# 模式识别引论

An Introduction to Pattern Recognition

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

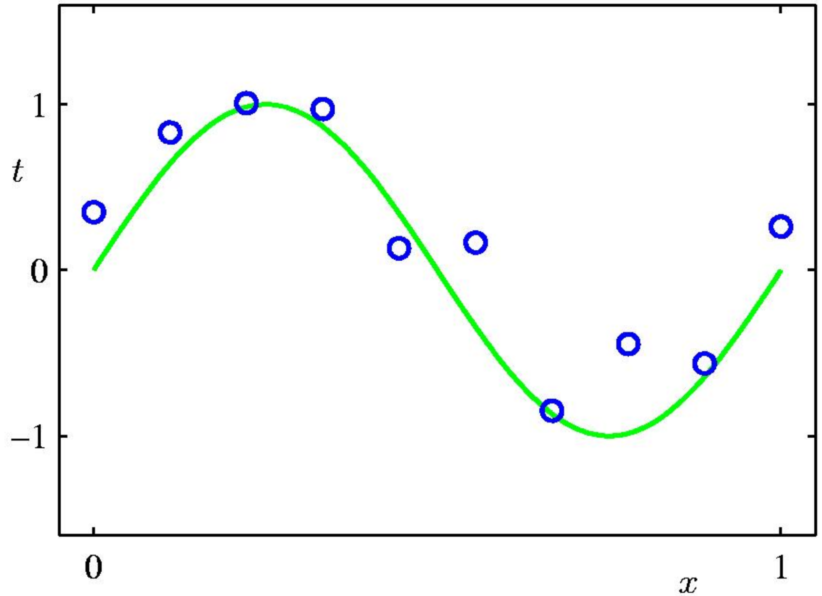
网络搜索教研中心 信息与通信工程学院 北京邮电大学

# 广义线性模型 内容提要

- 引子: 曲线拟合问题
- 最大似然估计
  - **最小二乘**
- 最大后验概率估计
  - **正则化最小二乘**
- 贝叶斯估计
- 方法比较:
  - 最大似然估计**(MLE) vs.** 最大后验概率估计**(MAP) vs.** 完全贝叶斯法

# 引例：多项式曲线拟合

- 给定N个训练数据：(x, t)
  - 数据其中绿色曲线为生成训练的真实曲线




- 多项式曲线拟合
  - 使用多项式模型去构造
  - 定义为

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

# 引例：多项式曲线拟合

- 学习任务:
  - 根据训练数据，估计模型中的参数

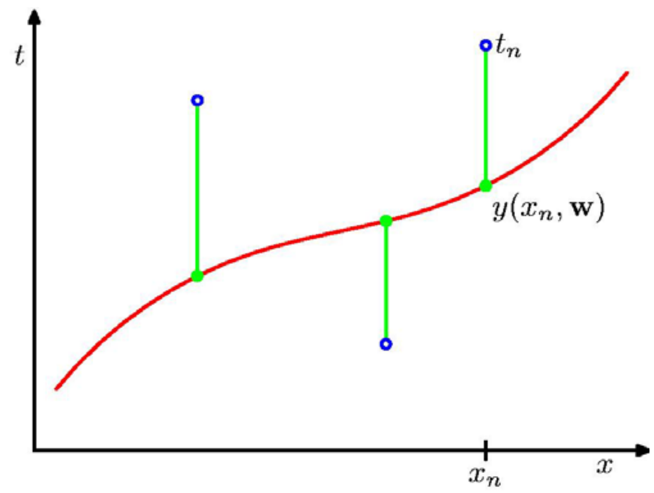
$$y(x,\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

- 根据目标函数的不同，分为:
  - 最小二乘法

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$$

  - 正则化最小二乘法

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$



# 广义线性模型内容提要

- 引子: 曲线拟合问题
- 最大似然估计
  - **最小二乘**
- 最大后验概率估计
  - **正则化最小二乘**
- 贝叶斯估计
- 方法比较:
  - 最大似然估计(**MLE**) **vs.** 最大后验概率估计(**MAP**) **vs.** 完全贝叶斯法

# 最小二乘法

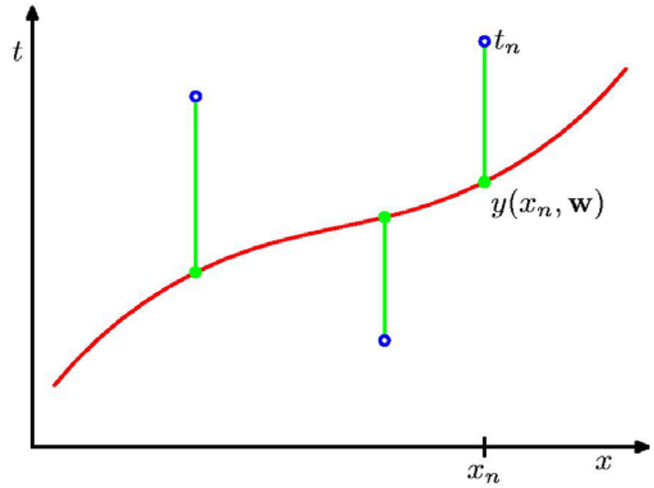
2乘? Why not 1、3、4、5乘。。。

- 学习任务:
  - 根据训练数据，估计模型中的参数

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

- 使用平方误差函数，则得到最小二乘法

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$$



# 最小二乘法的求解-1

- 写成矩阵向量形式

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

其中 $\boldsymbol{\varphi}(x)^T = (\phi_0(x), \phi_1(x), \cdots, \phi_M(x))$

$$= \boldsymbol{\varphi}(x)^T \mathbf{w}$$

- 目标函数变为

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 = \frac{1}{2}\sum_{n=1}^{N}\left\{\boldsymbol{\varphi}(x_n)^T \mathbf{w} - t_n\right\}^2$$

$$= \frac{1}{2}(\Phi\mathbf{w} - \mathbf{t})^T (\Phi\mathbf{w} - \mathbf{t})$$

$$= \frac{1}{2}\left(\mathbf{w}^T\Phi^T\Phi\mathbf{w} - 2\mathbf{w}^T\Phi^T\mathbf{t} + \mathbf{t}^T\mathbf{t}\right)$$

– 其中

$$\Phi = \begin{pmatrix} \boldsymbol{\varphi}(x_1)^T \\ \boldsymbol{\varphi}(x_2)^T \\ \vdots \\ \boldsymbol{\varphi}(x_N)^T \end{pmatrix}$$

# 最小二乘法的求解-2

- 优化问题变为：

$$\min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2}\left(\mathbf{w}^T\Phi^T\Phi\mathbf{w} - 2\mathbf{w}^T\Phi^T\mathbf{t} + \mathbf{t}^T\mathbf{t}\right)$$

