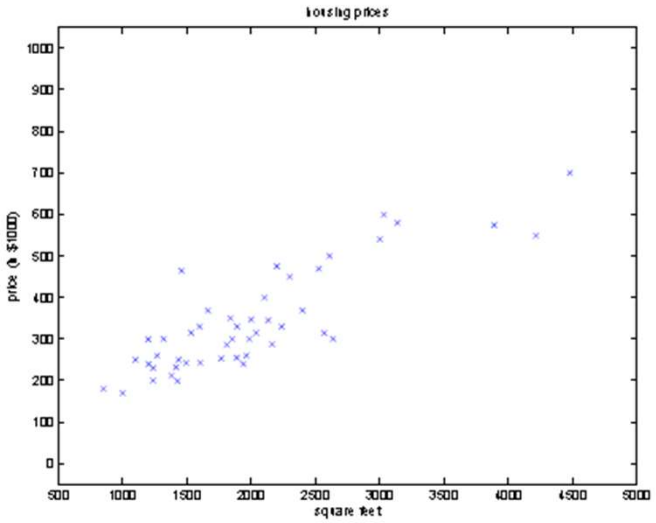
回归的线性模型
Linear Models for Regression

Regression: An Example

- If we have a dataset giving the living areas and prices of 47 houses as showing in the left-hand table, we can plot this data as in the right-hand plot.

Living area (feet2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540



Given data like this, how can we predict the prices of other houses as the function of the size of their living areas?

回归的定义

- 什么是回归(Regression)？
  在给定输入D维向量$\mathbf{x}$的取值条件下，去预测一个或多个连续目标变量t的值。
- 什么是线性回归模型？
  模型是可调节参数的线性函数
  - 最简单的线性回归模型：不仅是可调节参数的线性函数，也是输入变量的线性函数
  - **基函数**：输入变量的非线性函数

Solutions

- 输入：给定由N条观察数据 $\{x_n\}$，其中$n = 1, \ldots, N$，以及相应的目标值$\{t_n\}$
- 目的：为新$\mathbf{x}$值来预测t值

- 方法一、直接构造一个恰当的函数$y(\mathbf{x})$
- 方法二、对预测分布$p(t|\mathbf{x})$进行建模，然后对t进行预测，使得损失函数的期望达到最小

Outline

- 线性基函数模型(Linear Basis Function Model)
- 鲁棒线性回归(Robust Linear Regression)
- 偏差-方差分解(Bias-Variance Decomposition)
- 贝叶斯线性回归(Bayesian Linear Regression)
- 贝叶斯模型比较(Bayesian Model Comparison)
- 证据近似Evidence Approximation
- 固定基函数的局限性

Outline

- 线性基函数模型(Linear Basis Function Model)
- 鲁棒线性回归(Robust Linear Regression)
- 偏差-方差分解(Bias-Variance Decomposition)
- 贝叶斯线性回归(Bayesian Linear Regression)
- 贝叶斯模型比较(Bayesian Model Comparison)
- 证据近似Evidence Approximation
- 固定基函数的局限性

Outline

- 线性基函数模型(Linear Basis Function Model)
  - 极大似然和最小二乘
  - 最小二乘法的几何意义
  - 顺序学习
  - 多个输出
- 偏差-方差分解(Bias-Variance Decomposition)
- 贝叶斯线性回归(Bayesian Linear Regression)
- 贝叶斯模型比较(Bayesian Model Comparison)
- 证据近似Evidence Approximation
- 固定基函数的局限性

线性回归模型
Linear Regression Model

- 线性回归模型是输入变量的线性组合

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

其中$\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$

- 关键特性：它是参数$w_0, \ldots, w_D$的一个线性函数，也是输入变量的一个线性函数

线性基函数模型
Linear Basis Function Models

- 输入向量的固定非线性函数的线性组合：

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

- 定义一个哑“基函数”$\phi_0(\mathbf{x}) = 1$，则

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

其中$\mathbf{w} = (w_0, \ldots, w_{M-1})^T$而$\phi = (\phi_0, \ldots, \phi_{M-1})^T$

基函数
Basis Functions

- 多项式(Polynomial)基函数

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

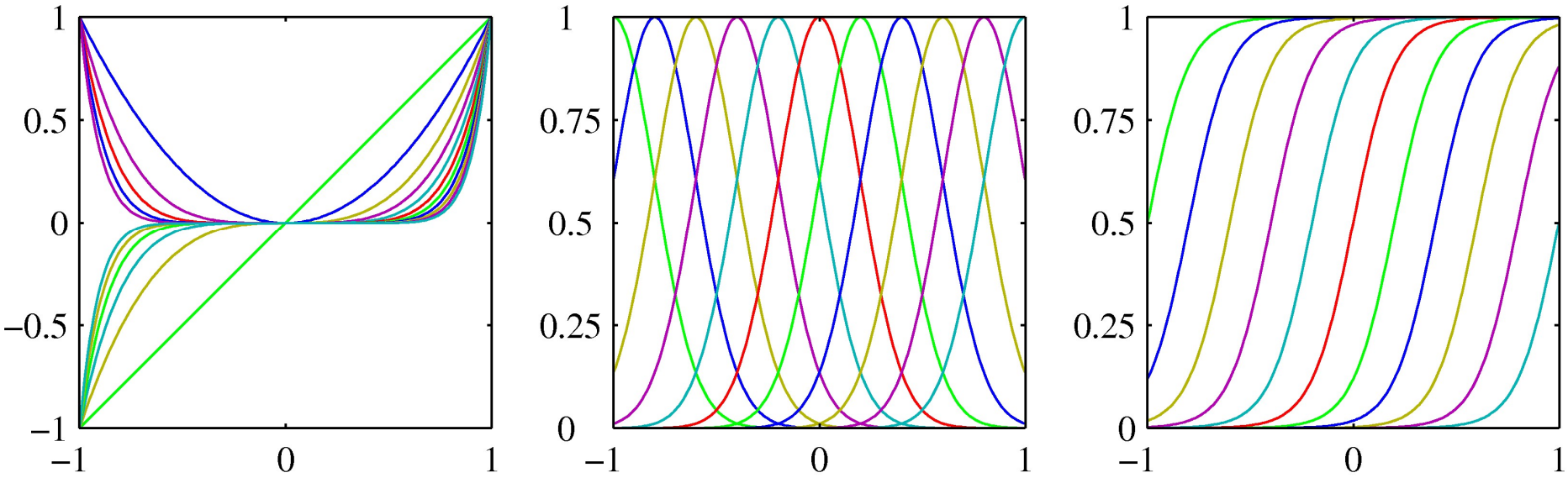
- 高斯(Gaussian)基函数

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

- Sigmoid基函数

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

基函数
示例



最大似然估计
Maximum Likelihood Estimation (MLE)

参数的最大似然估计，它定义为：

$$\widehat{\boldsymbol{\theta}} \triangleq \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta})$$

