机器学习
Machine Learning

第二讲 机器学习的统计与优化基础

概率基础、概率函数举例、最大似然、贝叶斯方法、决策理论、信息理论概述、优化基础

$p(x)$：对离散表示概率函数
对连续表示概率密度函数

1. 概率复习-概率的基本关系

边际概率公式（和公式）

$$p(x)=\sum_y p(y,x)$$

全概率公式（积公式）

$$p(x,y)=p(x|y)\,p(y)=p(y|x)\,p(x)$$

贝叶斯公式

$$p(y|x)=\frac{p(x,y)}{p(x)}=\frac{p(x|y)\,p(y)}{p(x)}$$

随机变量的基本特征

均值（1阶特征）

$$\mu = E[X] = \int xp(x)\mathrm{d}x$$

方差（2阶特征）

$$\sigma^2 = E\left[\left(X - E\left(X\right)\right)^2\right] = \int (x-\mu)^2 p(x)\mathrm{d}x$$

随机向量特征

$$\boldsymbol{\mu}_x = E[\mathbf{x}]$$ 均值向量

$$\boldsymbol{R}_{\mathbf{xx}} = E\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right]$$ 自相关矩阵

$$\boldsymbol{C}_{\mathbf{xx}} = E[(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^{\mathrm{T}}] = \boldsymbol{R}_{\mathbf{xx}} - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^{\mathrm{T}}$$ 自协方差矩阵

2. 函数期望的逼近
（蒙特卡罗方法）

$p(\boldsymbol{x})$，通过对该PDF采样产生一组独立样本

$$\{\boldsymbol{x}_n, \quad n = 1, 2, \cdots, N\}$$

PDF逼近为

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N_s}\sum_{i=1}^{N}\delta(\boldsymbol{x} - \boldsymbol{x}_n)$$

函数 $\boldsymbol{g}(\boldsymbol{x})$ 的期望为：

$$E[\boldsymbol{g}(\boldsymbol{x})] = \int \boldsymbol{g}(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{g}(\boldsymbol{x}_n)$$

例如： $L\left(f(\boldsymbol{x};\boldsymbol{\theta}),y\right)$ 表示每个样本的损失函数

风险函数是函数期望

$$J^*(\boldsymbol{\theta}) = \mathrm{E}_{(\boldsymbol{x},y)\sim p_{data}}\left\{L\left(f(\boldsymbol{x};\boldsymbol{\theta}),y\right)\right\}$$

p_{data} 表示数据的生成分布

则原理上经验风险是风险函数的蒙特卡洛逼近

$$J(\boldsymbol{\theta}) = \mathrm{E}_{(\boldsymbol{x},y)\sim \hat{p}_{data}}\left\{L\left(f(\boldsymbol{x};\boldsymbol{\theta}),y\right)\right\}$$

$$= \frac{1}{\mathrm{N}}\sum_{i=1}^{N} L\left(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}),y^{(i)}\right)$$

$\hat{p}_{data}$ 表示训练集的经验分布

3 概率函数举例

- 二元分布和二项分布
- 多元分布和多项分布
- 高斯分布和混合高斯分布
- 指数分布

3.1 二元变量（Binary Variables）

- Coin flipping: heads=1, tails=0

$$p(x=1|\mu)=\mu$$

- **Bernoulli** Distribution

$$\begin{aligned}\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu)\end{aligned}$$

二项式分布

- N coin flips:

$$p(m \text{ heads}|N,\mu)$$

- **Binomial** Distribution

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) = N\mu$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m-\mathbb{E}[m])^2\,\mathrm{Bin}(m|N,\mu) = N\mu(1-\mu)$$

3.2 多元变量（Multinomial Variables）

1-of-K coding scheme: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}$$

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

多项分布（The Multinomial Distribution

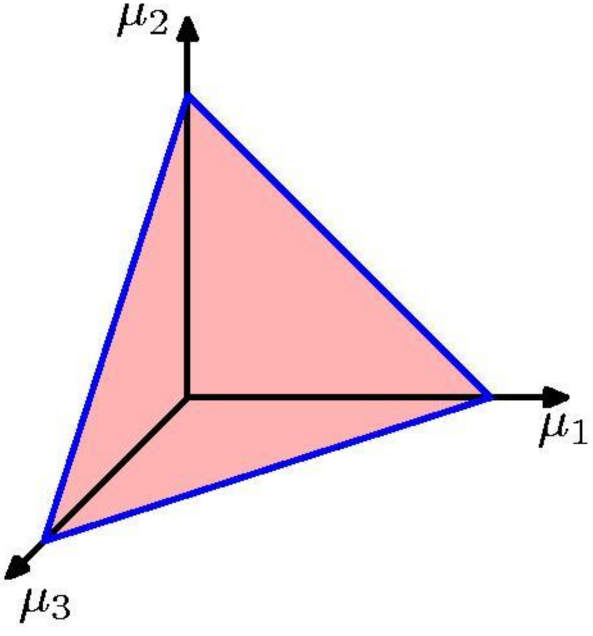
$$
\begin{aligned}
\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\
\mathbb{E}[m_k] &= N\mu_k \\
\mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\
\mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k
\end{aligned}
$$

The Dirichlet Distribution

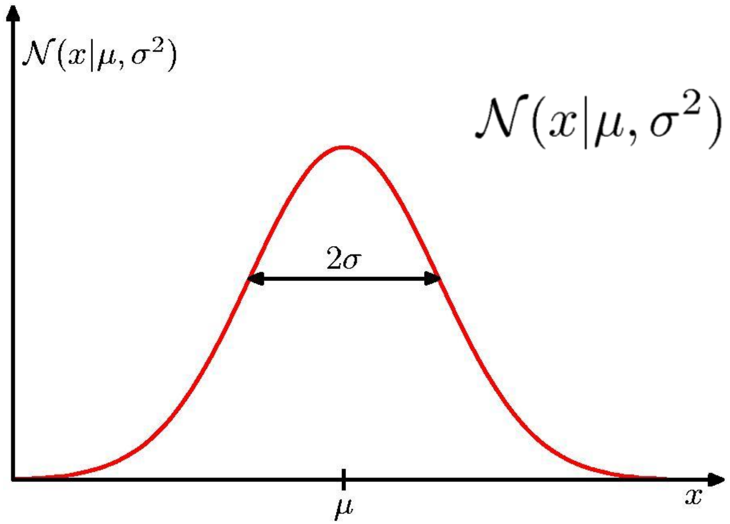
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$

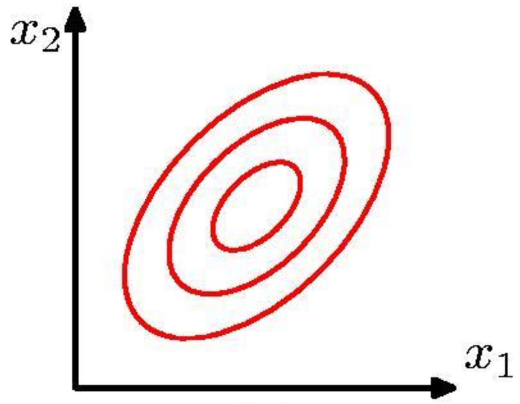
Conjugate prior for the multinomial distribution.



