我们并没有考虑联合最大化似然函数

- 现在考虑

$$\frac{\partial}{\partial\sigma^2}\mathrm{L}\left(\mathcal{D};\hat{\mathbf{w}},\sigma^2\right)=\frac{1}{2\sigma^4}\sum_{i=1}^{N}\left(y_i-\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{x}_i\right)^2-\frac{N}{2\sigma^2}=0$$

  - 方程两边乘上$2\sigma^4/N$ 就可得到解

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)^2$$

交叉验证和对数似然

- 余一交叉验证对数似然函数

$$\mathrm{CV} = \sum_{i=1}^{N} \log p\left(y_i \middle| \mathbf{x}_i, \hat{\mathbf{w}}^{-i}, \left(\hat{\sigma}^2\right)^{-i} \right)$$

其中： $\hat{\mathbf{w}}^{-i}$ 和 $(\hat{\sigma}^2)^{-i}$ 是除去第 i 个样本 $(\mathbf{x}_i, y_i)$ 计算的最大似然估计。

凸性

- 定义：如果对于任意 $\theta,\theta' \in S$，存在

$$\lambda\theta+(1-\lambda)\theta' \in S \quad \forall\lambda\in[0,1]$$

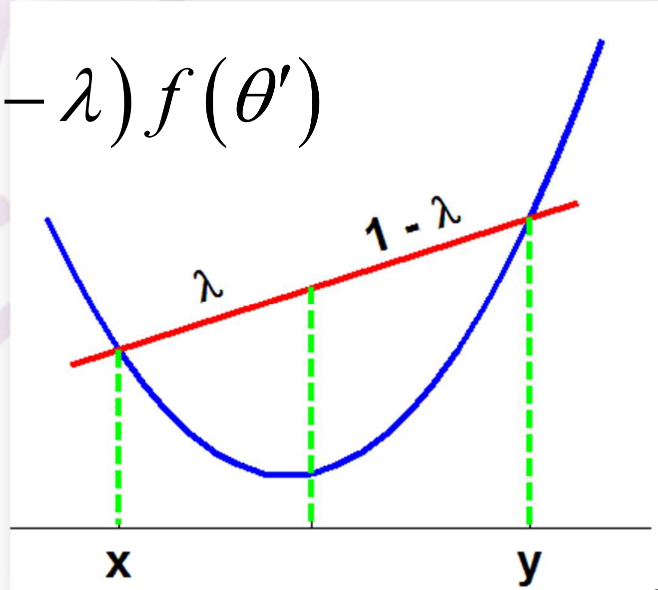
  则称集合 S 是凸的(convex)

- 定义：如果某个函数 $f(\theta)$ 上的点集(epigraph)定义了一个凸集，则该函数 $f(\theta)$ 是凸的

- 定义：如果某个函数定义在凸集上，且对于任意的 $\theta,\theta' \in S$ 和 $0\le\lambda\le 1$，有

$$f\left(\lambda\theta+(1-\lambda)\theta'\right)\le\lambda f(\theta)+(1-\lambda)f(\theta')$$

  则称函数 $f(\theta)$ 是凸的



凸性

- 定义：如果函数 $-f(\theta)$ 是凸的，则函数 $f(\theta)$ 是凹的(concave)
  - 标量凸函数：$\theta^2, e^\theta, \theta\log\theta(\text{for } \theta>0)$
  - 标量凹函数：$\log\theta, \sqrt{\theta}$
- 严格的凸函数呈碗状，在碗底具有唯一全局最小值 θ^*
- 定义：当且仅当，对任意多变量函数 $f(\theta)$，其 Hessian 矩阵是正定的且二阶连续可导，则该函数 $f(\theta)$ 是凸的。
- 期望模型的 NLL 是凸的，故总可以发现全局最优的 MLE

Hessian矩阵

- 定义：Hessian矩阵是一个自变量为矢量的实数函数的二阶偏导数组成的方阵
- 如果实数函数为$f(x_1,x_2,\ldots,x_n)$，则其 Hessian 矩阵为

