第2.1节 数据回归

Data regression

中国科学院大学 叶齐祥

qxye@ucas.ac.cn

主要内容:

线性回归

- 线性回归
- 局部加权的线性回归

非线性回归

- 带有非线性基的回归
- 欠拟合与过拟合
- **Logistic** 回归

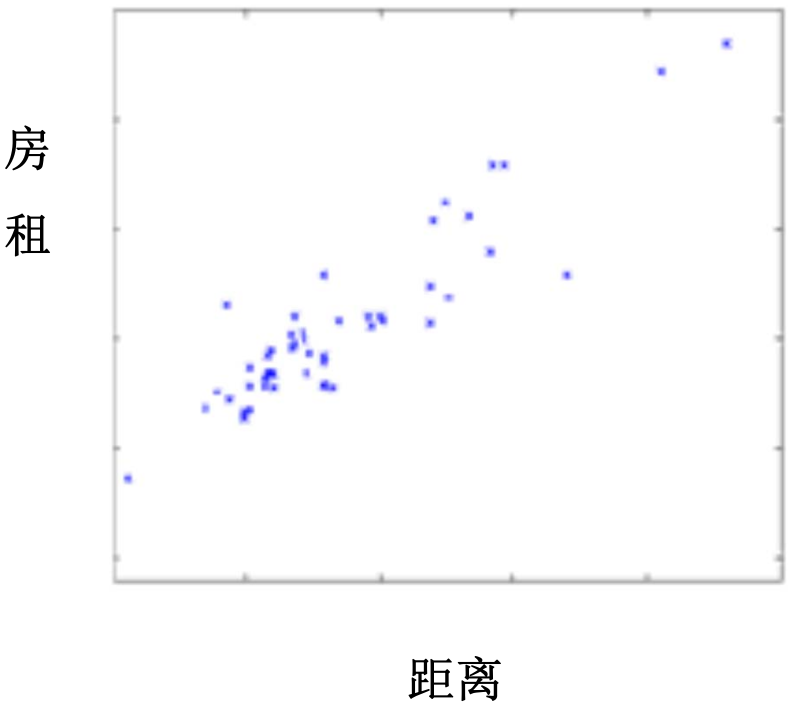
数据回归介绍

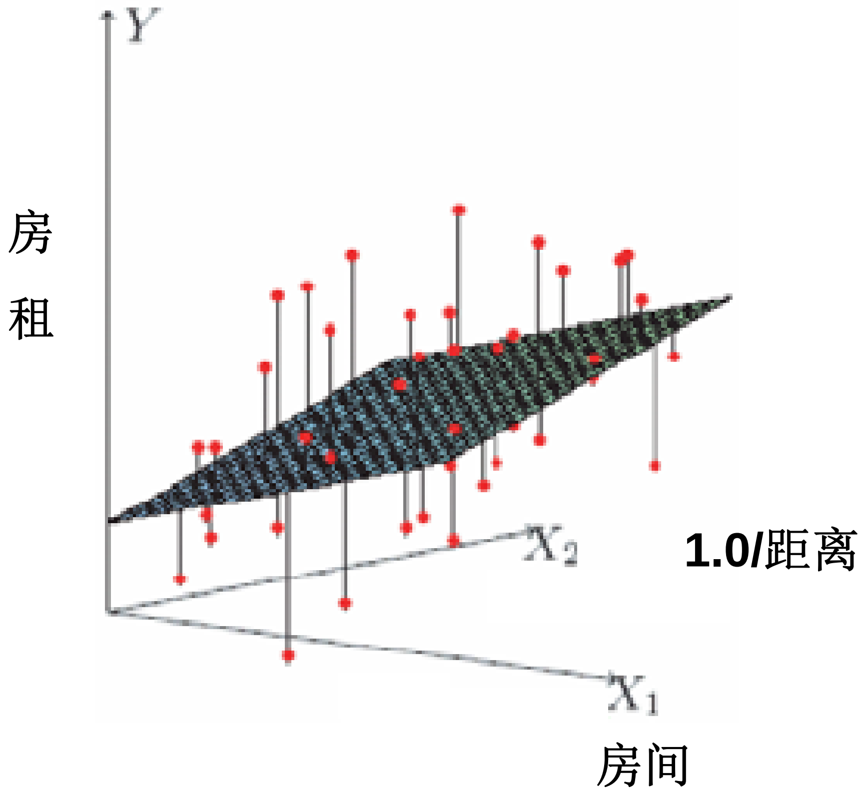
- 例子
  - 假如你刚刚搬到学校，需要知道在你学校周围的房价，设计一个数据回归程序。

距离学校的距离	卧室数目	房租
2.30km	1	1600
5.06km	2	2000
4.33km	2	2100
1.09km	1	1500
…		
1.50km	1	?
2.70km	1.5	?

