



数据挖掘导论

第五章：回归

王静远

北京航空航天大学

线性回归的概率解释

贝叶斯定理

$$\Pr(A, O) = \Pr(A)\Pr(O|A) = \Pr(O)\Pr(A|O)$$



$$\Pr(A|O) = \frac{\Pr(A)\Pr(O|A)}{\Pr(O)}$$

后验
Posterior 先验
Prior 似然
likelihood



Thomas Bayes
1701—1761
英国牧师、
业余数学家

2009
博洛尼亚
国际儿童书展
优秀童书奖

2010
第十五届
NOMA图画书
插画奖亚军

2011
新闻出版总署
向全国青少年推荐的
百种优秀图书之一



猜 猜 看



• 统计与概率 •



等一下，昨天妈妈只买了面包吗？

菜篮子里好像还有紫菜和蟹棒，

那么今天早上就可以做紫菜包饭了。

紫菜包饭？三明治？

肯定是这两个中的一个。

三明治？

紫菜包饭？



呵呵，好香的味道啊，
不是烤火腿的味道吗？
今天早上可以吃火腿三明治了，
妈妈昨天好像买了面包，
肯定没错。



闻到了烤火腿的味道

三明治的似然概率大！

P(烤火腿的味道|三明治)

周一到周五
吃三明治
周末吃紫菜
包饭

三明治的先
验概率大

P(三明治)

星期一 星期二 星期三 星期四 星期五



星期六 星期日



我们家的早餐主要是三明治，

只有星期六和星期日是紫菜包饭，

今天是星期五，所以三明治

比紫菜包饭的可能性更大，

火腿三明治，就是这个！





$P(\text{三明治}|\text{烤火腿的味道}) \propto P(\text{烤火腿的味道}|\text{三明治})P(\text{三明治})$

贝叶斯公式的应用

A 对应6:00交通拥堵状况

B 对应5:00交通拥堵状况

问：今天6:00堵不堵？

$$\Pr(A|O) = \frac{\Pr(A)\Pr(O|A)}{\Pr(O)}$$

后验
Posterior

先验
Prior

似然
likelihood

- 先验

- 6:00交通{拥堵}的概率0.8, {不拥堵}的概率0.2;

- 似然

- 6:00交通{拥堵}时, 5:00也{拥堵}的概率0.7, {不拥堵}概率0.3
 - 6:00交通{不拥堵}时, 5:00也{拥堵}的概率0.1, {不拥堵}概率0.9

- 后验, 如果5:00交通{拥堵}

- 6:00时交通{拥堵}的概率 = $0.8 \times 0.7 = 0.56$
 - 6:00时交通{不拥堵}的概率 = $0.2 \times 0.1 = 0.06$

- 后验, 如果5:00交通{不拥堵}

- 6:00时交通{拥堵}的概率 = $0.8 \times 0.3 = 0.24$
 - 6:00时交通{不拥堵}的概率 = $0.2 \times 0.9 = 0.18$

朴素贝叶斯分类
Naive Bayesian classification



Naïve Bayesian Classification

- 假设样本各特征条件独立

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^d P(x_i \mid c)$$

- 其中 d 为属性数目, x_i 为 \mathbf{x} 在第 i 个属性上的取值。

- 朴素贝叶斯分类器的表达式子

$$h_{nb}(\mathbf{x}) = \operatorname{argmax}_{c \in \mathcal{Y}} P(c) \prod_{i=1}^d P(x_i \mid c)$$

使得朴素贝叶斯模型后验概率最大的分类标签 c 就是预测结果

最大似然估计

- 数据 $D = \{x_1, x_2, \dots, x_M\}$ 独立同分布地采样自一个概率分布 $N_D(\theta)$ — 可观测事件 待推断事件
- 假设：待估计参数 θ 是客观存在的固定的值，只是未知而已

$$\Pr(\theta|\mathcal{D}) \propto \Pr(\mathcal{D}|\theta) = \Pr(x_1, x_2, \dots, x_M|\theta) = \prod_{i=1}^M \Pr(x_i|\theta)$$

$$\log \Pr(\theta|\mathcal{D}) \propto \sum_{i=1}^M \log \Pr(x_i|\theta)$$

$$\theta^* = \arg \max_{\theta} \sum_{i=1}^M \log \Pr(x_i|\theta)$$

最大后验估计

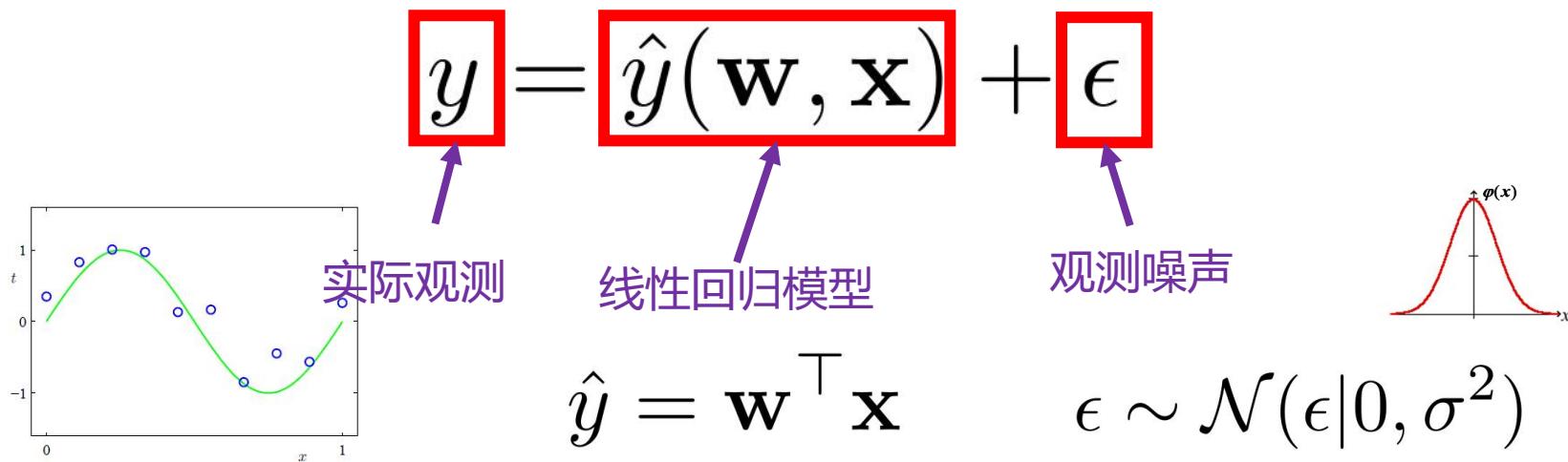
- 数据 $D = \{x_1, x_2, \dots, x_M\}$ 独立同分布地采样自一个概率分布 $N_D(\theta)$ —— 可观测事件 待推断事件
- 假设：待估计参数 θ 也采样自一个概率分布 N_θ

$$\Pr(\theta|D) \propto \Pr(\theta)\Pr(D|\theta) = \Pr(\theta)\Pr(x_1, x_2, \dots, x_M|\theta) = \Pr(\theta) \prod_{i=1}^M \Pr(x_i|\theta)$$

$$\log \Pr(\theta|D) \propto \log \Pr(\theta) + \sum_{i=1}^M \log \Pr(x_i|\theta)$$

$$\theta^* = \arg \max_{\theta} \left(\log \Pr(\theta) + \sum_{i=1}^M \log \Pr(x_i|\theta) \right)$$

