回归模型(2)

估计的性质

- 假设任意输入 x 的平均响应确实可以使用真实参数 $\mathbf{w}^*$ 的线性函数建模

$$\mathbb{E}\left\{y|x\right\} = f\left(x;\mathbf{w}^*\right) = (\mathbf{w}^*)^{\mathrm{T}}\begin{bmatrix}1\\ x\end{bmatrix}$$

- 问题：基于有限训练样本集估计的参数 $\hat{\mathbf{w}}$ 在任何意义上都接近参数 $\mathbf{w}^*$ 吗?

偏差

- 偏差(bias)：测量对于参数 $\mathbf{w}^*$ 的任何系统偏离

$$bias = \mathbb{E}\{\hat{\mathbf{w}}\} - \mathbf{w}^*$$

  - 对相同规模的重新采样训练集计算期望值，且
  - 训练集中每个样本对 (x, y) 都是来自分布 P 的独立样本
- 在线性回归中，估计 $\hat{\mathbf{w}}$ 是无偏的(unbiased)，即

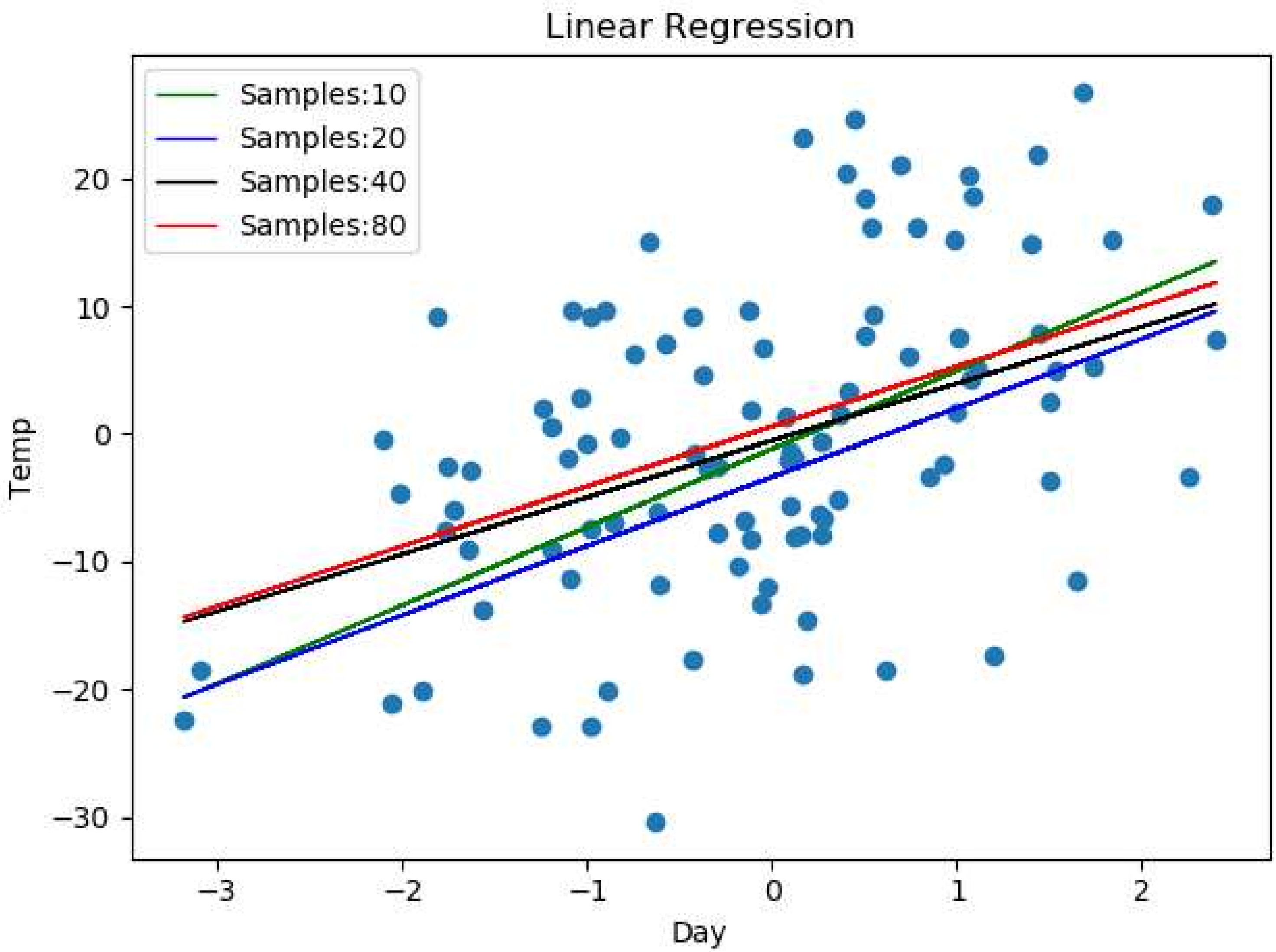
$$\mathbb{E}\{\hat{\mathbf{w}}\} - \mathbf{w}^* = 0$$

- 预测也是无偏的

$$\mathbb{E}\{f(x;\hat{\mathbf{w}})\} = \mathbb{E}\left\{\hat{\mathbf{w}}^{\mathrm{T}}\begin{bmatrix}1\\x\end{bmatrix}\right\} = \mathbf{w}^{*\mathrm{T}}\begin{bmatrix}1\\x\end{bmatrix} = f(x;\mathbf{w}^*)$$

泛化

❖ 问题：随着训练样本(x, y)数量增加，线性预测是如何改进的?



泛化

- 假设：样本 (x, y) 来自于某个未知分布 P；使用训练样本估计参数 $\hat{\mathbf{w}}$
  - 训练样本：$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - 测试样本：$\{(x_{n+1}, y_{n+1}), (x_{n+2}, y_{n+2}), \ldots, (x_{n+N}, y_{n+N})\}$
- 误差类型

$$\text{mean training error} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{w}_0 - \hat{w}_1 x_i\right)^2$$

$$\text{mean test error} = \frac{1}{N}\sum_{i=n+1}^{n+N}\left(y_i - \hat{w}_0 - \hat{w}_1 x_i\right)^2$$

$$\text{generalization error} = \mathbb{E}_{(x,y)\sim P}\left\{\left(y - \hat{w}_0 - \hat{w}_1 x\right)^2\right\}$$

  - 注意：$\hat{w}_0$ 和 $\hat{w}_1$ 本身是与训练集合有关的随机变量

泛化

- 将泛化误差

$$\mathbb{E}_{(x,y)\sim P}\left\{\left(y-\hat{w}_0-\hat{w}_1 x\right)^2\right\}$$

分解为两项：

- 最好预测值的误差

$$\mathbb{E}_{(x,y)\sim P}\left\{\left(y-w_0^*-w_1^* x\right)^2\right\}=\min_{w_0,w_1}\mathbb{E}_{(x,y)\sim P}\left\{\left(y-w_0-w_1 x\right)^2\right\}$$

- 如何估计最好预测值

$$\mathbb{E}_{(x,y)\sim P}\left\{\left(\left(w_0^*+w_1^* x\right)-\left(\hat{w}_0+\hat{w}_1 x\right)\right)^2\right\}$$