数据回归介绍

- 例子
  - 假如你刚刚搬到学校，需要知道在你学校周围的房价，设计一个数据回归程序。

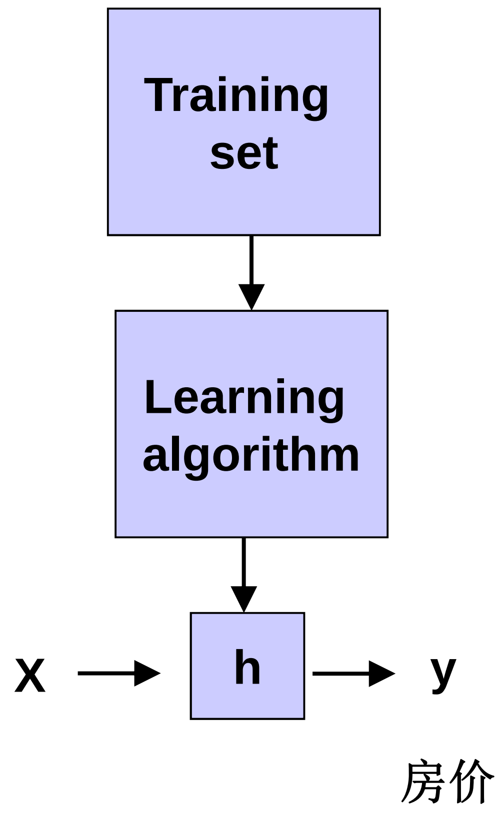




数据回归介绍

- 问题描述
  - 特征：
    - 居住面积、房间数、距离…$X=\{x_1, x_2, ..., x_K\}$
  - 训练集合
  - 回归目标

$$\begin{bmatrix} x_1^{(1)}, x_2^{(1)}, ..., x_K^{(1)} \\ x_1^{(2)}, x_2^{(2)}, ..., x_K^{(2)} \\ \cdots \\ x_1^{(n)}, x_2^{(n)}, ..., x_K^{(n)} \end{bmatrix} \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix}$$



线性回归

- 假设目标(Y)是特征的线性方程

$$\tilde{y} = \theta_0 + \theta_1 x_1 + \ldots, \theta_j x_j + \ldots, \theta_K x_K$$

- 如何求取参数 θ
  - 直观的方法是最小化

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{N}\left(\tilde{y}^{(i)}(x^{(i)}) - y^{(i)}\right)^2$$

线性回归—最小均方差（LMS）求解

- 目标方程 $J(\theta)=\frac{1}{2}\sum_{i=1}^{N}\left(X^{(i)}\theta-y^{(i)}\right)^{2}$

- 梯度方法求解

$$\theta_j^{t+1}=\theta_j^{t}-\alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

- 获得梯度下降迭代

$$\theta_j^{t+1}=\theta_j^{t}-\alpha\sum_{i=1}^{N}\left(X^{(i)}\theta^{t}-y^{(i)}\right)\cdot x_j^{(i)}$$

线性回归—最小均方差（LMS）求解

- 目标方程

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{N}\left(X^{(i)}\theta-y^{(i)}\right)^{2}$$