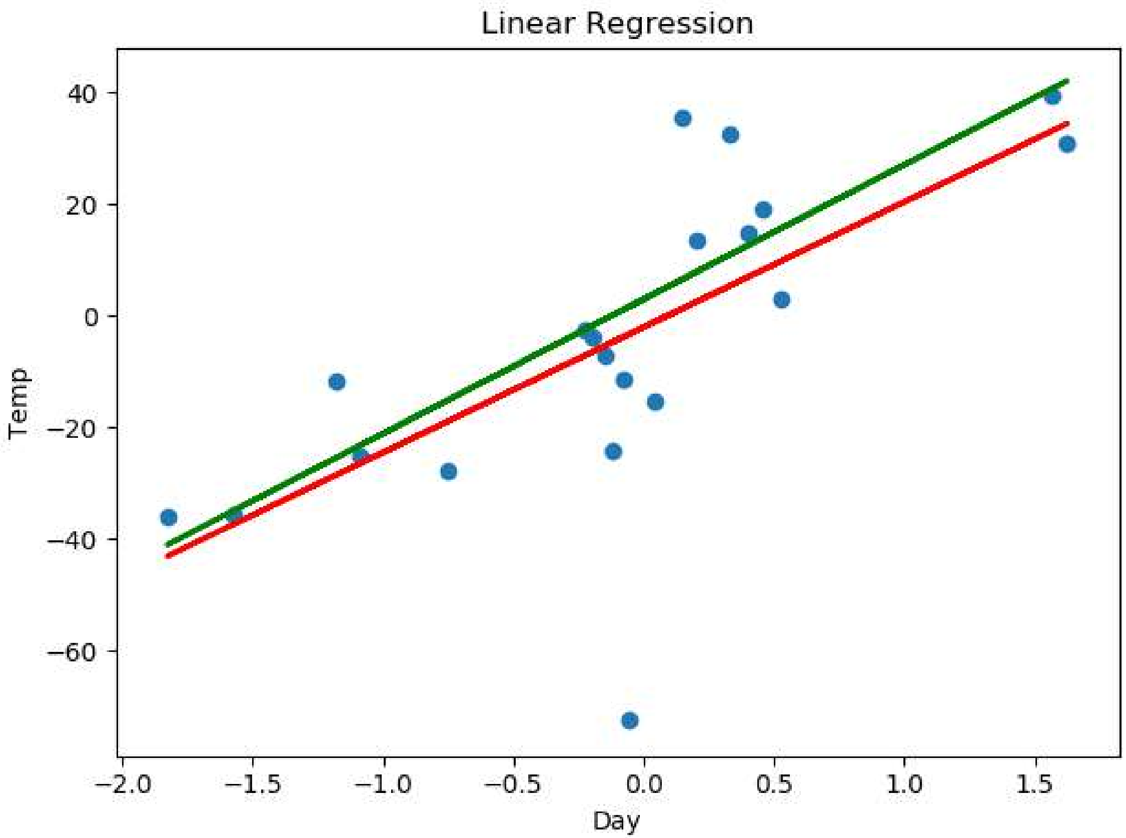
$$\mathbf{H}(f)=\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

鲁棒线性回归

- 噪声模型：零均值高斯分布

$$\varepsilon_i = y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i \sim \mathcal{N}(0,\sigma^2)$$