3.3 The Gaussian Distribution



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

高斯分布的划分
Partitioned Gaussian Distributions

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

划分的条件和边际分布
Partitioned Conditionals and Marginals

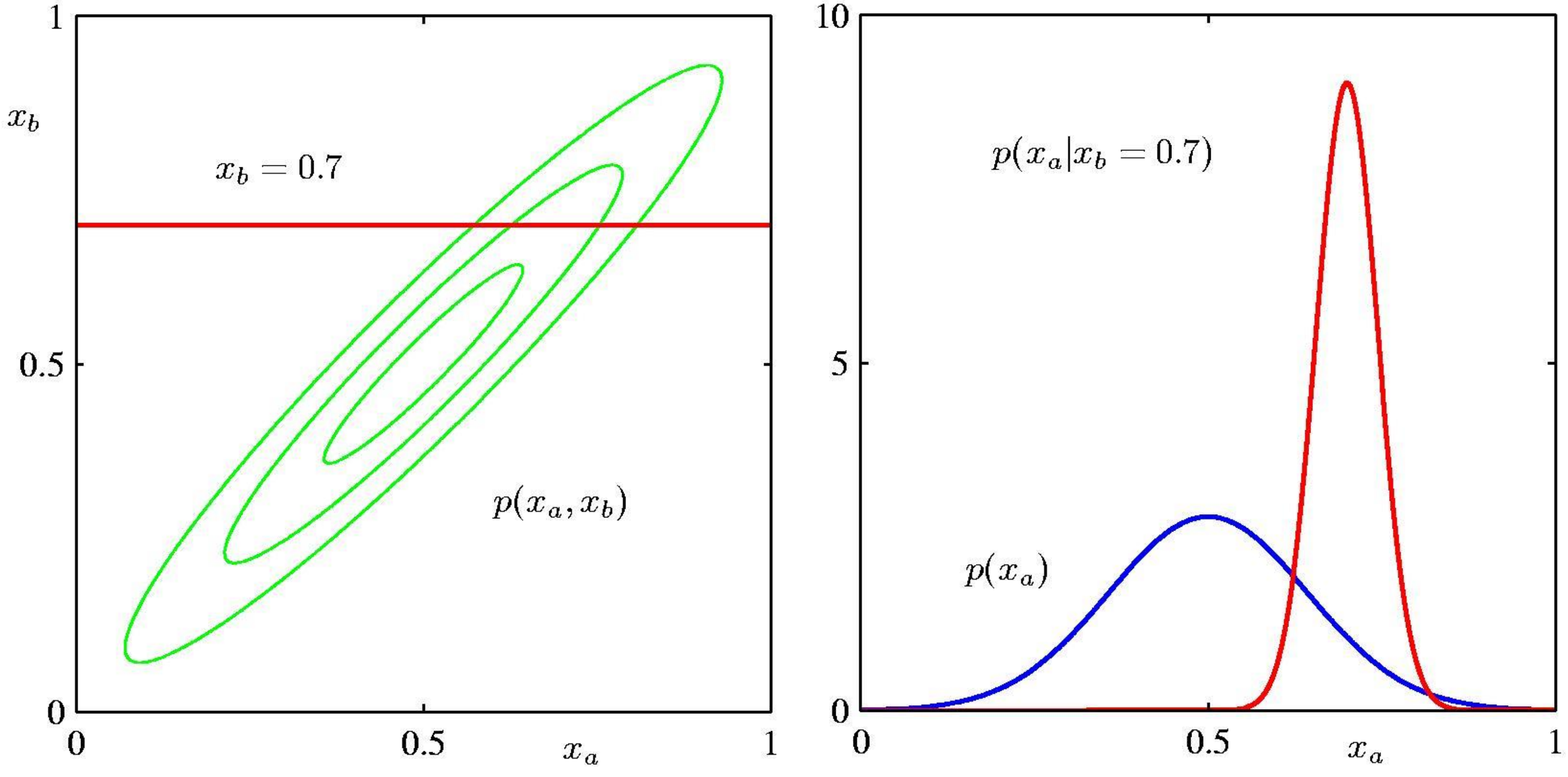
条件分布 $$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} \\
&= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}$$

边际分布

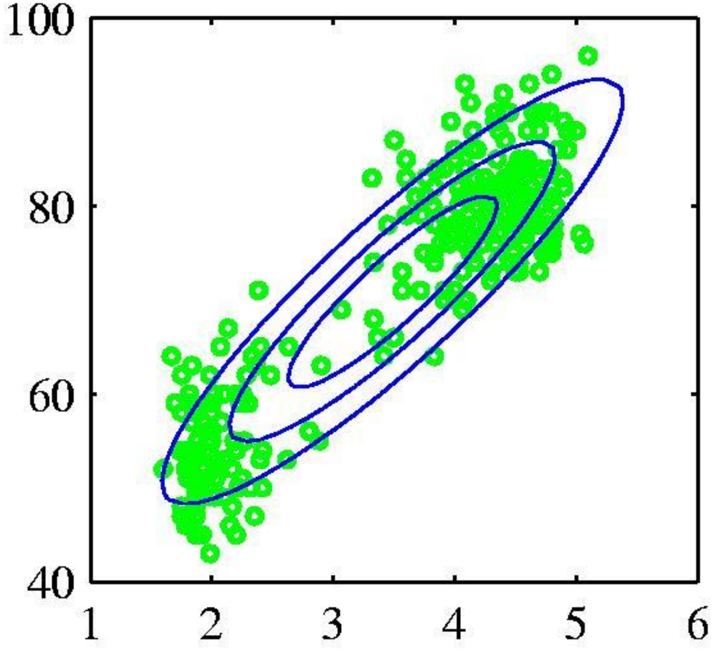
$$\begin{aligned}
p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \\
&= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})
\end{aligned}$$