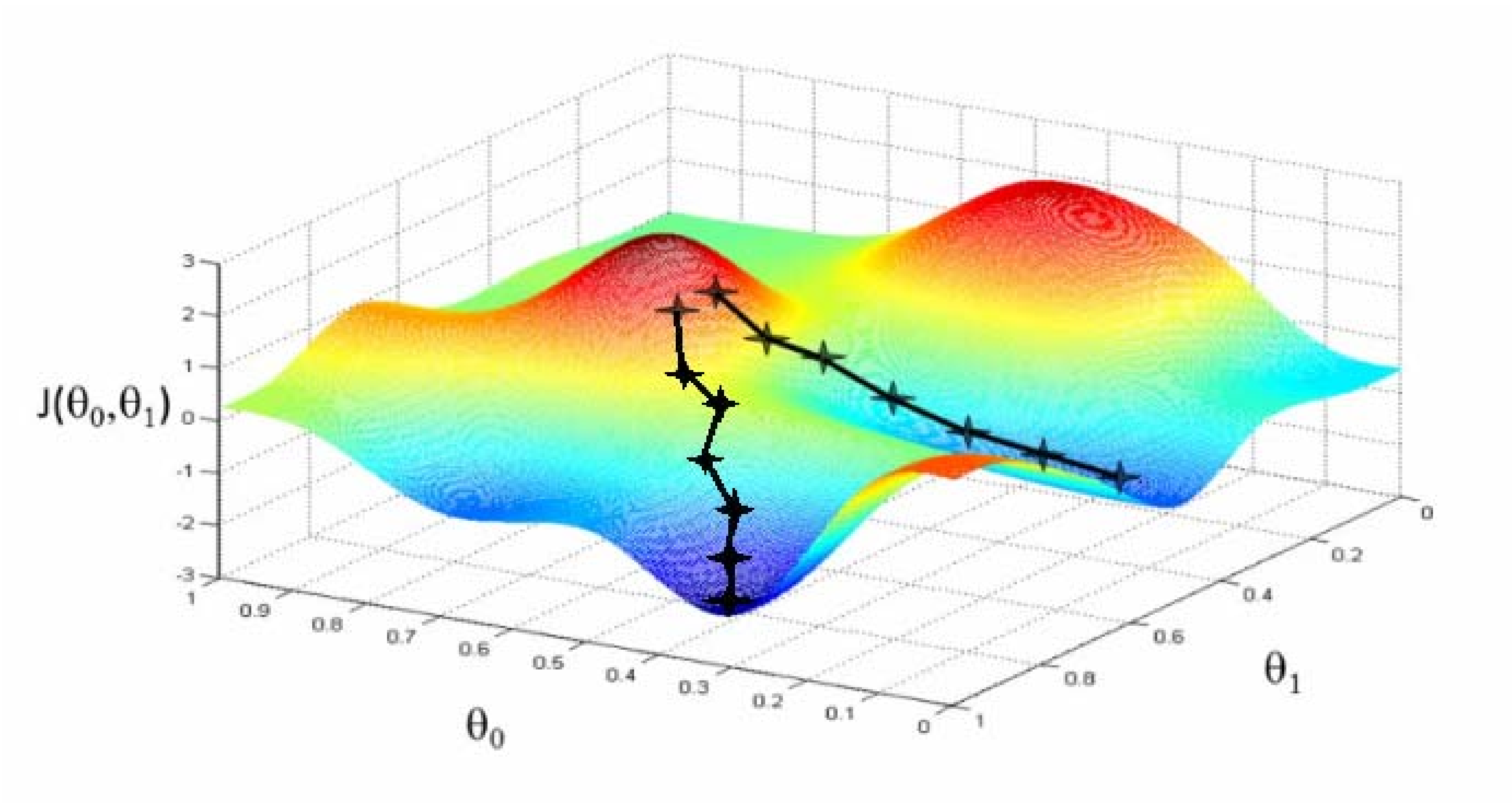
- Steepest下降方法求解

$$\nabla_{\theta}J=\left[\frac{\partial}{\partial\theta_1}J,\ldots,\frac{\partial}{\partial\theta_k}J\right]^{T}=\sum_{i=1}^{N}(X^{(i)}\theta-y^{(i)})X^{(i)}$$

- 获得Steepest decent迭代

$$\theta^{t+1}=\theta^{t}-\alpha\sum_{i=1}^{N}\left(X^{(i)}\theta^{t}-y^{(i)}\right)X^{(i)}$$

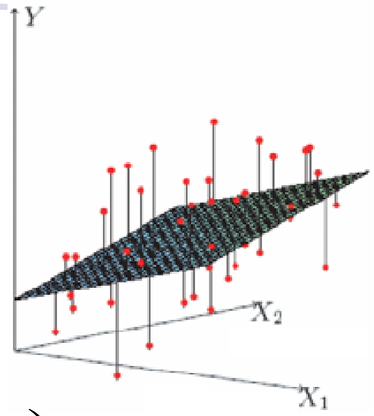
线性回归—梯度下降(Gradient descend)求解



线性回归—概率角度理解LMS

- 假如目标值与输入之间的关系为

$$y^{(i)} = \theta^T X^{(i)} + \varepsilon^{(i)}$$



- 其中$\varepsilon^{(i)}$表示符合正态分布的随机噪声

$$P(y^{(i)} \mid X^{(i)}, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T X^{(i)})^2}{2\sigma^2} \right)$$