- 结果：最大似然估计等价于最小残差平方和估计
- 缺点：对含有离群点(outlier)的数据拟合较差



鲁棒线性回归

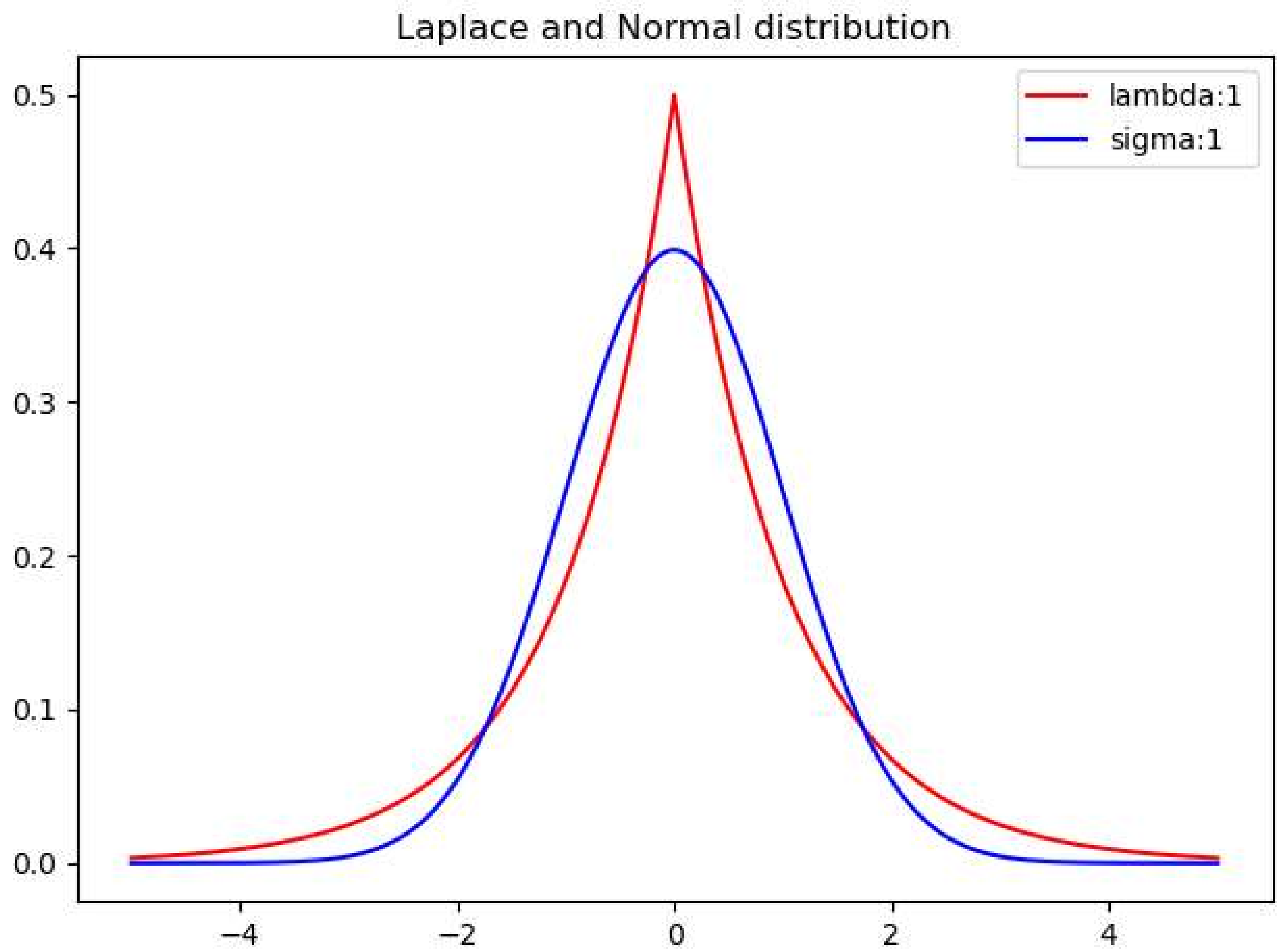
- 原因：平方误差是二次惩罚项，误差大的数据点对拟合性能的影响比误差小的数据点的影响大得多

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)^2$$

- 解决思路：降低离群点（相对误差大的数据点）对拟合性能的影响
- 如何实现?

鲁棒线性回归

- 方法一：使用重尾分布代替正态分布，让离群点具有更高的似然值



鲁棒线性回归

- 例：使用Laplace分布代替正态分布
  - Laplace分布

$$\mathrm{Laplace}(\mathbf{x})=\frac{1}{2\lambda}\exp\left(-\frac{1}{\lambda}|\mathbf{x}-\boldsymbol{\mu}|\right)$$

  - 似然函数

$$p(y|\mathbf{x};\mathbf{w})=\mathrm{Laplace}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x})\propto\exp\left(-\frac{1}{b}|y-\mathbf{w}^{\mathrm{T}}\mathbf{x}|\right)$$

  - 负对数似然

$$\mathrm{NLL}=\ell(\mathbf{w})=\sum_{i=1}^{N}|r_i(\mathbf{w})|=\sum_{i=1}^{N}|y_i-\mathbf{w}^{\mathrm{T}}\mathbf{x}_i|$$

  - 缺点：NLL是非线性函数，难于优化。怎么办?

