机器学习
Machine learning

第五章 回归分析
Regression

授课人：周晓飞
zhouxiaofei@iie.ac.cn
2020-11-12

第五章 回归分析

5.1 概述

5.2 最小二乘估计

5.3 最大似然估计

5.4 最大后验估计

5.5 扩展的非线性模型

5.6 误差分析

第五章 回归分析

5.1 概述

5.2 最小二乘估计

5.3 最大似然估计

5.4 最大后验估计

5.5 扩展的非线性模型

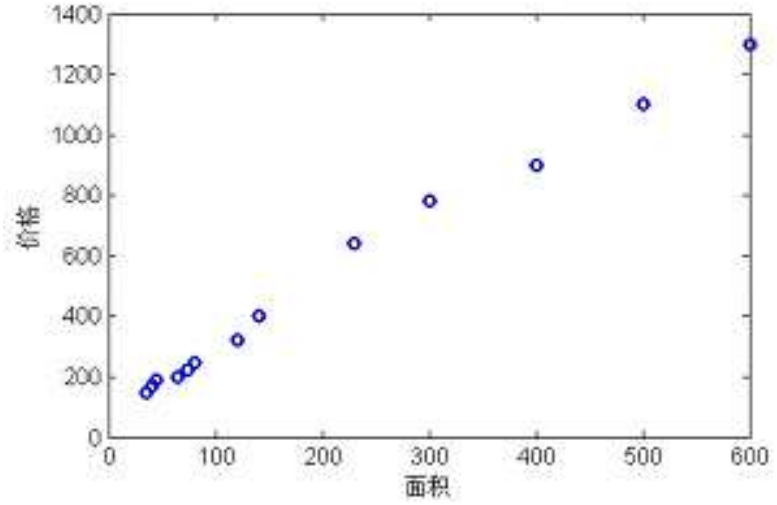
5.6 误差分析

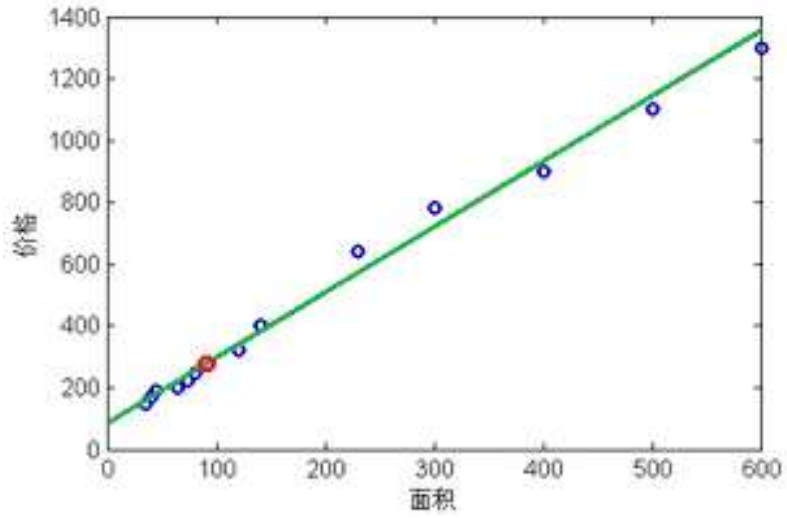
概述

例子

(房屋面积 m^2, 价钱(万元))

35	150
40	170
45	190
65	200
74	224
80	245
120	320
140	400
230	640
300	780
400	900
500	1100
600	1300
90	**?**





$y=ax+b$

概述

回归问题

定义：

根据给定的训练集

$$T=\{(\boldsymbol{x}_1,y_1),\cdots,(\boldsymbol{x}_l,y_l)\},$$

其中 $\boldsymbol{x}_i\in C=R^n$，$y_i\in \mathrm{Y}=R$，$i=1,2,...,l$，(预测的结果 y 是连续函数值)

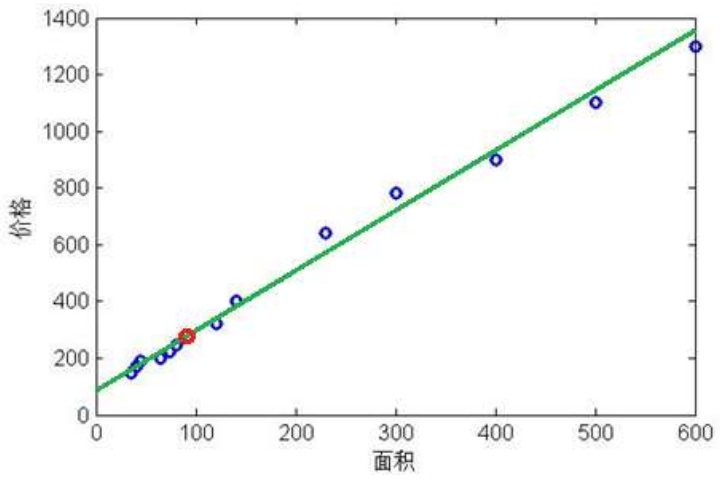
要求寻找 C 上的决策函数 $f(\boldsymbol{x}):\mathrm{C}\to\mathrm{Y}$

概述

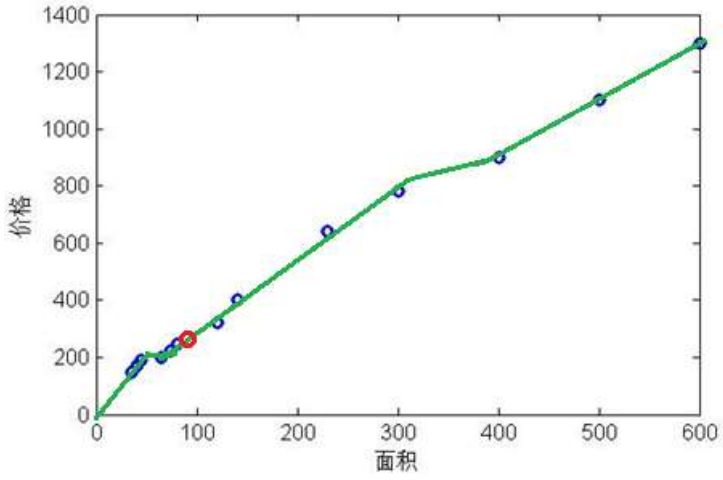
回归问题

线性与非线性

35	150
40	170
45	190
65	200
74	224
80	245
120	320
140	400
230	640
300	780
400	900
500	1100
600	1300
90	**?**