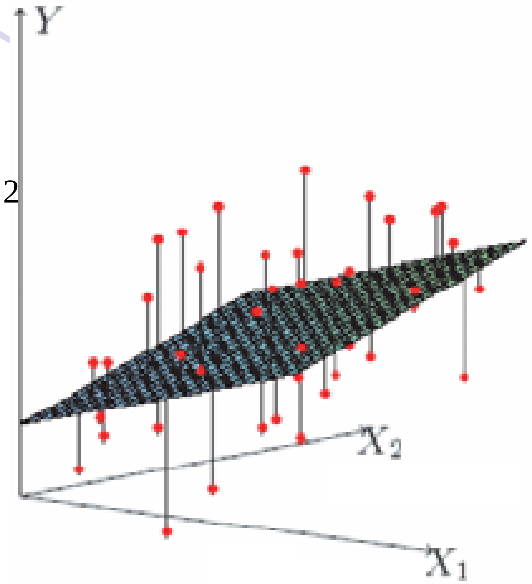
- 做样本独立性假设，得到

$$L(\theta) = \prod_{i=1}^{N} P(y^{(i)} \mid X^{(i)}, \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^N \exp\left(-\frac{\sum_{i=1}^{N} (y^{(i)} - \theta^T X^{(i)})^2}{2\sigma^2} \right)$$

线性回归—概率角度理解LMS

- 计算Log似然

$$L(\theta)=N\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{N}(y^{(i)}-\theta^T X^{(i)})^2$$



- 上式第二项即是LMS

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{N}\left(X^{(i)}\theta-y^{(i)}\right)^2$$

- 结论：**求最小均方差LMS和极大似然估计MLE是等效的**

主要内容:

线性回归

- 线性回归
- 欠拟合与过拟合
- 局部加权的线性回归

非线性回归

- 带有非线性基的回归
- 欠拟合与过拟合
- **Logistic** 回归

线性回归—局部加权

- 加权
  - 原来的目标函数：

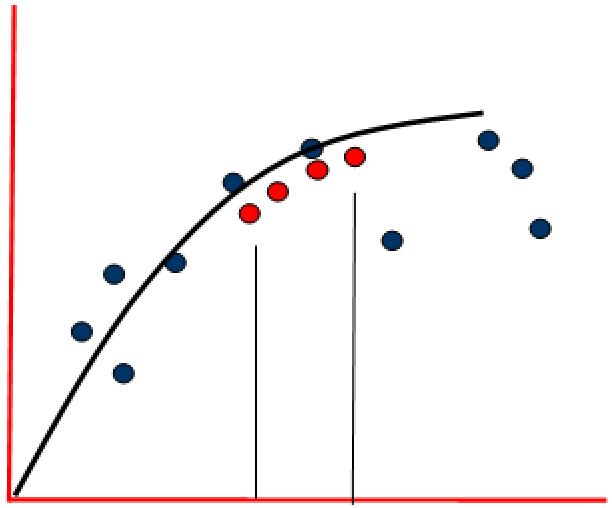
$$J(\theta)=\frac{1}{2}\sum_{i=1}^{N}\left(X^{(i)}\theta-y^{(i)}\right)^{2}$$

  - 加权函数：

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{N}w^{(i)}\cdot\left(X^{(i)}\theta-y^{(i)}\right)^{2}$$

  - 权值：

$$w^{(i)}=\exp\left(-\frac{(X^{(i)}-\bar{X})^{2}}{2\tau^{2}}\right)$$



- 其中**X**靠近测量值或者测量值均值变量

主要内容:

线性回归

- 线性回归
- 局部加权的线性回归

非线性回归

- 带有非线性基的回归
- 欠拟合与过拟合
- **Logistic** 回归

非线性回归—非线性基

- 设计出非线性特性

$$y=\theta_0+\sum_{j=1}^{m}\theta_j\cdot\phi_j(x)=\theta^T\cdot\phi(x)$$

  - 其中 $\phi_j(x)$ 是基，例如

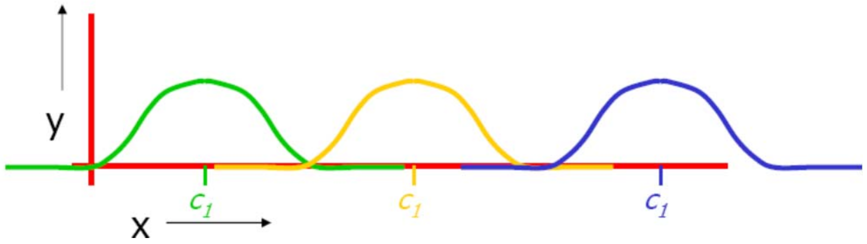
多项式： $\phi_j(x)=x^{j-1}$

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

Sigmod: $\phi_j(x)=\frac{1}{1+e^{-x}}$

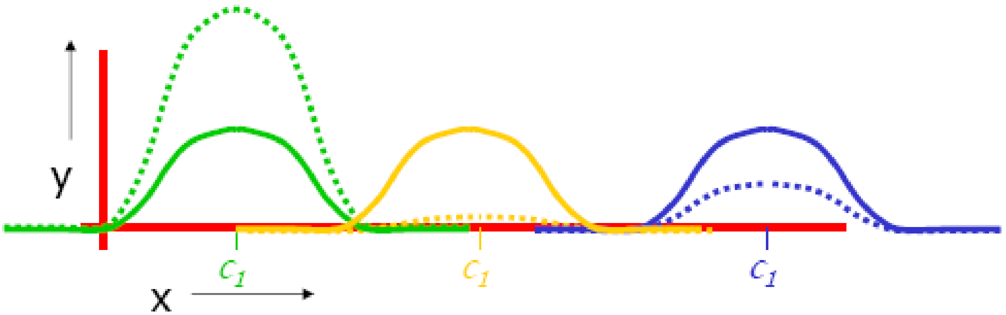
回归—1D and 2D RBFs

- 1D RBF



$$y^{est} = \beta_1 \phi_1(x) + \beta_2 \phi_2(x) + \beta_3 \phi_3(x)$$

- 拟合后

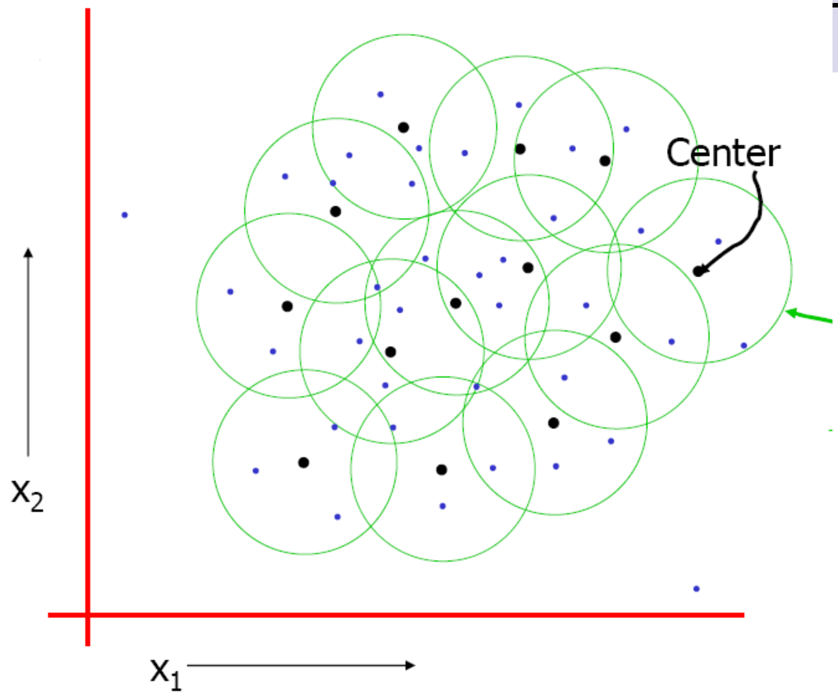


$$y^{est} = 2\phi_1(x) + 0.05\phi_2(x) + 0.5\phi_3(x)$$

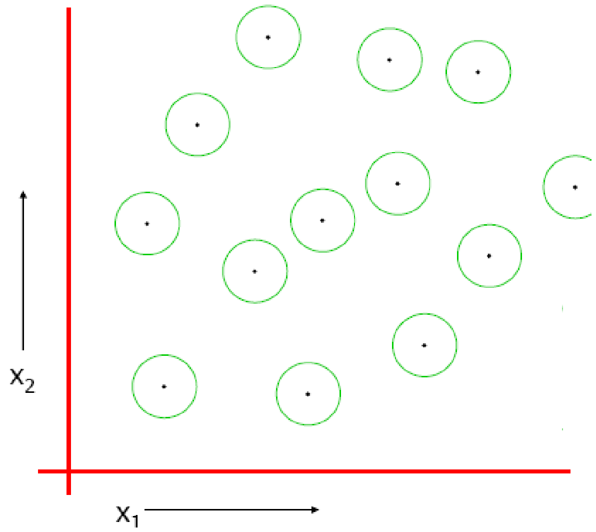
$$\phi_i(x) = KernelFunction(\,|\,x - c_i\,|\,/\,KW)$$