- 适合于分布 P 描述的所有输入/输出关系

泛化

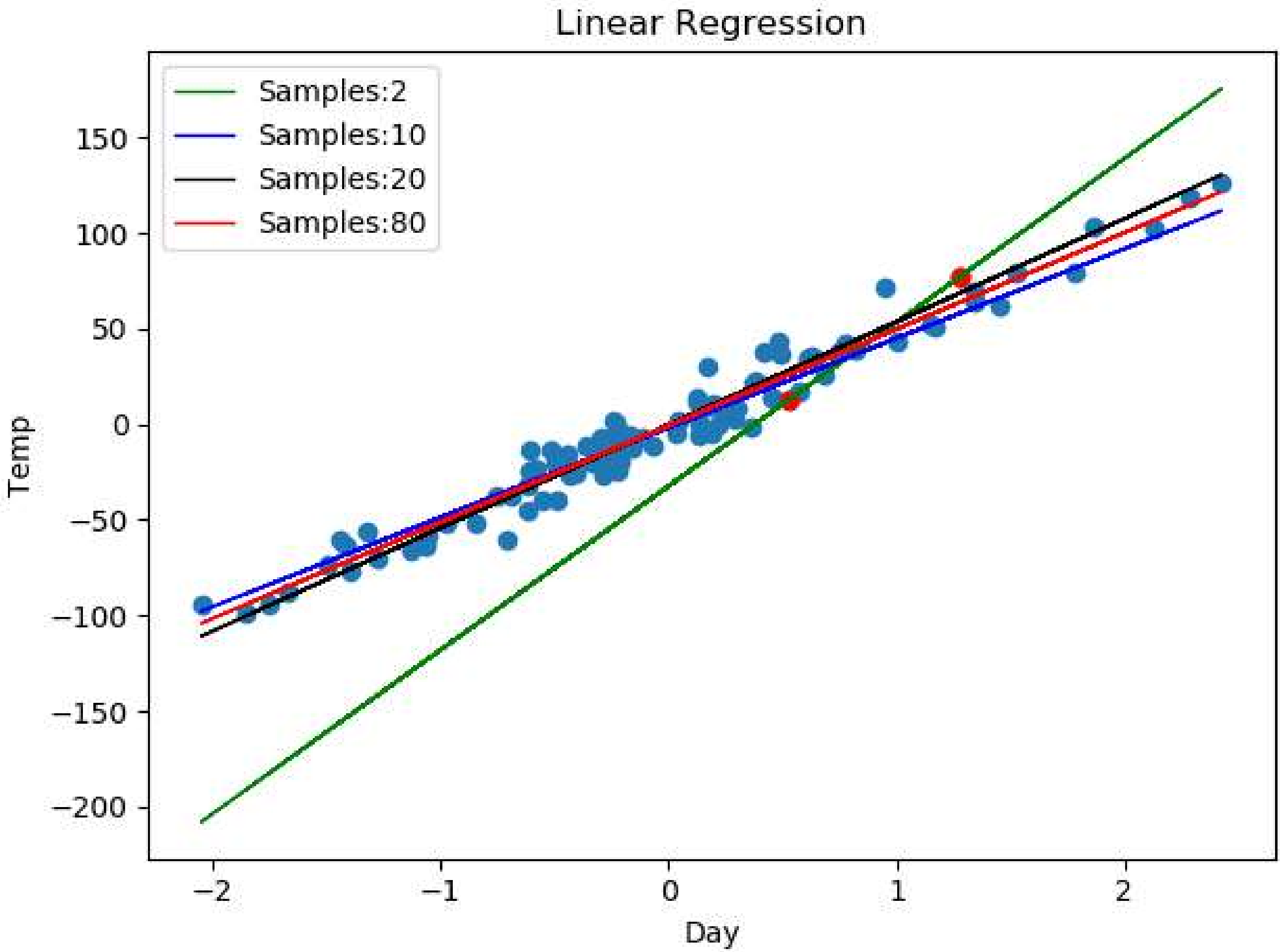
- 推导

$$
\begin{aligned}
\left(y-\hat{w}_0-\hat{w}_1x\right)^2 &= \left(\left(y-\left(w_0^*+w_1^*x\right)+\left(w_0^*+w_1^*x\right)-\left(\hat{w}_0+\hat{w}_1x\right)\right)\right)^2 \\
&= \left(y-\left(w_0^*+w_1^*x\right)\right)^2+ \\
&\boxed{2\left(y-\left(w_0^*+w_1^*x\right)\right)\left(\left(w_0^*+w_1^*x\right)-\left(\hat{w}_0+\hat{w}_1x\right)\right)+} \\
&\left(\left(w_0^*+w_1^*x\right)-\left(\hat{w}_0+\hat{w}_1x\right)\right)^2
\end{aligned}
$$

- 当我们对 $(x,y)\sim P$ 取期望时，交叉项消失。

过度拟合

- 由于训练样本少，线性回归模型可以达到零训练误差，但无论怎样，都会有很大的泛化误差。



- 当训练误差不再与泛化误差有任何关系时，模型过拟合数据

交叉验证

- 交叉验证帮助我们只利用训练集合估计泛化误差
- 例：余一交叉误差

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\hat{w}_0^{-i} + \hat{w}_1^{-i} x_i\right)\right)^2$$

其中：$(\hat{w}_0^{-i}, \hat{w}_1^{-i})$ 是除去第 i 个训练样本计算的最小二乘解

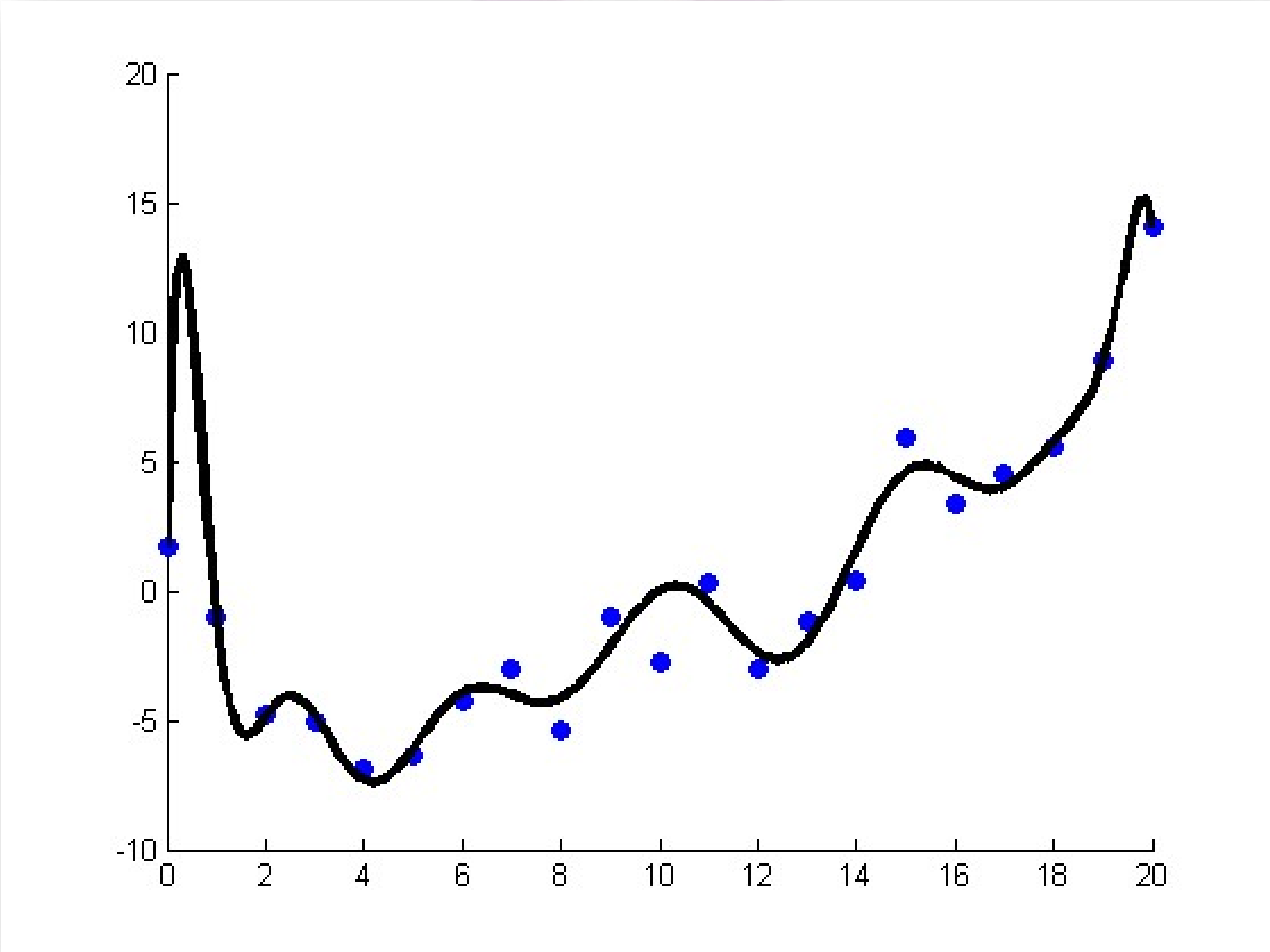
岭回归

- 问题：使用最大似然估计会造成过度拟合
- 原因
  - 最大似然估计使用最佳拟合训练数据的参数值建模
  - 如果训练数据包含噪声，则将导致模型复杂化
- 例
  - 使用 14 阶多项式拟合 21 个数据点
  - 使用最小二乘法求出最优解
  - 结果为

$$6.560,\ -36.934,\ -109.255,\ 543.452,\ 1022.561,\ -3046.224,\ -3768.013,\ 8524.540,\ 6607.897,\ -12640.058,\ -5530.188,\ 9479.730,\ 1774.639,\ -2821.526$$