- 对w求梯度

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \Phi^T\Phi\mathbf{w} - \Phi^T\mathbf{t}$$

  - 令梯度为0， $\nabla_{\mathbf{w}} E(\mathbf{w}) = \Phi^T\Phi\mathbf{w} - \Phi^T\mathbf{t} = 0$
  - 得出 $\Phi^T\Phi\mathbf{w} = \Phi^T\mathbf{t}$

如果 $\Phi$ 列满秩

$$\Rightarrow \mathbf{w} = \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{t} \Rightarrow y(x,\mathbf{w}) = \boldsymbol{\varphi}(x)^T\mathbf{w}$$

$$= \boldsymbol{\varphi}(x)^T\left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{t}$$

# 基于一般基函数的回归模型

- 使用一般的基函数

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

  - 其中 $\phi_j(x)$被称作基函数（basis functions ）
  - 特别的， $\boldsymbol{\phi}_{\mathbf{0}}(\mathbf{x}) = \mathbf{1}$, $\mathbf{w_0}$ 当作偏置**(bias)**
  - 最简单的形式——线性基函数，即

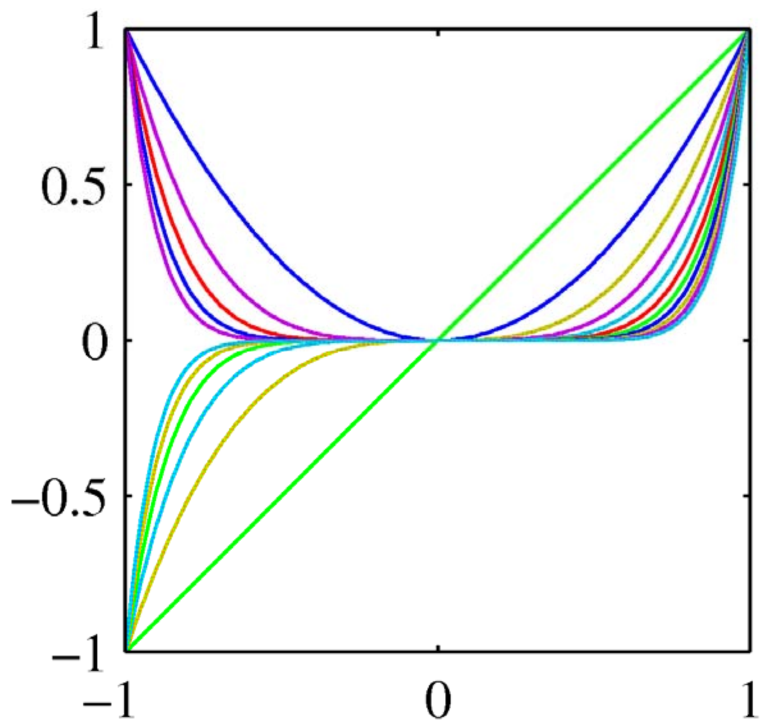
$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$$

# 常用的一般基函数

- 多项式基函数(Polynomial Basis Function)

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

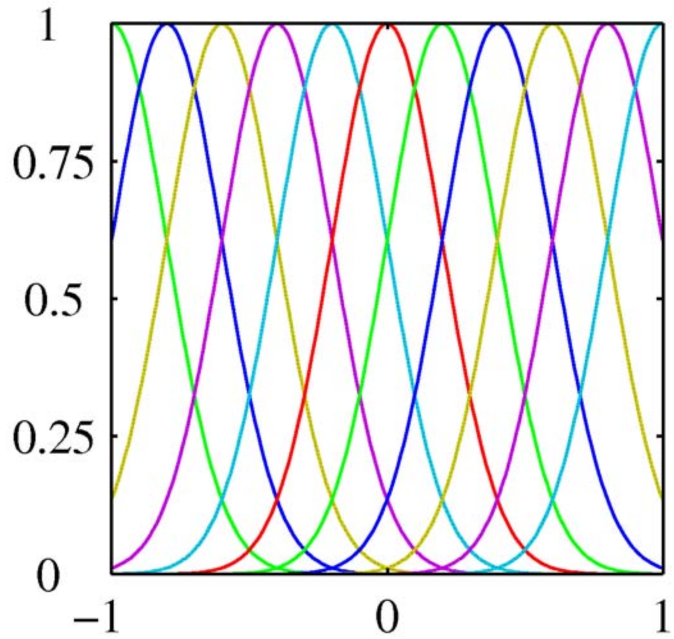
  – 非局部化基函数



# 常用的一般基函数

- 高斯基函数(Gaussian Basis Function)

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



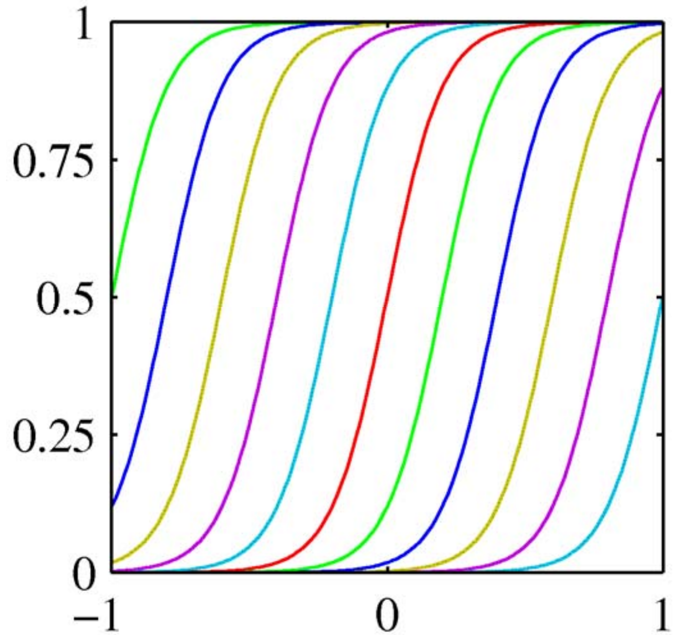

  - 局部基函数，x的变化仅影响邻近的基函数
  - 径向基函数的一个特例

# 常用的一般基函数

- Sigmoid基函数

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

$$\sigma(a) = \frac{1}{1+\exp(-a)}.$$




  – 局部基函数，**x**的变化仅影响邻近的基函数
  – 神经网络中最常用的激活函数

# 最小二乘法里的几个为什么

- 最小二乘法

$$\min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2$$

2乘? Why not 1、3、4、5乘。。。

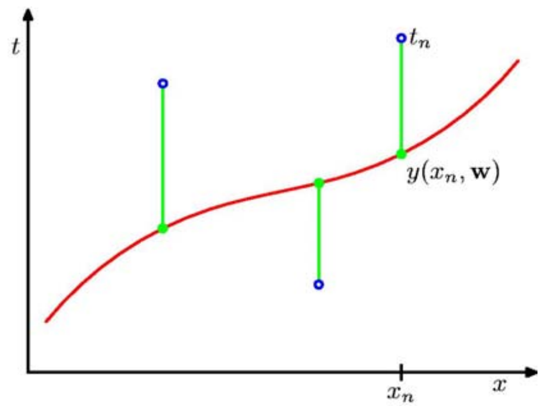
  - 请问几个为什么...