线性



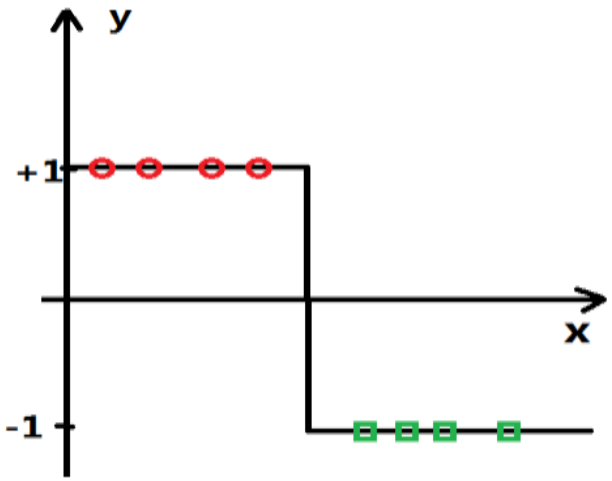
非线性

概述

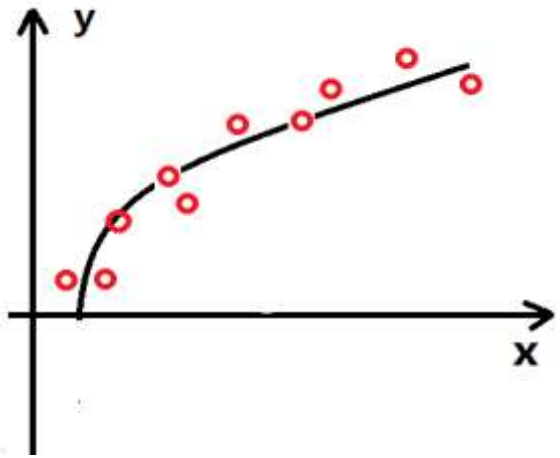
回归问题

vs.分类问题

预测值域不同，将预测值作为新增加维度，可示意：



分类



回归

回归问题

性能评价

回归任务最常用的性能度量是“均方误差”(mean squared error)

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2 \ .$$

更一般的, 对于数据分布 $\mathcal{D}$ 和概率密度函数 $p(\cdot)$, 均方误差可描述为

$$E(f;\mathcal{D})=\int_{\boldsymbol{x}\sim\mathcal{D}}\left(f\left(\boldsymbol{x}\right)-y\right)^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \ .$$

回归问题

偏差与方差

$$E(f;D) = bias^2(\boldsymbol{x}) + var(\boldsymbol{x}) + \varepsilon^2$$

泛化误差可分解为偏差、方差与噪声之和.

线性回归原理

线性函数来预测数据的分布

- 数学的描述：预测的“平均结果”与已有数据呈线性关系

$$E\{y|\mathbf{x}\} = \theta^T \mathbf{x} + \theta_0$$

- 对于(x, y), y 与 x 的关系

$$y = \theta^T \mathbf{x} + \theta_0 + \epsilon$$

- 回归问题：求取最优的线性函数

$$y = \theta^T \mathbf{x} + \theta_0$$

线性回归原理

一元、二元、多元回归

N 元对应变量 x 的维度.

$$h_\theta(x) = \theta_0 + \theta_1 x_1$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

… …

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

第五章 回归分析

5.1 概述

5.2 最小二乘估计

5.3 最大似然估计

5.4 最大后验估计

5.5 扩展的非线性模型

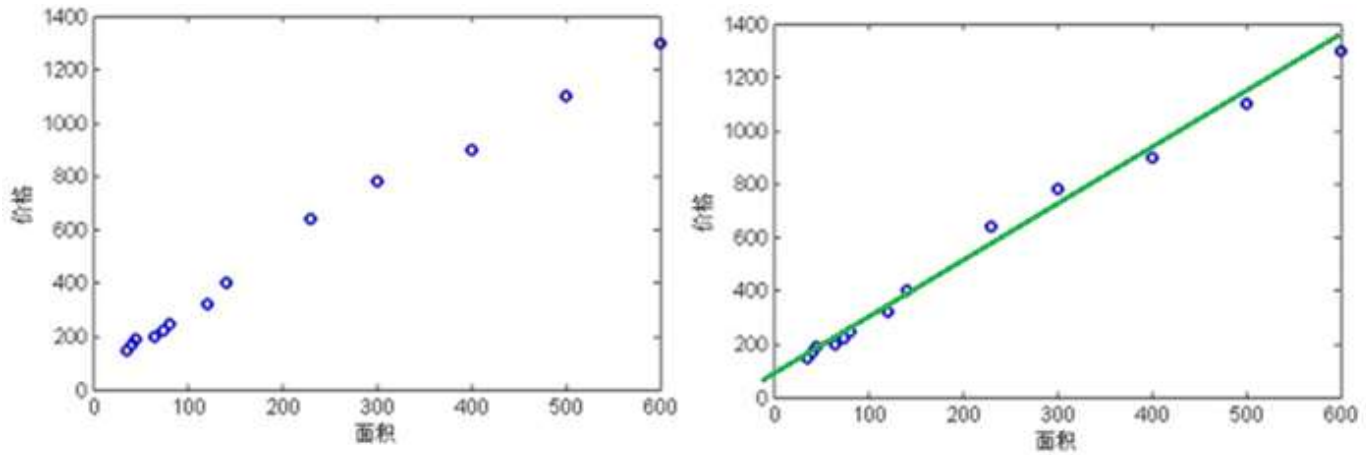
5.6 误差分析

最小二乘估计

Least Squares（最小二乘估计法）

目标函数：最小误差平方和

$$\min_{\theta} \sum_{t=1}^{n} (y_t - \theta^T \mathbf{x}_t - \theta_0)^2$$



最小二乘估计

Least Squares（最小二乘估计法）

求解过程

$$\begin{aligned}
\sum_{t=1}^{n}\left(y_t-\begin{bmatrix}\theta\\ \theta_0\end{bmatrix}^T\begin{bmatrix}\mathrm{x}_t\\ 1\end{bmatrix}\right)^2 &= \sum_{t=1}^{n}\left(y_t-[\mathrm{x}_t^T,1]\begin{bmatrix}\theta\\ \theta_0\end{bmatrix}\right)^2\\
&= \left\|\begin{bmatrix}y_1\\ \cdots\\ y_n\end{bmatrix}-\begin{bmatrix}\mathrm{x}_1^T,1\\ \cdots\\ \mathrm{x}_n^T,1\end{bmatrix}\begin{bmatrix}\theta\\ \theta_0\end{bmatrix}\right\|^2\\
&= \left\|\mathrm{y}-\mathrm{X}\begin{bmatrix}\theta\\ \theta_0\end{bmatrix}\right\|^2\\
&= \mathrm{y}^T\mathrm{y}-2\begin{bmatrix}\theta\\ \theta_0\end{bmatrix}^T\mathrm{X}^T\mathrm{y}+\begin{bmatrix}\theta\\ \theta_0\end{bmatrix}^T\mathrm{X}^T\mathrm{X}\begin{bmatrix}\theta\\ \theta_0\end{bmatrix}
\end{aligned}$$

where $\mathrm{y}=[y_1,\ldots,y_n]^T$ is a vector of training responses.

$$\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix}=(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y} \qquad (5.1)$$

第五章 回归分析

5.1 概述

5.2 最小二乘估计

5.3 最大似然估计

5.4 最大后验估计

5.5 扩展的非线性模型

5.6 误差分析

最大似然估计

正态分布假设的似然函数

误差服从正态分布

$$y_t = \theta^T \mathbf{x}_t + \theta_0 + \epsilon_t, \quad t = 1, \ldots, n$$

where $e_t \sim N(0, \sigma^2)$ and e_i is independent of e_j for any $i \neq i$.