假设训练样例独立同分布（iid），对数似然为：

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

概率解释
Probabity Interpretion

- 假设目标变量t是由附加了高斯噪音的确定性函数$y(\mathbf{x}, \mathbf{w})$来给出的

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

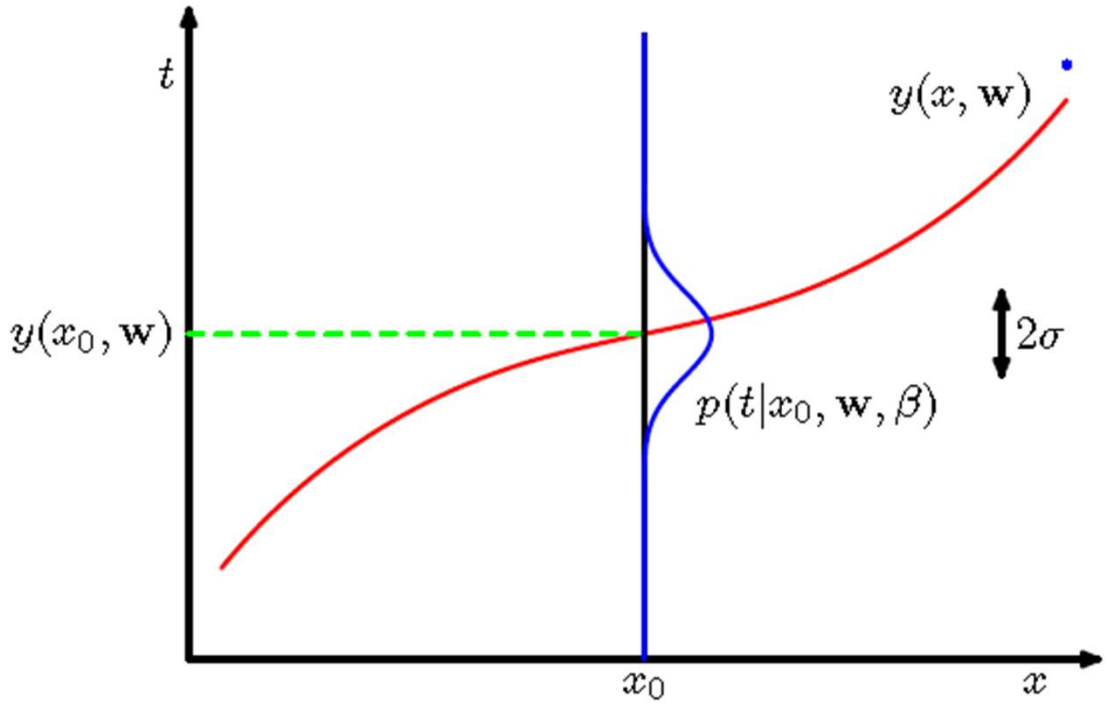
  - 其中ϵ是一个均值为0精度为β的高斯随机变量

- 因此：

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

线性回归的概率方法

- Schematic illustration of a Gaussian condition distribution for t given x.



似然函数
Likelihood Function

- 输入：$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ 及相应的目标值$\mathbf{t} = \{t_1, \ldots, t_N\}$
- 假设：数据独立地从分布中抽取出来
- 似然函数：

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

对数似然
Log Likelihood

取似然函数的对数，可得如下的对数似然：

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

$$= \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \beta E_D(\mathbf{w})$$

其中

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

w的最大似然解
Maximum likelihood

设计矩阵Design Matrix

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

对数似然函数对于**w**的导数为：

$$\nabla \ln p(\mathbf{t}|\mathbf{w},\beta) = \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}\phi(\mathbf{x}_n)^{\mathrm{T}}$$

导数设置为0得到

$$0 = \sum_{n=1}^{N} t_n \phi(\mathbf{x}_n)^{\mathrm{T}} - \mathbf{w}^{\mathrm{T}} \left(\sum_{n=1}^{N} \phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^{\mathrm{T}} \right)$$

$\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$　　$\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$

求解**w**，我们得到

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

伪逆Pseudo Inverse

β的最大似然解

针对噪音精度参数β来最大化对数似然，得到

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

噪音精度的逆（倒数）是由目标值围绕在回归函数周围的残留方差来给出的。

最小二乘法的几何解释

- 考虑一个N维空间，它的轴由t_n给定
  - $\mathbf{t} = (t_1, \ldots, t_N)^T$是这个空间中的一个向量
  - $\varphi_j = [\phi_j(x_1), \phi_j(x_2), \ldots, \phi_j(x_N)]^T$ 也是！

- 线性回归模型定义

$$\mathbf{y} = \begin{bmatrix} y(\mathbf{x}_1, \mathbf{w}) \\ y(\mathbf{x}_2, \mathbf{w}) \\ \vdots \\ y(\mathbf{x}_N, \mathbf{w}) \end{bmatrix} = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix}$$

$\boldsymbol{\Phi}$ $\mathbf{w}$

$$= [\varphi_0 \quad \varphi_1 \quad \cdots \quad \varphi_{M-1}] \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix}$$

最小二乘法的几何解释

- 最小化残差$\mathbf{t}-\mathbf{y}$的范数，要求残差向量与$\mathbf{\Phi}$的每一列都正交：

$$\varphi_j^T(\mathbf{t}-\mathbf{y})=0\ (j=1{:}D)$$

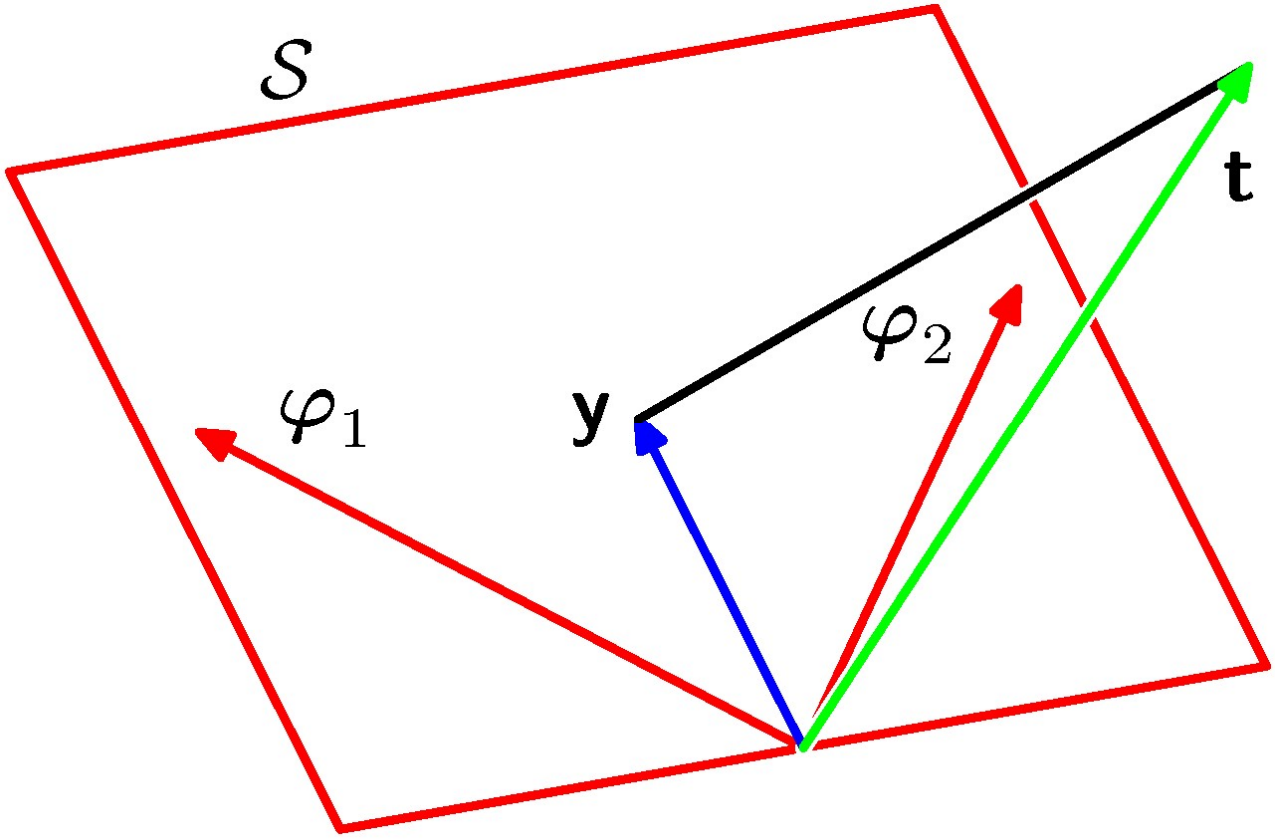
$$\mathbf{\Phi}^T(\mathbf{t}-\mathbf{y})=\mathbf{0}$$

$$\mathbf{\Phi}^T(\mathbf{t}-\mathbf{\Phi w})=\mathbf{0}$$

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

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

最小二乘的几何解释



顺序学习
Sequential Learning

- 批量处理技术(batch techniques)一次处理整个训练集

- 顺序算法（也称为在线算法Online Algorithm）一次只考虑一个数据点，而模型参数在每次看到一条数据点后将进行更新
  - 适用于大数据集
  - 适用于数据以一种连续数据流的形式达到的应用