鲁棒线性回归

- **问题转换**：将非线性目标函数转换为线性目标函数
  - 定义 $r_i = r_i^+ - r_i^-$，然后强加不等式约束 $r_i^+ \geq 0, r_i^- \geq 0$
  - 目标函数转换为

$$\min_{\mathbf{w},\mathbf{r}^+,\mathbf{r}^-} \sum_{i=1}^{N} \left(r_i^+ - r_i^-\right)$$

$$\text{s.t.} \quad r_i^+ \geq 0, r_i^- \geq 0, \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + r_i^+ - r_i^- = y_i$$

  - D+2N 个未知量和 3N 个约束的线性规划问题
- **求解**：线性规划问题
  - 标准形式

$$\min_{\boldsymbol{\theta}} \mathbf{f}^{\mathrm{T}}\boldsymbol{\theta}$$

$$\text{s.t.} \quad \mathbf{A}\boldsymbol{\theta} \leq \mathbf{b}, \mathbf{A}_{eq}\boldsymbol{\theta} = \mathbf{b}_{eq}, \mathbf{l} \leq \boldsymbol{\theta} \leq \mathbf{u}$$

  - 使用任何线性规划算法求解
  - 因为是凸优化问题，故具有唯一解。

鲁棒线性回归

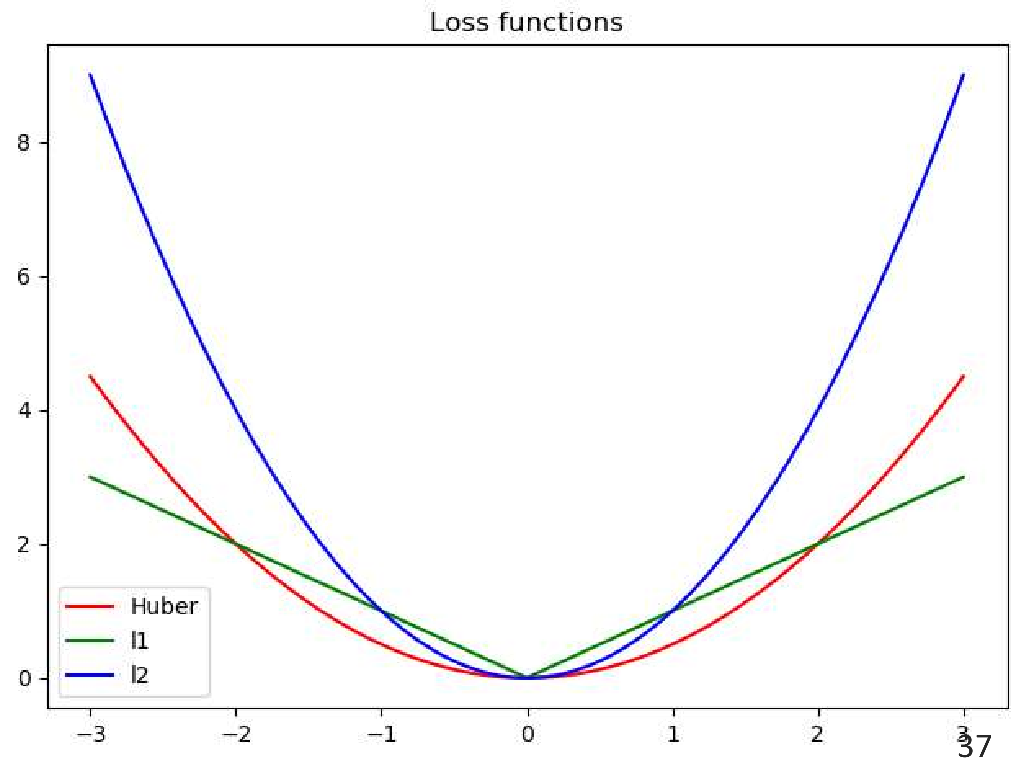
❖ 方法二：最小化 Huber 损失函数

$$L_{\text{Huber}}(r,\delta)=\begin{cases} r^2/2 & \text{if } |r|\le\delta \\ \delta|r|-\delta^2/2 & \text{otherwise} \end{cases}$$