岭回归

- 参数绝对值很大，不稳定
- 一旦输入发生变化，响应将会发生很大变化



岭回归

- 解决办法：使用零均值高斯先验概率鼓励绝对值较小的参数
  - 先验概率

$$p(\mathbf{w})=\prod_j \mathcal{N}\left(w_j \middle| 0,\tau^2\right)$$

    其中 $1/\tau^2$ 控制先验概率的强度
  - 后验概率估计

$$\arg\max_{\mathbf{w}} \sum_{i=1}^{n} \log \mathcal{N}\left(y_i \middle| w_0+\mathbf{w}^{\mathrm{T}}\mathbf{x}_i,\sigma^2\right)+\sum_{j=1}^{D}\log \mathcal{N}\left(w_j \middle| 0,\tau^2\right)$$

  - 等价于

$$\mathrm{J}(\mathbf{w})=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\left(w_0+\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)\right)^2+\lambda\|\mathbf{w}\|_2^2$$

    其中 $\lambda=\sigma^2/\tau^2$ 称为收缩系数，$\|\mathbf{w}\|_2^2=\sum_j w_j^2=\mathbf{w}^{\mathrm{T}}\mathbf{w}$ 是L2范数的平方

岭回归

- 上述方法称为岭回归或补偿最小二乘法
  - 目标函数

$$\mathrm{J}(\mathbf{w})=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\left(w_0+\mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)\right)^2+\lambda\|\mathbf{w}\|_2^2$$

  - 最优解

$$\hat{\mathbf{w}}_{\text{ridge}}=\left(\lambda\mathbf{I}+\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

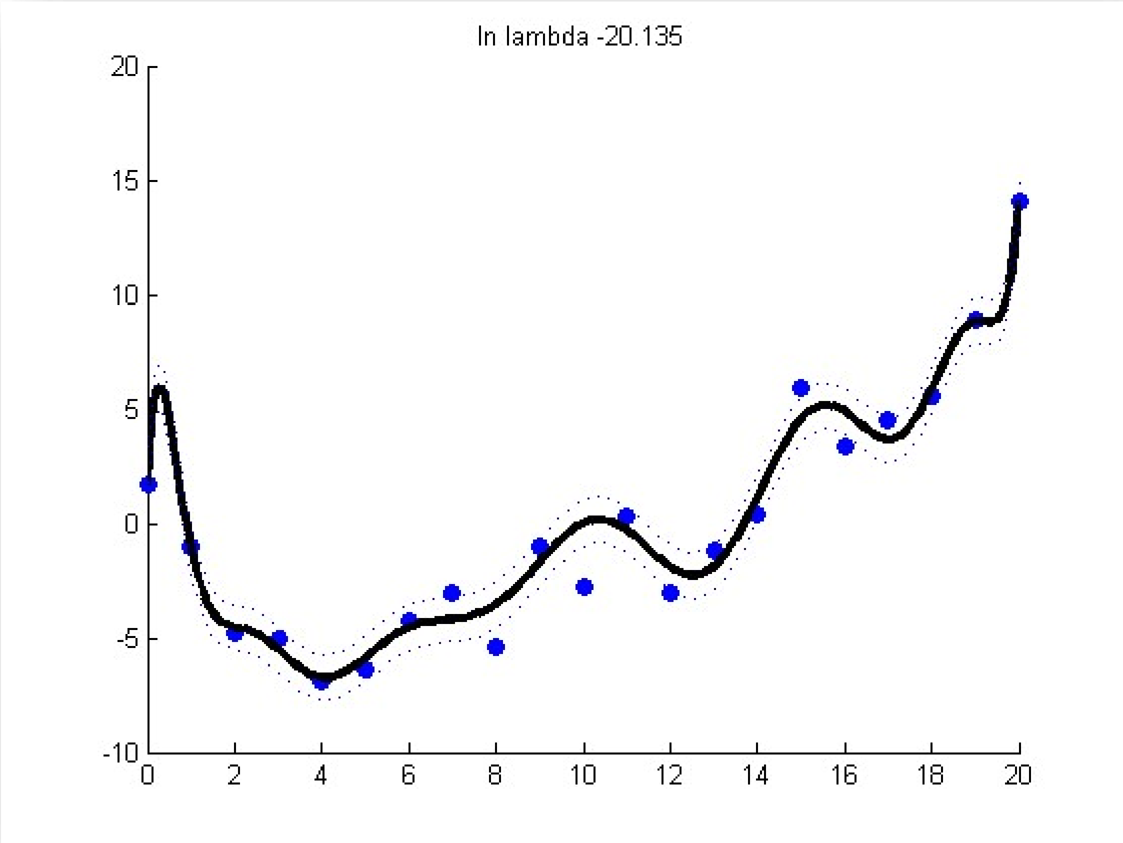
  - 通过给模型参数添加高斯先验概率来鼓励其取绝对值较小的数值，称为L2正则化或权重衰减
  - 使用 $\lambda=10^{-3}$ 计算时，得到的系数为

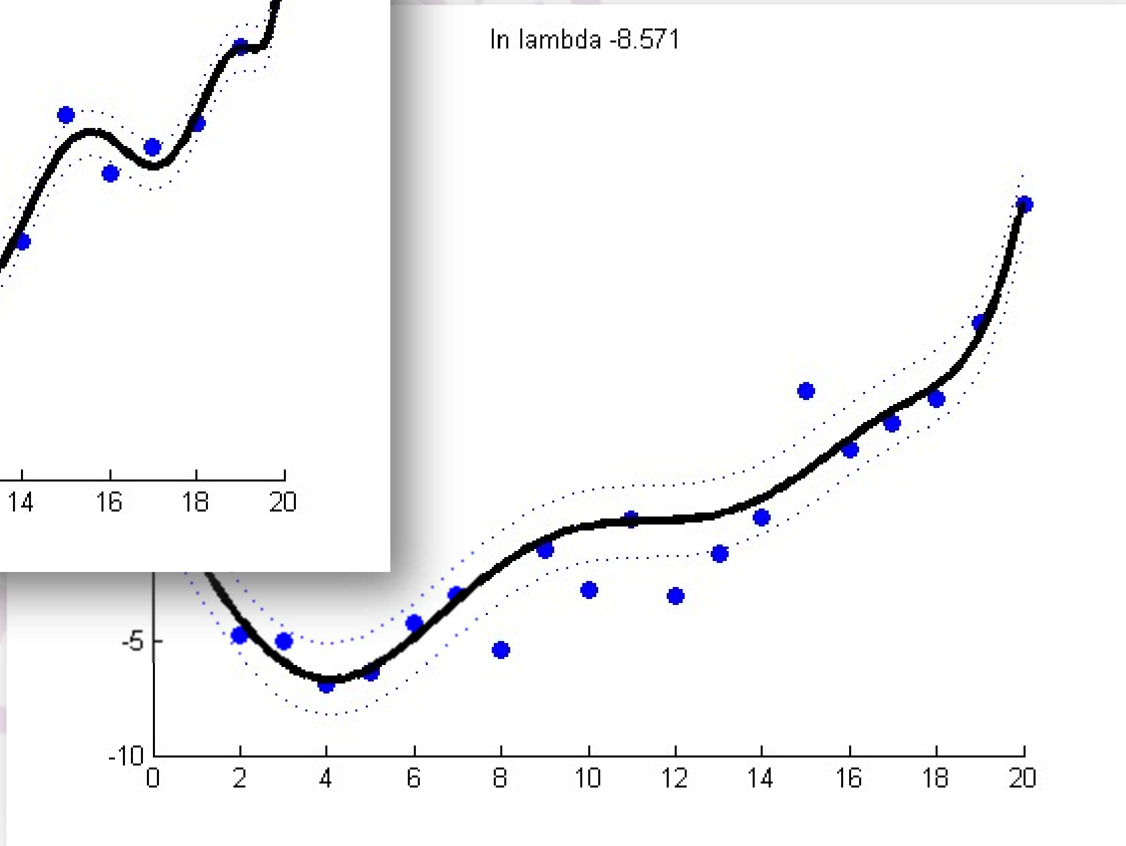
$$2.128,\ 0.807,\ 16.457,\ 3.704,\ -24.948,\ -10.472,\ -2.625,$$