Partitioned Conditionals and Marginals

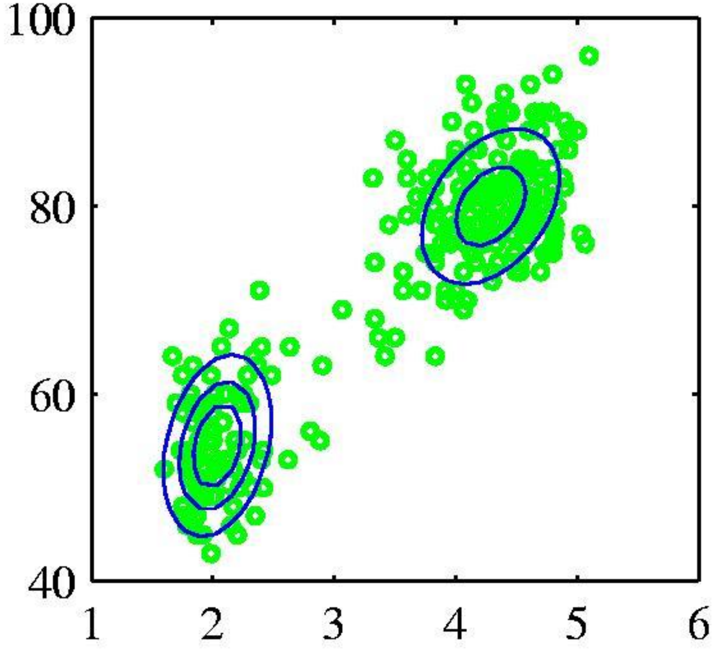


3.4 Mixtures of Gaussians

- Old Faithful data set



Single Gaussian



Mixture of two Gaussians

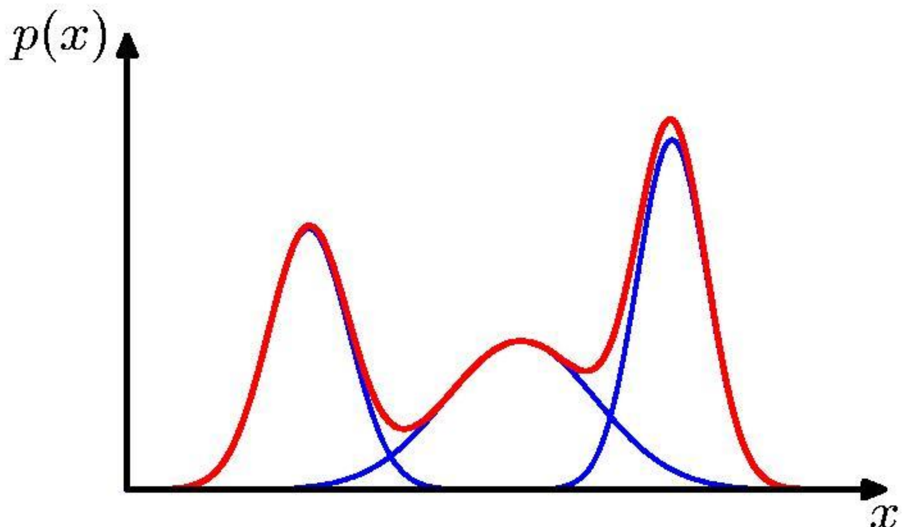
Mixtures of Gaussians (2)

- Combine simple models into a complex model:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{Component}}$$

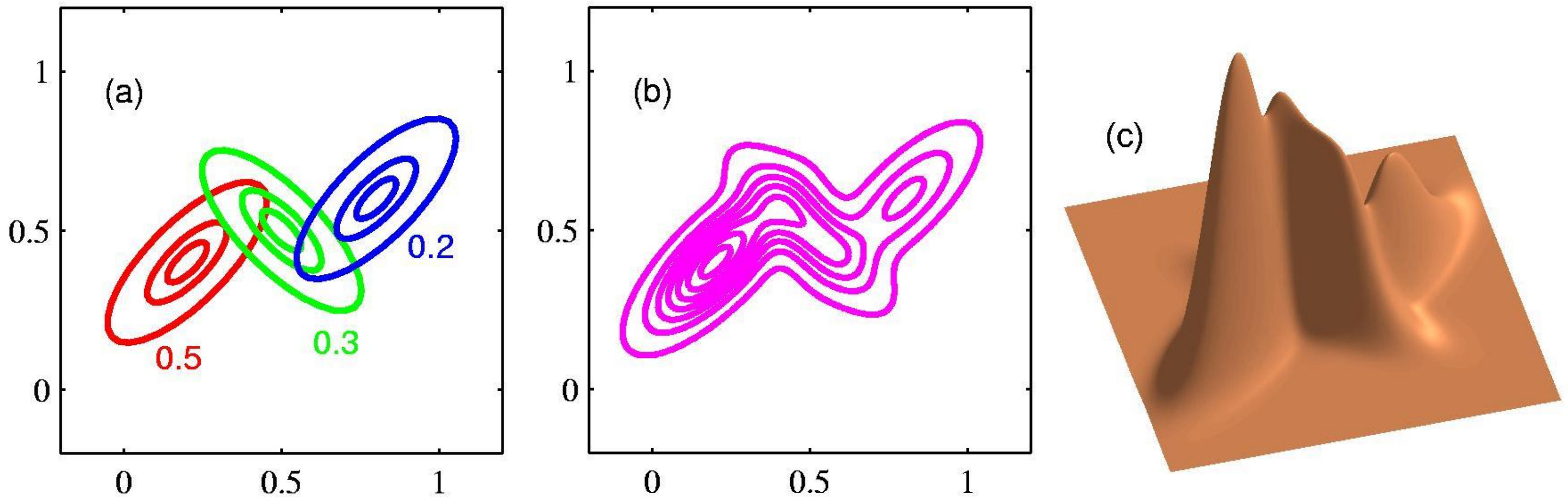
Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



K=3

Mixtures of Gaussians



3.5指数族
The Exponential Family

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

- Where η is the *natural parameter* and

$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x} = 1$$

- so g(η) can be interpreted as a normalization coefficient.

指数族的一般性

- 前面的高斯、二元、二项式、多元、多项式等分布均是指数分布的特例。
- 指数分布的进一步讨论参考Bishop，Chp2.

4. 最大似然准则

似然函数（Likelihood Function）：

概率密度函数 $p(\boldsymbol{x}|\theta)$ 中的 $\boldsymbol{x}$ 固定，由 θ 变化的函数，则称为似然函数，可表示似然函数为

$$L(\theta|\boldsymbol{x}) = p(\boldsymbol{x}|\theta)$$

最大似然准则：（Maximum Likelihood）

$$\hat{\theta} = \arg\max_{\theta\in\Omega} \left\{ L(\theta|\boldsymbol{x}) \right\}$$

$$= \arg\max_{\theta\in\Omega} \left\{ p(\boldsymbol{x}|\theta) \right\}$$

ML续

若存在**IID**样本

最大似然=负对数似然最小化

$$\boldsymbol{X} = \left\{ \boldsymbol{x}_n, \quad n = 1, 2, \cdots, N \right\}$$

似然函数为

$$L\left(\boldsymbol{\theta} \middle| \boldsymbol{X}\right) = \prod_{n=1}^{N} p\left(\boldsymbol{x}_n \middle| \boldsymbol{\theta}\right)$$

对数似然函数为（更常用）

$$l\left(\boldsymbol{\theta} \middle| \boldsymbol{X}\right) = \sum_{n=1}^{N} \log\left(p\left(\boldsymbol{x}_n \middle| \boldsymbol{\theta}\right)\right)$$

参数解为

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta \in \Omega} \left\{ \sum_{n=1}^{N} \log\left(p\left(\boldsymbol{x}_n \middle| \boldsymbol{\theta}\right)\right) \right\}$$

例1. 二元分布参数的最大似然估计

设有IID数据集：$\mathcal{D} = \{x_1, \ldots, x_N\}$

则联合分布

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

对数似然函数

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) =$$

$$= \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

ML解为 $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$

例2. 高斯向量概率密度的参数估计

$$\boldsymbol{x} = [x_1, x_2, \cdots, x_M]^T$$

概率密度函数

$$p_{\boldsymbol{x}}(\boldsymbol{x}|\boldsymbol{\mu}_x, \boldsymbol{C}_{xx}) = \frac{1}{(2\pi)^{M/2}\det^{1/2}(\boldsymbol{C}_{xx})}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_x)^{\mathrm{T}}\boldsymbol{C}_{xx}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_x)\right)$$

样本集 $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$ 是I.I.D的

似然函数为

$$L(\boldsymbol{\mu}_x, \boldsymbol{C}_{xx}|\boldsymbol{X}) = \prod_{n=1}^{N} p_{\boldsymbol{x}}(\boldsymbol{x}_n|\boldsymbol{\mu}_x, \boldsymbol{C}_{xx})$$

对数似然函数为

$$\ln L(\boldsymbol{\mu}_x, \boldsymbol{C}_{xx}|\boldsymbol{X})$$

$$= -\frac{NM}{2}\ln 2\pi - \frac{N}{2}\ln|\boldsymbol{C}_{xx}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_x)^{\mathrm{T}}\boldsymbol{C}_{xx}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_x)$$

例2.（续）

均值向量和自协方差矩阵的MLE

$$\hat{\boldsymbol{\mu}}_x = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$$

$$\hat{\boldsymbol{C}}_{xx} = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_x\right)\left(\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_x\right)^{\mathrm{T}}$$