# 最小二乘法里的几个为什么

$$\min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$$

2乘? Why not 1、3、4、5乘。。。

- 为什么不直接令 y(x,w)=t 去解w？
- 为什么使用平方误差函数？
- 为什么把所有的误差加起来, 而不是连乘起来?
- 为什么使用作差—— y(x,w) - t 来定义误差，而不是两者作除法?
  - 为什么求和到**N?** 为什么有**1/2?**
  - 。。。
  - 为什么我没有上面这些为什么。。。
  - 为什么会有上面这些为什么。。。



# 最小二乘法从哪里来？

$$\min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$$

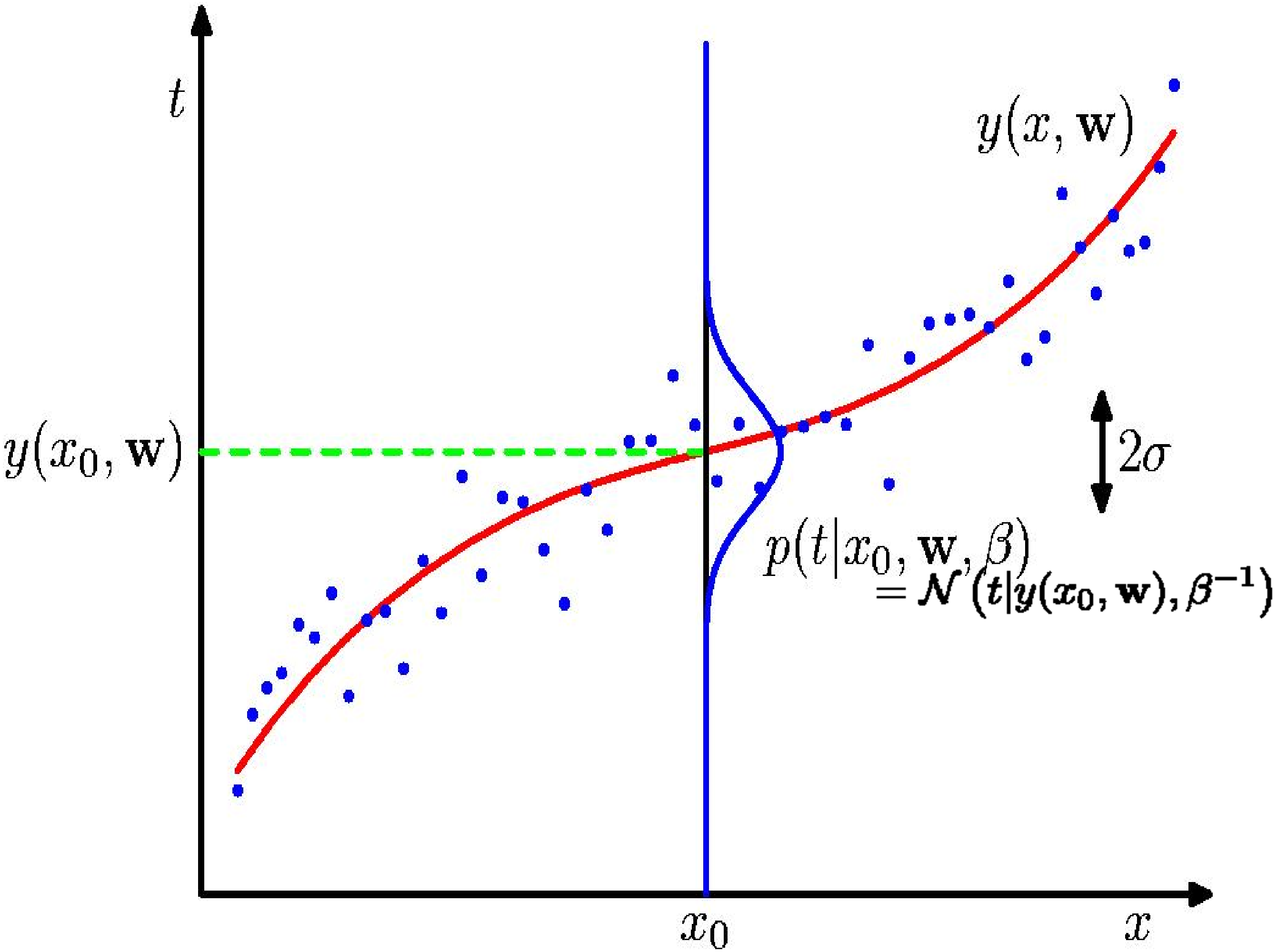
2乘？Why not 1、3、4、5乘。。。

- 最小二乘法来源于参数的最大似然估计法
  - 假设观测数据来采样于一个确定性(deterministic)函数，同时观测数据中存在i.i.d.加性高斯噪声(additive Gaussian noise)

$$\boxed{t = y(\mathbf{x},\mathbf{w}) + \epsilon \quad \text{where} \quad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0,\beta^{-1})}$$

$$\Rightarrow \quad p(t|\mathbf{x},\mathbf{w},\beta) = \mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1}).$$

# 多项式曲线拟合中i.i.d.数据的图示

- 

# 参数w的最大似然估计

- 给定观测数据 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, 目标输出为 $\mathbf{t} = [t_1, \ldots, t_N]^{\mathrm{T}}$
- 计算似然函数(Likelihood function)

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}).$$

- **Taking the logarithm, we get**

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w}) \end{aligned}$$

where

$$\boxed{E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2}$$

**Sum-of-squares error**
**误差的平方和**

# 参数w的最大似然估计

- 最大似然估计

$$
\begin{aligned}
\max_{\mathbf{w}} \quad \ln p(\mathbf{t}|\mathbf{w},\beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\
&= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})
\end{aligned}
$$

**误差的平方和**

- 等价于最小化误差的平方和

$$
\min_{\mathbf{w}} \quad E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2
$$