线性回归模型的数据分布假设



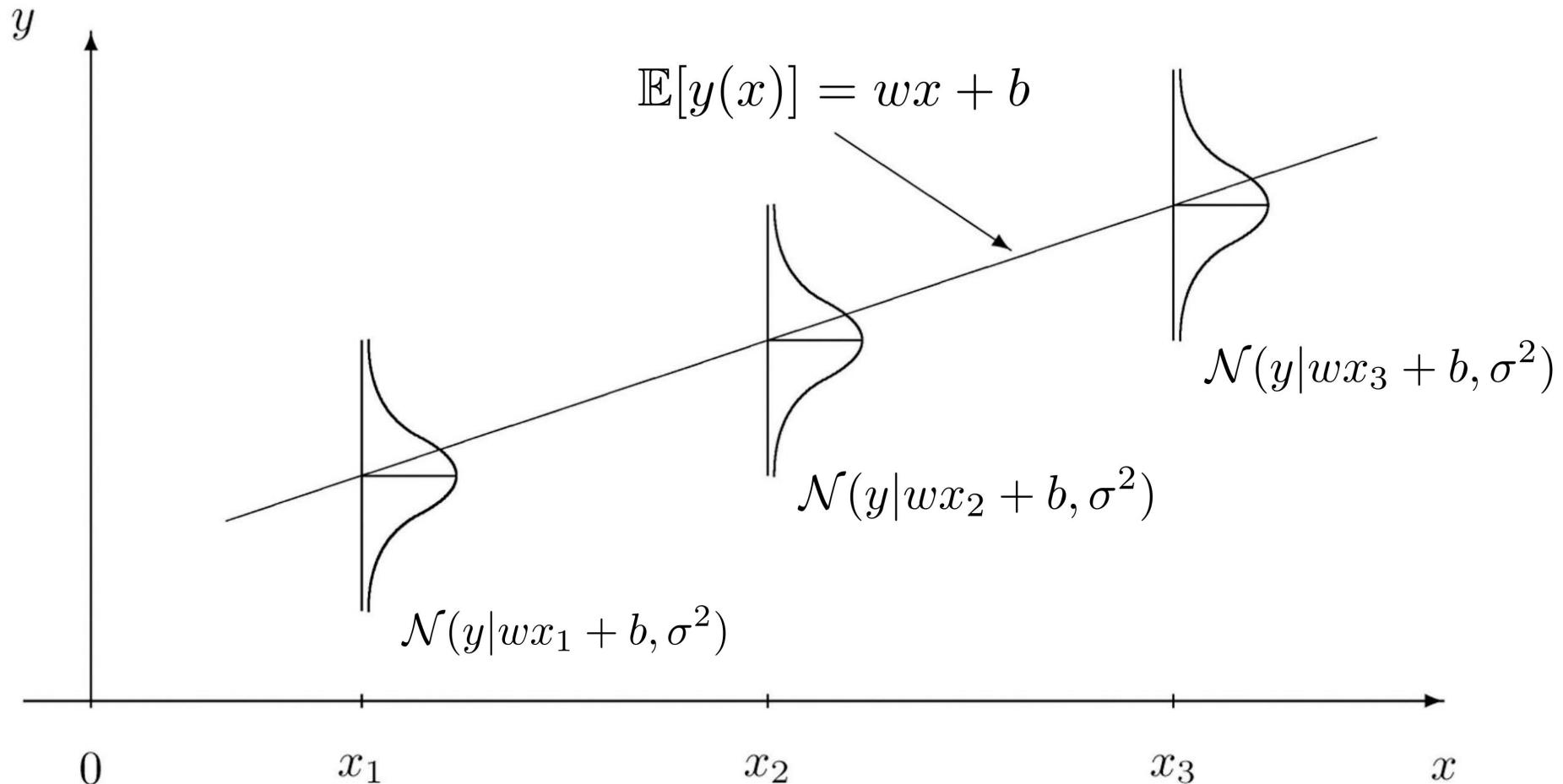
数据 y 取值的概率分布：

$$y \sim \mathcal{N}(y | \hat{y}(\mathbf{x}, \mathbf{w}), \sigma^2)$$

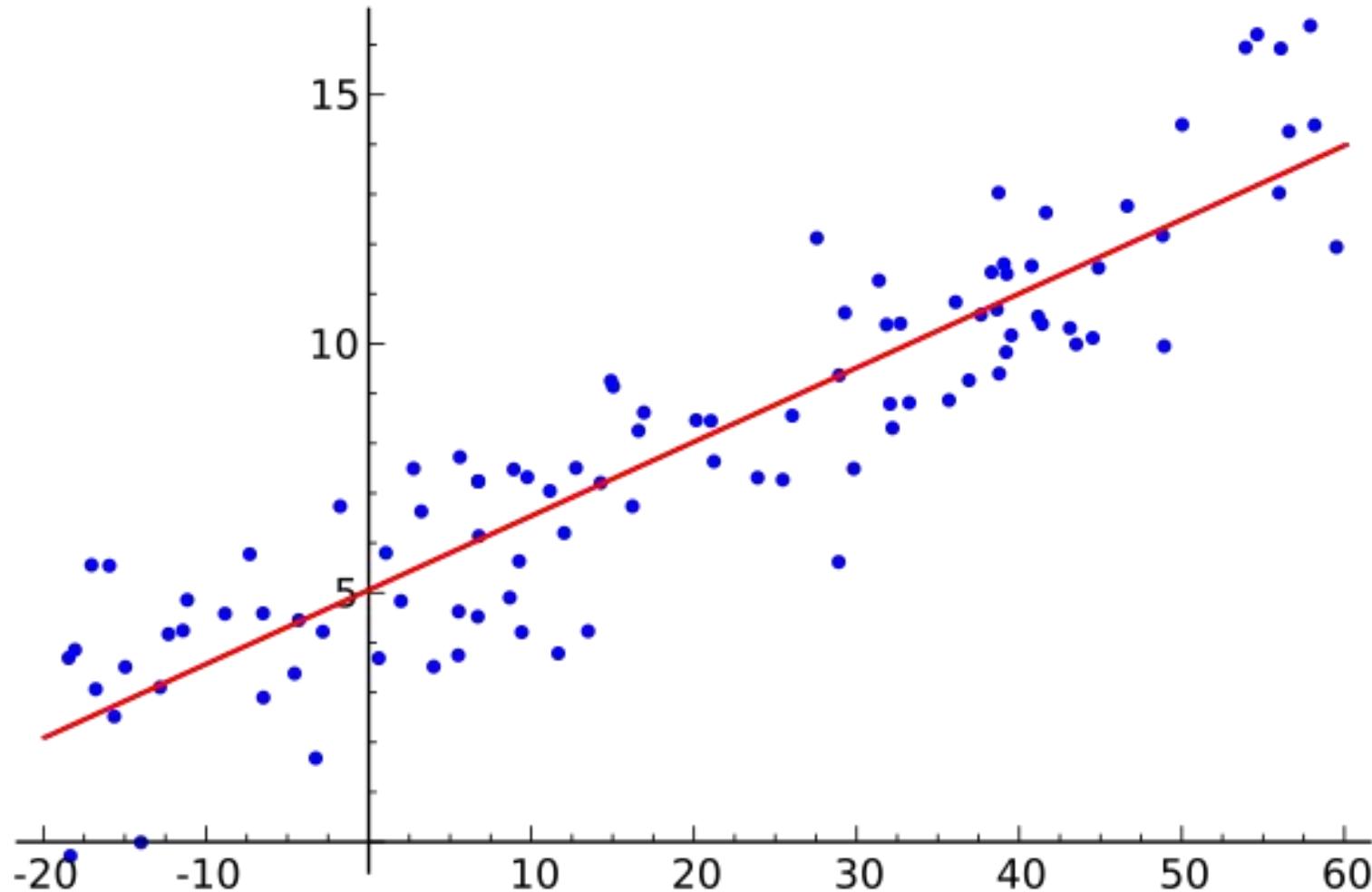
对于给定 x 值， y 的期望满足：

$$\mathbb{E}[y(\mathbf{x})] = \hat{y}(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \mathbf{x}, \quad \forall \mathbf{x}$$