例3. 模型参数估计的例子

考虑只有两个样本点的简单例子

样本集为I.I.D $\{(x_i, y_i)\}_{i=1}^{2} = \{(2,1),(3,0)\}$

设回归模型

$$\hat{y}(x, w) = w_1 x + w_0$$

设模型与标注存在逼近误差（高斯分布误差）

$$y_i = \hat{y}(x_i, w) + \varepsilon_i = w_1 x_i + w_0 + \varepsilon_i$$

设：$\varepsilon_i \sim N(0, \sigma^2)$ 则有

$$p(y_i|w) = N\left(y_i|\hat{y}(x_i, w), \sigma^2\right)$$

例3（续）由于样本集是I.I.D的，故

$$p(y|w) = \prod_{i=1}^{2} N\left(y_i | \hat{y}(x_i, w), \sigma^2\right)$$

$$= \prod_{i=1}^{2} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \hat{y}(x_i, w)\right)^2\right]$$

可简化为

$$\ln p(y|w) = -\sum_{i=1}^{2}\left(y_i - \hat{y}(x_i, w)\right)^2 + C$$

$$= -\sum_{i=1}^{2}(y_i - w_1 x_i - w_0)^2 + C$$

可解

$$\begin{cases} \dfrac{\partial J(\mathbf{w})}{\partial w_1} = 0 \\ \dfrac{\partial J(\mathbf{w})}{\partial w_0} = 0 \end{cases} \Rightarrow \begin{cases} 13w_1 + 5w_0 = 2 \\ 5w_1 + 2w_0 = 1 \end{cases}$$

解得：$w_1 = -1, w_0 = 3$

模型为：

$$\hat{y}(x, w) = -x + 3$$

5. 贝叶斯（Bayesian）框架

$p_\theta(\theta)$ 待学习参数的先验分布**(Piror distribution)**

由联合分布

$$p(\boldsymbol{x},\boldsymbol{\theta}) = p_\theta(\boldsymbol{\theta})\, p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\, p_x(\boldsymbol{x})$$

得到后验分布**(posterior distribution)**

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) = \frac{p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\, p_\theta(\boldsymbol{\theta})}{p_x(\boldsymbol{x})}$$

其中

$$p_x(\boldsymbol{x}) = \int p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\, p_\theta(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

贝叶斯点估计：MAP估计

$$\hat{\theta} = \arg\max_{\theta \in \Omega} p(\theta \mid \boldsymbol{x})$$

等价形式

$$\hat{\theta} = \arg\max_{\theta \in \Omega} \left\{ p(\boldsymbol{x} \mid \boldsymbol{\theta}) p_{\theta}(\boldsymbol{\theta}) \right\}$$

对数形式

$$\hat{\theta} = \arg\max_{\theta \in \Omega} \left\{ \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) + \log p_{\theta}(\boldsymbol{\theta}) \right\}$$

样本集下的MAP估计

若存在IID样本

思考题：将ML的数值例子推广到MAP

$$X = \left\{ \boldsymbol{x}_n, \quad n = 1, 2, \cdots, N \right\}$$

MAP参数解为

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta \in \Omega} \left\{ \sum_{n=1}^{N} \log\left(p\left(\boldsymbol{x}_n \middle| \boldsymbol{\theta} \right) \right) + \log p_{\theta}(\boldsymbol{\theta}) \right\}$$

小样本集时，先验分布起到较明显作用
但随样本集趋于很大时，先验分布作用减小渐近被忽略

以均方误差作为消耗函数的贝叶斯参数学习

$$\hat{\theta} = E_{\theta|x}(\boldsymbol{\theta} \mid \mathbf{x})$$

后验期望方法

在ML中，可以讨论更一般的全贝叶斯学习框架

例1. 简单说明贝叶斯参数学习

高斯分布：$N\left(\mu,\sigma_x^2\right)$ ，σ_x^2已知

估计：μ，其先验分布为

$$p_\mu(\mu) = \frac{1}{\sqrt{2\pi\sigma_o^2}} e^{\frac{-(\mu-\mu_o)^2}{2\sigma_o^2}}$$

由

$$p(\boldsymbol{x}|\mu) = \frac{1}{\left(2\pi\sigma_x^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma_x^2}\sum_{n=0}^{N-1}(x_n-\mu)^2\right]$$

例1（续），因此

$$p(\boldsymbol{x}|\mu)P_\mu(\mu)=\frac{1}{(2\pi\sigma_x^2)^{\frac{N}{2}}}\frac{1}{(2\pi\sigma_o^2)^{\frac{1}{2}}}\exp\left[-\frac{1}{2\sigma_x^2}\sum_{n=0}^{N-1}(x_n-\mu)^2\right]\exp\left[-\frac{1}{2\sigma_o^2}(\mu-\mu_o)^2\right]$$

上式两边取对数，并求最大值点，得

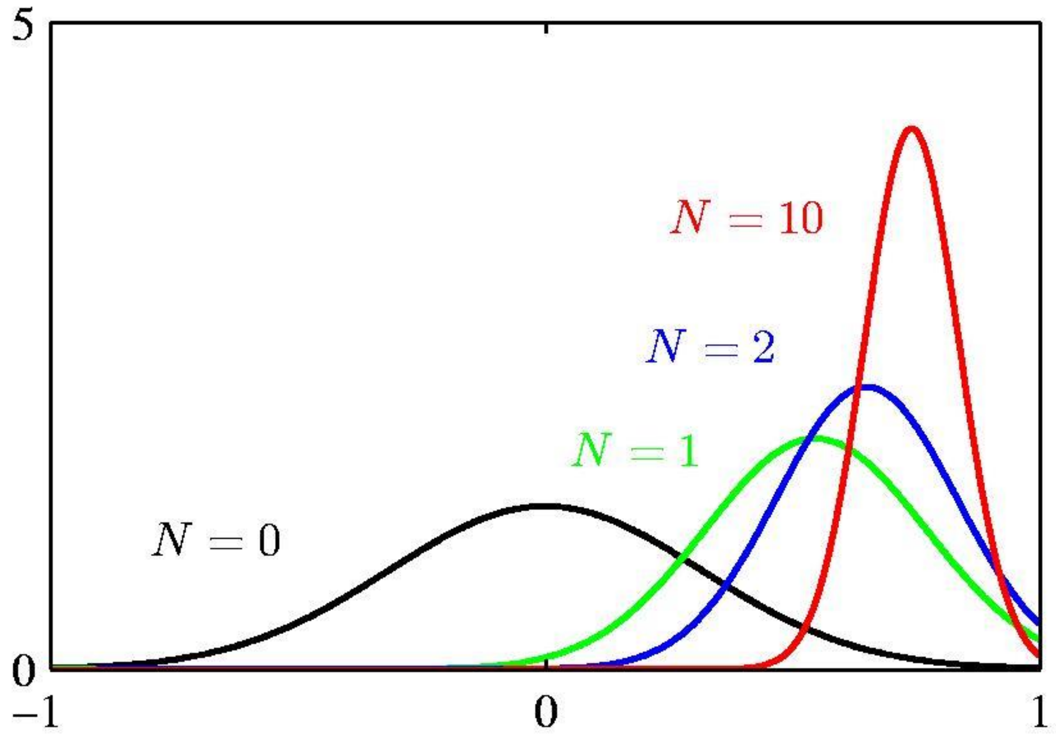
$$\hat{\mu}_{map}=\frac{\sigma_o^2}{\sigma_o^2+\sigma_x^2/N}\frac{1}{N}\sum_{n=1}^{N}x_n+\frac{\sigma_x^2/N}{\sigma_o^2+\sigma_x^2/N}\mu_o$$

当：$N\to\infty$

$$\hat{\mu}_{map}\to\frac{1}{N}\sum_{n=0}^{N-1}x_n$$

例1（续）：另一方面，随N增加后验概率的变化

- Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$ for N = 0, 1, 2 and 10.