批量梯度下降
Batch Gradient Descent

- 批量梯度下降
  - 训练数据集D上的损失$E_D = \sum_n E_n$是$\mathbf{w}$的二次函数
  - 首先，设定$\mathbf{w}$的初始猜测值，而后不断地更新$\mathbf{w}$使得E_D变得更小，直到其收敛，从而最小化E_D
- 迭代更新公式：

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \frac{\partial}{\partial w} E_D(\mathbf{w})$$

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \sum_{n=1}^{N} \{t_n - \mathbf{w}^{(\tau)\mathrm{T}} \phi(x_n)\} \phi(x_n)$$

If this term is small, it means that last w is enough good and we only have to make a slight change to w.

随机梯度下降
Stochastic Gradient Descent

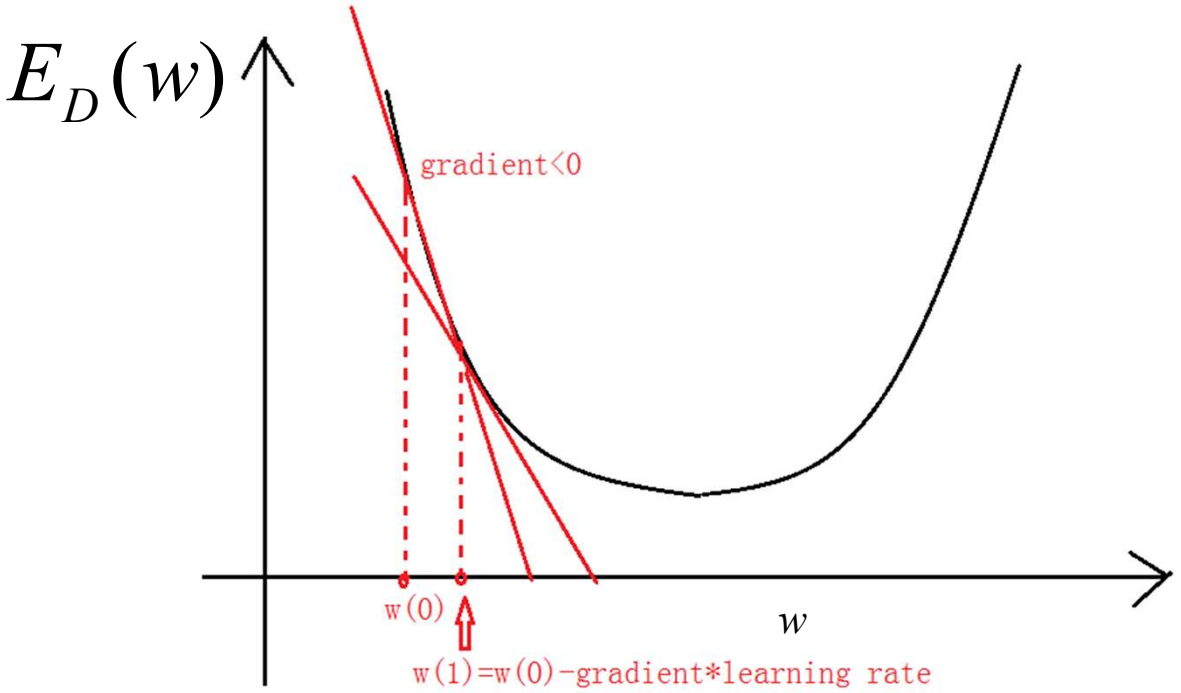
随机梯度下降在看到第n条数据（损失为E_n）即更新参数向量$\mathbf{w}$:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n$$

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \left\{ t_n - \mathbf{w}^{(\tau)\mathrm{T}} \phi(x_n) \right\} \phi(x_n)^{\mathrm{T}}$$

梯度下降：学习率

- 学习率太小=>收敛很慢
- 学习率太大=>错过最小点



正则化的最小二乘法
Regularized Least Squares

- 正则化技术：向误差函数中添加正则化项来控制过拟合

$$E_D(\mathbf{w}) + \lambda E_w(\mathbf{w})$$

- 一种最简单的正则化项:

$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

权重衰减
(weight decay)

- 误差函数仍然是二次函数，因此其精确的最小点存在闭式解析解

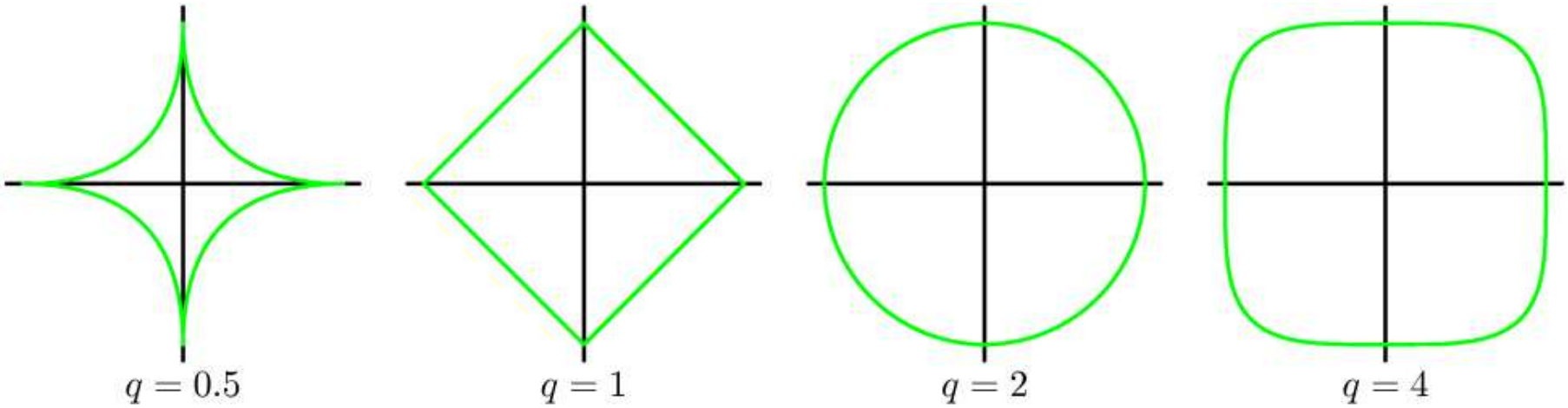
$$\mathbf{w} = (\lambda\mathbf{I} + \mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