线性回归模型的数据分布假设

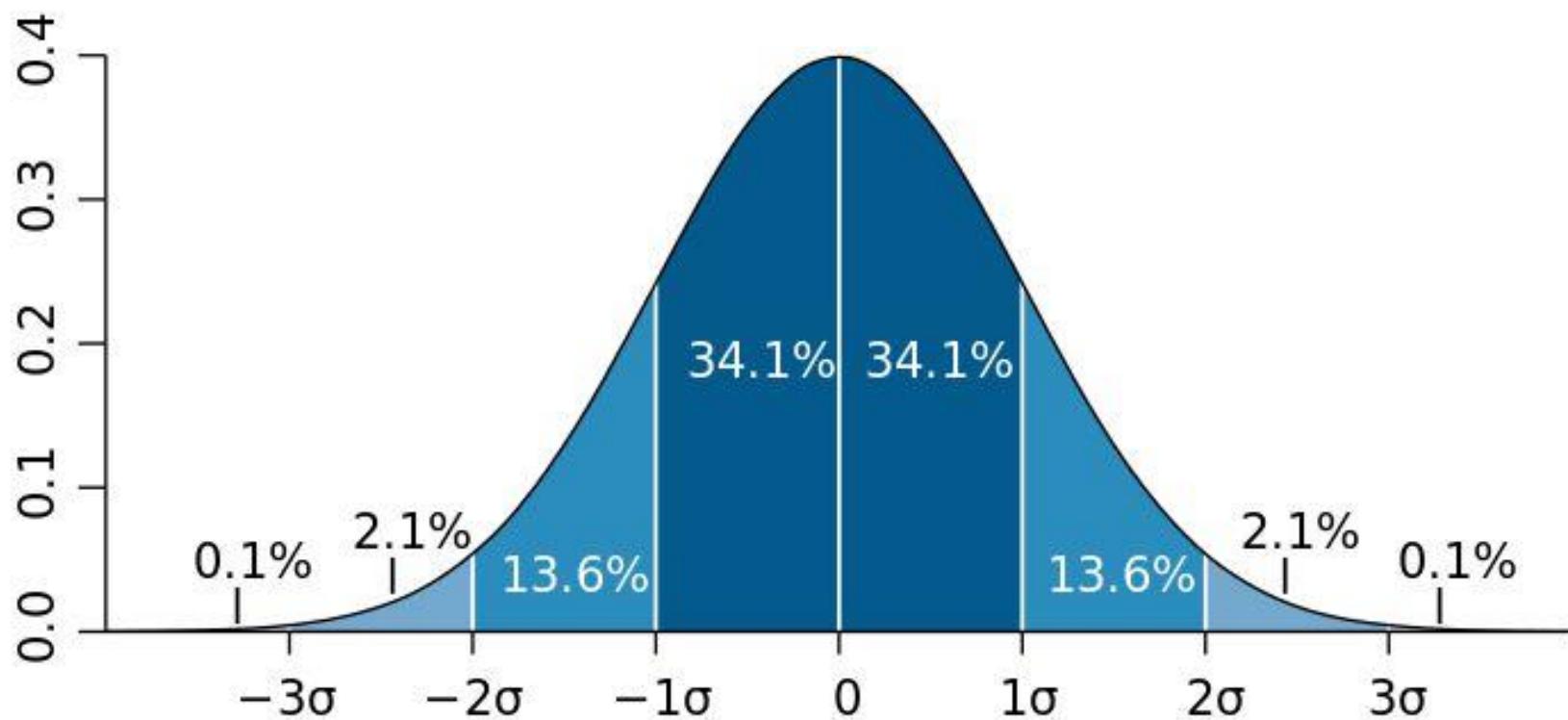


线性回归模型的数据分布假设



高斯分布

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



线性回归模型的参数求解

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\epsilon_i = y_i - \mathbf{w}\mathbf{x}_i$$

条件一：

给定样本和对应的特征，
可以生成一个预测误差。

条件二：

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

预测误差符合高斯分布

一个给定参数 w 的线性回归模型，生成观测误差 ϵ 的概率为：

$$P\{\epsilon = \epsilon_i | \mathbf{w}\} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mathbf{w}\mathbf{x}_i)^2}{2\sigma^2}\right)$$

对于一组数据集，其对应观测误差生成概率为：

$$\max \prod_{i=1}^n P(\epsilon_i | \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mathbf{w}\mathbf{x}_i)^2}{2\sigma^2}\right)$$

线性回归模型的参数求解

• 线性回归的目标函数

$$\max \prod_{i=1}^n P(\epsilon_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(y_i - \mathbf{w}\mathbf{x}_i)^2}{2\sigma^2} \right)$$

$$\max \ln \prod_{i=1}^n P(\epsilon_i) \propto - \sum_{i=1}^n (y_i - \mathbf{w}\mathbf{x}_i)^2$$

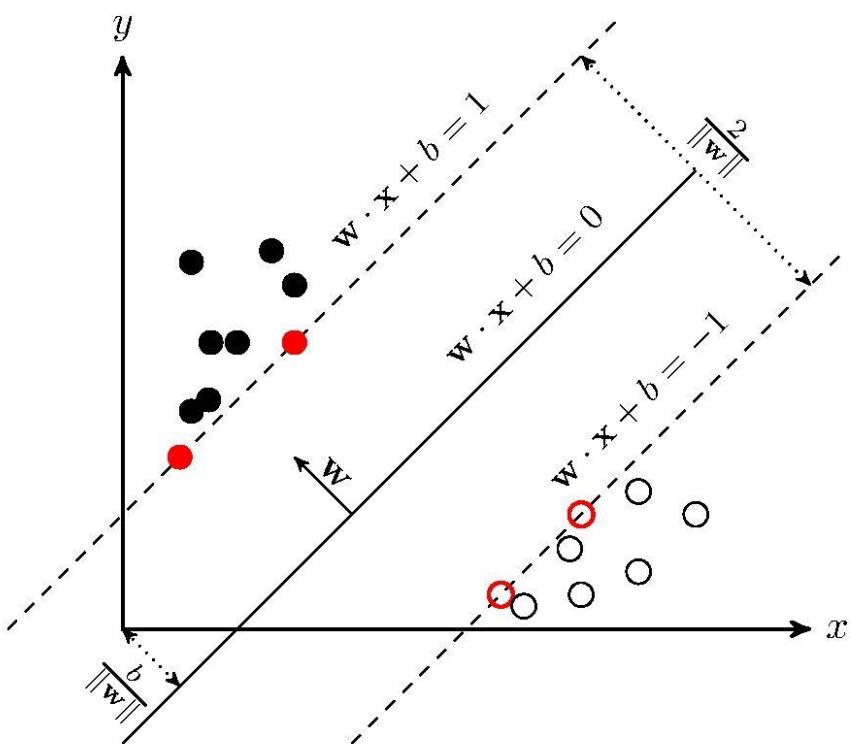
$$\min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}\mathbf{x}_i)^2$$

最小二乘目标函数

残差
平方和

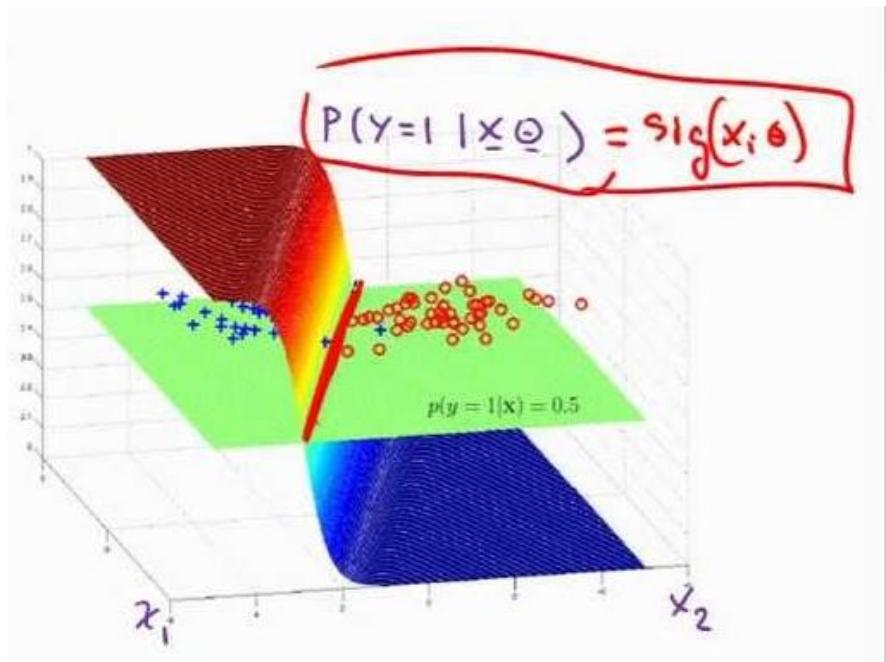
逻辑回归的概率解释

逻辑回归回顾



SVM的分类方式

$$\{x_1, x_2, \text{sigmoid}(w_1x_1+w_2x_2)\}$$



逻辑回归的分类方式

逻辑回归模型的数据分布假设

• 总体假设

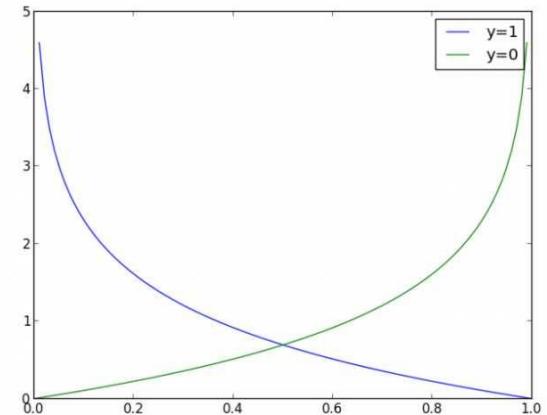
• 伯努利分布 (0-1分布)