$$4.360,\ 13.711,\ 10.063,\ 8.716,\ 3.966,\ -9.349,\ -9.232$$

岭回归

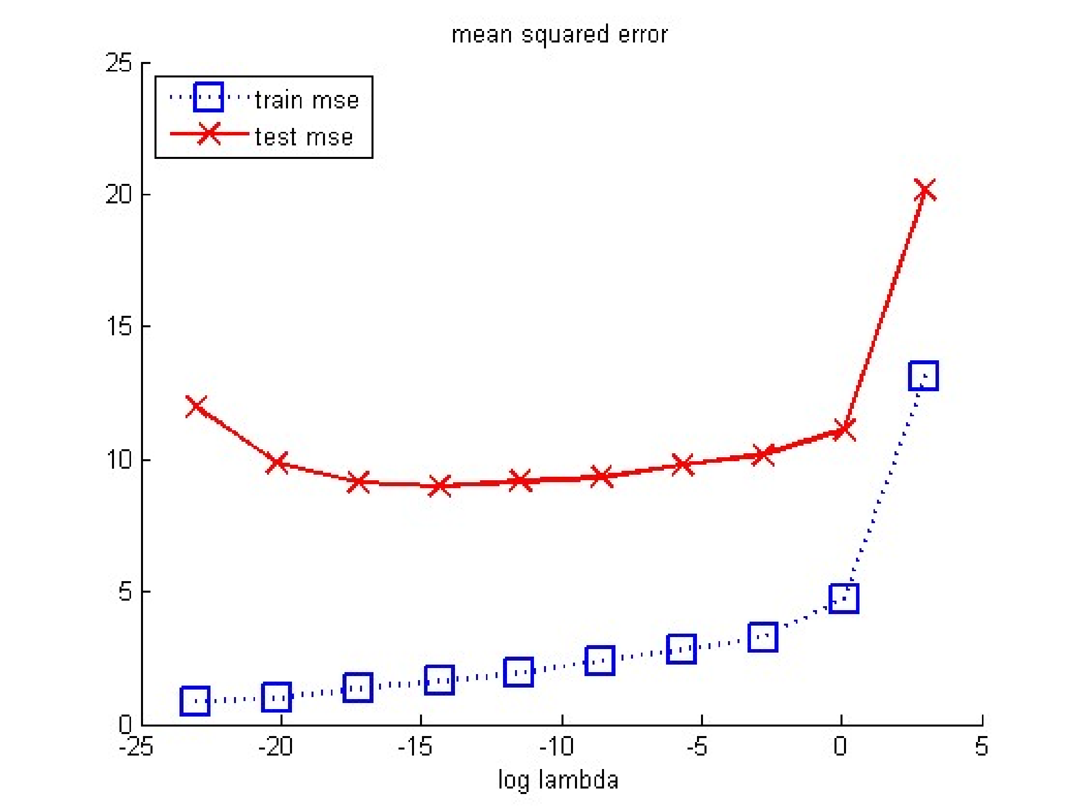
- 示意图（随着 λ 值增加拟合函数将更平滑，系数变得更小）

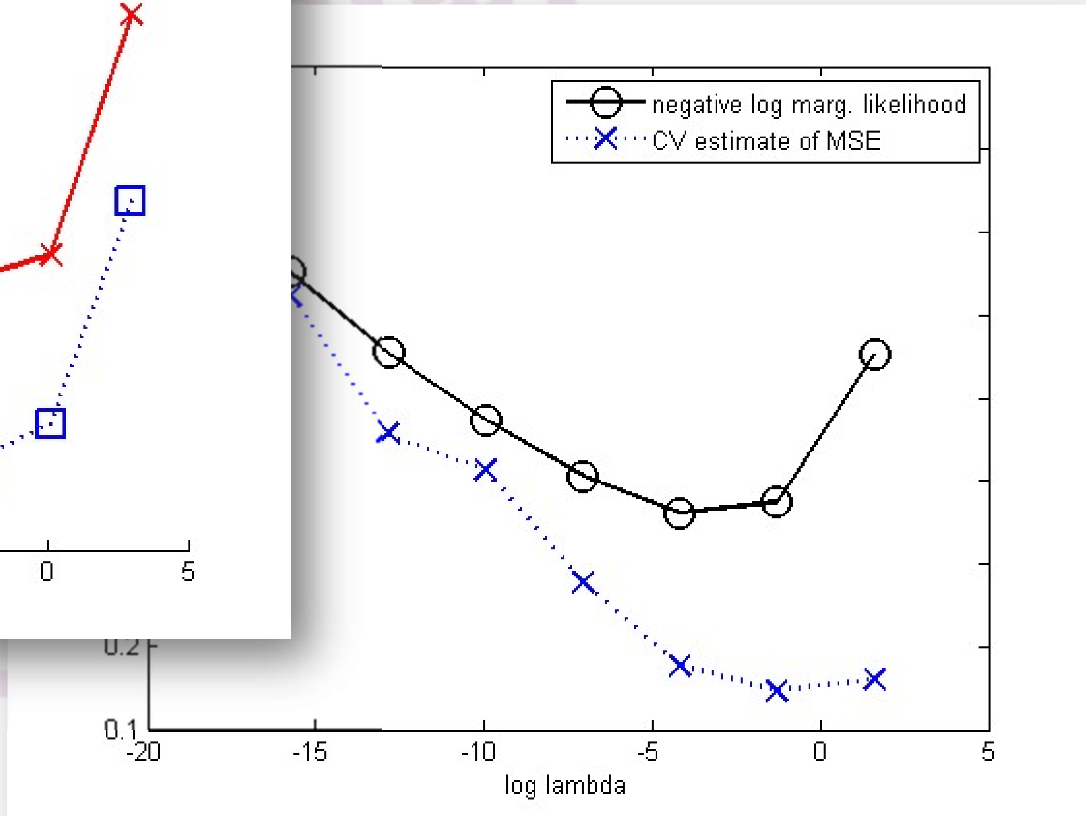




岭回归

- 随着λ 值增加，训练样本误差增加，测试样本误差呈 U 型，对应于过拟合和欠拟合。
  - 使用交叉验证样本确定λ 值。





套索回归(Lasso)

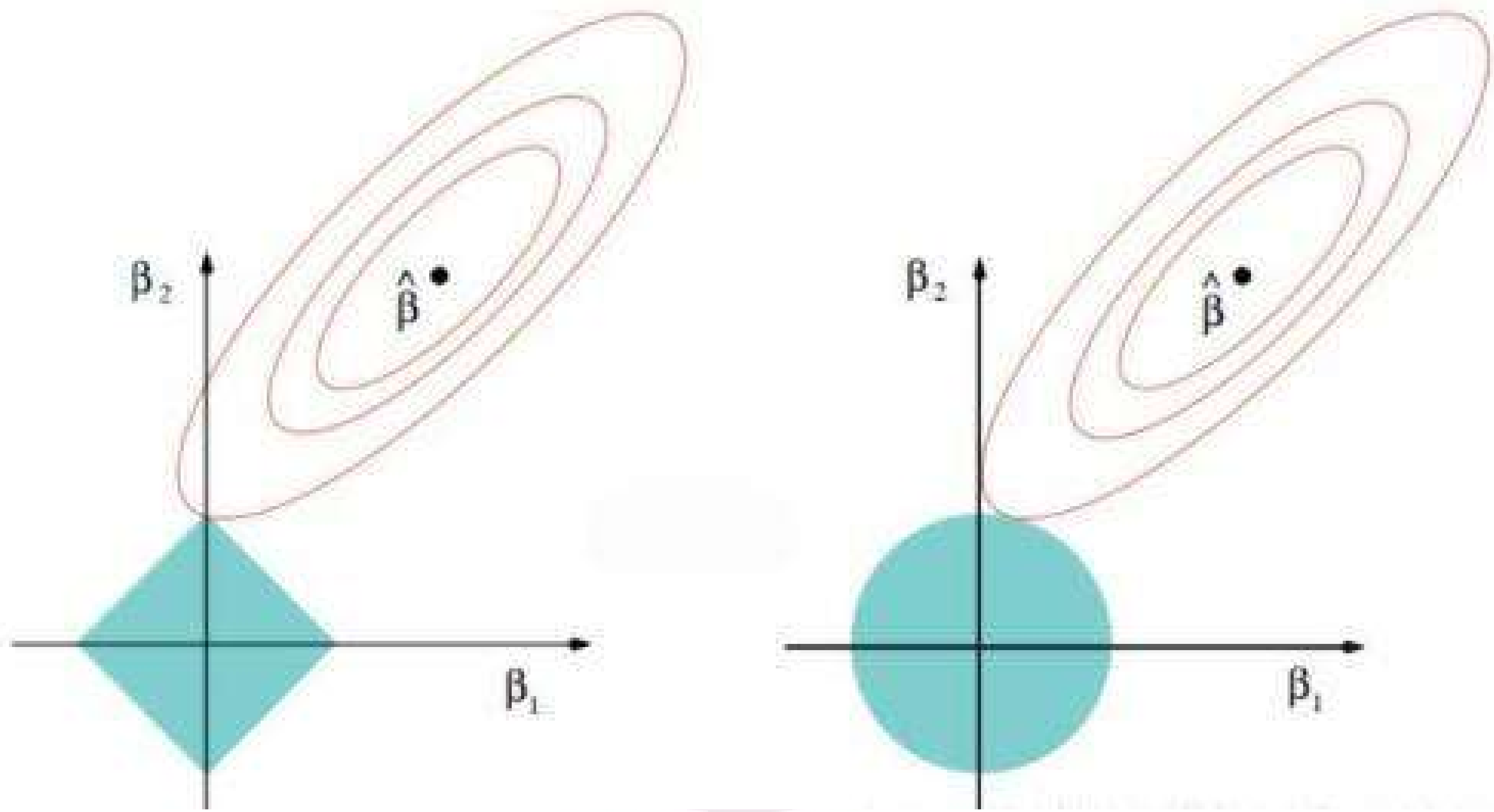
- 类似于岭回归
- 优化目标

$$J(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)\right)^2 + \lambda\|\mathbf{w}\|_1$$