# 参数w的最大似然估计

- 计算对数似然函数的梯度，并令其为零

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w},\beta) = \beta \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} = \mathbf{0}.$$

- 得出：

$$\mathbf{w}_{\mathrm{ML}} = \overbrace{\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}}^{\text{The Moore-Penrose pseudo-inverse, } \boldsymbol{\Phi}^{\dagger}.}\mathbf{t}$$

– 其中

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

# 参数$w_0$和$\beta$的最大似然估计

- 估计参数$w_0$
  - 对$\mathbf{w_0}$单独求偏导，令之为**0**，则得到

$$
\begin{aligned}
w_0 &= \bar{t} - \sum_{j=1}^{M-1} w_j \overline{\phi_j} \\
&= \frac{1}{N}\sum_{n=1}^{N} t_n - \sum_{j=1}^{M-1} w_j \frac{1}{N}\sum_{n=1}^{N} \phi_j(\mathbf{x}_n).
\end{aligned}
$$

- 估计参数$\beta$

$$
\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2
$$

# 最小二乘的代数解释

- 考虑前面介绍过的多项式曲线拟合问题

$$\min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{\boldsymbol{\varphi}(x_n)^T \mathbf{w} - t_n\right\}^2 = \frac{1}{2}(\Phi\mathbf{w}-\mathbf{t})^T(\Phi\mathbf{w}-\mathbf{t}) = \frac{1}{2}\|\Phi\mathbf{w}-\mathbf{t}\|_2^2$$

求得 $$\mathbf{w}_{ML} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\Phi\mathbf{w}-\mathbf{t}\|_2^2$$

从而获得目标输入的最佳近似

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w}_{\mathrm{ML}} = [\varphi_1, \ldots, \varphi_M]\,\mathbf{w}_{\mathrm{ML}}.$$

其中

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

$\mathbf{y} \in \mathcal{S} \subseteq \mathcal{T}$ $\quad\quad$ $\mathbf{t} \in \mathcal{T}$

N-dimensional

M-dimensional

$\mathcal{S} = \mathrm{span}\{\varphi_1, \cdots, \varphi_M\}$

# 最小二乘的几何解释

- 考虑前面介绍过的多项式曲线拟合问题

$$\min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{\boldsymbol{\varphi}(x_n)^T \mathbf{w} - t_n\right\}^2 = \frac{1}{2}(\Phi\mathbf{w}-\mathbf{t})^T(\Phi\mathbf{w}-\mathbf{t}) = \frac{1}{2}\|\Phi\mathbf{w}-\mathbf{t}\|_2^2$$

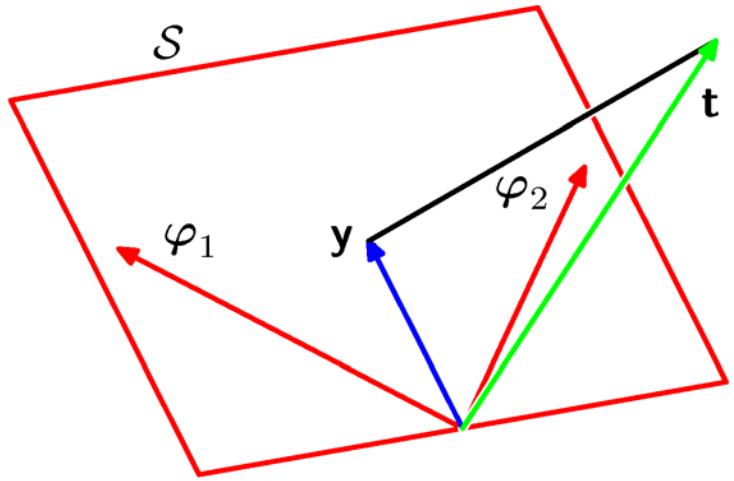
$$\mathbf{w}_{ML} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\Phi\mathbf{w}-\mathbf{t}\|_2^2$$

$$\mathbf{y} = \Phi\mathbf{w}_{\mathrm{ML}} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M]\,\mathbf{w}_{\mathrm{ML}}.$$

$$\mathbf{y} \in \mathcal{S} \subseteq \mathcal{T} \qquad \mathbf{t} \in \mathcal{T}$$

N-dimensional

M-dimensional




- **S is spanned by** $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M$

  **$\mathbf{w}_{\mathrm{ML}}$ minimizes the distance between t and its orthogonal projection on S, i.e. y**

# 广义线性模型内容提要

- 引子: 曲线拟合问题
- 最大似然估计
  - **最小二乘**
- 最大后验概率估计
  - **正则化最小二乘**
- 贝叶斯估计
- 方法比较:
  - 最大似然估计**(MLE) vs.** 最大后验概率估计**(MAP) vs.** 完全贝叶斯法

# 正则化的最小二乘法

- 学习任务:
  – 根据训练数据，估计模型中的参数

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

- 使用平方误差函数和权值的$L_2$范数正则化项，则得到正则化的最小二乘法

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

# 正则化最小二乘法的求解

- 优化问题变为：

$$\min_{\mathbf{w}} E(\mathbf{w}) = \frac{1}{2}\left(\mathbf{w}^T\left(\Phi^T\Phi + \lambda\cdot \mathrm{I}\right)\mathbf{w} - 2\mathbf{w}^T\Phi^T\mathbf{t} + \mathbf{t}^T\mathbf{t}\right)$$

- 对w求梯度

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \left(\Phi^T\Phi + \lambda\cdot \mathrm{I}\right)\mathbf{w} - \Phi^T\mathbf{t}$$

  - 令梯度为0， $\nabla_{\mathbf{w}} E(\mathbf{w}) = \left(\Phi^T\Phi + \lambda\cdot \mathrm{I}\right)\mathbf{w} - \Phi^T\mathbf{t} = 0$
  - 得出 $\left(\Phi^T\Phi + \lambda\cdot \mathrm{I}\right)\mathbf{w} = \Phi^T\mathbf{t}$