似然函数

$$L(\theta, \theta_0, \sigma^2) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_t - \theta^T \mathbf{x}_t - \theta_0)^2\right)$$

最大似然估计

正态分布假设的似然函数

Log-likelihood

$$
\begin{aligned}
l(\theta, \theta_0, \sigma^2) &= \sum_{t=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_t - \theta^T \mathbf{x}_t - \theta_0)^2 \right) \right] \\
&= \sum_{t=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_t - \theta^T \mathbf{x}_t - \theta_0)^2 \right] \\
&= \text{const.} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \theta^T \mathbf{x}_t - \theta_0)^2
\end{aligned}
$$

最大似然估计

最大似然估计

最大化 Log-likelihood：

$$\max_{\theta} -\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t-\theta^T\mathbf{x}_t-\theta_0)^2$$

等价于最小二乘估计：

$$\min_{\theta}\sum_{t=1}^{n}(y_t-\theta^T\mathbf{x}_t-\theta_0)^2$$

$\frac{\partial l}{\partial \sigma^2}=0$ 求得平方误差 $\hat{\sigma}^2=\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{\theta}^T\mathbf{x}_t-\hat{\theta}_0)^2$

结论：高斯误差的最大似然估计 = 最小二乘估计

最大似然估计

优化学习

（1）5.2 的最小二乘矩阵方法

（2）梯度下降优方法

$$h_\theta(x) = \theta_0 + \theta_1 x_1$$

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

Batch gradient descent

$$\theta_j := \theta_j + \alpha\sum_{i=1}^{m}\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$$

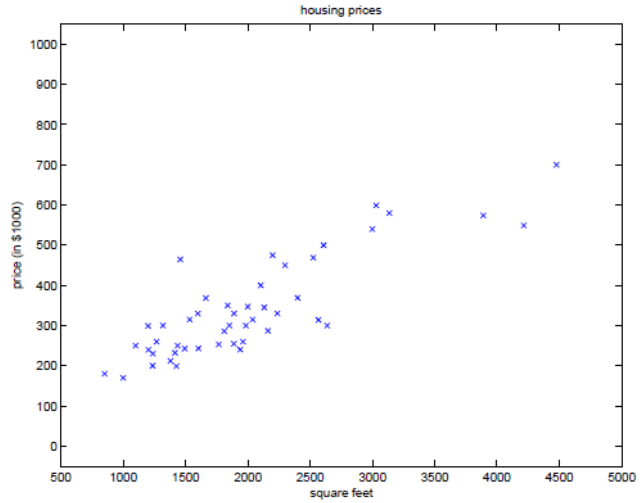
stochastic gradient descent

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$$

最大似然估计

例子

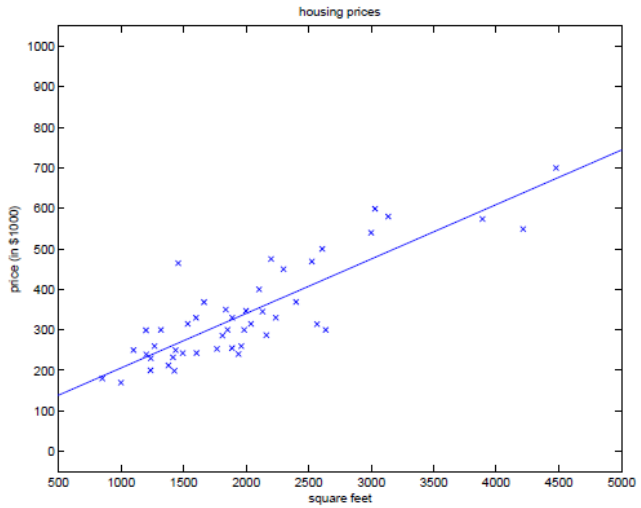
Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



一元线性回归：

$$h_\theta(x) = \theta_0 + \theta_1 x_1$$

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

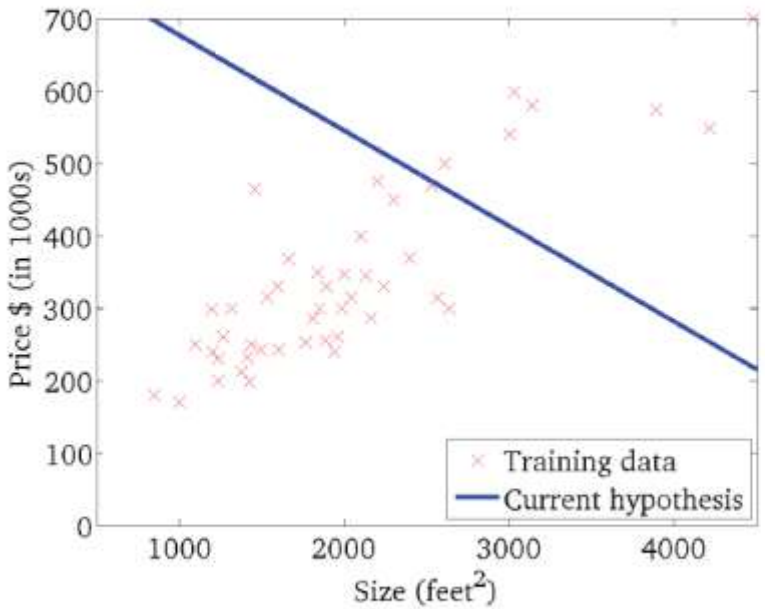


最大似然估计

例子

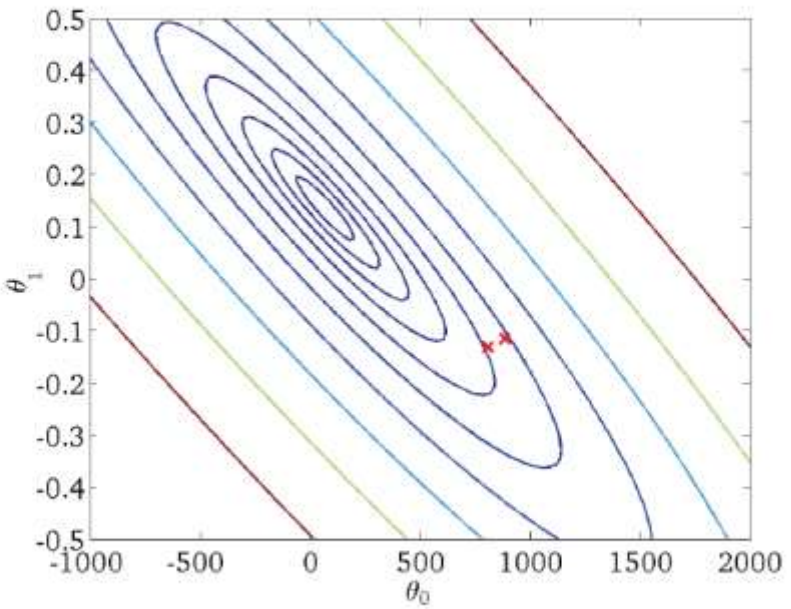
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)