- 变量选择：由于约束域为正方形，故会存在与坐标轴的切点，使得部分维度特征权重为零，因此很容易产生稀疏的结果。
- 求解方法：目标函数存在不可导处，故梯度下降法等优化算法失效。
  - 坐标轴下降法(coordinate descent)
  - 最小角回归法(Least Angle Regression, LARS)

岭回归和套索回归

- 两种方法的差异
  - 等高线与约束域的切点就是目标函数的最优解



弹性网络回归

- 岭回归和套索回归的混合体
- 优化目标

$$\mathrm{J}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x}_i\right)\right)^2 + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2$$

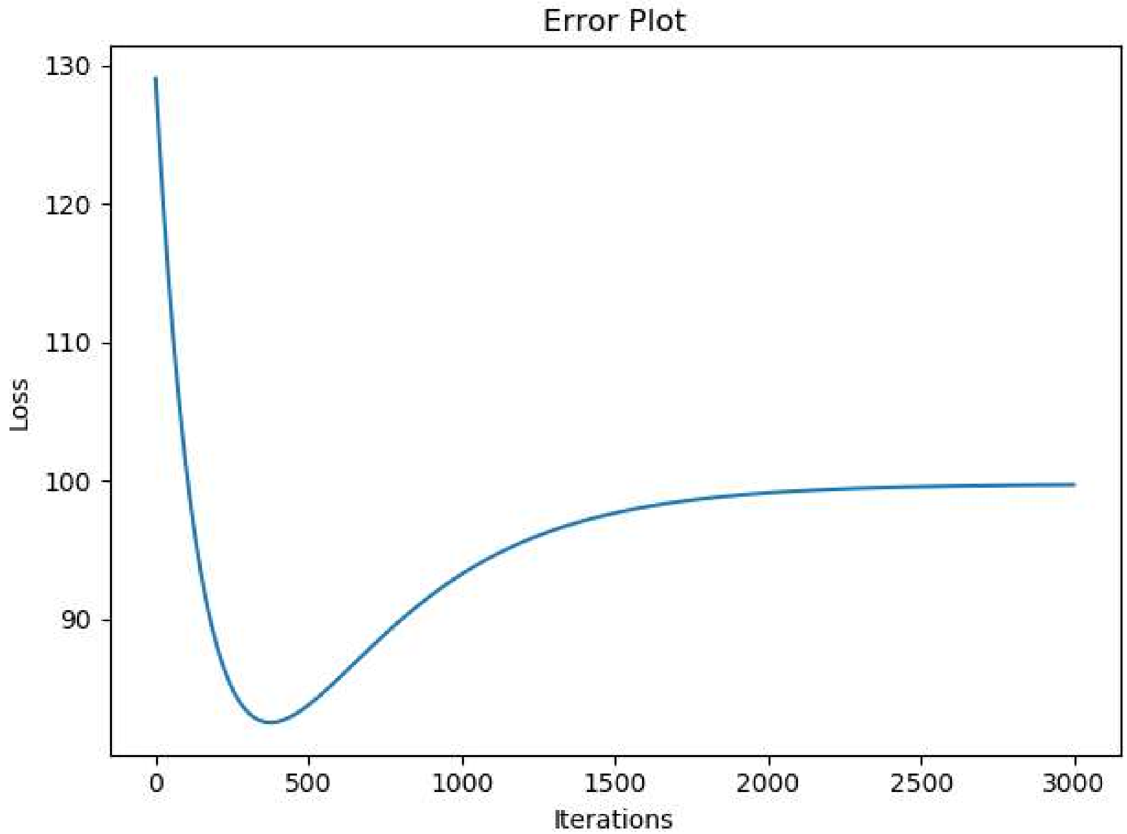
- 特点
  - 在高度相关变量的情况下，会产生群体效应
  - 选择变量的数目没有限制
  - 承受双重收缩

梯度下降算法

- 假设：目标函数为 $J(\mathbf{w})$，学习率为 α
- 梯度下降优化法
  - repeat until convergence {

$$\mathbf{w} \Leftarrow \mathbf{w} - \alpha \frac{\partial}{\partial \mathbf{w}} J(\mathbf{w})$$

  - }



数值计算技巧

- 问题：为了数值稳定计算，最好避免逆矩阵的计算

$$\hat{\mathbf{w}}_{\text{ridge}} = \left(\lambda\mathbf{I} + \mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- 技巧：
  - 令先验概率 $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{0}, \Lambda^{-1}\right)$，其中 Λ 是精度矩阵。
  - 在岭回归中 $\Lambda = \left(1/\tau^2\right)\mathbf{I}$。
  - 使用来自于先验概率的“虚拟数据”扩充原始数据，得到

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}/\sigma \\ \sqrt{\Lambda} \end{pmatrix}, \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}/\sigma \\ \mathbf{0} \end{pmatrix}$$

其中，$\Lambda = \sqrt{\Lambda}\sqrt{\Lambda}^{\mathrm{T}}$ 是 Λ 的 Cholesky 分解，$\tilde{X}$ 矩阵维数为 $(N + D) \times D$，额外行是伪数据。

Cholesky分解

- 定义：把一个对称正定矩阵表示为一个下三角矩阵和其转置矩阵的乘积，称为 Cholesky 分解。

$$\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$$

- 对称正定矩阵的性质
  - 如果 $\mathbf{A}$ 是对称正定的，则 $\mathbf{A}^{-1}$ 亦是对称正定的，且 $a_{ii} > 0$；
  - $\mathbf{A}$ 的顺序主子阵 $\mathbf{A}_k$ 亦是对称正定的；
  - $\mathbf{A}$ 的特征值 $\lambda_i > 0$；
  - $\mathbf{A}$ 的全部顺序主子式 $\det(\mathbf{A}_k)$（$\mathbf{A}$ 可做 Cholesky 分解的充分条件）
- 可以证明：最大后验概率估计为

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

  - 令 $\tilde{\mathbf{X}} = \mathbf{Q}\mathbf{R}$ 是 $\mathbf{X}$ 的 QR 分解，其中 $\mathbf{Q}$ 是正交的，$\mathbf{R}$ 是上三角矩阵，有

$$\left(\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \left(\mathbf{R}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\mathbf{R}\right)^{-1} = \left(\mathbf{R}^{\mathrm{T}}\mathbf{R}\right)^{-1} = \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}$$

Cholesky分解

从此

$$\hat{\mathbf{w}}_{\text{ridge}} = \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\tilde{\mathbf{y}} = \mathbf{R}^{-1}\mathbf{Q}\tilde{\mathbf{y}}$$