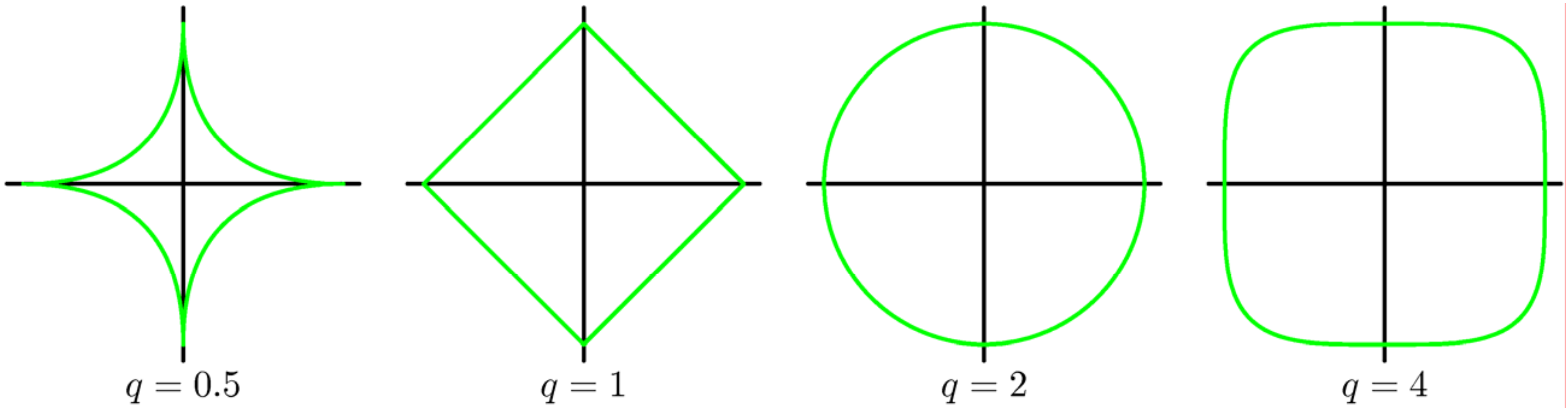
$$\Rightarrow \quad \mathbf{w} = \left(\Phi^T\Phi + \lambda\cdot \mathrm{I}\right)^{-1}\Phi^T\mathbf{t}$$

$$\Rightarrow \quad y(x,\mathbf{w}) = \boldsymbol{\varphi}(x)^T\mathbf{w} = \boldsymbol{\varphi}(x)^T\left(\Phi^T\Phi + \lambda\cdot \mathrm{I}\right)^{-1}\Phi^T\mathbf{t}$$

# 正则化项的形式

- 更一般的正则化项

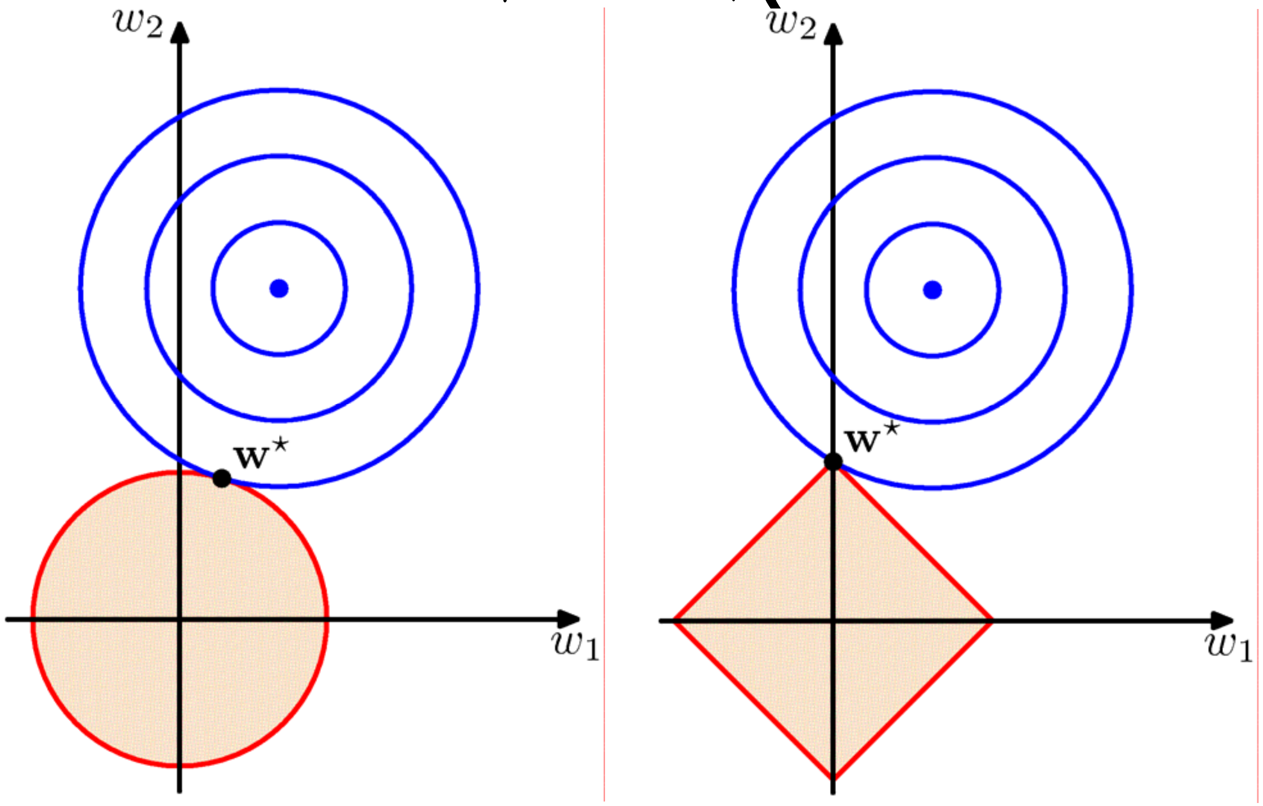
$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 \boxed{+ \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q}$$




Lasso　　Quadratic

# Lasso

- 如果使用w的L1范数作为正则化项，则变成Lasso模型
  - **L1范数倾向于生成比较稀疏的解(与L2范数相比)**



# 正则化最小二乘法从哪里来？

$$\min_{\mathbf{w}} \widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

- 正则化最小二乘法来源于最大后验概率估计法
  - 假设观测数据X来采样于一个确定性(deterministic) 函数，观测数据中存在i.i.d.加性高斯噪声(additive Gaussian noise)
  - 假设参数w服从零均值固定方差的高斯分布