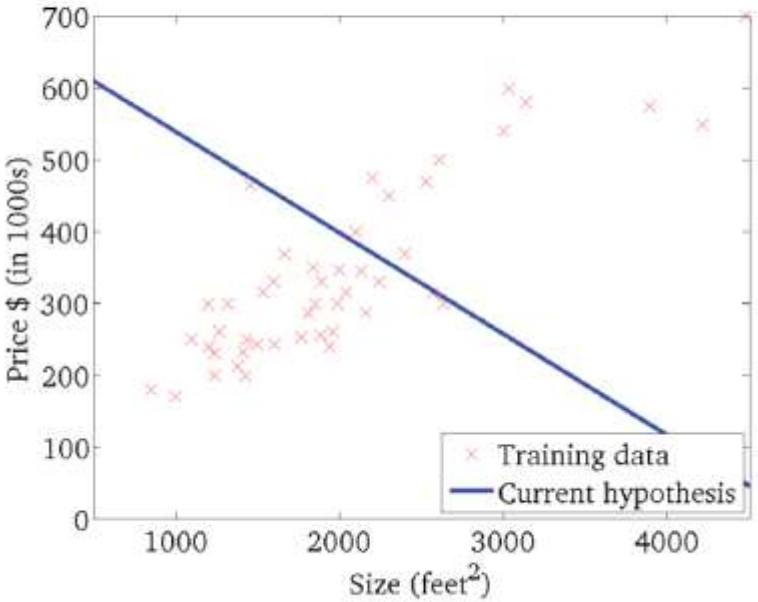


最大似然估计

例子

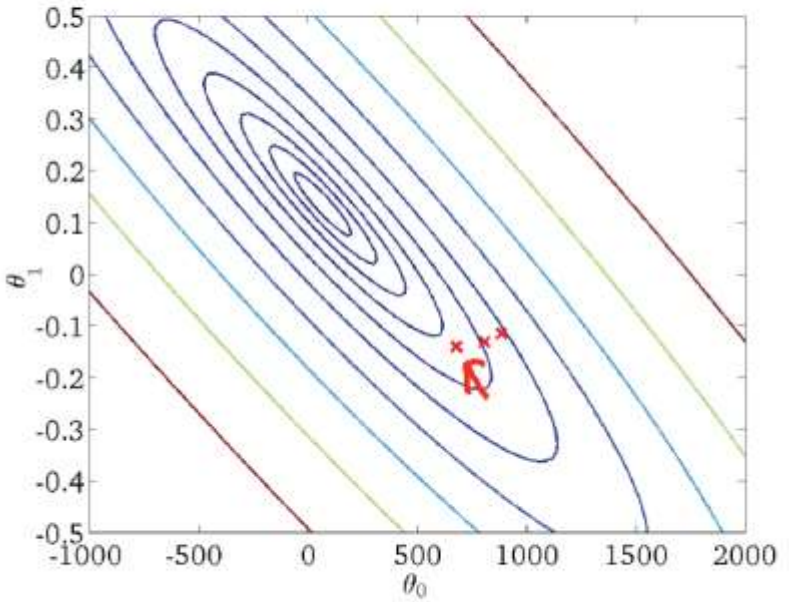
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)

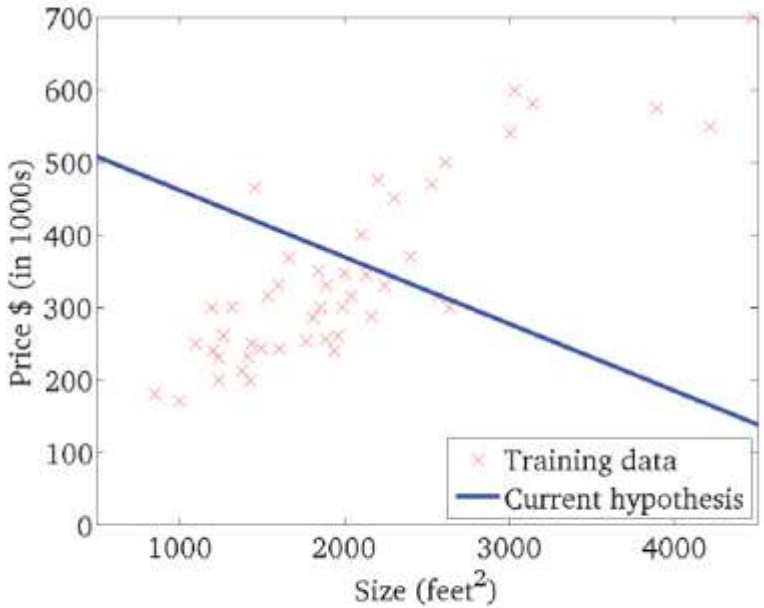


最大似然估计

例子

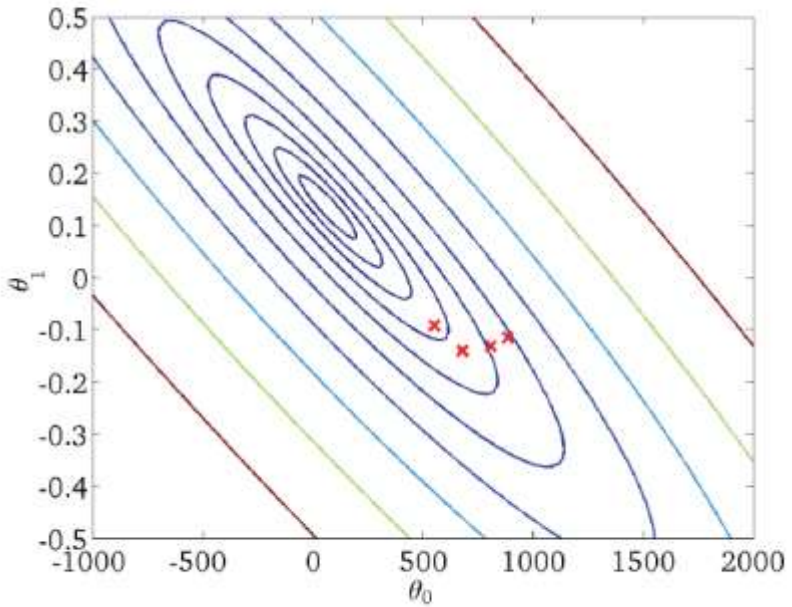
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)