$$\Pr(y = 1) = p, \quad \Pr(y = 0) = 1 - p$$

其中：

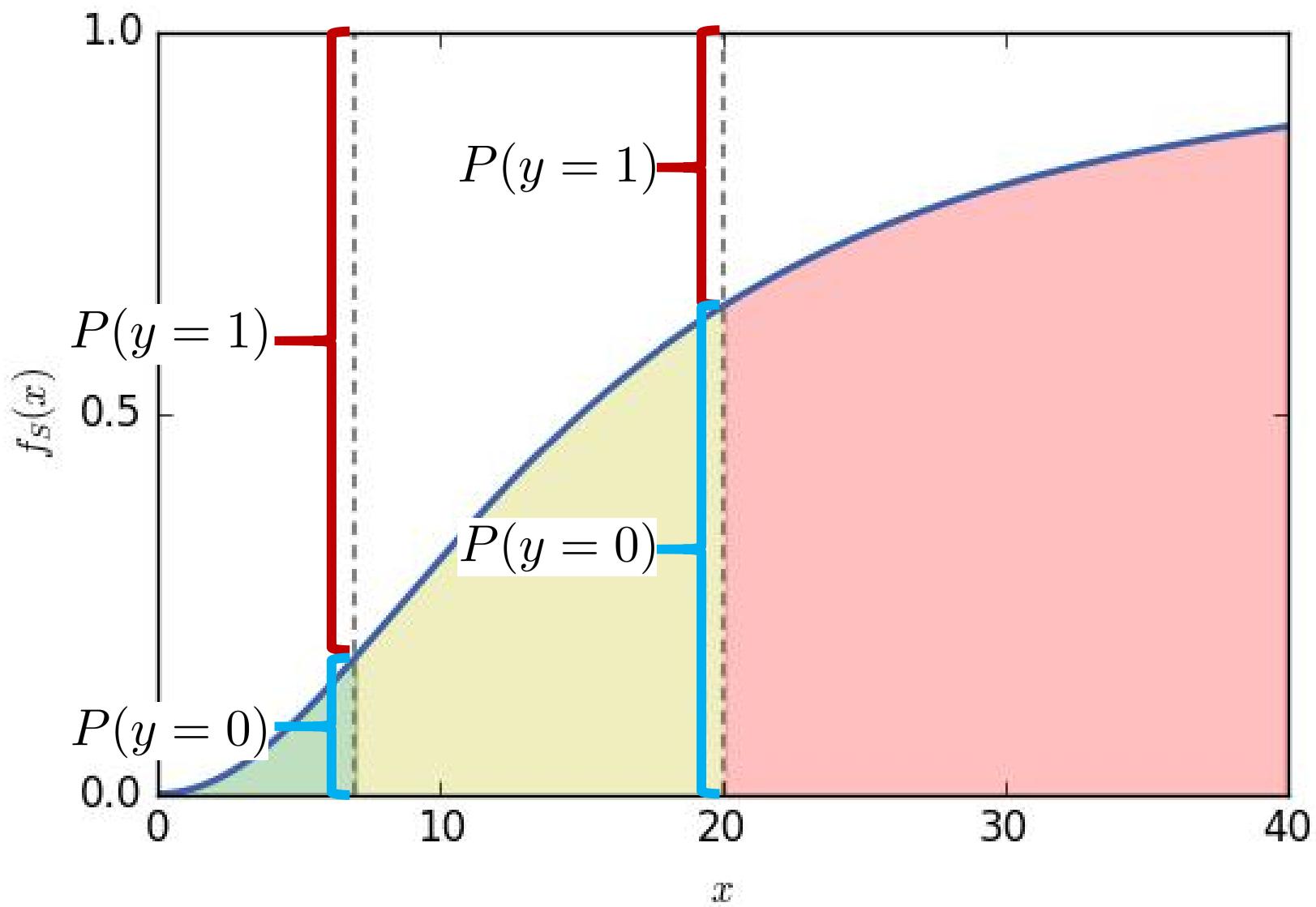
$$p(\theta, x) = \frac{1}{1 + e^{-(\theta^\top \tilde{x})}}$$

$$1 - p(\theta, x) = \frac{e^{-(\theta^\top \tilde{x})}}{1 + e^{-(\theta^\top \tilde{x})}}$$



- 给定 θ 之后，对于每一个 x_i 取值， y_i 是一个确定概率的伯努利分布；
- 回忆一下：对于线性回归， y_i 是一个确定均值高斯分布。

逻辑回归模型的数据分布假设



逻辑回归的概率解释

- **最大似然估计**

- 根据样本推断参数 θ 就是计算后验概率 $Pr(\theta | Y; X)$

$$Pr(\theta | Y; X) \propto Pr(Y | \theta; X) = \prod_{m=1}^M Pr(y_m | \theta; x_m)$$

$$Pr(y_m | \theta; x_m) = p(\theta, x_m)^{y_m} (1 - p(\theta, x_m))^{1-y_m}$$

$$\log Pr(\theta | Y; X) \propto \log Pr(Y | \theta; X)$$

$$= \sum_{m=1}^M (y_m \log p(\theta, x_m) + (1 - y_m) \log (1 - p(\theta, x_m)))$$

$$= \boxed{\sum_{m=1}^M \left(y_m \log \frac{1}{1 + e^{-\theta^\top \tilde{x}_m}} + (1 - y_m) \log \frac{e^{-\theta^\top \tilde{x}_m}}{1 + e^{-\theta^\top \tilde{x}_m}} \right)}$$

负的交叉熵损失函数

相对熵 (KL散度)

- 样本 y_m 和其预测值 \hat{y}_m

- Q 为 Bernoulli(y_m)，即对于随机变量 $z \sim Q$, $\Pr_Q(z = 1) = y_m$, $\Pr_Q(z = 0) = 1 - y_m$ 。
- Q' 为 Bernoulli(\hat{y}_m)，即对于随机变量 $z \sim Q'$, $\Pr_{Q'}(z = 1) = \hat{y}_m$, $\Pr_{Q'}(z = 0) = 1 - \hat{y}_m$ ；

- 度量概率分布 Q 和 Q' 之间的差异程度

$$\begin{aligned} D_{KL}(Q\|Q') &= \sum_{z_i=0}^1 \Pr_Q(z_i) \log \left(\frac{\Pr_Q(z_i)}{\Pr_{Q'}(z_i)} \right) \\ &= \sum_{z_i=0}^1 \Pr_Q(z_i) \log \Pr_Q(z_i) - \sum_{z_i=0}^1 \Pr_Q(z_i) \log \Pr_{Q'}(z_i) \\ &= -H(Q) - [\Pr_Q(z_i = 0) \log \Pr_{Q'}(z_i = 0) + \Pr_Q(z_i = 1) \log \Pr_{Q'}(z_i = 1)] \\ &= -H(Q) - [(1 - y_m) \log (1 - \hat{y}_m) + y_m \log \hat{y}_m] \end{aligned}$$

多项逻辑回归的概率解释

多项逻辑回归回顾

- SoftMax 回归
- 多分类数据集

其中 $x_m \in \mathbb{R}^N$, $y'_m \in \{1, 2, \dots, K\}$, $m = 1, 2, \dots, M$

- 对于 $y'_m = k$, 独热编码构造一个向量 $y_m \in R^K$,
其中第 k 个分量 $y_{m,k} = 1$, 其他分量为 0
- 新数据集

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_M, y_M)\}$$

多项逻辑回归的回归函数

- 对每一个样本 x_m , 使用 K 个线性组合对其进行变换, 生成 K 个得分

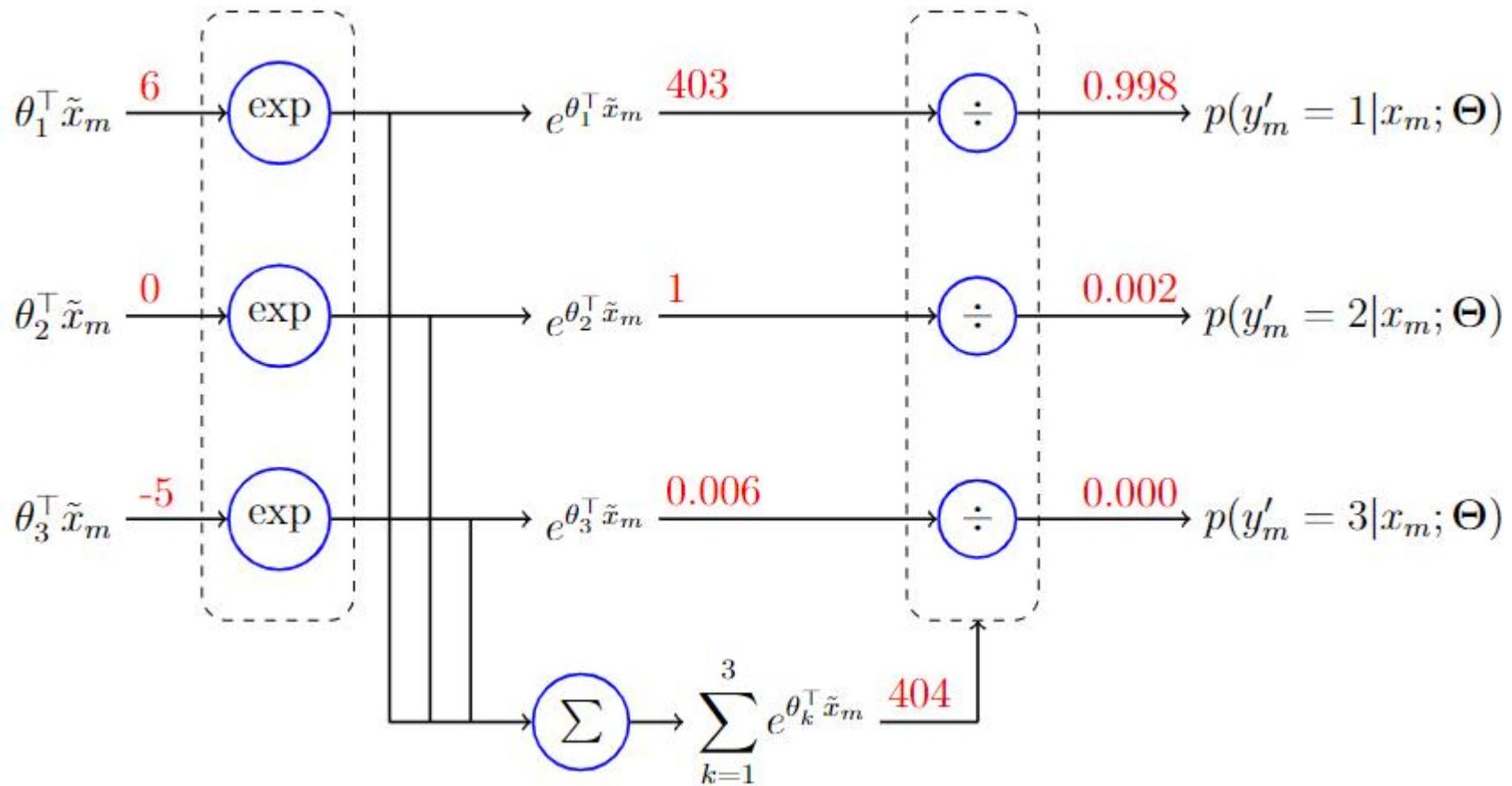
$$\begin{pmatrix} z_{m,1} \\ \vdots \\ z_{m,k} \\ \vdots \\ z_{m,K} \end{pmatrix} = \begin{pmatrix} \theta_1^\top \tilde{x}_m \\ \vdots \\ \theta_k^\top \tilde{x}_m \\ \vdots \\ \theta_K^\top \tilde{x}_m \end{pmatrix} \quad \hat{y}'_m = \arg \max_k \{z_{m,1}, \dots, z_{m,k}, \dots, z_{m,K}\}$$