更一般的正则化项

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

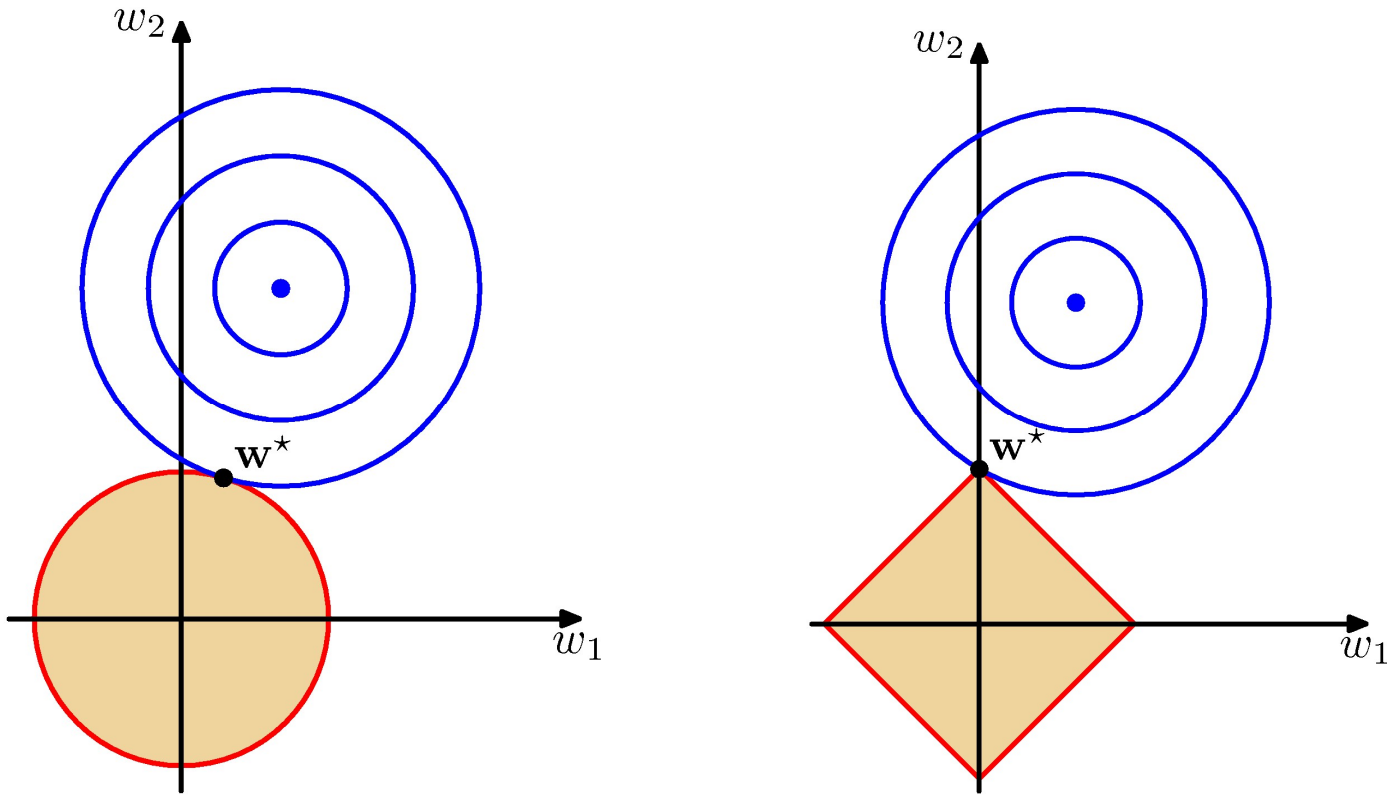
当$q=2$时，就对应于二次正则化项

Contours of the Regularization Term



Lasso

Lasso具有如下特性：如果λ足够大，某些系数w_j将被驱使为0，从而得到一个**稀疏模型**



Outline

- 线性基函数模型(Linear Basis Function Model)
- 鲁棒线性回归(Robust Linear Regression)
- 偏差-方差分解(Bias-Variance Decomposition)
- 贝叶斯线性回归(Bayesian Linear Regression)
- 贝叶斯模型比较(Bayesian Model Comparison)
- 证据近似Evidence Approximation
- 固定基函数的局限性