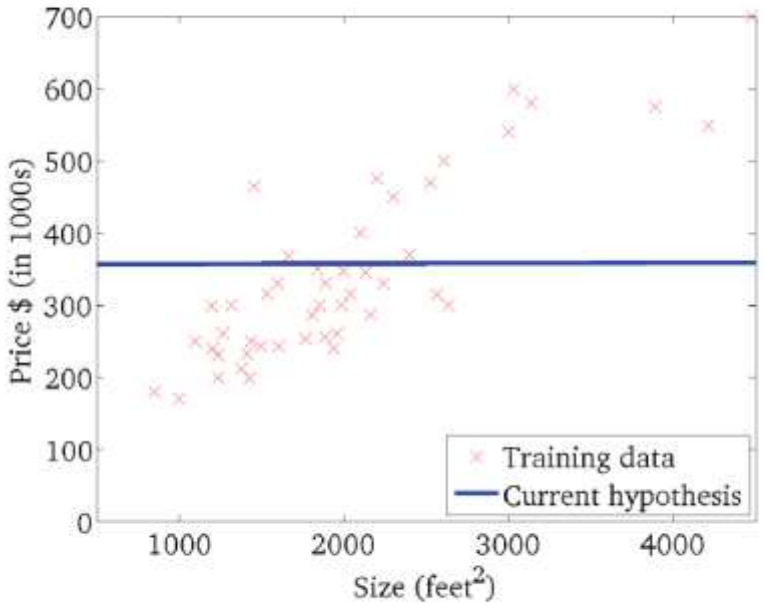


最大似然估计

例子

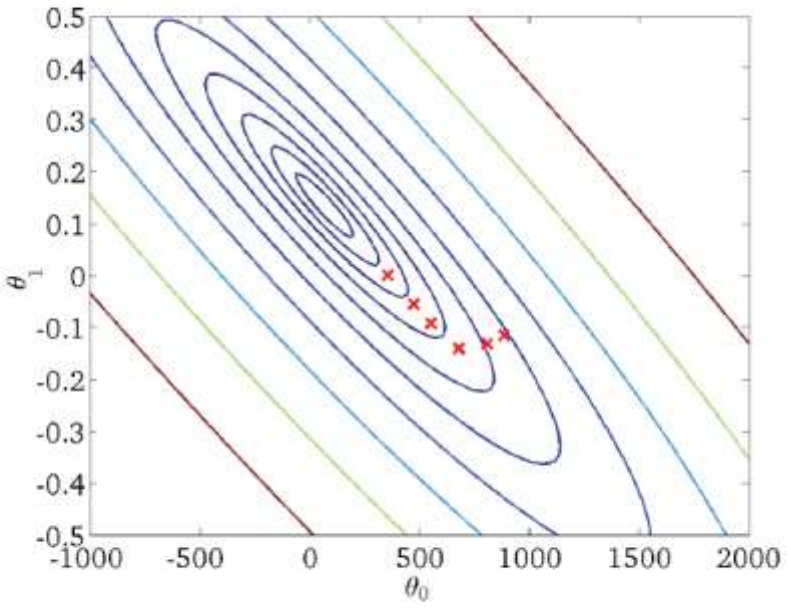
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)

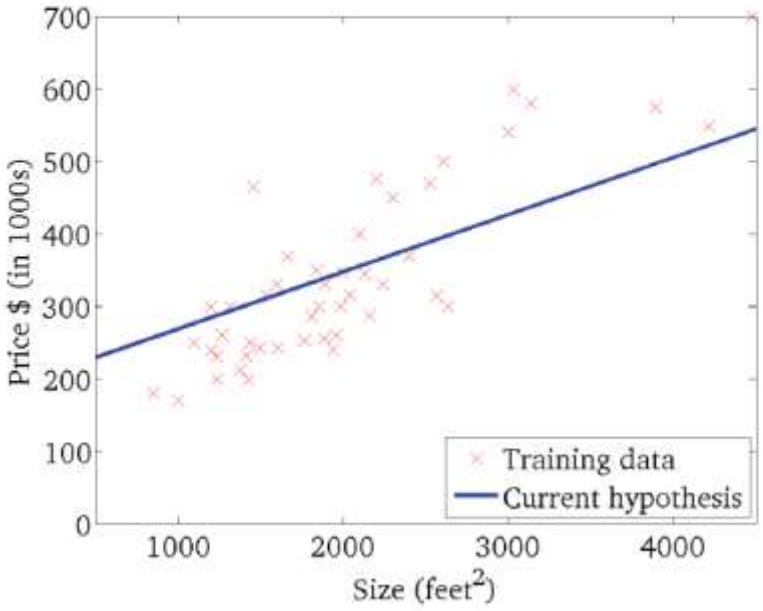


最大似然估计

例子

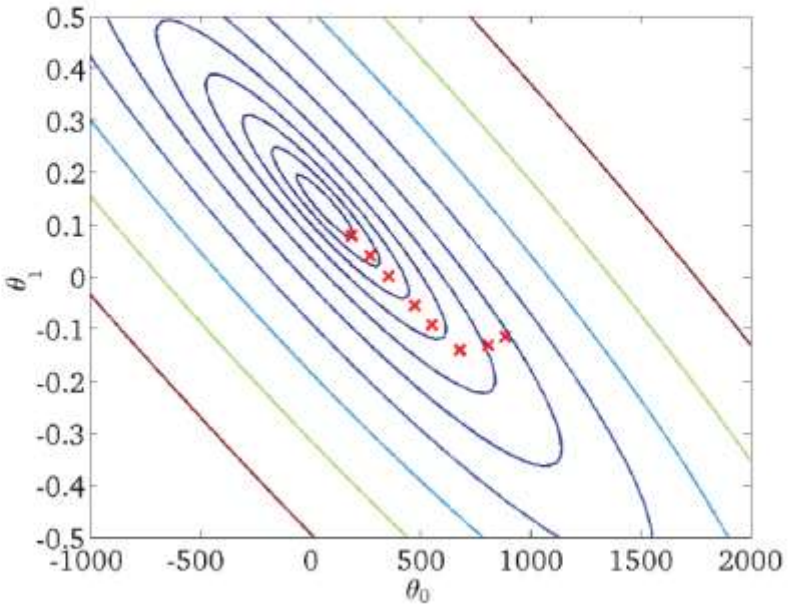
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)

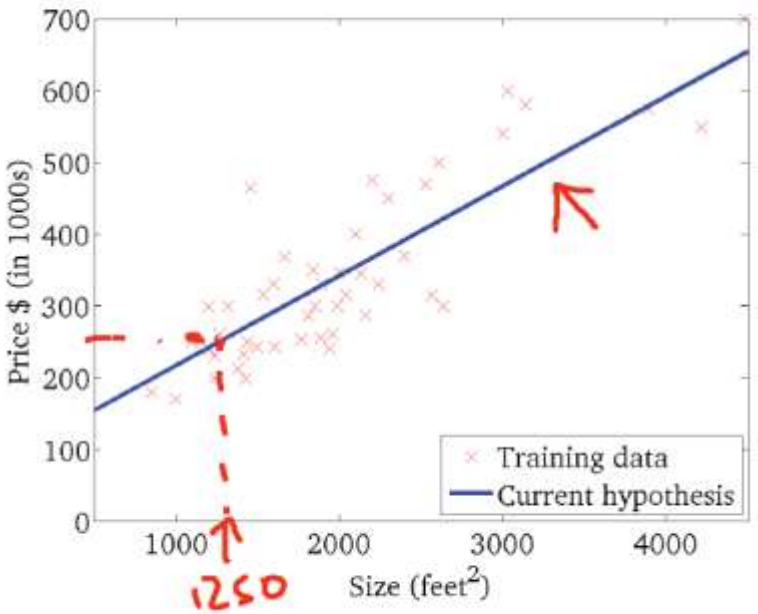


最大似然估计

例子

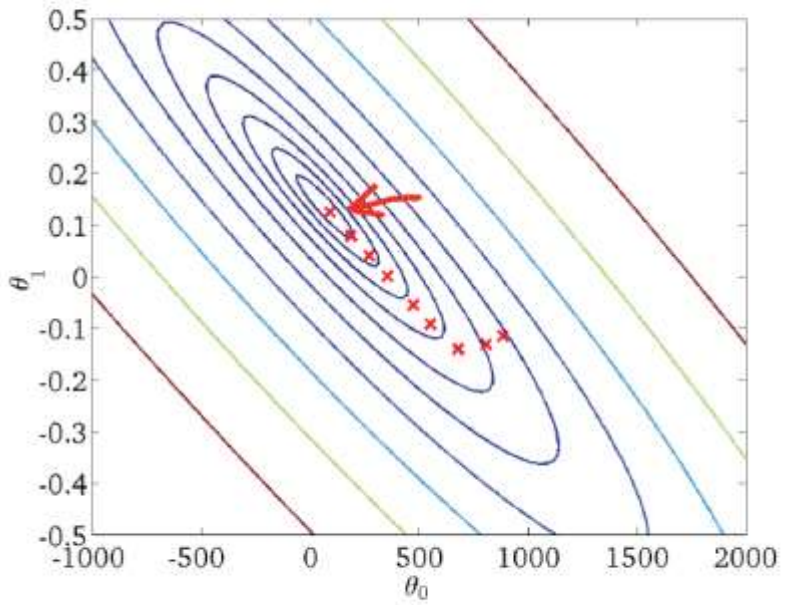
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