归一化

$$\hat{y}_m = \begin{pmatrix} \hat{y}_{m,1} \\ \vdots \\ \hat{y}_{m,k} \\ \vdots \\ \hat{y}_{m,K} \end{pmatrix} = \begin{pmatrix} \frac{e^{\theta_1^\top \tilde{x}_m}}{\sum_{j=1}^K e^{\theta_j^\top \tilde{x}_m}} \\ \vdots \\ \frac{e^{\theta_k^\top \tilde{x}_m}}{\sum_{j=1}^K e^{\theta_j^\top \tilde{x}_m}} \\ \vdots \\ \frac{e^{\theta_K^\top \tilde{x}_m}}{\sum_{j=1}^K e^{\theta_j^\top \tilde{x}_m}} \end{pmatrix}$$

$\hat{y}_{m,k}$ 可以被视为 $y'_m = k$ 的概率

多项逻辑回归的解释示例



多项逻辑回归的损失函数

- 在多分类问题中，我们依然采用交叉熵和来度量 y_m 与 \hat{y}_m 之间的误差

$$\mathcal{L}(\Theta) = -\frac{1}{M} \sum_{m=1}^M \left(\sum_{k=1}^K y_{m,k} \log \hat{y}_{m,k} (\theta_k, x_m) \right)$$

其中 $\hat{y}_{m,k} (\theta_k, x_m)$ 表示 $\hat{y}_{m,k}$ 是 θ_k 和 x_m 的函数

多项逻辑回归的概率解释

•多项分布Multi(n, p_1, \dots, p_K)

将 n 个小球随机放入 K 个桶，其中 p_k 表示每个小球进入桶 k 的概率
样本 $\mathbf{y} = (y_1, \dots, y_k, \dots, y_K)$ ，其中 y_k 表示最终第 k 个桶中小球的个数

$$\sum_{k=1}^K p_k = 1, \text{ and } \sum_{k=1}^K y_k = n$$

概率质量函数为

$$\Pr(y_1, \dots, y_K) = \frac{n!}{y_1! y_2! \dots y_K!} p_1^{y_1} p_2^{y_2} \dots p_K^{y_K}$$

多项逻辑回归的概率解释

• 总体假设

$\mathbf{y} = (y_1, \dots, y_k, \dots, y_K)$ 服从 $n = 1$ 的多项分布，即

$$\mathbf{y} \sim \text{Multi}(1, p_1, \dots, p_K; \Theta, \mathbf{x})$$

其中

$$\ln p_k(\boldsymbol{\theta}_k, \mathbf{x}) = \boldsymbol{\theta}_k^\top \tilde{\mathbf{x}} - \ln Z$$

这里 $\ln Z$ 是为了确保约束 $\sum p_k = 1$ 成立而引入的一个变量

$$p_k = \frac{1}{Z} e^{\boldsymbol{\theta}_k^\top \tilde{\mathbf{x}}}$$

$$1 = \sum_{k=1}^K \frac{1}{Z} e^{\boldsymbol{\theta}_k^\top \tilde{\mathbf{x}}} \Rightarrow Z = \sum_{k=1}^K e^{\boldsymbol{\theta}_k^\top \tilde{\mathbf{x}}}$$

$$\Pr(y_1, \dots, y_K) = \prod_{k=1}^K \left(\frac{e^{\boldsymbol{\theta}_k^\top \tilde{\mathbf{x}}}}{\sum_{j=1}^K e^{\boldsymbol{\theta}_j^\top \tilde{\mathbf{x}}}} \right)^{y_k}$$

多项逻辑回归的概率解释

• 最大似然估计

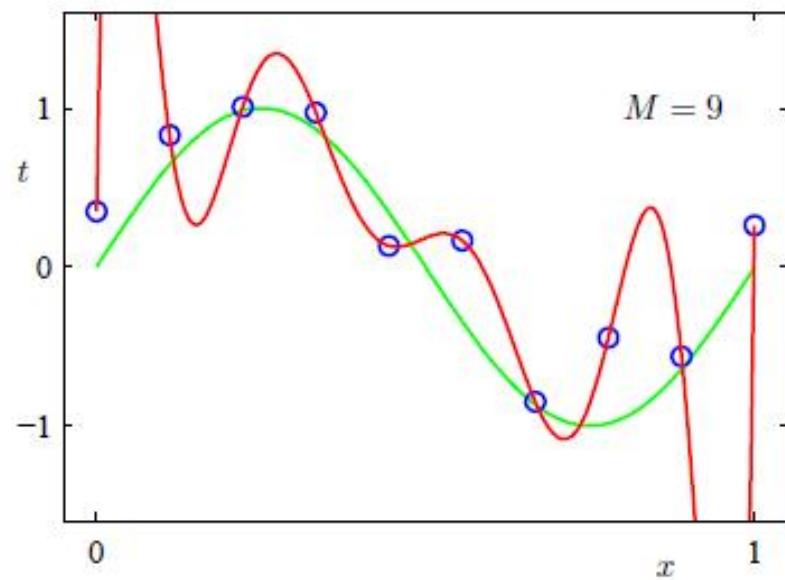
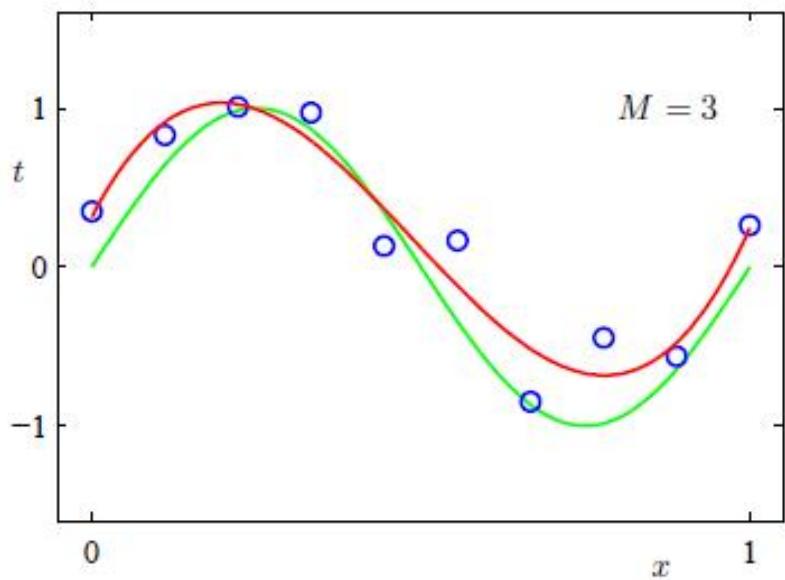
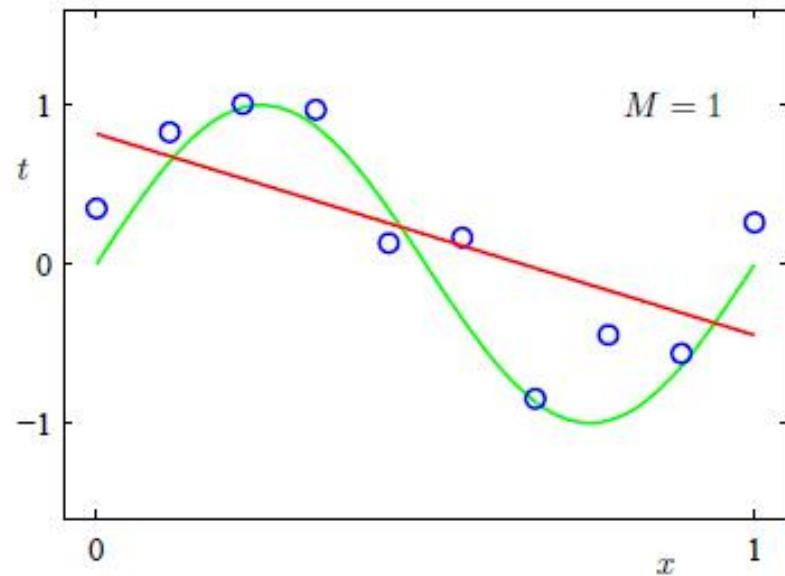
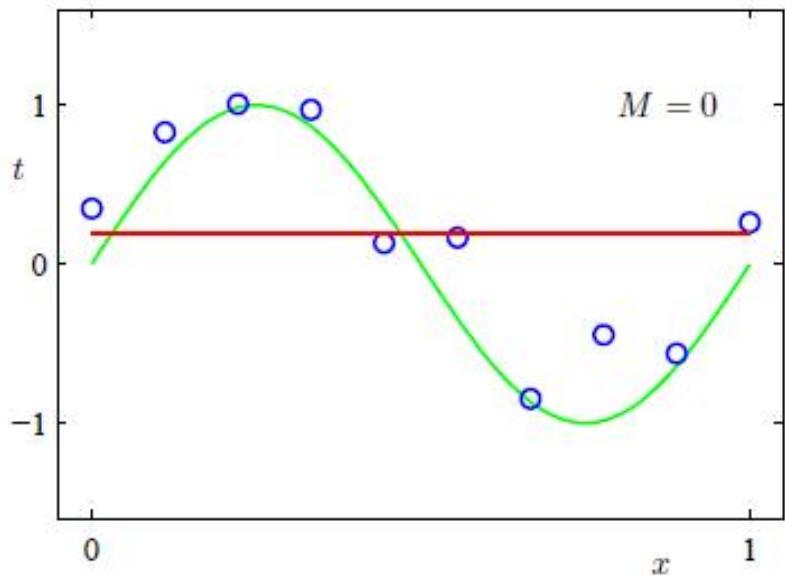
- 根据样本推断参数 θ 就是计算后验概率 $Pr(\theta | Y; X)$