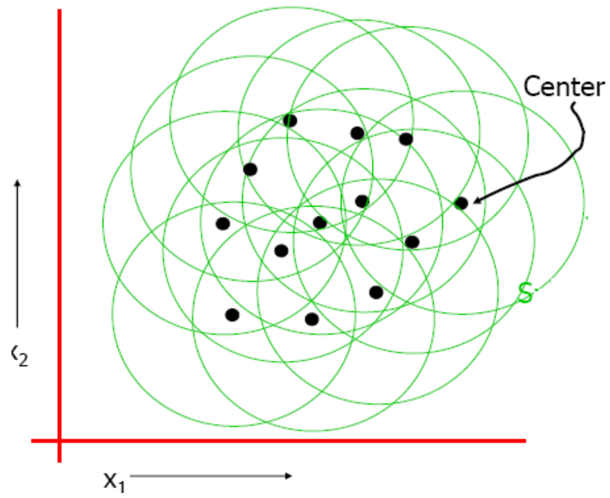
回归—1D and 2D RBFs

- Good 2D RBF



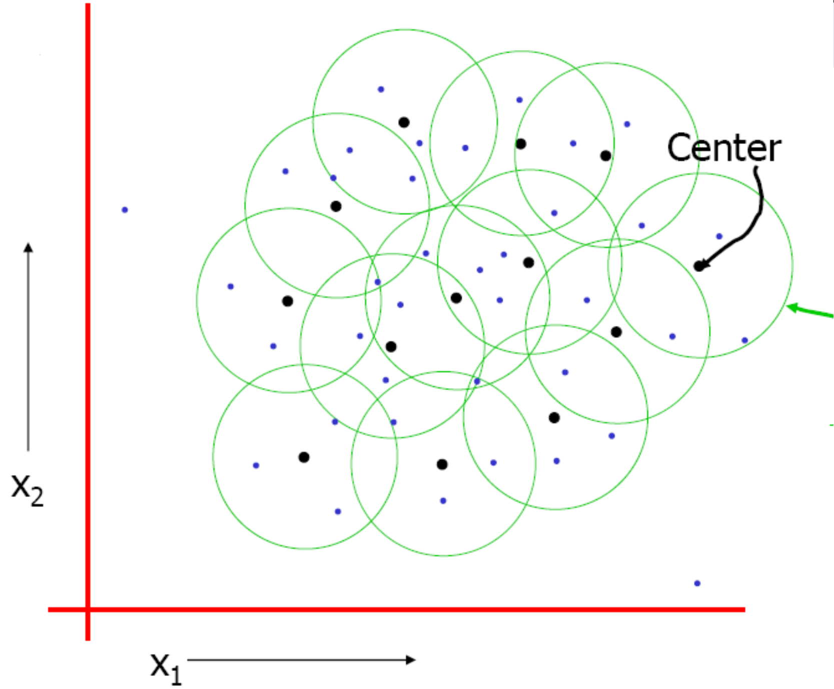
- Bad 2D RBF



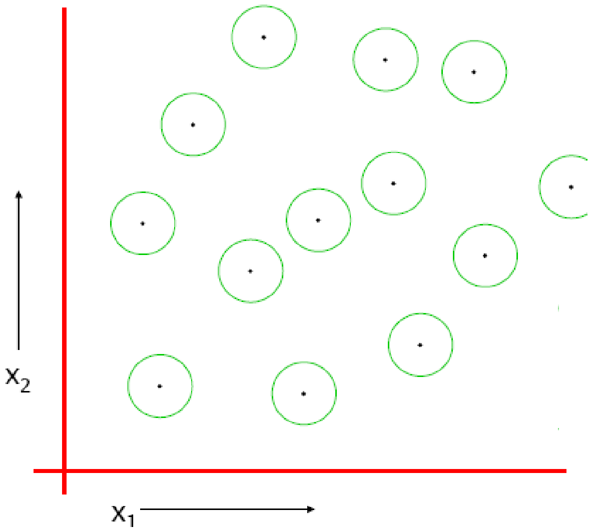


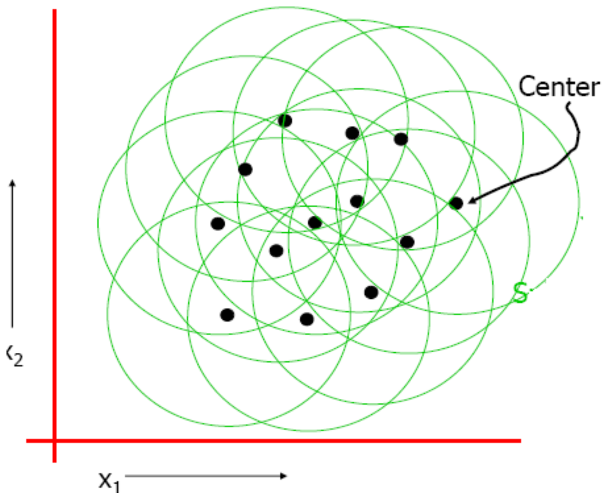
回归—1D and 2D RBFs

- Good 2D RBF



- Bad 2D RBF





主要内容:

线性回归

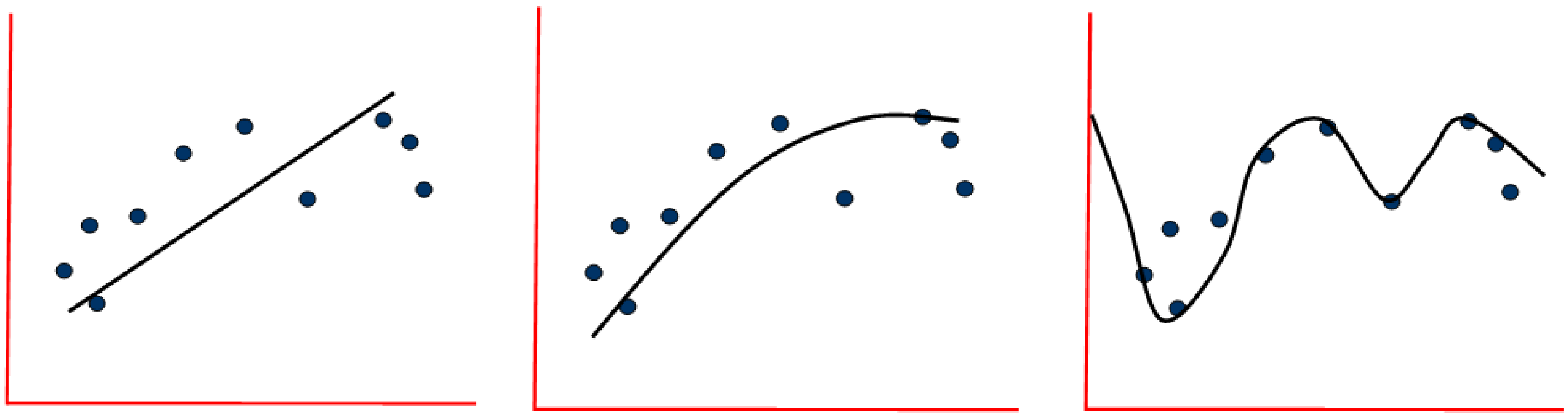
- 线性回归
- 局部加权的线性回归

非线性回归

- 带有非线性基的回归
- 欠拟合与过拟合
- **Logistic** 回归

数据回归—欠拟合与过拟合

- 欠拟合与过拟合



$$y=\theta_0+\theta_1 x \qquad y=\theta_0+\theta_1 x+\theta_2 x^2 \qquad y=\sum_{j=0}^{5}\theta_j x^j$$

主要内容:

线性回归

- 线性回归
- 局部加权的线性回归

非线性回归

- 带有非线性基的回归
- 欠拟合与过拟合
- Logistic 回归

Logistic 回归

回归分析可用来分析一个/多个自变量与一个因变量的关系，模型中因变量Y是边连续性随机变量，并要求呈正态分布。

但在医学研究中，常碰到因变量的取值仅有两个，如药物实验中，动物出现死亡或生存，P和X的关系显然不能用一般线性回归模型$P = B_0 + B_1X$来表示。这时可用Logistic回归分析。