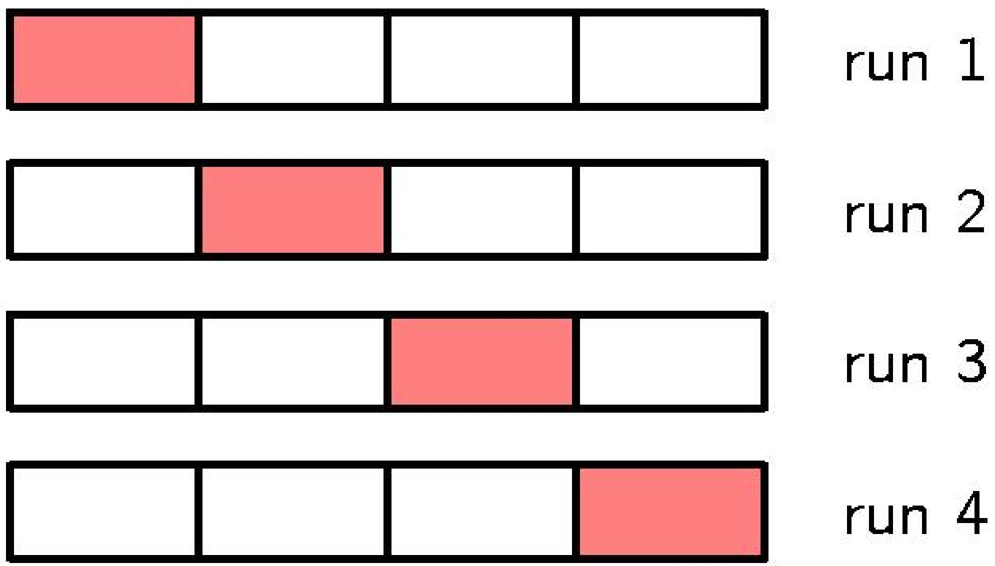
$$\boxed{t = y(\mathbf{x},\mathbf{w}) + \epsilon \text{ where } p(\epsilon|\beta) = \mathcal{N}(\epsilon|0,\beta^{-1}),\ p(\mathbf{w}) = N(\mathbf{w}\,|\,0,\sigma^2 I)}$$

$$\Rightarrow p(\mathbf{w}\,|\,D) \propto p(D\,|\,\mathbf{w})\,p(\mathbf{w}) \Rightarrow p(\mathbf{w}\,|\,D) \propto p(\mathbf{t}\,|\,X,\mathbf{w},\beta)\,p(\mathbf{w})$$

$$\Rightarrow \propto \prod_{n=1}^{N} N\left(t_n\,|\,\mathbf{w}^T\boldsymbol{\varphi}(x_n),\beta^{-1}\right) N\left(\mathbf{w}\,|\,0,\sigma^2 I\right)$$

# 模型选择(Model Selection)

- 交叉验证(Cross-Validation)
  – 把训练集划分为估计子集和验证子集
    - 在估计子集上训练模型
    - 在验证子集上评价模型的性能



# Q / A

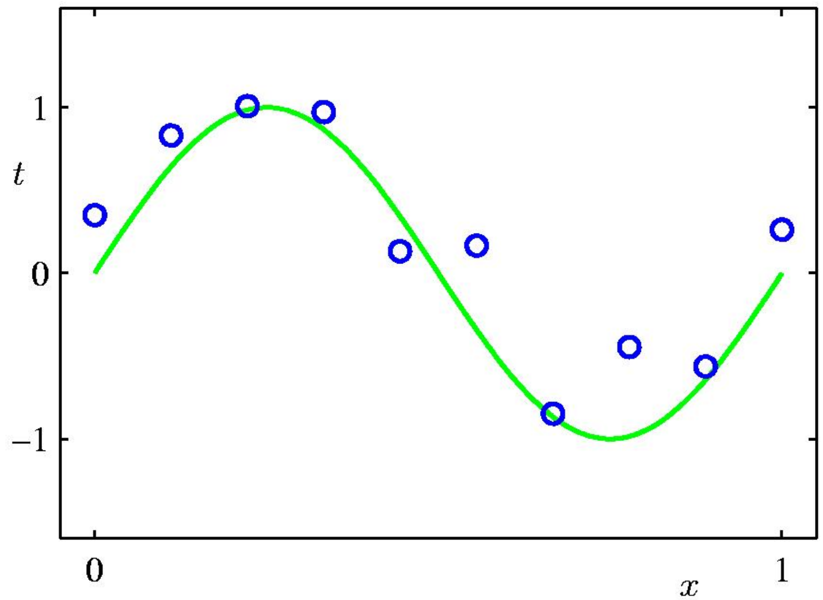
- Any Questions...

# 广义线性模型内容提要

- 引子: 曲线拟合问题
- 最大似然估计
  - **最小二乘**
- 最大后验概率估计
  - **正则化最小二乘**
- 贝叶斯估计
- 方法比较:
  - 最大似然估计**(MLE) vs.** 最大后验概率估计**(MAP) vs.** 完全贝叶斯法

# 例：多项式曲线拟合——完全贝叶斯法

- 给定N个训练数据：(x, t)
  - 数据其中绿色曲线为生成训练的真实曲线



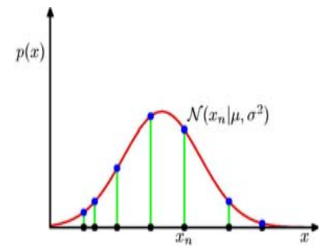

- 多项式曲线拟合
  - 使用多项式模型去构造回归模型
  - 定义为

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

# 从Bayes定理到MLE和MAP

- 把数据集记为D，分布中的待估计参数记为w，考虑到数据中的不确定性，则给定数据D，估计参数w的问题表示为:

$$p(\mathbf{w}\,|\,D) = \frac{p(D\,|\,\mathbf{w})\,p(\mathbf{w})}{p(D)}$$



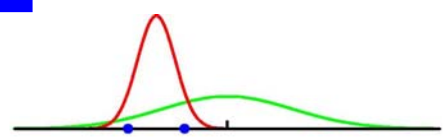

- 参数w的估计问题即优化问题 $\arg\max_{\mathbf{w}} p(\mathbf{w}\,|\,D)$
  - 如果关于**w** 没有任何已知信息，则

$$\boxed{\begin{aligned}\arg\max_{\mathbf{w}} p(\mathbf{w}\,|\,D) &= \arg\max_{\mathbf{w}} p(D\,|\,\mathbf{w}) \\ &= \arg\max_{\mathbf{w}} L(\mathbf{w}\,|\,D) = \arg\max_{\mathbf{w}} \log L(\mathbf{w}\,|\,D)\end{aligned}}$$

    - 最大似然估计法(Maximal Likelihood Estimation)
  - 如果关于**w**有一些已知信息，**e.g., w**是高斯分布, **w**是**Laplacian**分布

$$\boxed{\arg\max_{\mathbf{w}} p(\mathbf{w}\,|\,D) = \arg\max_{\mathbf{w}} p(D\,|\,\mathbf{w})\,p(\mathbf{w})}$$

    - 最大后验概率估计(Maximum A Posterior Estimation)