$$\begin{aligned} \Pr(\Theta | Y; X) &\propto \Pr(Y | \Theta; X) = \prod_{m=1}^M \Pr(y_m | \Theta; x_m) \\ &= \prod_{m=1}^M \Pr(y_{m,1}, y_{m,2}, \dots, y_{m,K} | \Theta; x_m) \\ &= \prod_{m=1}^M \prod_{k=1}^K \left(\frac{e^{\theta_k^\top \tilde{x}_m}}{\sum_{j=1}^K e^{\theta_j^\top \tilde{x}_m}} \right)^{y_{m,k}} \end{aligned}$$

$$\log \Pr(\Theta | Y; X) \propto \sum_{m=1}^M \sum_{k=1}^K y_{m,k} \log \frac{e^{\theta_k^\top \tilde{x}_m}}{\sum_{j=1}^K e^{\theta_j^\top \tilde{x}_m}}$$

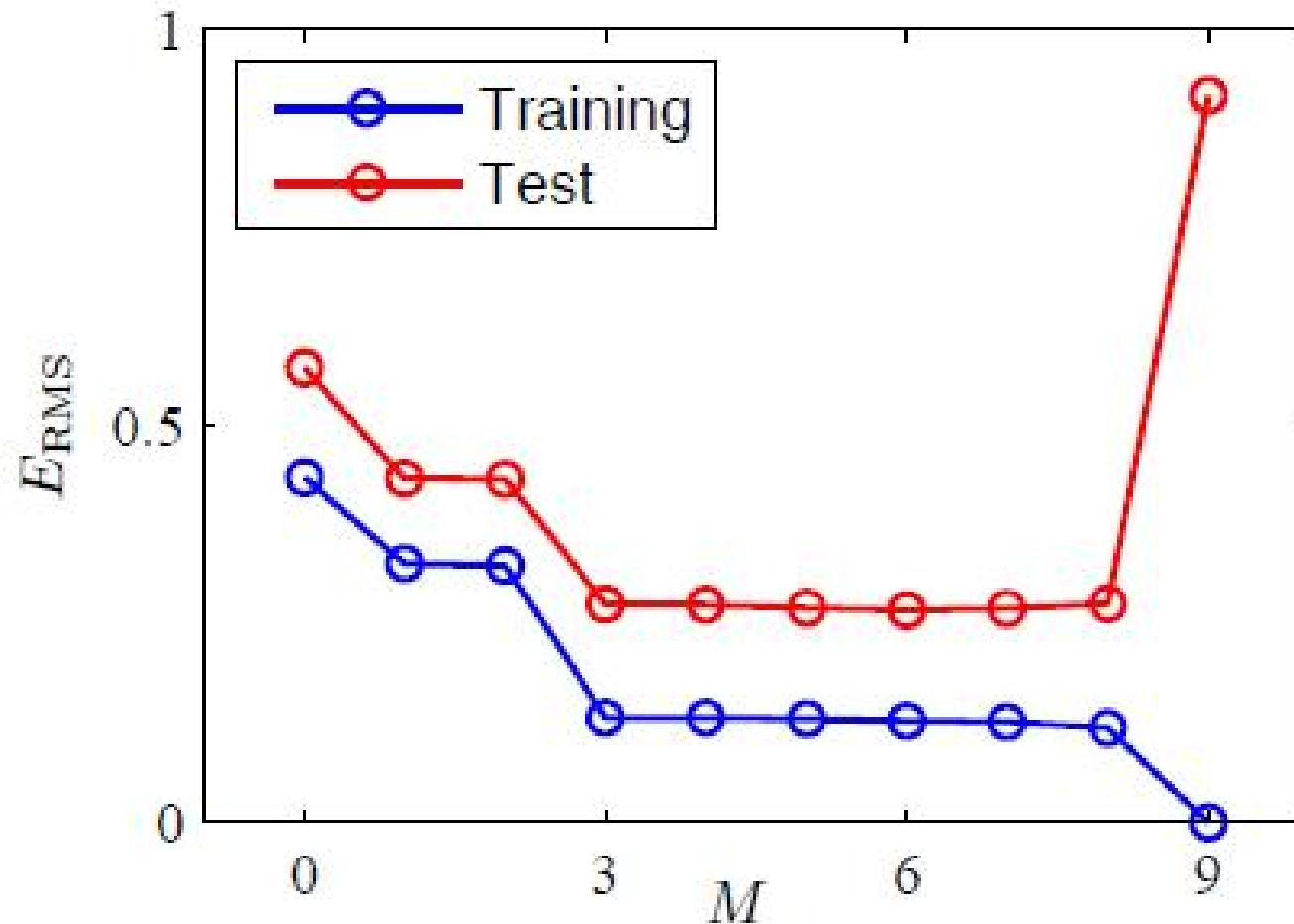
负的交叉熵损失函数

概率视角下的正则项



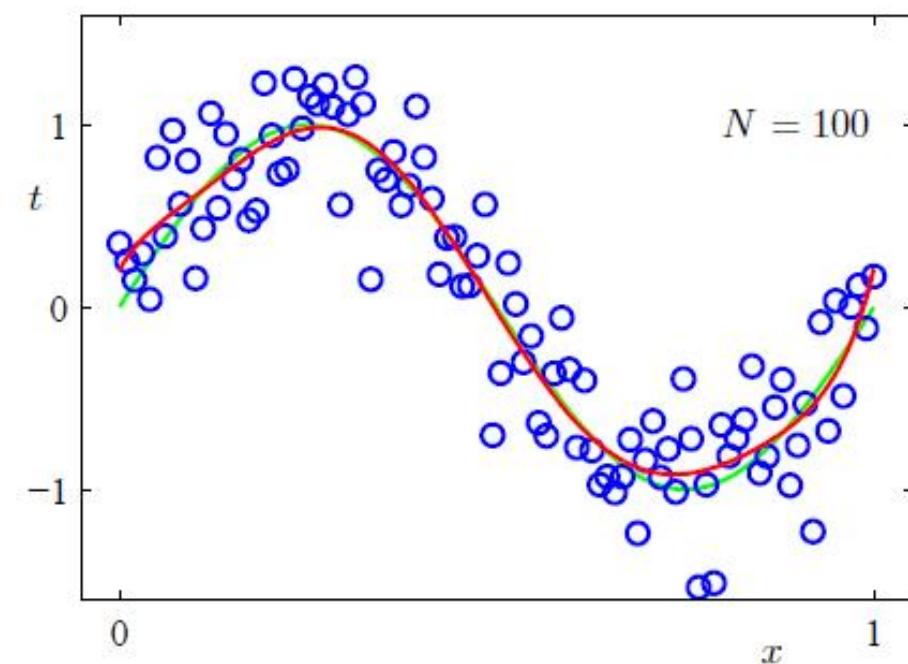
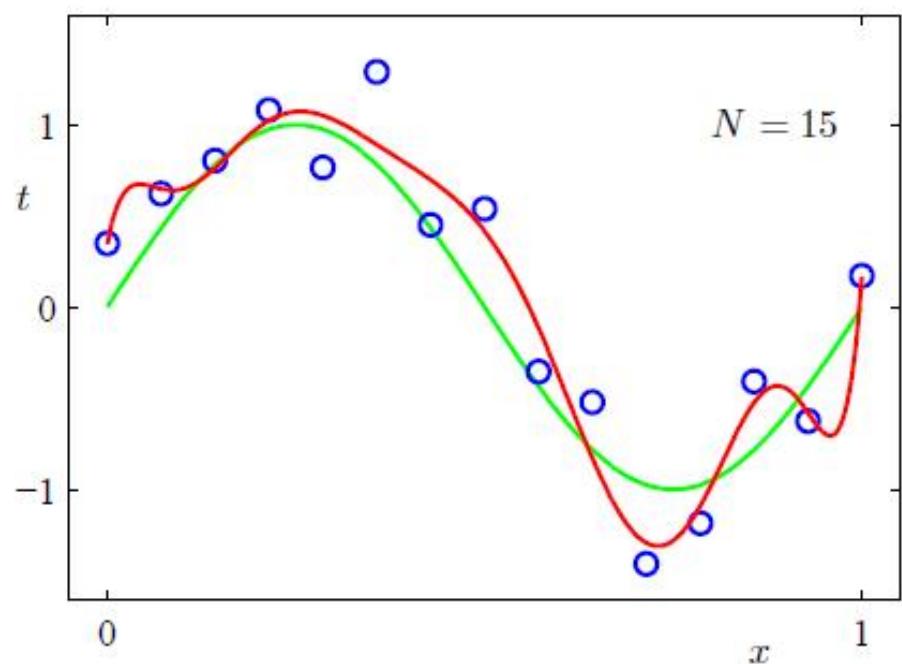
过拟合

- 模型过拟合的表现



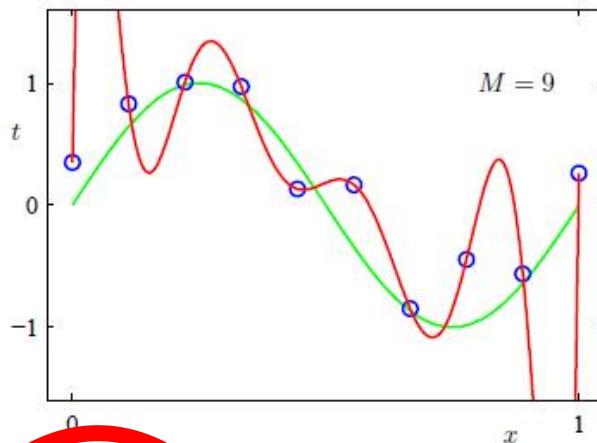
克服过拟合的方法

- 通过增加训练数据来对抗过拟合



过拟合

- 模型过拟合的表现



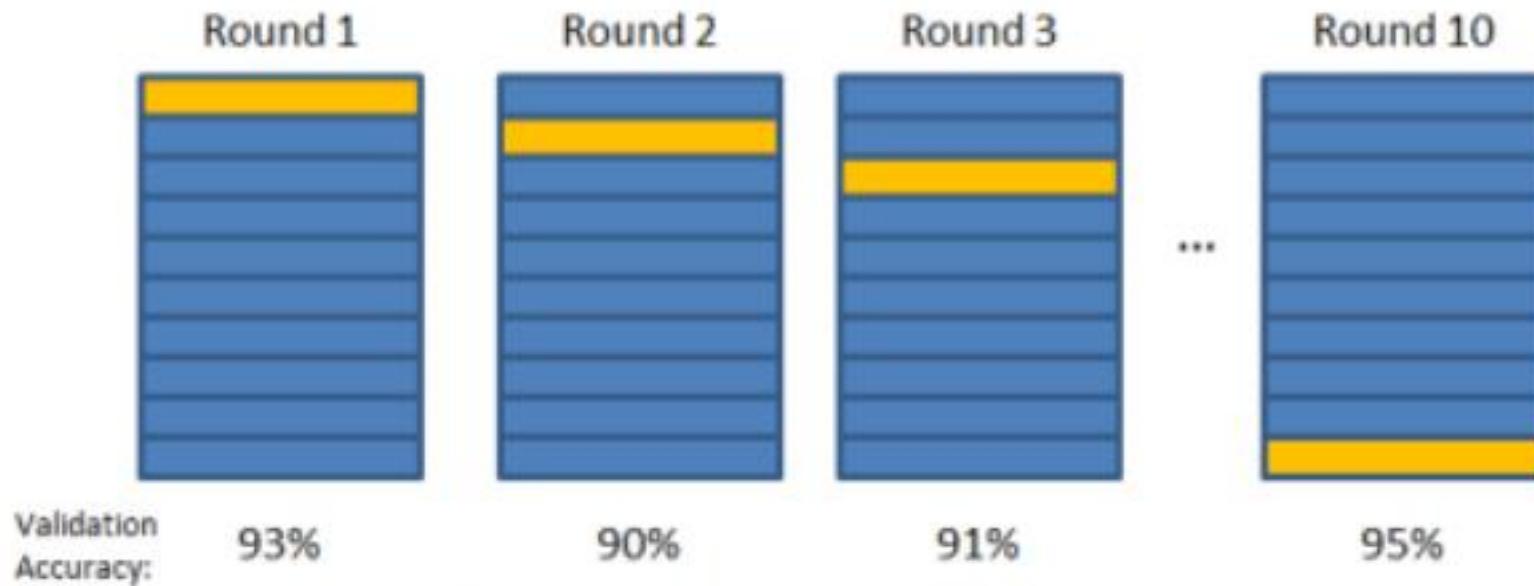