第五章 回归分析

5.1 概述

5.2 最小二乘估计

5.3 最大似然估计

5.4 最大后验估计

5.5 扩展的非线性模型

5.6 误差分析

最大后验估计

正态分布的先验

- 似然函数：

$$p(y|\mathbf{x},\theta,\theta_0)=L(\theta,\theta_0,\sigma^2)=\prod_{t=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y_t-\theta^T\mathbf{x}_t-\theta_0)^2\right)$$

- 参数先验：

$$\begin{aligned}\pi(\theta,\theta_0)&=\frac{1}{(\sqrt{2\pi}\sigma_\theta)^M}\prod_{k=0}^{M}\exp\left(-\frac{\theta_k^2}{2\sigma_\theta^2}\right)\\&=\frac{1}{(\sqrt{2\pi}\sigma_\theta)^M}\exp\left(-\frac{1}{2\sigma_\theta^2}\sum_{k=0}^{M}\theta_k^2\right)\\&=\frac{1}{(\sqrt{2\pi}\sigma_\theta)^M}\exp\left(-\frac{1}{2\sigma_\theta^2}\|\theta\|^2\right)\end{aligned}$$

- 参数后验：

$$p(\theta,\theta_0|\mathbf{x},y)\propto\pi(\theta,\theta_0)\,p(y|\mathbf{x},\theta,\theta_0)$$

最大后验估计

最大后验估计

- log 参数后验：

$$\log\ \pi(\theta,\theta_0)\ p(y|\mathbf{x},\theta,\theta_0)$$

$$= -\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t-\theta^T\mathbf{x}_t-\theta_0)^2\ -\frac{1}{2\sigma_\theta^2}\|\theta\|^2\ +\ \text{const.}$$

- 目标函数

$$\min_{\theta}\frac{1}{2}\sum_{t=1}^{n}(y_t-\theta^T\mathbf{x}_t-\theta_0)^2\ +\frac{\lambda}{2}\|\theta\|^2$$

$$\lambda=\frac{\sigma^2}{\sigma_\theta^2}$$

结论：高斯分布的最大后验估计 = 正则化最小二乘估计

最大后验估计

正则化最小二乘估计解

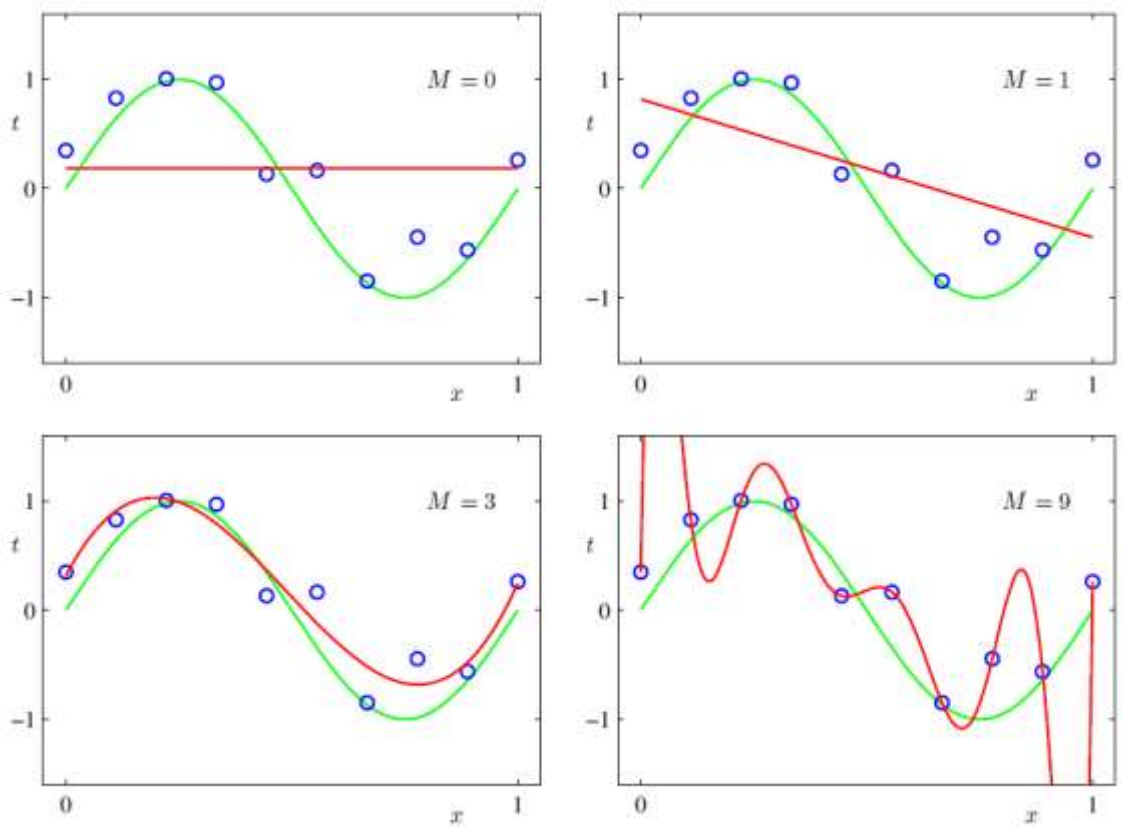
$$\min_{\theta} \frac{1}{2}\sum_{t=1}^{n}(y_t - \theta^T \mathbf{x}_t - \theta_0)^2 + \frac{\lambda}{2}\|\theta\|^2$$

$$\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- 正则项解决过拟合问题

最大后验估计

过拟合问题

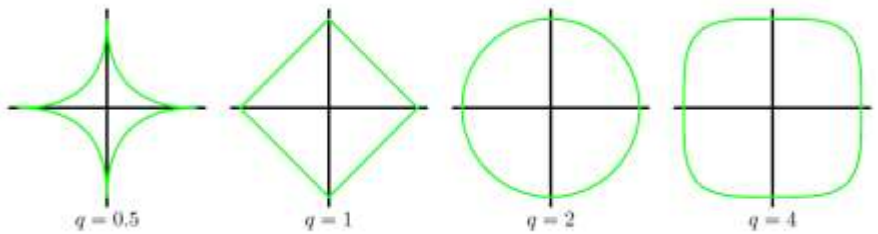


最大后验估计

过拟合问题

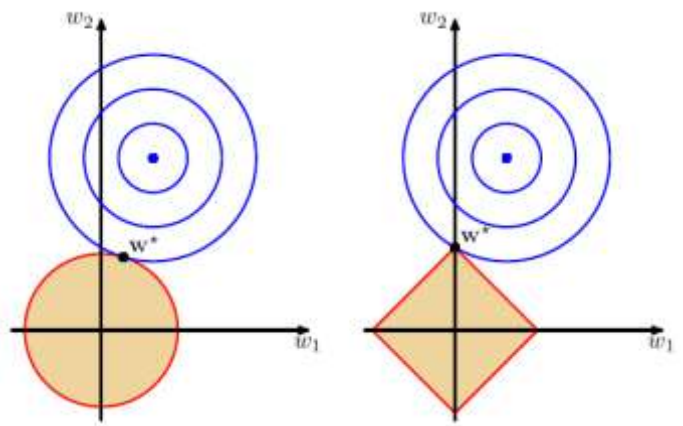
不同范数的正则项

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$



1 范数正则求得 w 具有稀疏性:

$$\sum_{j=1}^{M}|w_j|^q \leqslant \eta$$



第五章 回归分析

5.1 概述

5.2 最小二乘估计

5.3 最大似然估计

5.4 最大后验估计

5.5 扩展的非线性模型

5.6 误差分析

扩展的非线性模型

线性基函数回归

线性回归

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

扩展的非线性回归

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

where $\phi_j(\mathbf{x})$ are known as *basis functions*.