# 最大似然估计法(MLE)

- 似然函数为

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

- 对数似然函数为

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

- 通过最大化似然函数估计参数$w_{ML}$和$\beta_{ML}$

$$\boxed{\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}}$$

$$\boxed{\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2}$$

# 最大后验概率(MAP)估计法

- 先验分布密度

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

- 似然函数

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n | y(x_n, \mathbf{w}), \beta^{-1}\right)$$

- 后验分布密度

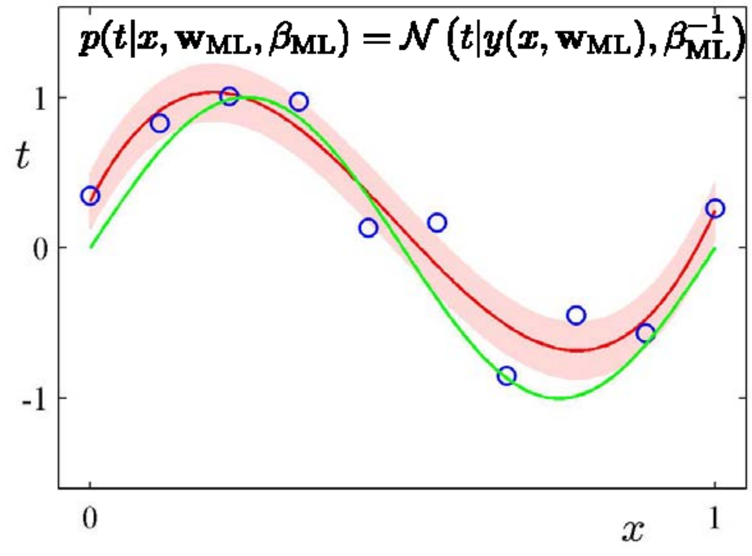
$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

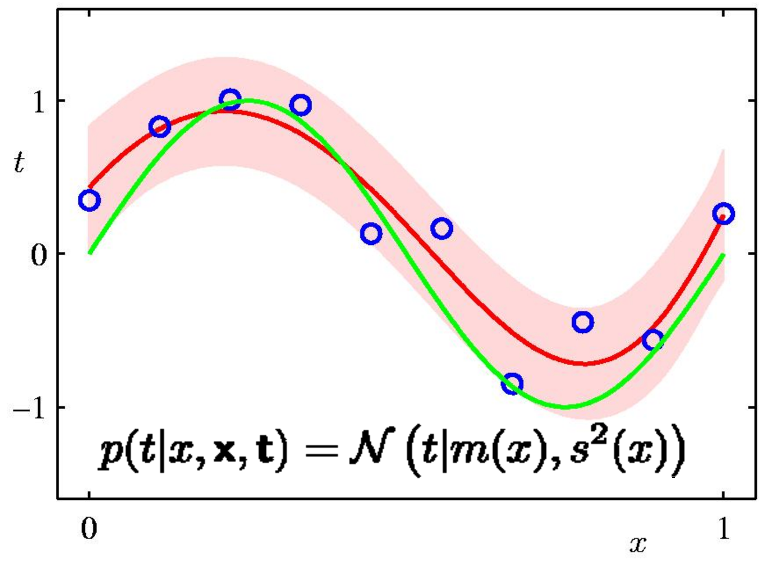
- 等价于正则化的最小二乘法

$$\boxed{\beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}}$$

# 例：多项式曲线拟合——贝叶斯法

- 贝叶斯法
  - 确定参数的先验分布
    - 根据先验知识
  - 通过推理，计算参数的后验分布
    - 根据训练数据和先验分布
  - 决策过程
    - 利用参数的后验分布加权回归分布密度函数



# 贝叶斯估计法(Bayesian Estimation)

- 先验分布密度

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

- 似然函数

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right)$$

- 后验分布密度

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

- 回归模型的预测性分布密度

$$\boxed{p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})\,\mathrm{d}\mathbf{w} = \mathcal{N}\left(t|m(x), s^2(x)\right)}$$

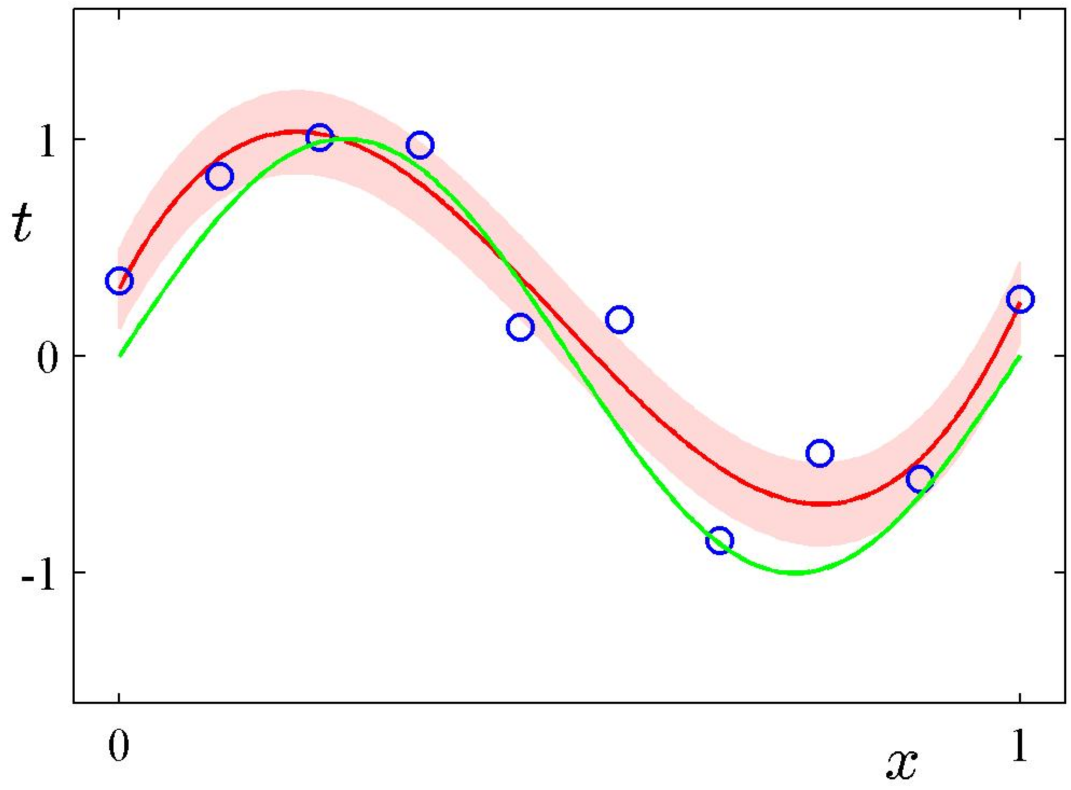
其中

$$m(x) = \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n \qquad s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}} \qquad \boldsymbol{\phi}(x_n) = \left(x_n^0, \ldots, x_n^M\right)^{\mathrm{T}}$$