高斯分布贝叶斯参数估计：一般向量情况

假设 $p(\mathbf{x}|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \Sigma)$, $\quad p(\boldsymbol{\mu}) = N(\boldsymbol{\mu}_0, \Sigma_0)$

样本集 $\mathbf{D} = \{\mathbf{x}_n\}_{n=1}^{N}$

参数后验概率为： $p(\boldsymbol{\mu}|\mathbf{D}) = N(\boldsymbol{\mu}|\boldsymbol{\mu}_N, \Sigma_N)$

则可证明参数 $\boldsymbol{\mu}$ 贝叶斯估计为：

$$\boldsymbol{\mu}_N = \Sigma_0\left(\Sigma_0 + \frac{1}{N}\Sigma\right)^{-1}\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n + \frac{1}{N}\Sigma\left(\Sigma_0 + \frac{1}{N}\Sigma\right)^{-1}\boldsymbol{\mu}_0$$

和

$$\Sigma_N = \Sigma_0\left(\Sigma_0 + \frac{1}{N}\Sigma\right)^{-1}\frac{1}{N}\Sigma$$

6. 决策论（Decision Theory）

- 一般机器学习中（统计框架下），首先完成模型学习，对于概率模型，给出新输入，需要进行决策
- 推断（Inference step）
  - 确定后验（条件）概率或联合概率 $p(t|\mathbf{x})$、$p(\mathbf{x},t)$
- 决策（Decision step）
  - 给定**x**, 决定最优分类或回归结果 t.

在有监督机器学习中，相对讲推断（Inference）是困难的，决策（decision）是相对简单的工作。
决策是推断完成后的一步工作，学习过程主要完成推断过程

分类的决策

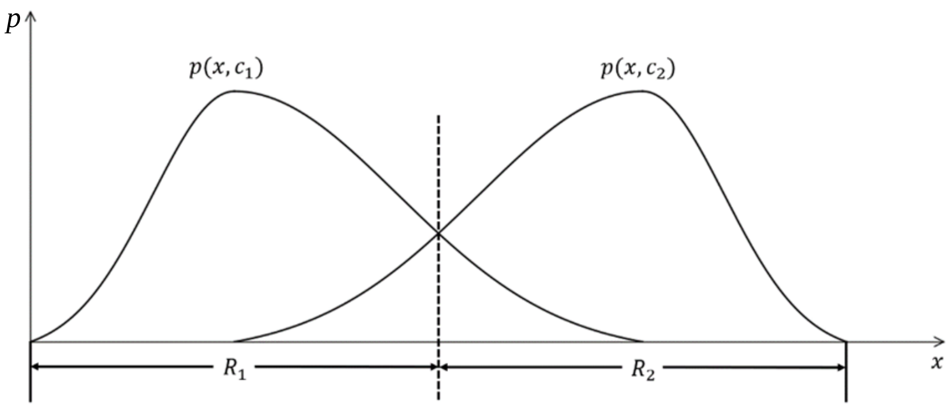
二类情况，最小错分类率准则
Minimum Misclassification Rate

多类情况？

误分类概率

$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}.
\end{aligned}
$$

分类准则：$p(C_1, \boldsymbol{x}) > p(C_2, \boldsymbol{x})$ 分类为 C_1



由 $p(C_1, \boldsymbol{x}) = p(C_1 | \boldsymbol{x})\, p(\boldsymbol{x})$

则

$$p(C_1 | \boldsymbol{x}) > p(C_2 | \boldsymbol{x})$$

Then C_1

Minimum Expected Loss

- Example: classify medical images as ‘cancer’ or ‘normal’

$$
\begin{array}{cc}
 & \text{Decision} \\
 & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\
\text{Truth} \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}
\end{array}
\qquad L_{kj}
$$

若发生错误的代价不同，
可利用加权矩阵对不同错误判决加权

定义：$L_{kj} = L\left(C_j \middle| C_k\right)$ 把 C_k 分类为 C_j 的损失（风险）

Minimum Expected Loss

总的期望损失 $\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k)\,\mathrm{d}\mathbf{x}$

总风险（条件损失）

$$R = \sum_j \sum_k L_{kj} p\left(C_k \middle| \boldsymbol{x}\right)$$

分类为 C_j 的风险

$$R\left(C_j \middle| \boldsymbol{x}\right) = \sum_k L_{kj} p\left(C_k \middle| \boldsymbol{x}\right)$$

为使总风险为最小，对于给定 $\boldsymbol{x}$ ，分类为 C_j 的准则为使

$$R\left(C_j \middle| \boldsymbol{x}\right) = \sum_k L_{kj} p\left(C_k \middle| \boldsymbol{x}\right) \qquad \text{最小}$$

Minimum Expected Loss

二类情况的例子

$$R(C_1|\mathbf{x}) = L_{11}p(C_1|\mathbf{x}) + L_{21}p(C_2|\mathbf{x})$$

$$R(C_2|\mathbf{x}) = L_{12}p(C_1|\mathbf{x}) + L_{22}p(C_2|\mathbf{x})$$

分类为 C_1 的准则是

$$R(C_1|\mathbf{x}) < R(C_2|\mathbf{x})$$

即满足

$$(L_{12} - L_{11})\,p(C_1|\mathbf{x}) > (L_{21} - L_{22})\,p(C_2|\mathbf{x})$$

若取： $L_{12} = L_{21} = 1, L_{22} = L_{11} = 0$

分类准则简化为 $p(C_1|\mathbf{x}) > p(C_2|\mathbf{x})$

Minimum Expected Loss
二类情况的例子（续）

用先验分布和类条件件分布替代后验分布，得到

由： $$\left(L_{12}-L_{11}\right)p\left(C_1\middle|\boldsymbol{x}\right)>\left(L_{21}-L_{22}\right)p\left(C_2\middle|\boldsymbol{x}\right)$$