	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

克服过拟合的方法

- 通过交叉验证对抗过拟合

- K-fold Cross validation

 Validation Set
 Training Set



$$\text{Final Accuracy} = \text{Average}(\text{Round 1}, \text{Round 2}, \dots)$$

正则项方法

先验呢？



贝叶斯公式展开

后验

似然

先验

$$P(\mathbf{w}|\epsilon) = \frac{P(\epsilon|\mathbf{w}) P(\mathbf{w})}{P(\epsilon)}$$

假设 \mathbf{w} 符合高斯分布的先验：

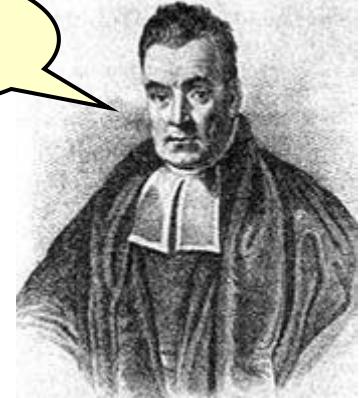
$$\mathbf{w} \sim \mathcal{N}(w|0, \sigma_w^2)$$

对于给定 \mathbf{w} , 其出现的先验概率为

$$P(\mathbf{w}) = \prod_{i=1}^m \mathcal{N}(w_i|0, \sigma_w^2) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left(-\frac{w_i^2}{2\sigma_w^2}\right)$$

正则项方法

先验呢？



贝叶斯公式展开

后验

似然 先验

$$P(\mathbf{w}|\epsilon) = \frac{P(\epsilon|\mathbf{w}) P(\mathbf{w})}{P(\epsilon)}$$