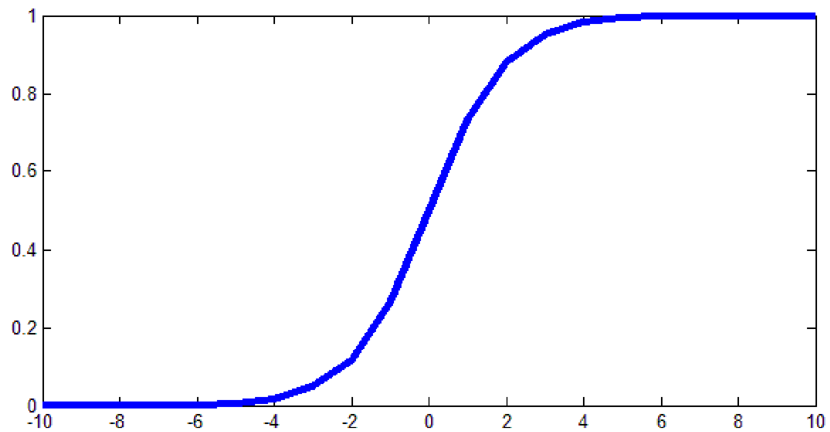
Logistic 回归

- 先引入Logistic分布函数，表达式为：

$$F(X)=\frac{e^{x}}{1.0+e^{x}}$$

X的取值在正负无穷大之间；F(x)则在0－1之间取值，并呈单调上升S型曲线。人们正是利用Logistic分布函数这一特征，将其应用到例如：

临床医学和流行病学中来描述事件发生的概率。



Logistic 回归

例子：以因变量Y＝1表示死亡， Y ＝0表示生存，以P(Y ＝ 1|X)表示使用药物剂量X的动物死亡的概率，设

$$P(Y=1|X)=\frac{e^{B_0+B_1X}}{1.0+e^{B_0+B_1X}}$$

记Logit(P)=ln[p/(1-p)],则上式可表示为：

$$Logit(P)=B_0+B_1X$$

这里X的取值仍是任意的， Logit(P)的值亦在正负无穷大之间，概率P的数值则必然在0－1之间。 p/(1-p)为事件的优势， Logit(P)为对数优势，故logistic回归又称对数优势线性回归

Logistic 回归

- 一般地，设某事件Y发生（Y＝1）的概率P依赖于多个自变量（$x_1, x_2, \cdots, x_p$），且

$$P(Y=1|X)=\frac{e^{B_0+B_1X_1+\ldots+B_pX_p}}{1.0+e^{B_0+B_1X_1+\ldots+B_pX_p}}$$

或者

$$Logit\left(P\right)=B_0+B_1X_1+\ldots,B_pX_p$$

则称该事件发生的概率与变量间关系符合多元Logistic回归或对数优势线性回归。

Logistic 回归—应用

- 优势比(odds ratio，OR)：
  - 某个自变量X_j 改变一个单位，造成的后验概率的比值的变化

$$P(Y=1)=\frac{e^X}{1.0+e^X} \qquad P(Y=0)=\frac{1.0}{1.0+e^X}$$

$$OR=\frac{P(Y=1)}{P(Y=0)}=e^X$$

$$OR_j=\frac{P(Y=1\mid x_j=x_j+1)}{P(Y=0\mid x_j=x_j+1)}=e^{B_0+,\dots,B_j\left(X_j+1\right)+\dots+B_pX_p}$$

例如某次统计分析：

令X_2~X_8保持不变，年龄X_1改变1个单位（10岁），
如年龄从50岁提高到60岁（X_1分别为2，3），患冠心病的
概率增加了$exp(0.6443\times(3-2))=1.9047\approx 2$倍

Logistic 回归—MLE参数估计

记得到一个样本观测值 $y^{(i)}(i=1,2,\cdots,N)$ 的概率为

$$\mathrm{P}(y^{(i)})=p_i^{y^i}(1-p_i)^{1-y^i}$$

则 $y_1,y_2,\cdots,y^{(N)}$ 的似然函数为 $L=\prod_{i=1}^{N}\mathrm{P}(y^i)=\prod_{i=1}^{N}p_i^{y^i}(1-p_i)^{1-y^i}$

两边取对数： $\ln L=\sum_{i=1}^{N}[y^i\ln p^i+(1-y^i)\ln(1-p^i)]=$

最后得到：

$$\sum_{i=1}^{N}[y^i\ln\frac{p^i}{1-p^i}+\ln(1-p^i)]$$

$$\ln L=\sum_{i=1}^{N}[y^i(\alpha+\beta_1x_1^i+\cdots+\beta_px_p^i)-\ln(1+\exp(\alpha+\beta_1x_1^i+\cdots+\beta_px_p^i))]$$

当使得 $\ln L$ 取得最大值时，参数估计值即为所求。

Thank you