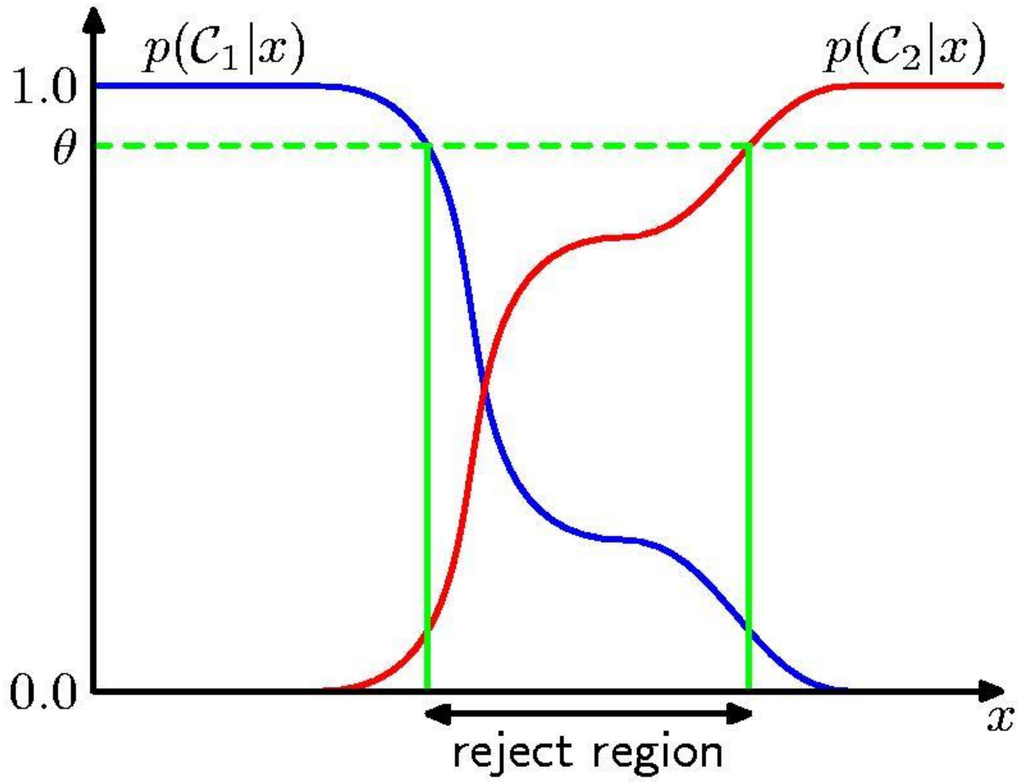
得： $$\left(L_{12}-L_{11}\right)p\left(\boldsymbol{x}\middle|C_1\right)p\left(C_1\right)>\left(L_{21}-L_{22}\right)p\left(\boldsymbol{x}\middle|C_2\right)p\left(C_2\right)$$

分类为 C_1 的准则是

$$\frac{p\left(\boldsymbol{x}\middle|C_1\right)}{p\left(\boldsymbol{x}\middle|C_2\right)}>\frac{\left(L_{21}-L_{22}\right)}{\left(L_{12}-L_{11}\right)}\frac{p\left(C_2\right)}{p\left(C_1\right)}$$

称为似然比准则

拒绝判决选择



在后验概率均小于预定门限时，拒绝判决。

回归的决策

前提：已得到 $p(t|\mathbf{x})$

定义损失函数为误差平方

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

期望损失为

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

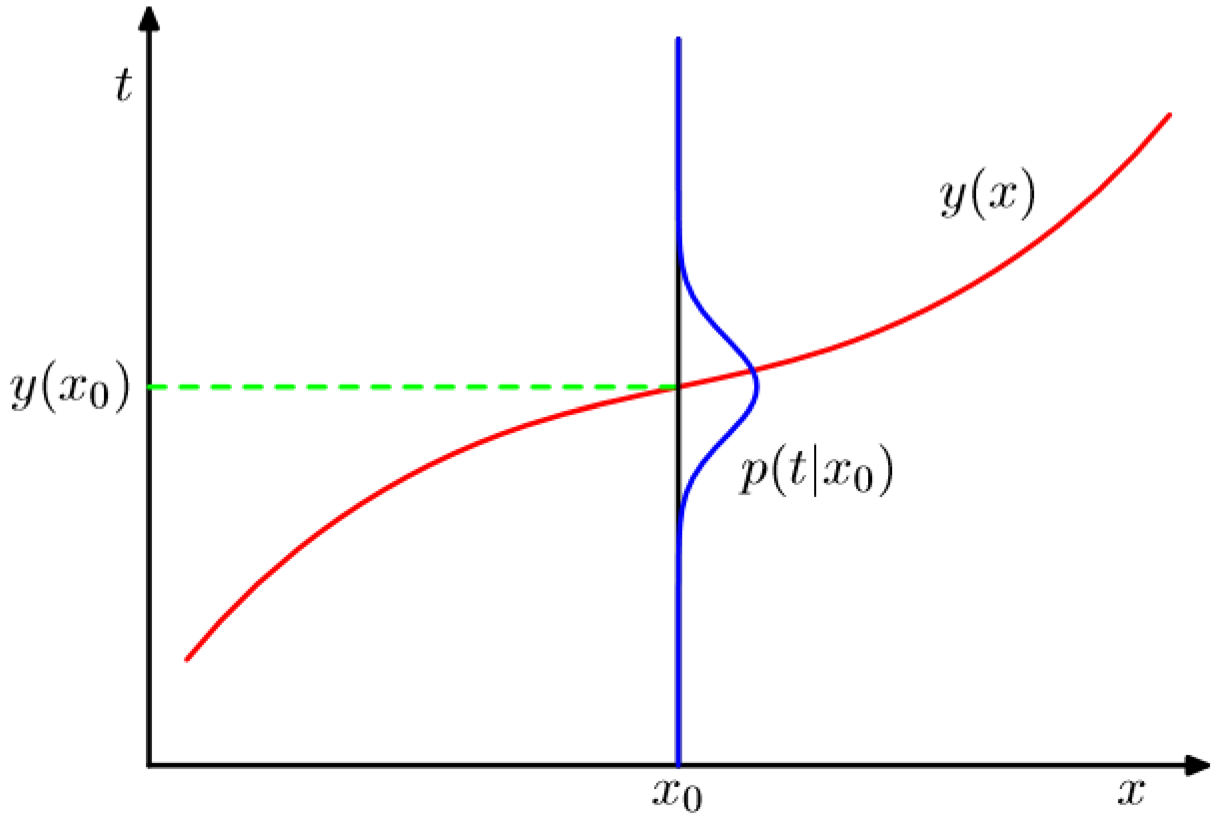
由：

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

得到：

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

回归决策的示意图



如果 $p(t|\mathbf{x})$ 是高斯分布

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

则回归输出：$y(\mathbf{x}) = \mathbb{E}_t[t|\mathbf{x}] = \mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x})$

为什么分开推断与决策过程

- 风险最小化需求 (不同情况下损失加权矩阵可变化)
- 拒绝判决可选择
- 不平衡的类先验概率下的权衡
- 多模型的组合

注：目前多数情况下，推断和决策分为两步，（采用概率统计方法）
也有一些方法是直接导出决策结果的。（采用非概率统计方法）

不同模型比较：生成模型VS鉴别模型（ Generative vs Discriminative ）

- 1.Generative approach:
- Model $p(t, \mathbf{x}) = p(\mathbf{x}|t)p(t)$
- Use Bayes' theorem

直接得到等号左右哪种形式都等价

$$p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$$

- 2.Discriminative approach:
- Model $p(t|\mathbf{x})$ directly
- 3.Discriminative Function

$$t = f(x)$$

<u>比较</u>

1.信息最全，训练复杂性最高；2.适中；3. 简单，丢失概率信息，例如：无法做Reject option和组合模型等

7. 熵（Entropy）

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x)$$

Important quantity in

- coding theory
- statistical physics
- machine learning

Differential Entropy

- Put bins of width Δ along the real line

$$\lim_{\Delta\to 0}\left\{-\sum_i p(x_i)\Delta \ln p(x_i)\right\} = -\int p(x)\ln p(x)\,\mathrm{d}x$$