扩展的非线性模型

线性基函数回归

基函数形式

- polynomial basis functions

  $$\phi_j(x) = x^j$$

- 'Gaussian' basis functions

  $$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

- sigmoidal basis function

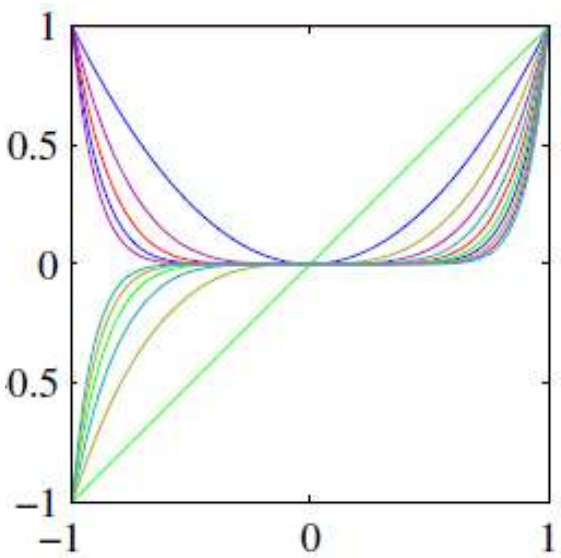
  $$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right), \qquad \sigma(a) = \frac{1}{1+\exp(-a)}$$

- 'tanh' function

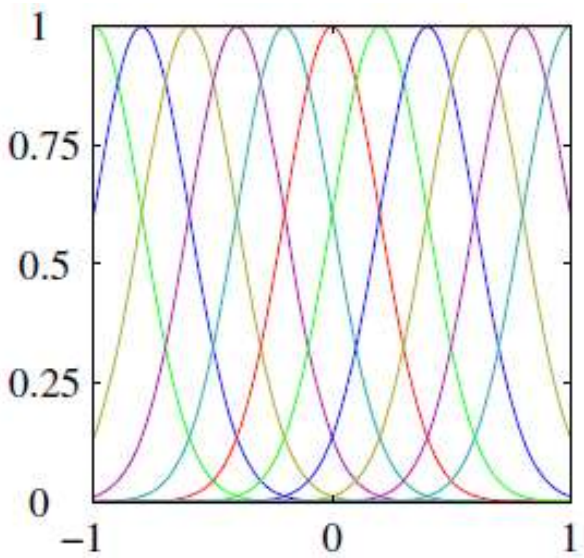
  $$\tanh(a) = 2\sigma(a) - 1$$

扩展的非线性模型

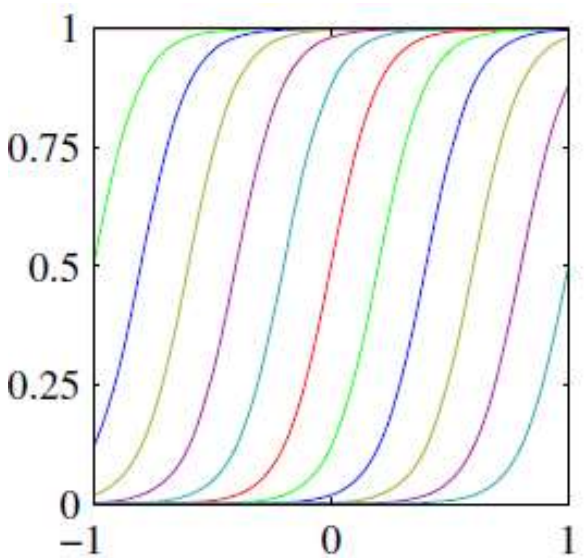
线性基函数回归



Poly basis



Gaussian basis



Sigmoidal basis

扩展的非线性模型

多项式回归

金融、数据分析、趋势预测中常用的非线性回归;

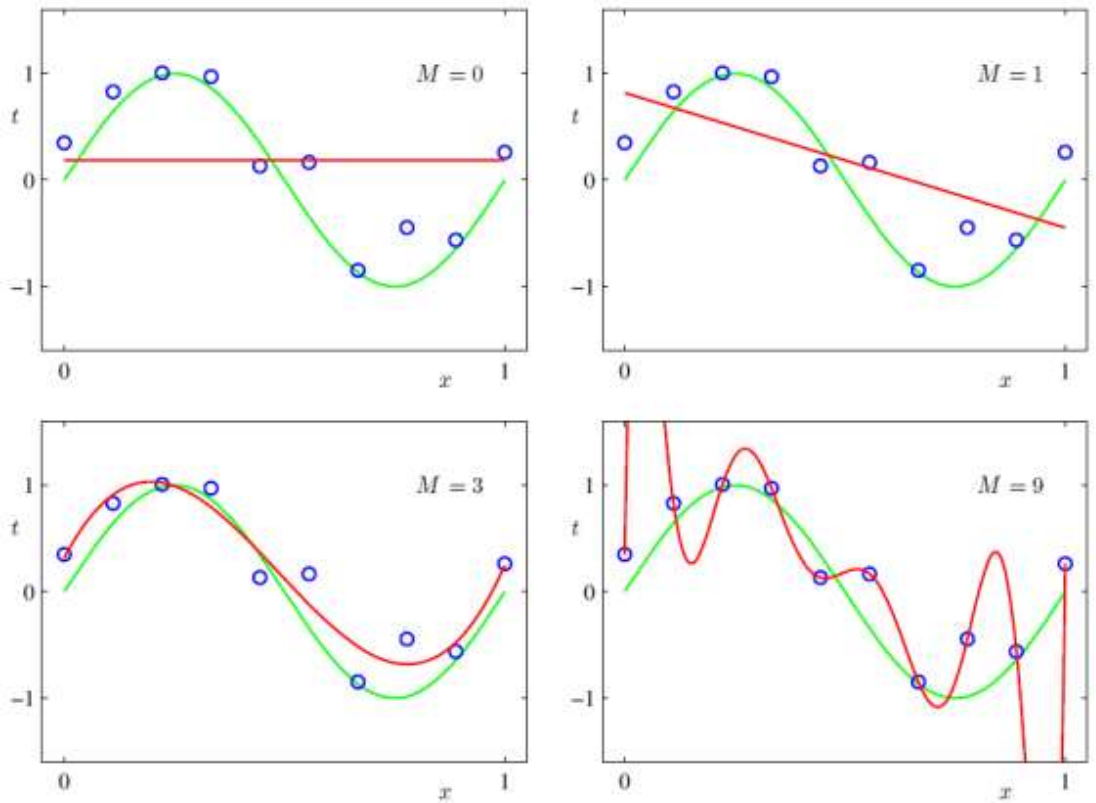
回归函数等价于 polynomial 基扩展的函数:

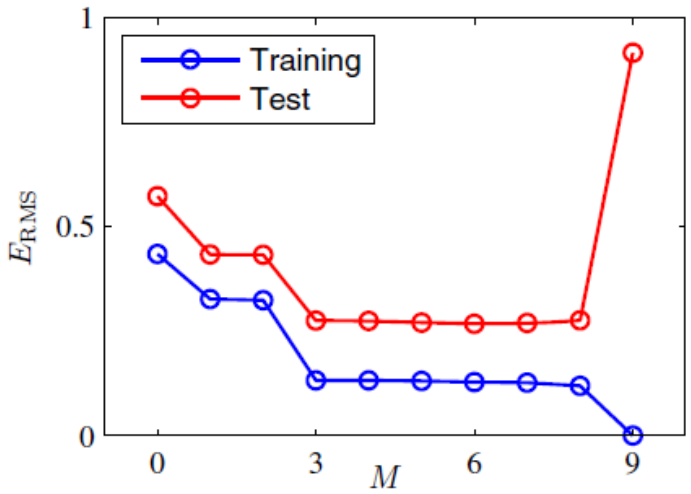
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

扩展的非线性模型

多项式回归

多项式参数与过拟合





第五章 回归分析

5.1 概述

5.2 最小二乘估计

5.3 最大似然估计

5.4 最大后验估计

5.5 扩展的非线性模型

5.6 误差分析

误差分析

Bias², Variance, Noise

$$h(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = \int y p(y|\mathbf{x})\,\mathrm{d}y$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - y\}^2 p(\mathbf{x},y)\,\mathrm{d}\mathbf{x}\,\mathrm{d}y$$

$$= \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \underbrace{\int \{h(\mathbf{x}) - y\}^2 p(\mathbf{x},y)\,\mathrm{d}\mathbf{x}\,\mathrm{d}y}_{\text{noise}}$$

$$\begin{aligned}&\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\ &= \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 \\ &\quad +2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.\end{aligned}$$

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right]$$

$$= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.$$

误差分析

正则项对 Bias, Variance 的影响

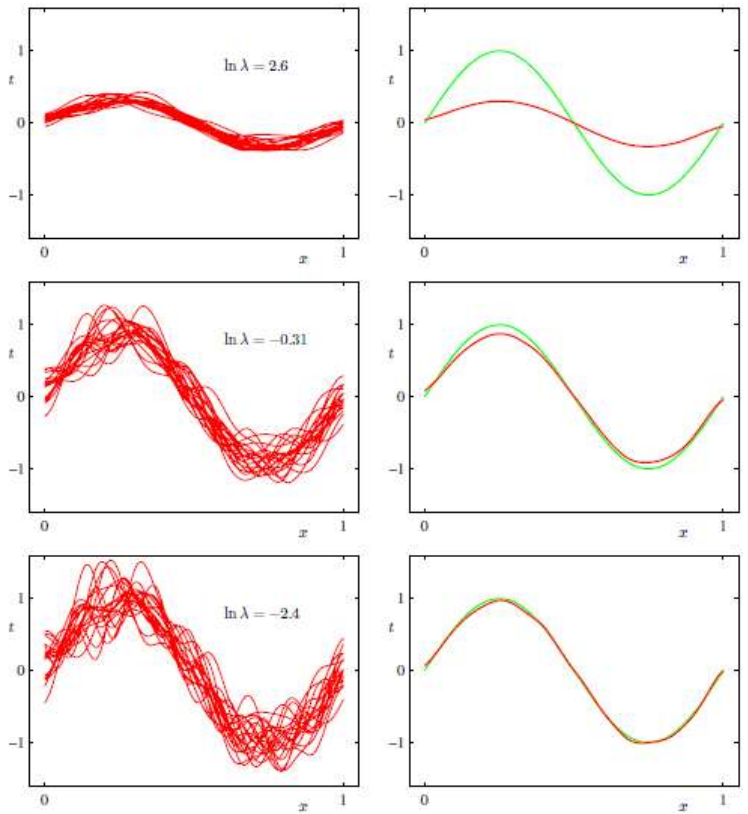
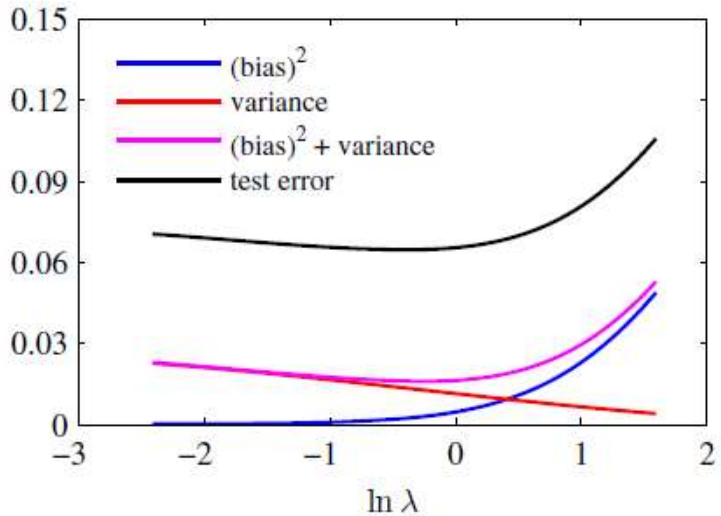


Figure 3.5 Illustration of the dependence of bias and variance on model complexity, governed by a regularization parameter λ, using the sinusoidal data set from Chapter 1. There are $L = 100$ data sets, each having $N = 25$ data points, and there are 24 Gaussian basis functions in the model so that the total number of parameters is $M = 25$ including the bias parameter. The left column shows the result of fitting the model to the data sets for various values of $\ln \lambda$ (for clarity, only 20 of the 100 fits are shown). The right column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the data sets were generated (green).

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right]$$
$$= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.$$



误差分析

参数估计

最小二乘估计是无偏估计

$$\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} \rightarrow \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X} \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} + \mathbf{e}$$

where $\mathbf{e} = [\epsilon_1, \ldots, \epsilon_n]^T$, $E\{\mathbf{e}\} = 0$ and $E\{\mathbf{e}\mathbf{e}^T\} = \sigma^{*2}\mathbf{I}$.

$\epsilon_t \sim N(0, \sigma^{*2})$.

将 y 代入 (5.1) $\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$$\begin{aligned} \begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} + \mathbf{e}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \\ &= \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e} \end{aligned}$$

$$E\{\begin{bmatrix} \hat{\theta} \\ \hat{\theta}_0 \end{bmatrix} |\mathbf{X}\} = \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E\{\mathbf{e}|\mathbf{X}\} = \begin{bmatrix} \theta^* \\ \theta_0^* \end{bmatrix}$$

误差分析

参数估计

正则化最小二乘估计是有偏估计

使得参数估计更加稳定

(1) 相当于增加正则项

(2) 相当于加入白噪声

小 结

1. 回归问题
2. 最小二乘估计、最大似然估计
3. 正则化最小二乘估计、最大后验估计
4. 基函数扩展的回归模型
5. 误差分析

参考文献

1. Chris Bishop. Pattern recognition and Machine Learning. Springer, 2006. （PR&ML）
2. 周志华，机器学习，清华大学出版社，2016.