- 结论：因为 $\mathbf{R}$ 是上三角矩阵，故容易求解。

加性模型

- 推广到参数 w 是线性的模型，而不一定输入 x 是线性的
  - 简单线性预测 $f:\mathcal{R}\to\mathcal{R}$

$$f(x;\mathbf{w}) = w_0 + w_1 x$$

  - m 阶多项式预测 $f:\mathcal{R}\to\mathcal{R}$

$$f(x;\mathbf{w}) = w_0 + w_1 x + \ldots + w_{m-1}x^{m-1} + w_m x^m$$

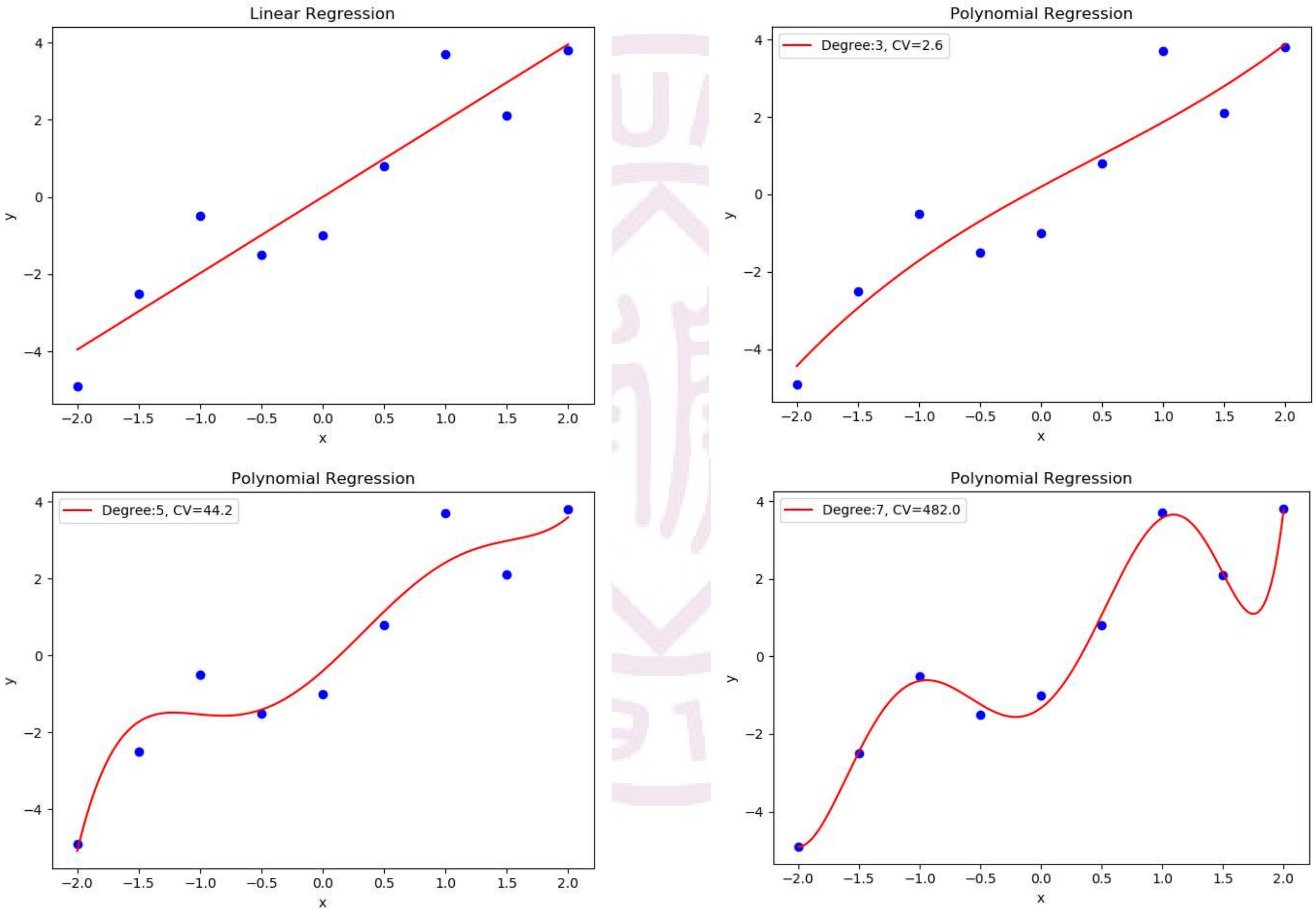
  - 多维线性预测 $f:\mathcal{R}^d\to\mathcal{R}$

$$f(\mathbf{x};\mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_{d-1}x_{d-1} + w_d x_d$$

其中： $\mathbf{x} = [x_1 x_2 \ldots x_{d-1} x_d]^{\mathrm{T}}, d = \dim(\mathbf{x})$

多项式回归

❖ 例



Linear Regression
Polynomial Regression
Degree:3, CV=2.6
Polynomial Regression
Degree:5, CV=44.2
Polynomial Regression
Degree:7, CV=482.0
x
y

加性模型

- 更一般地，基于基函数（特征）$\{\phi_1(\mathbf{x}),\ldots,\phi_m(\mathbf{x})\}$ 线性组合的预测，其中每个 $\phi_i(\mathbf{x}):\mathcal{R}^d \to \mathcal{R}$，且

$$f(\mathbf{x};\mathbf{w}) = w_0 + w_1\phi_1(\mathbf{x}) + \ldots + w_{m-1}\phi_{m-1}(\mathbf{x}) + w_m\phi_m(\mathbf{x})$$

- 例：
  - 如果 $\phi_i(x) = x^i, i = 1,\ldots,m$，则

$$f(x;\mathbf{w}) = w_0 + w_1 x + \ldots + w_{m-1}x^{m-1} + w_m x^m$$

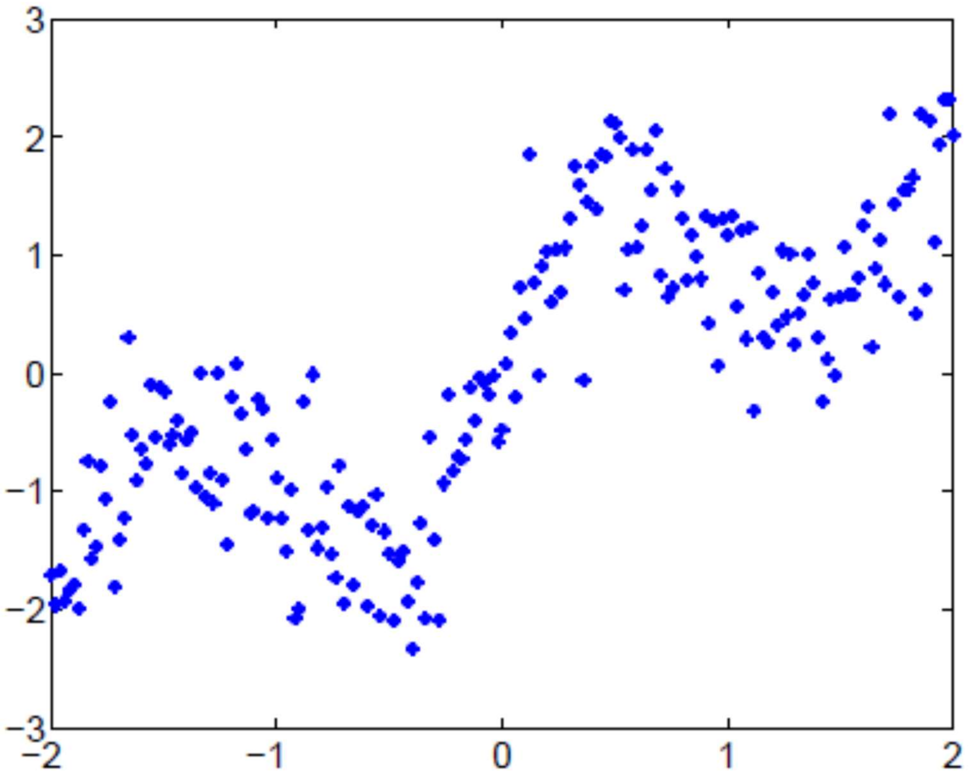
  - 如果 $m = d, \phi_i(x) = x_i, i = 1,\ldots,d$，则

$$f(x;\mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_{d-1}x_{d-1} + w_d x_d$$

加性模型

- 例：寻找原型输入矢量 $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m$ 通常是有用的。对于预测，这些矢量是不同“上下文”的典型样例。
  - 对每个原型定义一个基函数，度量输入向量 x 与原型的接近程度

$$\phi_k(\mathbf{x}) = \exp\left\{-\frac{1}{2}\left\|\mathbf{x}-\boldsymbol{\mu}_k\right\|^2\right\}$$