- Differential entropy maximized (for fixed σ^2) when

$$p(x) = \mathcal{N}(x|\mu,\sigma^2)$$

- in which case

$$\mathrm{H}[x] = \frac{1}{2}\left\{1+\ln(2\pi\sigma^2)\right\}.$$

条件熵（Conditional Entropy）

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}$$

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

离散情况下条件熵的更有直观意义的表示

$$H(Y|X) = -\sum_{x_i}\sum_{y_j} p(x_i, y_j)\log p(y_j|x_i)$$

$$= -\sum_{x_i}\sum_{y_j} p(y_j|x_i)\, p(x_i)\log p(y_j|x_i)$$

$$= -\sum_{x_i} p(x_i)\sum_{y_j} p(y_j|x_i)\log p(y_j|x_i)$$

$$= \sum_{x_i} p(x_i) H(Y|X = x_i)$$

KL散度
The Kullback-Leibler Divergence

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}\end{aligned}$$

$$\mathrm{KL}(p\|q) \geqslant 0 \qquad \mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N}\{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

（1）用q逼近p，当q=p时，KL散度最小
（2）当有N个样本，学习参数时，
最小KL散度等价于最大似然

互信息（ Mutual Information ）

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \end{aligned}$$

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$

互信息的一个作用是判断x和y是否独立?
或逼近独立。

8. 非参数方法
Nonparametric Methods

- 在概率模型估计中，首先假设一种数学形式表示的概率（密度）函数，例如高斯分布、混合高斯分布等，通过样本估计表征该概率函数的参数。但这种预先假设的模型是否成立，在实际中可能无法保证。
- 非参数方法（non-parametric method）没有预先假设，可处理任意概率分布

非参数方法估计概率

- 设观测取自概率分布p(x)并考虑包含x的小区域R，则

$$P = \int_{\mathcal{R}} p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

- 样本数N充分大，若有K个样本落在区域R，则概率P近似为

$$P \approx K/N$$

如果区域R的体积 V充分小，区域内p(x) 近似常数，则概率P

$$P \simeq p(\mathbf{x})V$$

两个P相等，故

$$\boxed{p(\mathbf{x}) = \frac{K}{NV}.}$$

核密度估计

- **核密度估计:** 固定 V, 从数据中估计 K. 设区域 R 是围绕x的超立方体，定义核 (Parzen window)

$$k((\mathbf{x}-\mathbf{x}_n)/h) = \begin{cases} 1, & |(x_i - x_{ni})/h| \leqslant 1/2, \qquad i = 1, \ldots, D, \\ 0, & \text{otherwise.} \end{cases}$$

- K可数为如下，因此PDF估计为

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right) \qquad p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right).$$

核密度估计

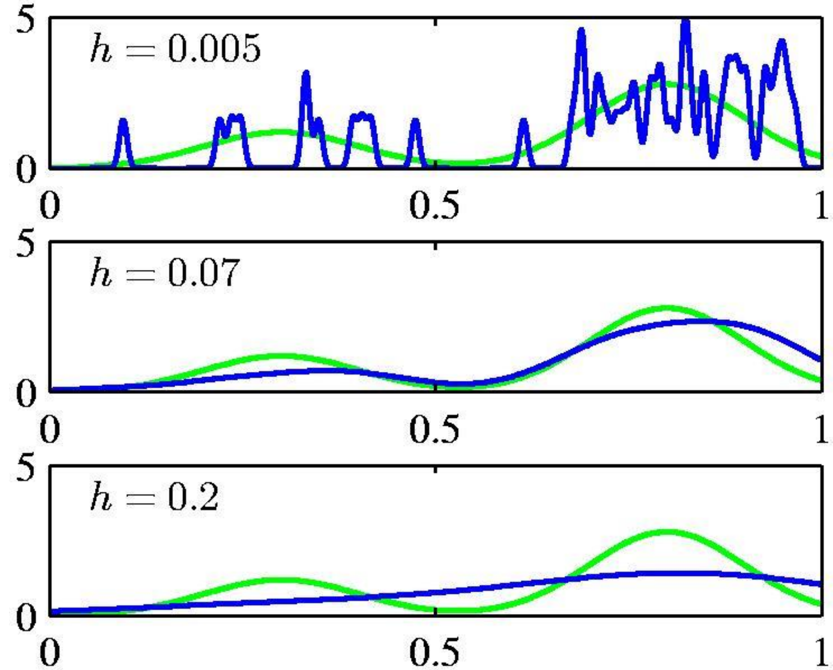
- 为避免不连续，可采用光滑窗函数，例如高斯窗

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{D/2}} \exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right\}$$

- 任意核函数需满足

$$\begin{aligned} k(\mathbf{u}) &\geqslant 0, \\ \int k(\mathbf{u})\,\mathrm{d}\mathbf{u} &= 1 \end{aligned}$$

- 核函数的尺度影响估计性能.



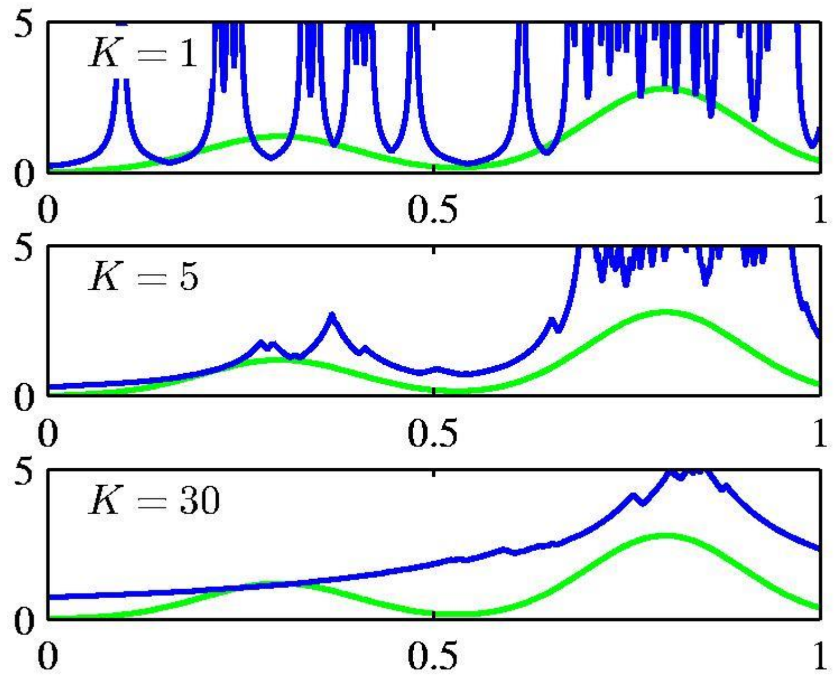
h acts as a smoother.

K近邻密度估计

- 固定K，估计体积V，是的围绕x的超球体包含K个样本。V^* 为包含K个样本的体积，故

$$p(\mathbf{x}) \simeq \frac{K}{NV^\star}.$$

- K近邻不是一个好的密度估计方法，但可以构成学习模型。



K acts as a smoother.