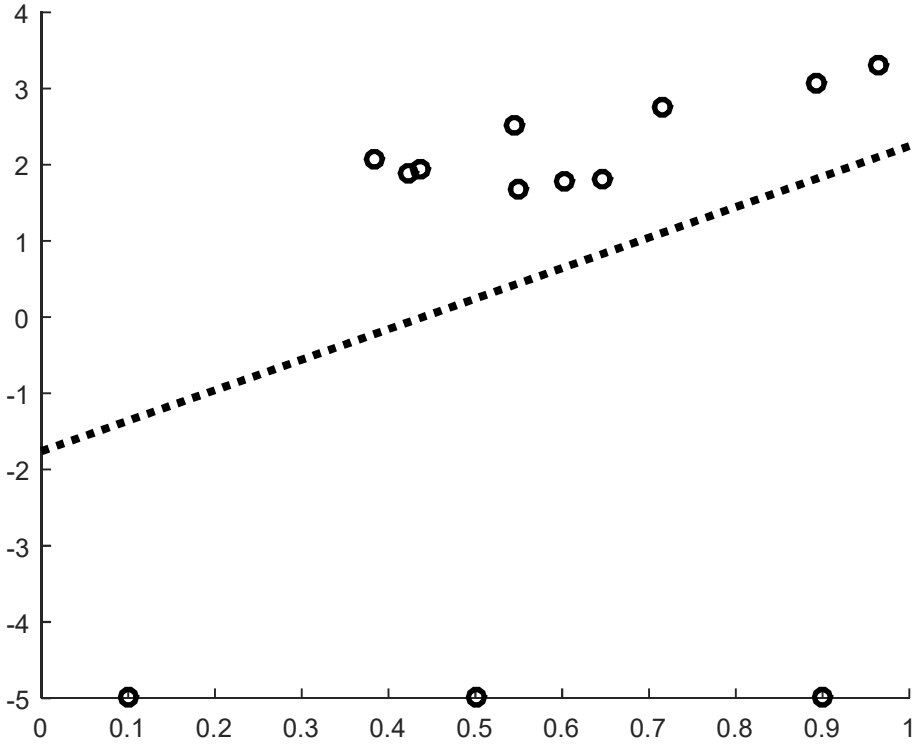
鲁棒线性回归
Robust Linear Regression

噪音建模：零均值和常量方差的高斯分布

$$(\epsilon_i = t_i - \mathbf{w}^T\mathbf{x}_i) \sim \mathcal{N}(0, \sigma^2)$$

最大似然⇔最小化平方误差和

异常/离群数据的存在可能会导致较差的拟合



鲁棒线性回归
拉普拉斯分布

鲁棒线性回归模型的基本思想：

用长尾分布（代替高斯分布）进行噪音建模

拉普拉斯分布是一种常见的长尾分布：

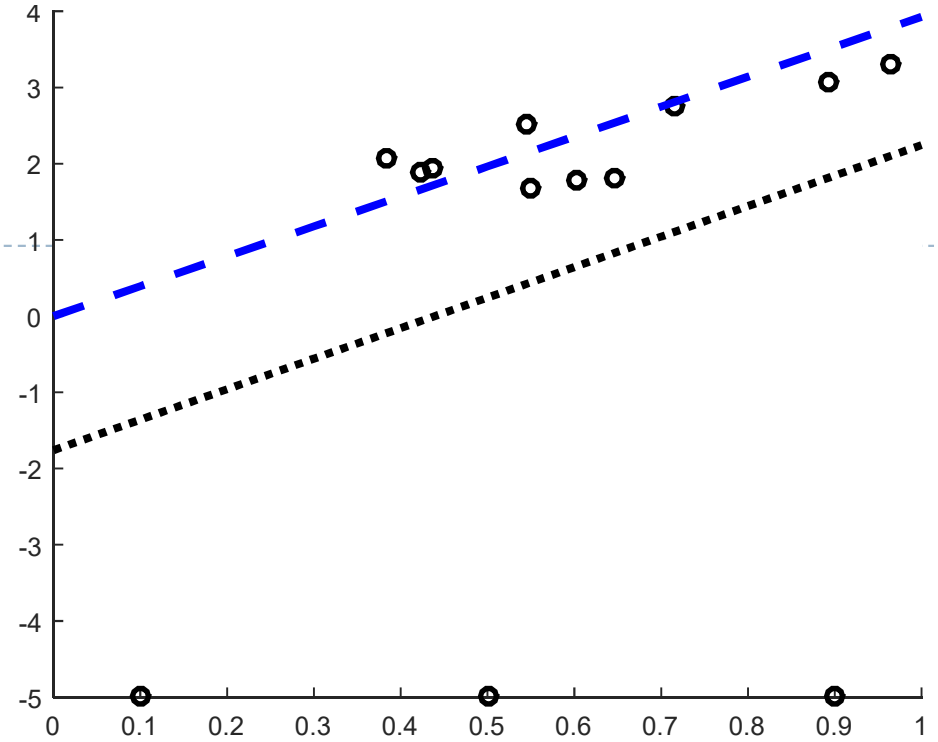
$$p(y|\mathbf{x},\mathbf{w},b) = Lap\left(y\middle|\mathbf{w}^T\mathbf{x},b\right) \propto \exp\left(-\frac{1}{b}|y-\mathbf{w}^T\mathbf{x}|\right)$$

负对数似然函数(NLL)：

$$\ell(\mathbf{w}) = \sum_i |r_i(\mathbf{w})| \qquad r_i \equiv y_i - \mathbf{w}^T\mathbf{x}_i \text{是第} i \text{个残差(residual)}$$

难以优化的非线性目标函数

Split Variable



$$r_i \triangleq r_i^+ - r_i^-$$
$$r_i^+ = \max(r_i, 0)$$
$$r_i^- = \max(-r_i, 0)$$

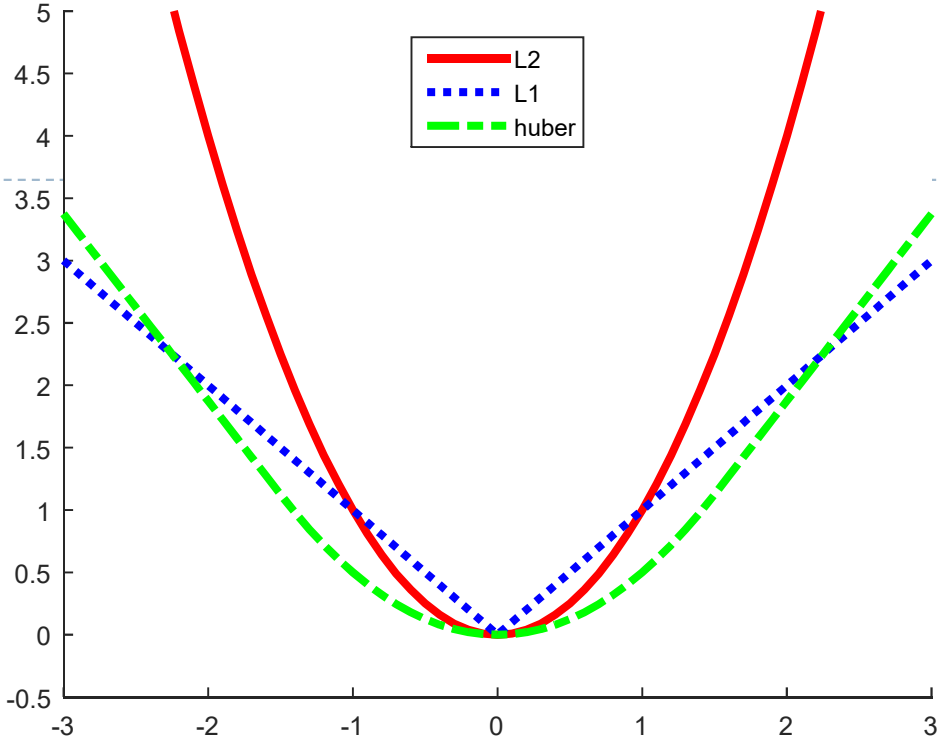
$$\min_{\mathbf{w},\mathbf{r}^+,\mathbf{r}^-} \sum_i \left(r_i^+ + r_i^-\right) \quad s.t. \;\; r_i^+ \geq 0, r_i^- \geq 0, \mathbf{w}^T\mathbf{x}_i + r_i^+ - r_i^- = y_i$$

线性程序（LP）的标准形式：

$$\min_{\boldsymbol{\theta}} \mathbf{f}^T\boldsymbol{\theta} \;\; s.t. \;\; \mathbf{A}\boldsymbol{\theta} \leq \mathbf{b}, \qquad \mathbf{A}_{eq}\boldsymbol{\theta} = \mathbf{b}_{eq}, \qquad \mathbf{l} \leq \boldsymbol{\theta} \leq \mathbf{u}$$

有：$\boldsymbol{\theta} = (\mathbf{w}, \mathbf{r}^+, \mathbf{r}^-)$, $\mathbf{f} = [\mathbf{0}, \mathbf{1}, \mathbf{1}]$, $\mathbf{A} = []$, $\mathbf{b} = []$, $\mathbf{A}_{eq} = [\mathbf{X}, \mathbf{I}, -\mathbf{I}]$, $\mathbf{b}_{eq} = \mathbf{y}$, $\mathbf{l} = [-\infty\mathbf{1}, \mathbf{0}, \mathbf{0}]$, $\boldsymbol{u} = []$

Huber损失
Huber Loss



Huber损失函数

$$L_H(r,\delta)=\begin{cases}\dfrac{r^2}{2} & if\ |r|\le\delta\\ \delta|r|-\dfrac{\delta^2}{2} & if\ |r|>\delta\end{cases}$$

处处可微：

$$\frac{d}{dr}L_H(r,\delta)\Big|_{r=\pm\delta}=\pm\delta$$

总结：线性回归
似然与先验

似然	先验	方法名称
高斯	均匀	最小二乘(Least Squares)
高斯	高斯	岭回归(Ridge Regression)
高斯	拉普拉斯	Lasso
拉普拉斯	均匀	鲁棒回归
Student	均匀	鲁棒回归

Outline

- 线性基函数模型(Linear Basis Function Model)
- 鲁棒线性回归(Robust Linear Regression)
- 偏差-方差分解(Bias-Variance Decomposition)
- 贝叶斯线性回归(Bayesian Linear Regression)
- 贝叶斯模型比较(Bayesian Model Comparison)
- 证据近似Evidence Approximation
- 固定基函数的局限性

回顾：期望损失

$$\mathbb{E}[L] = \int L\big(t, y(\mathbf{x})\big) p(\mathbf{x}, t) d\mathbf{x} dt$$

- 平方损失函数：

$$L\big(t, y(\mathbf{x})\big) = (y(\mathbf{x}) - t)^2$$

- 期望平方损失为：

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

需要最小化的目标函数

期望平方损失的最小化

$$\frac{\partial \mathbb{E}[L]}{\partial y(\mathbf{x})} = 2\int \{y(\mathbf{x}) - t\} p(\mathbf{x}, t) dt = 0$$

$$\Rightarrow \int y(\mathbf{x}) p(\mathbf{x}, t) dt = \int t p(\mathbf{x}, t) dt$$

$$\Rightarrow y(\mathbf{x}) p(\mathbf{x}) = \int t p(\mathbf{x}, t) dt$$

$$\Rightarrow y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) dt}{p(\mathbf{x})} = \int t p(t|\mathbf{x}) dt = \mathbb{E}_t[t|\mathbf{x}]$$