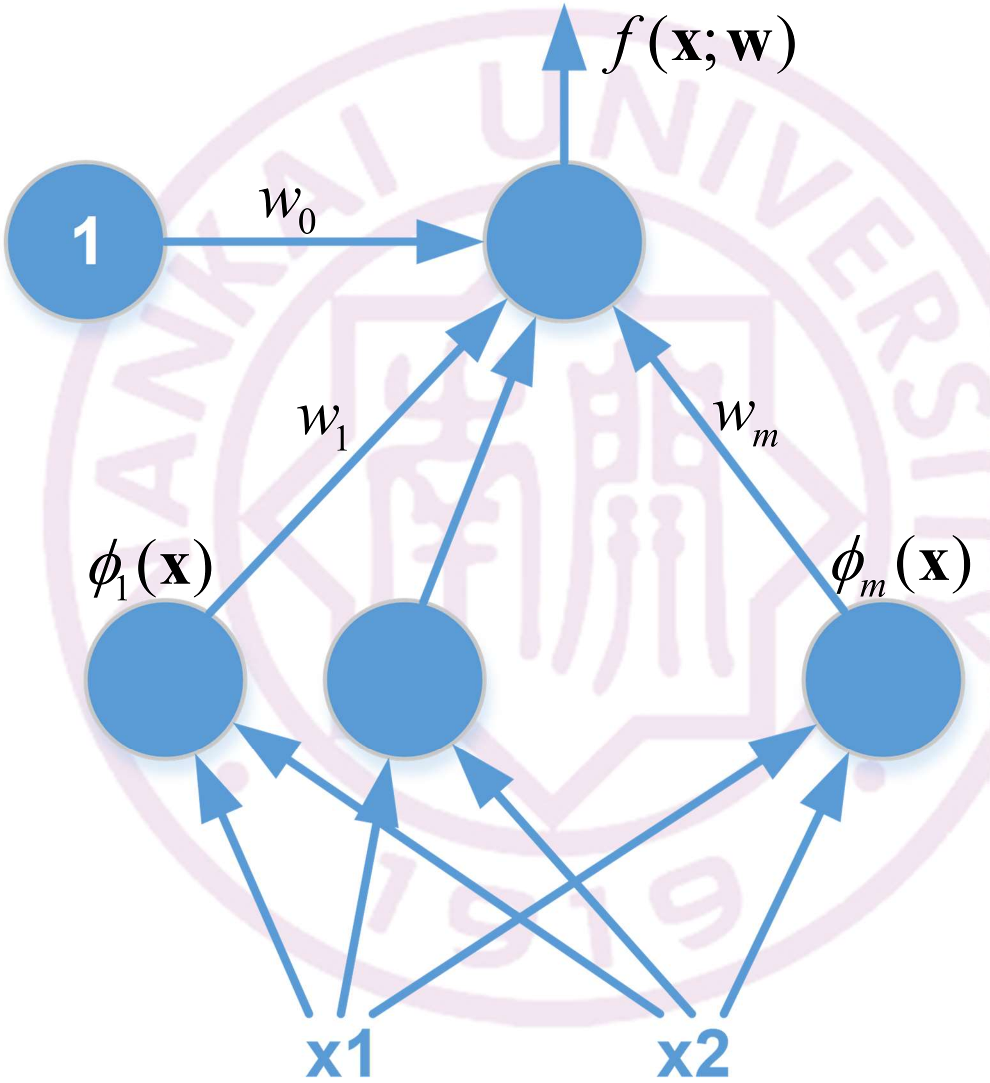


加性模型

- 基函数可以捕捉输入的各种（定性的）性质
- 例：根据文档描述给公司打分
  - x：文档（单词串）
  - $$\phi_i(\mathbf{x}) = \begin{cases} 1 & \text{if word } i \text{ appears in the document} \\ 0 & \text{otherwise} \end{cases}$$
  - $$f(\mathbf{x};\mathbf{w}) = w_0 + \sum_{i \in \text{words}} w_i \phi_i(\mathbf{x})$$

加性模型

- 图表示（神经网络）



选择重要特征

- 向前选择法
  - 从模型中最显著预测开始，然后每步添加特征
- 向后剔除法
  - 从模型所有特征预测开始，然后每步删除最小显著性特征
- 逐步筛选法
  - 基于指定标准添加/删除协变量来拟合模型

选择回归模型

- 可选择越多，选择正确的就越难
- 选择模型的关键因素
  - 数据探索是构建预测模型的必然组成部分
  - 通过将模型与所有可能的子模型进行对比，检查在你的模型中可能存在的偏差
  - 交叉验证是评估预测模型最好的方法
  - 如果数据集是多个混合变量，那么你就不应该选择自动模型选择方法
  - 它也将取决于你的目的
  - 回归正则化方法在高维和数据集变量之间多重共线情况下运行良好

主动学习

- 监督学习(supervised learning)
  - （输入，输出）样本来自于对未知联合分布 $P(x,y)$的采样
- 主动监督学习(active supervised learning)
  - 选择的（输入，输出）样本来自于对未知条件分布 $P(y\,|\,x)$ 的采样

- 原因：为何需要主动学习
  - 由于获取训练样本的成本很高，通常希望只需要较少的训练样本
- 问题：主动学习的危险
  - 因为是选择的样本，所以可能会是不重要的、罕见的或无效的样本

主动学习

- 需要决定
  - 函数类型：取决于我们的想法
  - 选择标准：哪个样本值得选择
  - 应用选择标准的方法：顺序的 或 批量的

- 函数类型：线性回归 或 多项式回归

$$y = w_0 + w_1 x + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

主动线性回归

- 选择输入以揭示“真实的”线性关系

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} w_0^* \\ w_1^* \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\varepsilon}$$

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

- 需要首先了解参数估计值 $\hat{\mathbf{w}}$ 与 $\mathbf{w}^*$ 之间的关系

主动线性回归

- 假设：与 X 中输入相对应的输出满足

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

- 参数估计：$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$
  - 根据相同输入 X 和采样的输出 y，参数估计满足

$$\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}^*, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$$

主动线性回归

- 两种类型的选择标准
  - 选择输入，以最小化参数中的不确定性
  - 选择输入，以最小化预测中的不确定性
- 选择标准的两种使用方法
  - 批量(batch)：观测到任何响应之前，选择所有输入
  - 顺序(sequential)：充分了解所有响应之后，选择下一个输入
- 批量选择，最小化参数不确定性
  - 期望找到 n 个输入 $x_1,\ldots,x_n$，以最小化结果参数 $\hat{\mathbf{w}}$ 的不确定性

$$\hat{\mathbf{w}} \sim \mathcal{N}(\mathbf{w}^*, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$$

  - 如：找到最小化协方差矩阵行列式的输入

$$\det[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]$$

行列式原则

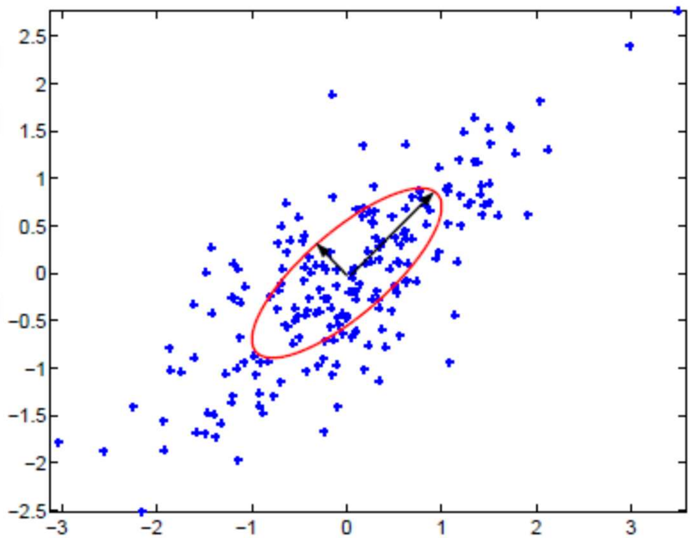
- 协方差矩阵的特征值分解

$$\mathbf{C}=\mathbf{R}\begin{bmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2\end{bmatrix}\mathbf{R}^{\mathrm{T}}$$

其中：正交旋转矩阵 R 指定方差的主轴，每个特征值σ_i^2给出沿着某个主要方向上的方差。

- 正态分布的“体量(volume)”是$\sigma_i^2, i=1,\ldots,m$ 的函数

$$\text{"volume"}\propto\prod_{i=1}^{m}\sigma_i=\sqrt{\det\mathbf{C}}$$



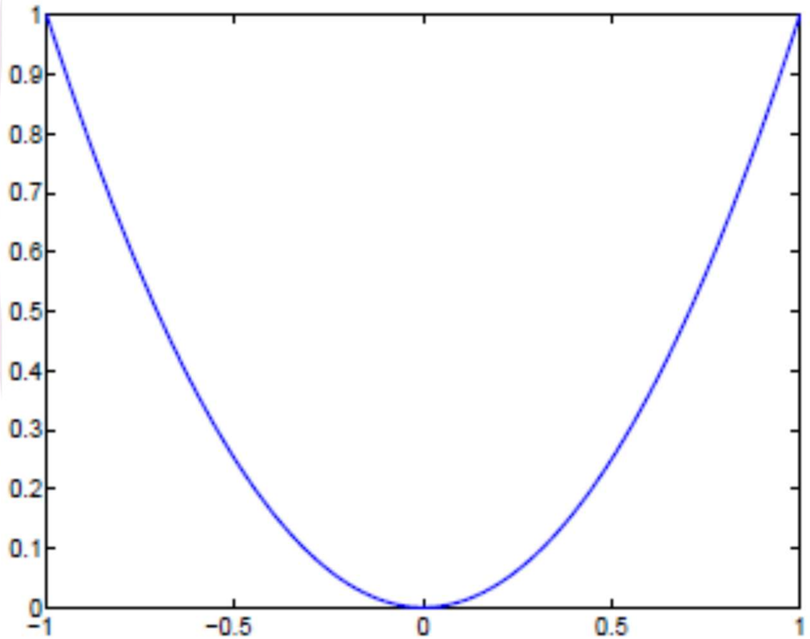
行列式原则

❖ 例:

- 一维问题，二阶多项式回归，$x \in [-1,1]$

$$f(x;\mathbf{w}) = w_0 + w_1 x + w_2 x^2$$

- 当 $n = 4$ 时，选择哪些点?