K-近邻分类器

- Given a data set with N_k data points from class C_k and $\sum_k N_k = N$, we have

$$p(\mathbf{x}) = \frac{K}{NV}$$

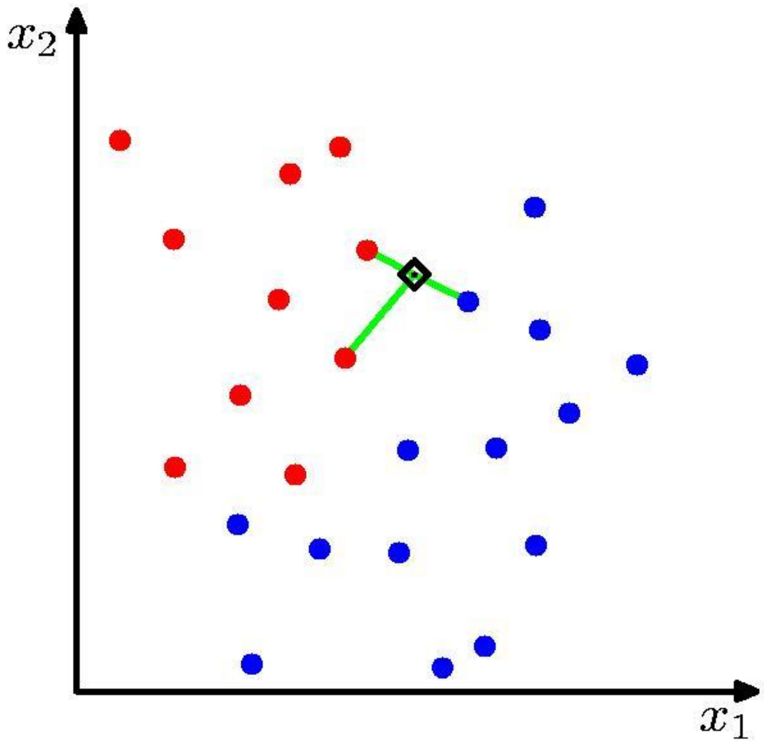
- and correspondingly

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}.$$

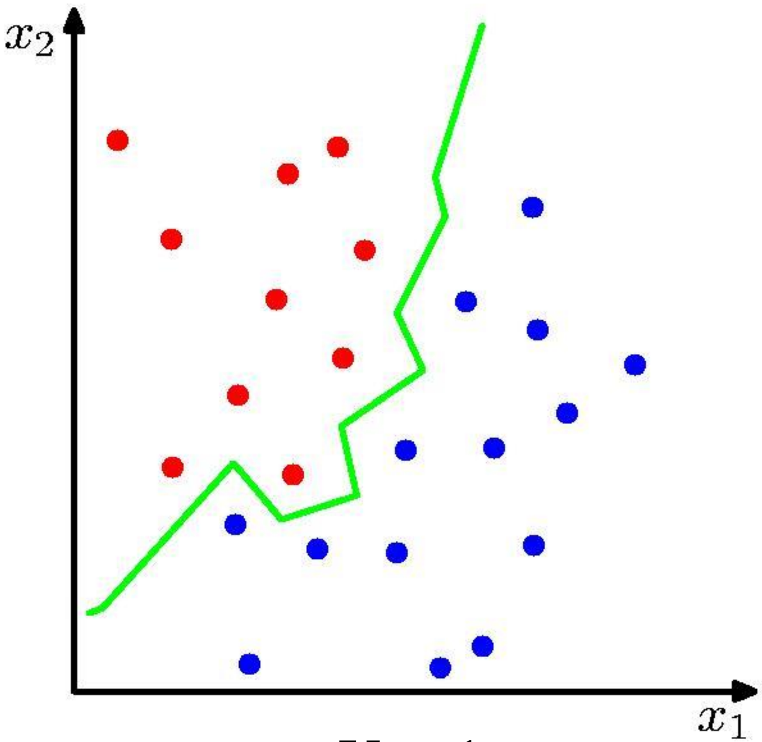
- Since $p(\mathcal{C}_k) = N_k/N$, Bayes' theorem gives

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K}.$$

K-Nearest-Neighbours for Classification



K = 3



K = 1

K近邻回归

样本集 $D=\{(x_n,y_n)\}_{n=1}^{N}$

K近邻样本集合为 $D_K(\boldsymbol{x})$

$$\hat{y}(\boldsymbol{x})=\frac{1}{K}\sum_{(\boldsymbol{x}_i,y_i)\in D_K(\boldsymbol{x})}y_i$$

是对$E(y|\boldsymbol{x})$的近似估计

9. 优化技术概述

最小化的数学形式描述为

$$g(\boldsymbol{w}^*) = \min_{\boldsymbol{w}} \{ g(\boldsymbol{w}) \}$$

定义函数梯度

$$\nabla g(\boldsymbol{w}) = \frac{\partial g(\boldsymbol{w})}{\partial \boldsymbol{w}} = \left[\frac{\partial g(\boldsymbol{w})}{\partial w_1}, \frac{\partial g(\boldsymbol{w})}{\partial w_2}, \cdots, \frac{\partial g(\boldsymbol{w})}{\partial w_M} \right]^{\mathrm{T}}$$

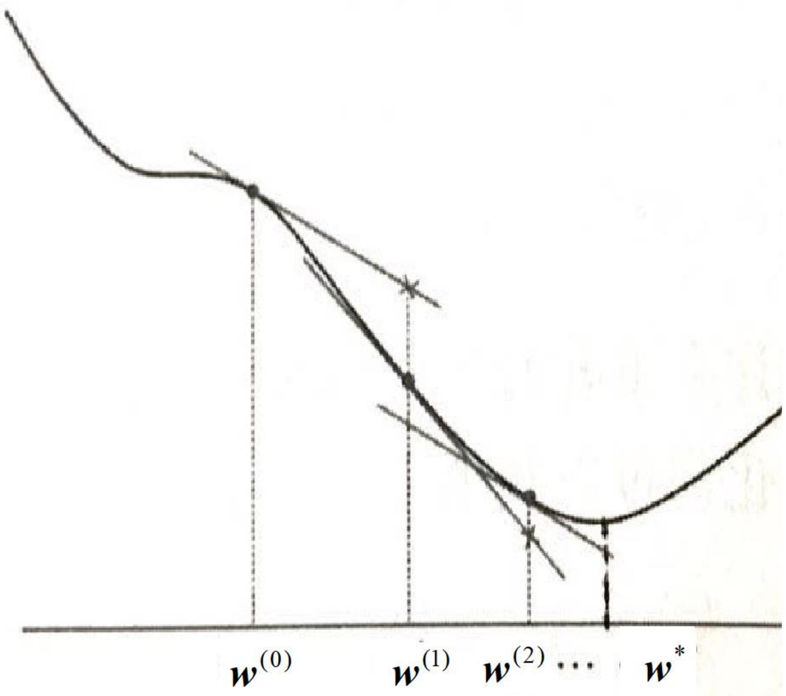
最优解的基本条件：最优点上梯度为0

$$\nabla g(\boldsymbol{w}^*) = \mathbf{0}$$

最基本的迭代算法：梯度下降算法

给出一个初始猜测值 $\mathbf{w}^{(0)}$ 按下式重复迭代直到收敛

$$\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} - \alpha_k \nabla g(\mathbf{w}^{(k-1)})$$



注：最大化问题的梯度解为梯度上升算法

$$\mathbf{w}^{(k)} = \mathbf{w}^{(k-1)} + \alpha_k \nabla g(\mathbf{w}^{(k-1)})$$

α_k 迭代步长参数

调整算法收敛性

凸函数与优化

定义：凸函数

函数：$g(s): \Omega \rightarrow R$　对任意：$\alpha \in [0,1]$满足：

$$g(\alpha s_1 + (1-\alpha)s_2) \leq \alpha g(s_1) + (1-\alpha)g(s_2)$$

定理：　对任意：$s_1, s_2 \in \Omega$满足

$$g(s_2) \geq g(s_1) + [\nabla g(s_1)]^T (s_2 - s_1)$$

或当且仅当Hessian矩阵 $\nabla^2 g(s)$ 是半正定的

则该函数是凸的。

优化问题： 若目标函数是严格凸函数，优化问题可以保证得到全局最小值。

对于非凸的目标函数，优化问题的解要困难得多。