# 回归模型的预测性(Predictive)分布密度

- $$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$



# 贝叶斯预测性分布密度

- $$p(t|x,\mathbf{x},\mathbf{t}) = \int p(t|x,\mathbf{w})p(\mathbf{w}|\mathbf{x},\mathbf{t})\,\mathrm{d}\mathbf{w} = \mathcal{N}\left(t|m(x), s^2(x)\right)$$

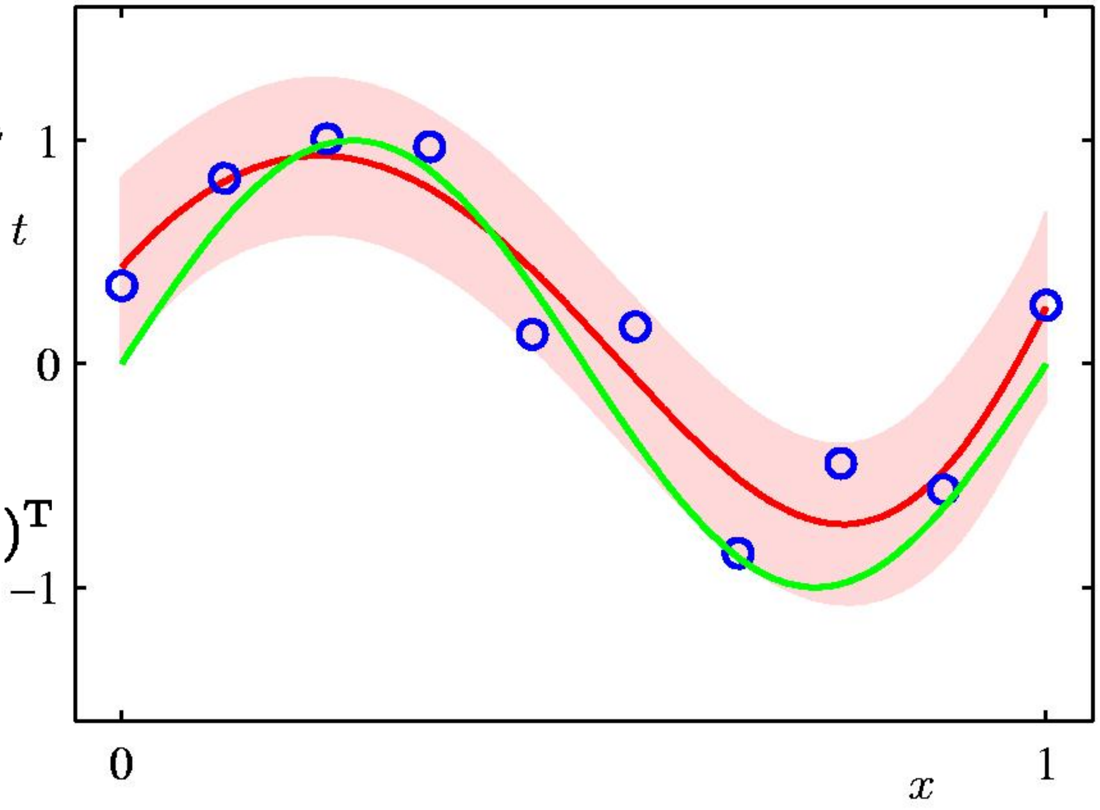
  – 其中

$$m(x) = \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}}$$

$$\boldsymbol{\phi}(x_n) = \left(x_n^0, \ldots, x_n^M\right)^{\mathrm{T}}$$



# 广义线性模型内容提要

- 引子: 曲线拟合问题
- 最大似然估计
  - **最小二乘**
- 最大后验概率估计
  - **正则化最小二乘**
- 贝叶斯估计
- 方法比较:
  - 最大似然估计**(MLE) vs.** 最大后验概率估计**(MAP) vs.** 完全贝叶斯法

# 从Bayes 定理到MLE\MAP\Bayesian方法

- 贝叶斯定理

$$P(\mathbf{w} \mid D) = \frac{P(D \mid \mathbf{w}) P(\mathbf{w})}{P(D)}$$

  – 最大似然(ML)法

$$\boxed{\max l(\mathbf{w} \mid D) = P(D \mid \mathbf{w})}$$

  – 最大后验概率(MAP)法

$$\boxed{\max P(\mathbf{w} \mid D) \propto P(D \mid \mathbf{w}) P(\mathbf{w})}$$

  – 贝叶斯方法
    - 不再估计参数，而是估计参数的后验分布 $p(\mathbf{w} \mid D)$
    - 不是构造回归函数，而是构造一个回归模型的分布密度 $p(t \mid x, \mathbf{w})$
    - 决策阶段利用参数的后验分布函数去加权回归模型的预测性分布密度函数

$$\boxed{p(t \mid x, D) = \int_{\mathbf{w}} p(t \mid x, \mathbf{w})\, p(\mathbf{w} \mid D)\, d\mathbf{w}}$$

# Q / A

- Any Questions...

# 广义线性模型内容小结

- 最大似然估计
  - **最小二乘**
- 最大后验概率估计
  - **正则化最小二乘**
- 贝叶斯估计
- 方法比较:
  - 最大似然估计**(MLE) vs.** 最大后验概率估计**(MAP) vs.** 完全贝叶斯法

# Q / A

- Any Questions...

# 模式识别引论 作业安排
An Introduction to Pattern Recognition

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

网络搜索教研中心 信息与通信工程学院 北京邮电大学

# 作业 1 推导正则化回归模型

- 带正则化项的方法一般源于最大后验概率估计。假设观测数据X来采样于一个确定性(**deterministic**) 的线性函数，观测数据中存在i.i.d.加性高斯噪声(**additive Gaussian noise**)，即

$$t = y(\mathbf{x}, \mathbf{w}) + \varepsilon \quad \text{where} \quad p(\varepsilon \mid \beta) = N(\varepsilon \mid 0, \beta^{-1})$$

1. 假设参数**w**服从**Gaussian**分布，即 $p(\mathbf{w}) = N(\mathbf{w} \mid 0, \sigma^2 I)$

2. 假设参数**w**服从**Laplacian**分布，即 $p(\mathbf{w}) = \frac{\lambda}{2} \exp(-\lambda \|\mathbf{w}\|_1)$

– 请分别推导**w**的估计公式(或优化模型)，并简单分析两者的异同.

– 截止期限: **11**月**6**日课上