\mathbf{w} 的后验概率为：

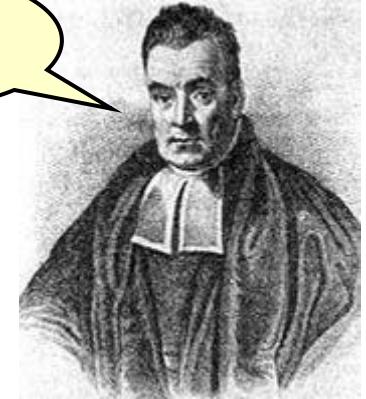
$$P(\mathbf{w}|\epsilon) \propto \prod_{i=1}^n P(\epsilon_i|\mathbf{w}) \prod_{i=1}^m P(w_i)$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mathbf{w}\mathbf{x}_i)^2}{2\sigma^2}\right) \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma_w} \exp\left(-\frac{w_i^2}{2\sigma_w^2}\right)$$

$$\ln P(\mathbf{w}|\epsilon) \propto -\sum_{i=1}^n (y_i - \mathbf{w}\mathbf{x}_i)^2 - \lambda \sum_{i=1}^m w_i^2 \quad \lambda = \frac{\sigma^2}{\sigma_w^2}$$

正则项方法

先验呢？



贝叶斯公式展开

后验

$$P(\mathbf{w}|\epsilon) = \frac{P(\epsilon|\mathbf{w}) P(\mathbf{w})}{P(\epsilon)}$$

似然 先验

最大似然



$$\min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}\mathbf{x}_i)^2$$

L2 正则
L2 regularization

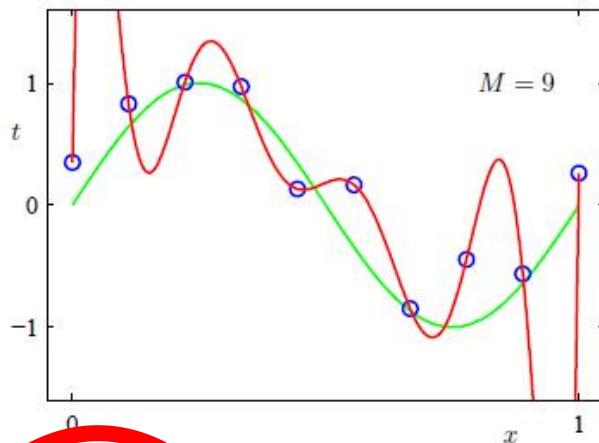
最大后验

$$\min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}\mathbf{x}_i)^2 + \lambda \sum_{i=1}^m w_i^2, \quad \lambda = \frac{\sigma^2}{\sigma_w^2}$$

L2正则是在目标函数中引入了 w 的高斯分布作为先验知识。

过拟合

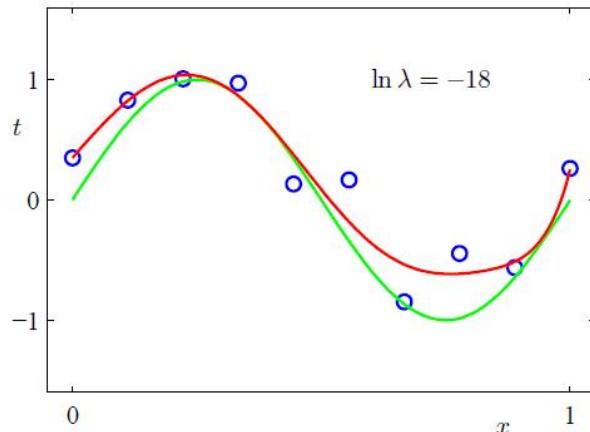
- 模型过拟合的表现



	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

正则项的效果

- M = 9的多项式

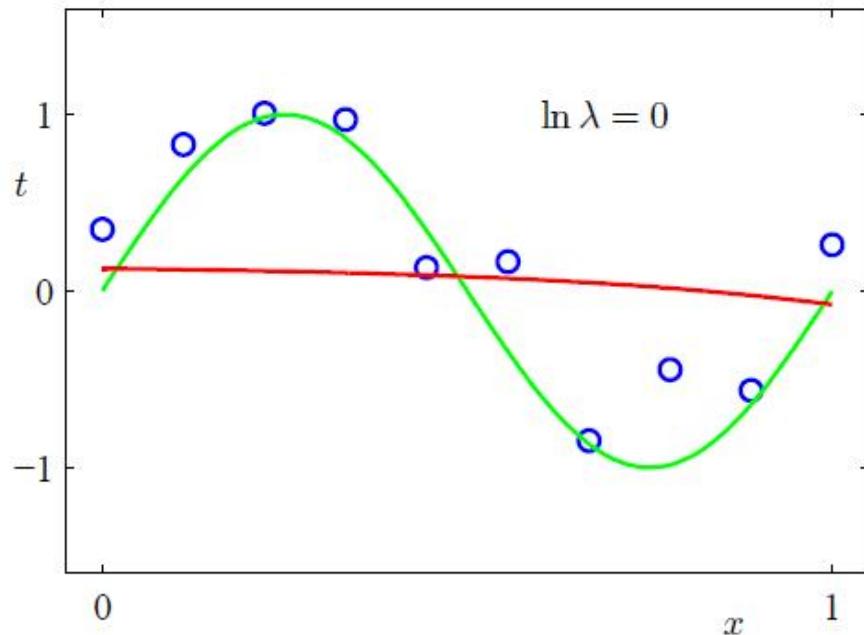
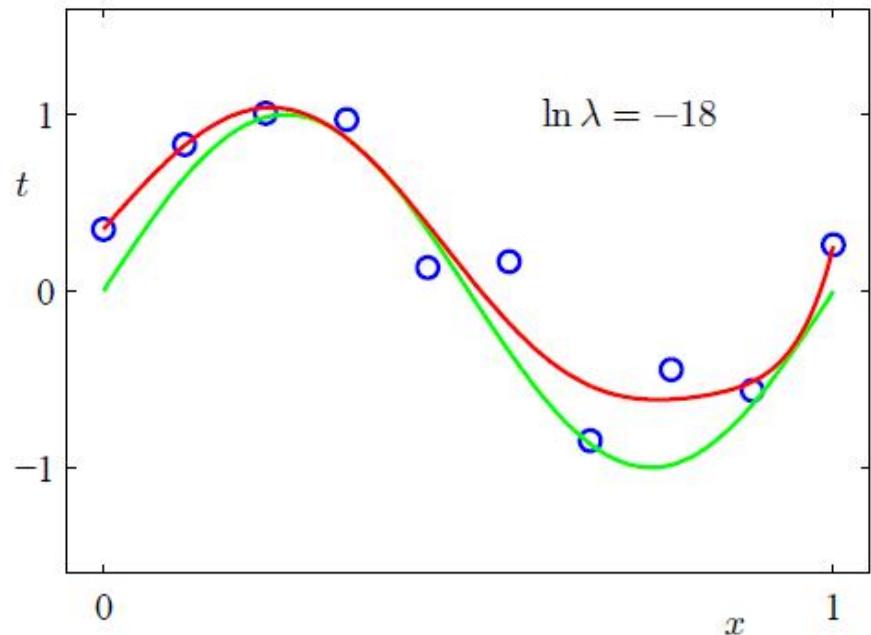
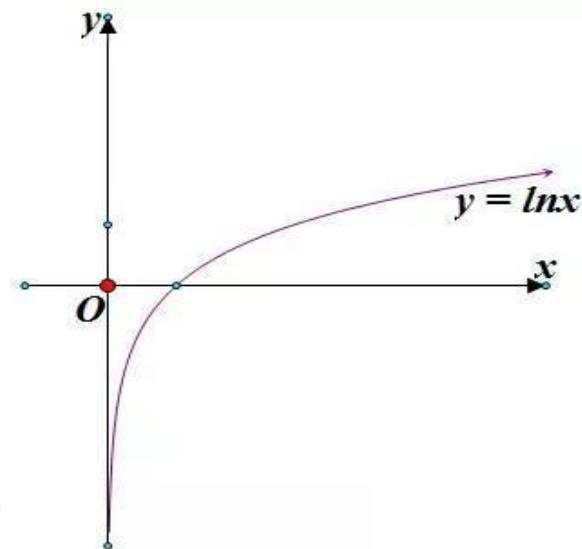


	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
w_0^*	0.35	0.35	0.13
w_1^*	232.37	4.74	-0.05
w_2^*	-5321.83	-0.77	-0.06
w_3^*	48568.31	-31.97	-0.05
w_4^*	-231639.30	-3.89	-0.03
w_5^*	640042.26	55.28	-0.02
w_6^*	-1061800.52	41.32	-0.01
w_7^*	1042400.18	-45.95	-0.00
w_8^*	-557682.99	-91.53	0.00
w_9^*	125201.43	72.68	0.01

正则项的效果

- M = 9的多项式

$$\min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}\mathbf{x}_i)^2 + \lambda \sum_{i=1}^m w_i^2, \lambda = \frac{\sigma^2}{\sigma_w^2}$$



谢 谢

E-mail: jywang@buaa.edu.cn

Weibo: @王静远BUAA