期望平方损失的分解

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

与$y(\mathrm{x})$无关的固有噪音

需要最小化该项：因此
取$y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$

记$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$，有

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

将平方损失函数进行展开

$$\{y(\mathbf{x})-t\}^2=\{y(\mathbf{x})-\mathbb{E}[t|\mathbf{x}]+\mathbb{E}[t|\mathbf{x}]-t\}^2$$
$$=\{y(\mathbf{x})-\mathbb{E}[t|\mathbf{x}]\}^2+2\{y(\mathbf{x})-\mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}]-t\}+\{\mathbb{E}[t|\mathbf{x}]-t\}^2$$

可以得到

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

$$=\iint\{y(\mathbf{x})-\mathbb{E}[t|\mathbf{x}]\}^2p(\mathbf{x},t)d\mathbf{x}dt+2\iint\{y(\mathbf{x})-\mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}]-t\}p(\mathbf{x},t)d\mathbf{x}dt$$

$$+\iint\{\mathbb{E}[t|\mathbf{x}]-t\}^2p(\mathbf{x},t)d\mathbf{x}dt$$

为0消去

因此

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

$$=\int\{y(\mathbf{x})-h(\mathbf{x})\}^2p(\mathbf{x})d\mathbf{x}+\int\{h(\mathbf{x})-t\}^2p(\mathbf{x},t)dtd\mathbf{x}$$

Bias-Variance Decomposition

- 使用参数化模型$y(\mathbf{x}, \mathbf{w})$来对$h(\mathbf{x})$建模
  - 贝叶斯观点：模型不确定性由$\mathbf{w}$的后验分布来表示
  - 频率主义者方法：基于数据集$\mathcal{D}$来对$\mathbf{w}$进行点估计，并试图去解释这个点估计的不确定性
    - 多个大小都是N数据集：从分布$p(t, \mathbf{x})$中独立抽取出来的。
    - 对任何一个数据集$\mathcal{D}$，运行学习算法得到一个预测函数$y(\mathbf{x}; \mathcal{D})$
    - 不同的数据集将给出不同的函数，因此得到不同的平方损失值

Bias-Variance Decomposition

- 考虑被积函数$\{y(\mathbf{x};\mathcal{D})-h(\mathbf{x})\}^2$，记$y(\mathbf{x};\mathcal{D})$在多个数据集上的平均值$\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]$，则

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

- 对这个表达式关于$\mathcal{D}$取期望值，得到$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D})-h(\mathbf{x})\}^2\right]$

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

期望平方损失的分解

$$expected\ loss = (bias)^2 + variance + noise$$

其中

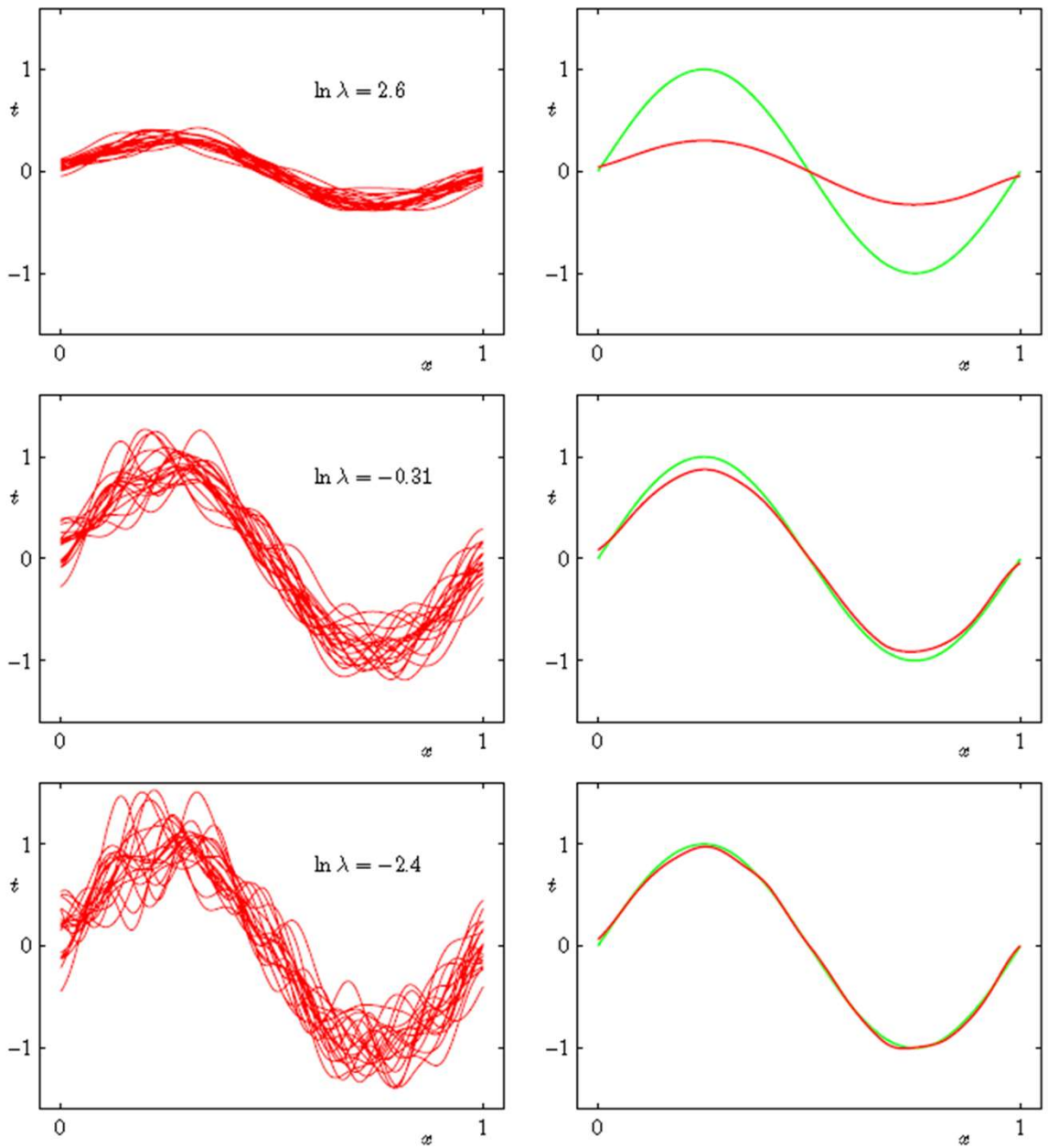
$$(bias)^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x}$$

$$variance = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x}) d\mathbf{x}$$

$$noise = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

偏差和方差之间的权衡

- 目标：最小化期望损失
  - 期望损失=平方偏差+方差+噪音
- 方差和偏差之间存在一种权衡
  - 非常灵活的模型：低偏差和高方差
  - 相对死板的模型：高偏差和低方差