- 误差绝对值小于等于 δ 时，等价于L2范数
- 误差绝对值大于 δ 时，等价于L1范数

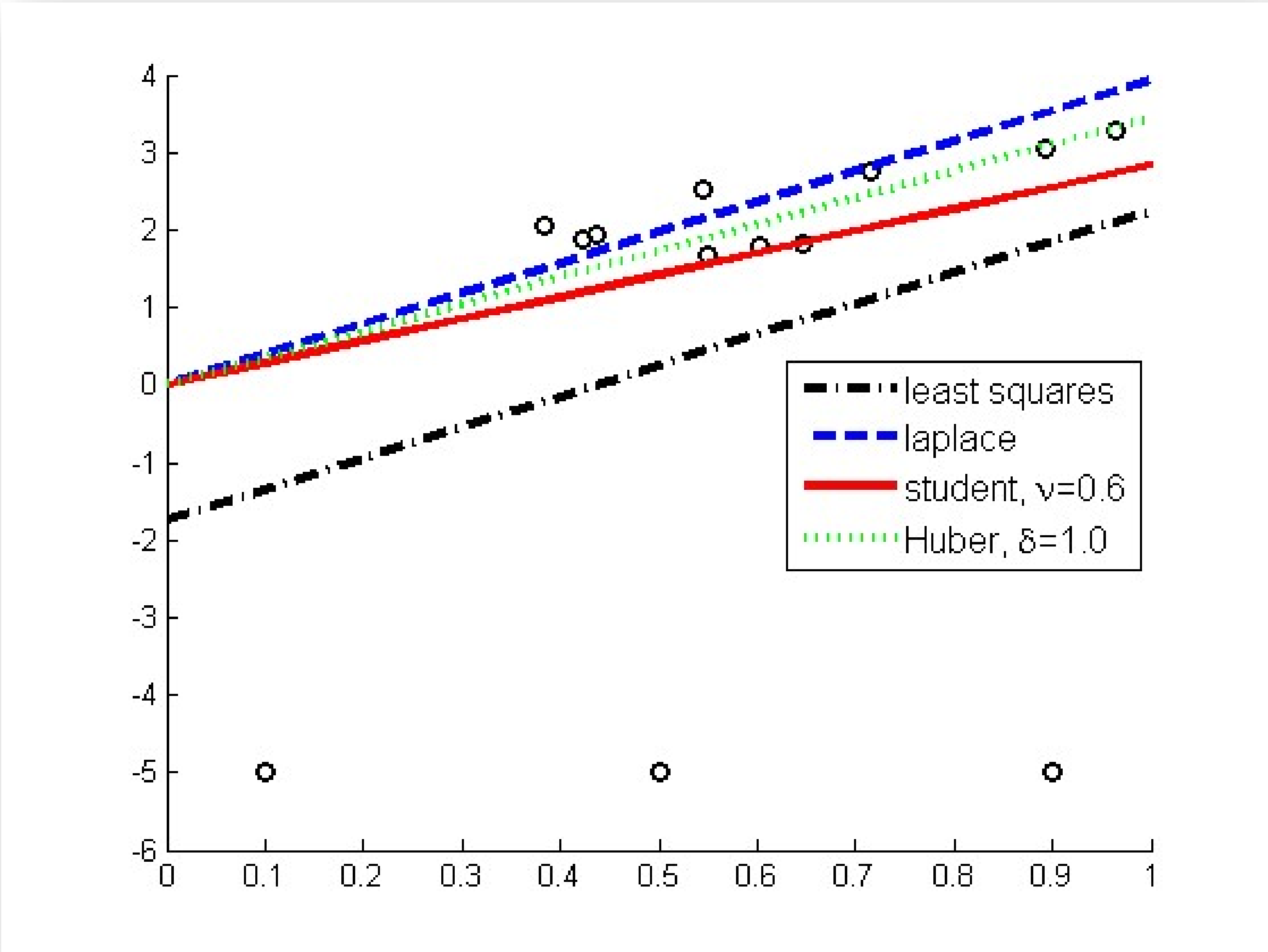
❖ 优点

- 损失函数处处可导
- 优化速度比 Laplace 似然方法快



鲁棒线性回归

❖ 几种不同方法的示意图



诚信　创新　实践