  - $x_1 = -1, x_2 = 0, x_3 = 0, x_4 = 1$

主动线性回归

- 顺序选择，最小化预测不确定性
  - 在现有信息的基础上选择下一个输入
  - 对新点 x 的预测

$$\hat{y}(x) = \hat{w}_0 + \hat{w}_1 x = \begin{bmatrix} 1 \\ x \end{bmatrix}^{\mathrm{T}} \hat{\mathbf{w}}$$

  - 预测中的方差

$$\mathrm{Var}\{\hat{y}(x)\} = \begin{bmatrix} 1 \\ x \end{bmatrix}^{\mathrm{T}} \mathrm{cov}(\hat{\mathbf{w}}) \begin{bmatrix} 1 \\ x \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} 1 \\ x \end{bmatrix}^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

    - 噪声方差 σ^2 只影响整体比例
    - 方差是已选输入的函数，与输出无关

主动线性回归

- 假设输入点在区间 X 内，选择新点来减少不确定预测的方差：

$$x^{\text{new}} = \arg\min_{x\in\mathbf{X}}\left\{\mathrm{Var}\left\{\hat{y}(x)\right\}\right\}$$

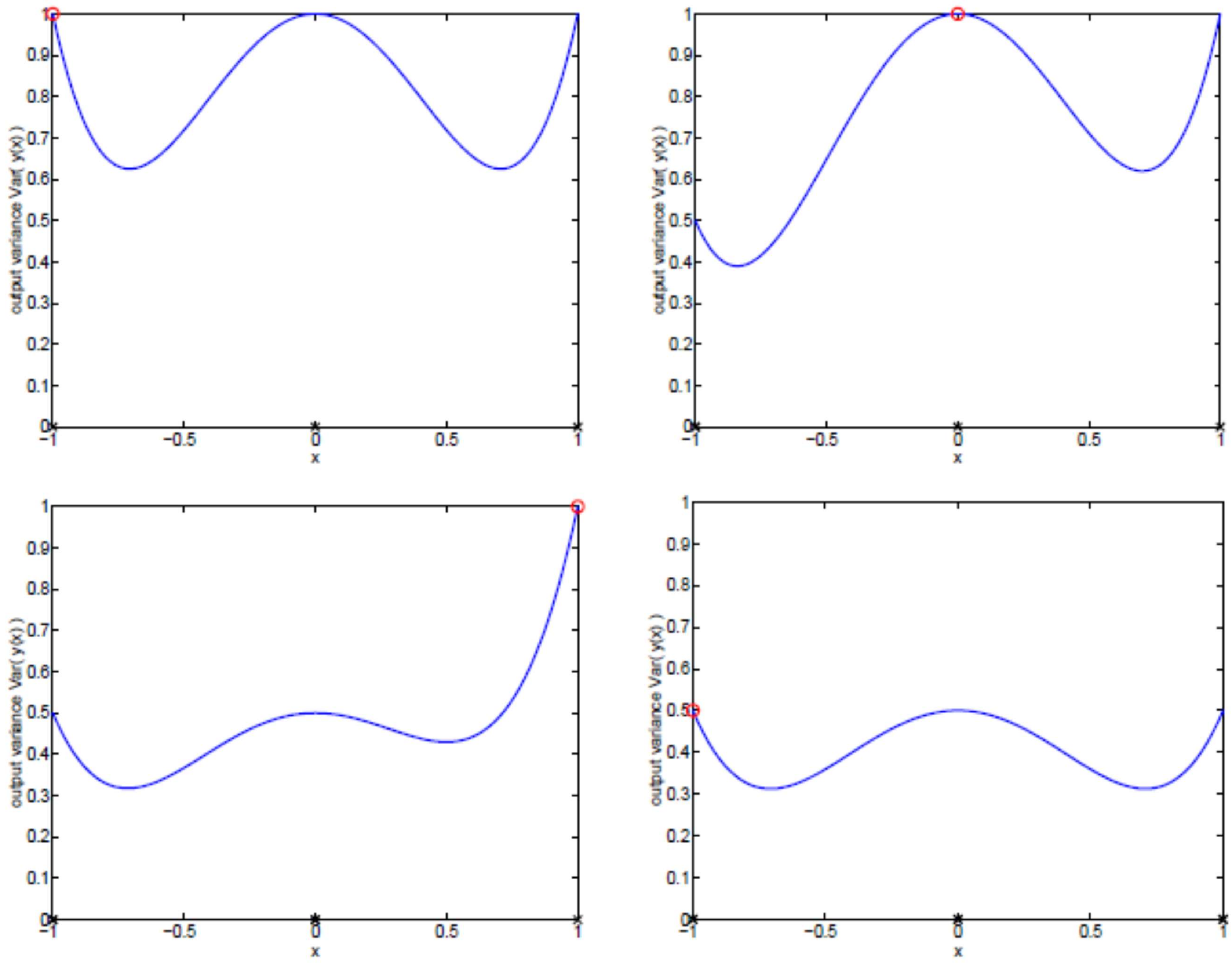
- 例：
  - 一维问题，二阶多项式回归，$x\in[-1,1]$

$$\hat{y}(x) = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2$$

  - 先前已选择输入 $x_1 = -1, x_2 = 0, x_3 = 1$

$$\mathrm{Var}\left\{\hat{y}(x)\right\} = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}, \text{ where } \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \cdots & \cdots & \cdots \end{bmatrix}$$

主动线性回归



顺序选择的性质

- 在线性/加性回归背景下，方差不能随着新点的增加而增加
  - $\hat{\mathbf{w}}$ 的协方差矩阵：$\mathbf{C}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$
  - 方差的逆：$\mathbf{A}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})$
  - 预测的方差

$$\mathrm{Var}\left\{\hat{y}\left(x\right)\right\}=\begin{bmatrix}1\\x\\x^2\end{bmatrix}^{\mathrm{T}}\mathbf{C}\begin{bmatrix}1\\x\\x^2\end{bmatrix}=\begin{bmatrix}1\\x\\x^2\end{bmatrix}^{\mathrm{T}}\mathbf{A}^{-1}\begin{bmatrix}1\\x\\x^2\end{bmatrix}$$

  - 结论：如果逆协方差矩阵 A 的特征值随着添加新点而增加（或保持不变），则任何点 x 的方差都不会增加。

更普遍的主动学习

- 为了进行主动学习，我们必须评估“新信息的价值”，即希望从查询另一个响应中获得多少信息。
- 这种计算能够在任何学习任务的上下文中完成。

简单推导

- 新点 x':

$$\mathbf{A}' = \begin{bmatrix} 1 & x' & x'^2 \\ \mathbf{X} & & \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 1 & x' & x'^2 \\ \mathbf{X} & & \end{bmatrix}$$

$$= \mathbf{X}^{\mathrm{T}}\mathbf{X} + \begin{bmatrix} 1 \\ x' \\ x'^2 \end{bmatrix} \begin{bmatrix} 1 \\ x' \\ x'^2 \end{bmatrix}^{\mathrm{T}}$$

$$= \mathbf{A} + \begin{bmatrix} 1 \\ x' \\ x'^2 \end{bmatrix} \begin{bmatrix} 1 \\ x' \\ x'^2 \end{bmatrix}^{\mathrm{T}}$$

- 将特征值均为非负的矩阵加到矩阵 A 上 $\Rightarrow$ 矩阵 A 的特征值是非负的

诚信 创新 实践