ln λ = 2.6
ln λ = −0.31
ln λ = −2.4
t
x
0
1
−1

偏差-方差的权衡

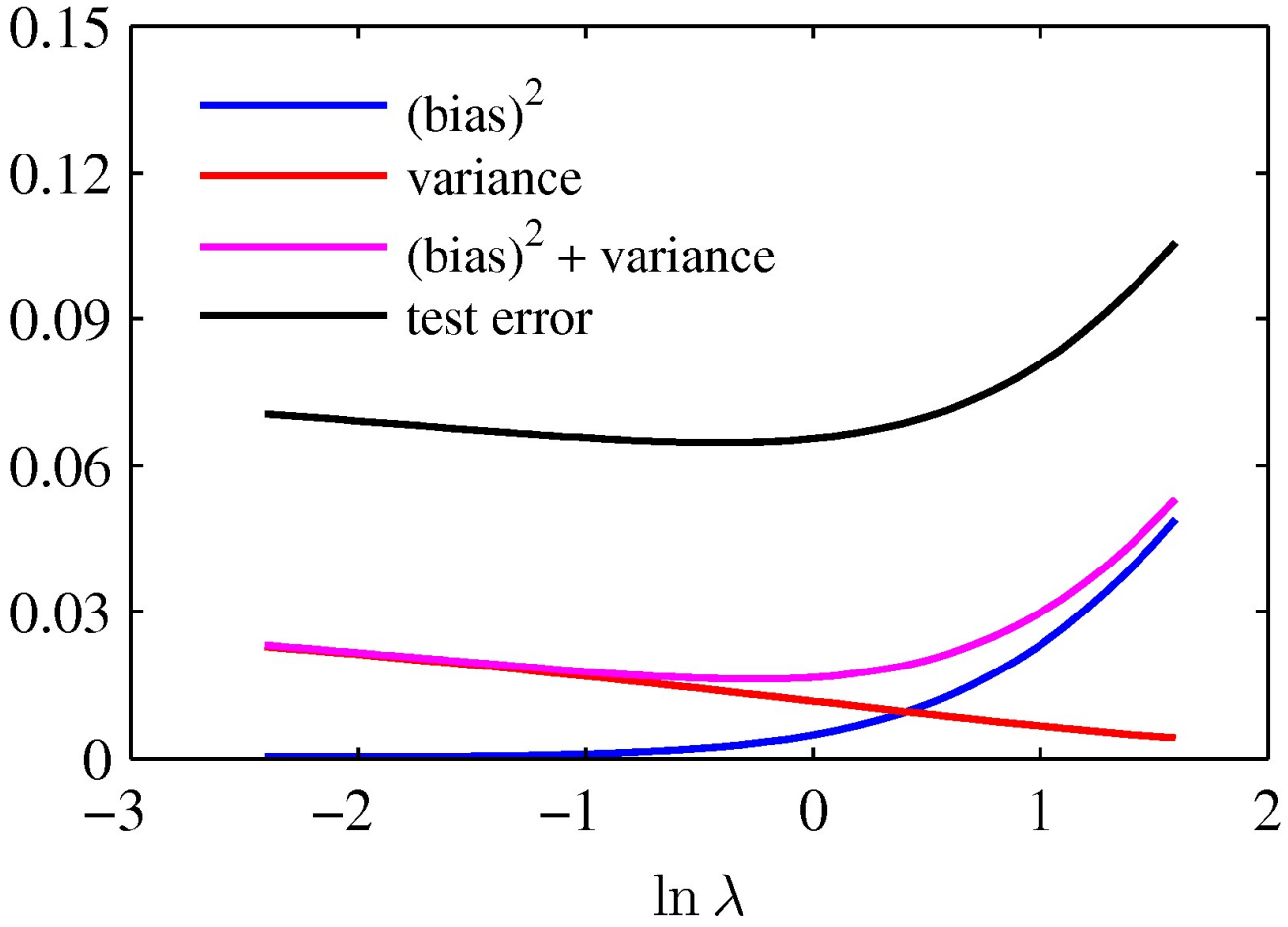
- 平均预测

$$\bar{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

- 积分的平方偏差和积分的方差分别为

$$(bias)^2 = \frac{1}{N}\sum_{n=1}^{N}\{\bar{y}(x_n) - h(x_n)\}^2$$

$$variance = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\{y^{(l)}(x_n) - \bar{y}(x_n)\}^2$$


0.15
0.12
0.09
0.06
0.03
0
−3
−2
−1
0
1
2
ln λ
(bias)2
variance
(bias)2 + variance
test error

局限性

- 使用价值有限：
- 偏差方差分解是基于对于一组数据集的平均
  - 然而我们实际上只具有单个的观察数据集
  - 即便有大量的独立训练集，将它们组合成一个单一的大型训练集将可以降低过拟合的程度。

Outline

- 线性基函数模型(Linear Basis Function Model)
- 鲁棒线性回归(Robust Linear Regression)
- 偏差-方差分解(Bias-Variance Decomposition)
- 贝叶斯线性回归(Bayesian Linear Regression)
- 贝叶斯模型比较(Bayesian Model Comparison)
- 证据近似Evidence Approximation
- 固定基函数的局限性

Outline

- 线性基函数模型(Linear Basis Function Model)
- 偏差-方差分解(Bias-Variance Decomposition)
- 贝叶斯线性回归(Bayesian Linear Regression)
  - 参数分布(Parameter Distribution)
  - 预测分布(Predictive Distribution)
  - 等价核(Equivalent Kernel)
- 贝叶斯模型比较(Bayesian Model Comparison)
- 证据近似Evidence Approximation
- 固定基函数的局限性

为什么需要贝叶斯线性回归?
Why Bayesian Linear Regression?

- 避免过拟合（与极大似然方法相比）：在模型参数上进行边缘化，而不是计算参数值的点估计
- 确定模型复杂度的自动方法（只使用训练数据）：
  - **正则化方法**则需要通过正则化系数来控制
  - **独立的验证集数据**可以计算复杂且浪费数据

参数分布(Parameter Distribution)
模型参数的似然函数与先验分布

似然函数：

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

共轭先验分布：

$$p(\mathbf{w})=\mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{S}_0)$$

参数分布(Parameter Distribution)

参数的后验分布

如果

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \Lambda^{-1})$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1})$$

则

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\Sigma\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\Lambda\mu\}, \Sigma)$$

其中

$$\Sigma = (\Lambda + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}.$$

先验： $p(\mathbf{w}) = \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0),$

似然： $p(\mathbf{t}|\mathbf{w}) = \mathcal{N}\left(\mathbf{w}^T\Phi(\mathbf{X}), \beta^{-1}\mathbf{I}\right)$

后验分布

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

其中

$$\mathbf{m}_N = \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right)$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

一种简单的高斯先验
先验与后验

高斯先验

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

均值为0，等方性(isotropic)

w的后验分布：

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

其中

$$\mathbf{m}_N = \beta\mathbf{S}_N\Phi^T\mathbf{t}$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\Phi^T\Phi$$

一种简单的高斯先验
对数后验

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(x_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + const$$

后验分布最大化就等价于对附加了二次正则化项的平方和误差函数的最小化，其中正则化系数为$\lambda = \alpha/\beta$

例子：参数分布

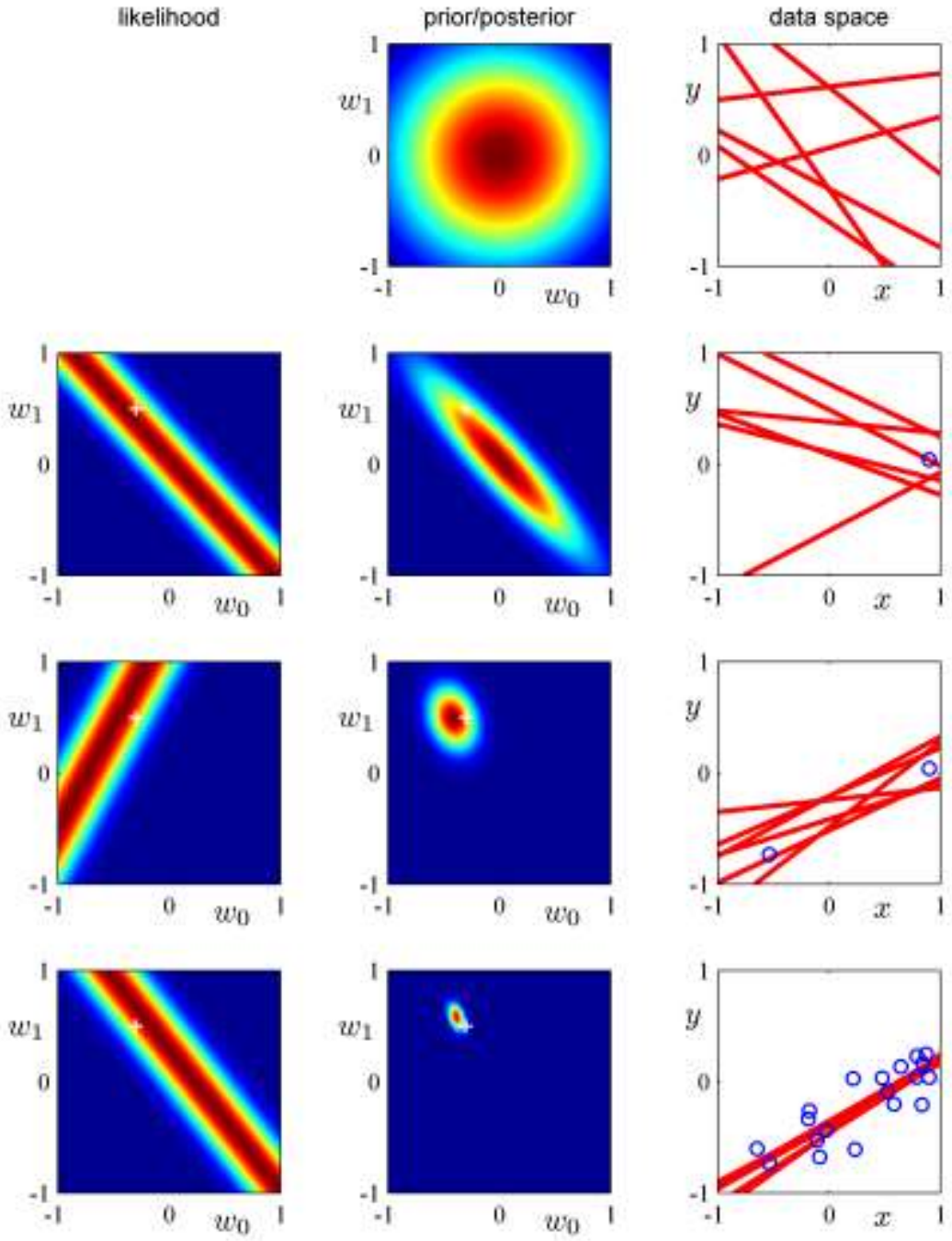
例子：函数$f(x, \mathbf{a}) = a_0 + a_1 x$，其中$a_0 = -0.3$而$a_1 = 0.5$

数据生成：

- 从均匀分布$\mathbf{U}(x| -1,1)$中选择x_n的值
- 计算$f(x_n, \mathbf{a})$的取值
- 添加标准差为0. 2的高斯噪音，得到目标值t_n

- 任务：根据这些数据找回a_0和a_1的值

$$\beta = (1/0.2)^2 = 25$$
$$\alpha = 2.0$$


likelihood
prior/posterior
data space
w_1
w_0
y
x
-1
0
1

贝叶斯预测分布
Bayesian Predictive Distribution

针对新的$\mathbf{x}$值做预测t

Posterior distribution over w

$$p(t|\mathbf{t},\alpha,\beta)=\int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)d\mathbf{w}$$

其中：

(1)目标变量的条件概率为

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

(2)后验权重分布为

$$p(\mathbf{w}|\mathbf{t})=\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$$

- $\mathbf{m}_N=\mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0+\beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right)$
- $\mathbf{S}_N^{-1}=\mathbf{S}_0^{-1}+\beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$

贝叶斯预测分布

如果

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1})$$

则

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}})$$

其中

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}.$$

既然

$$p(t|\mathbf{w},\beta) = \mathcal{N}\left(t\middle|y(\mathbf{x},\mathbf{w}),\beta^{-1}\right)$$
$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$$

因此

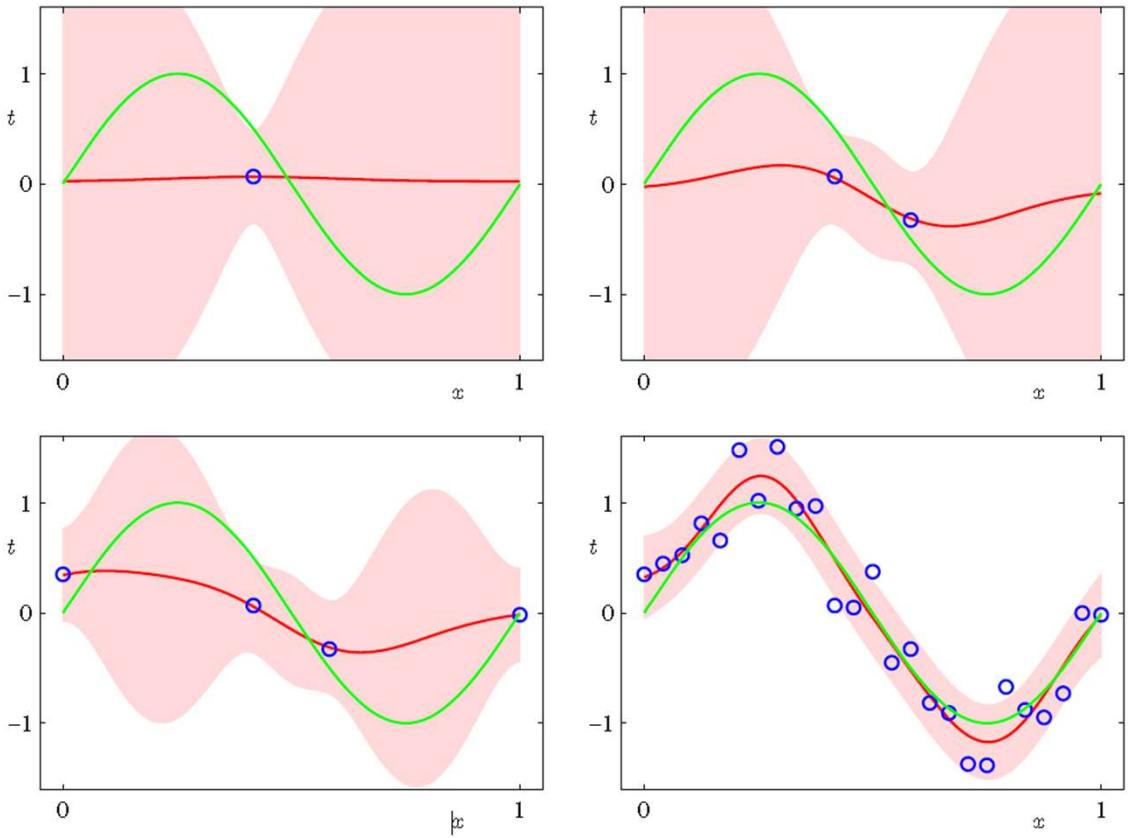
$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x}))$$

where the variance $\sigma_N^2(\mathbf{x})$ of the predictive distribution is given by

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}).$$

贝叶斯线性回归
Bayesian Linear Regression

- Example



Bayesian Linear Regression

- Example