Probability Distribution

王 星

中国人民大学统计学院
E-mail:wangxingwisdom@126.com

October 15, 2020

https://github.com/ctgk/PRML

数据是一种资源，自带不确定性，解读数据需要对不确定性的感知，稳定结构的认知，这就需要对不确定性进行测量和描述，不确定随着数据如何变化，分析影响到不确定的因素，进而提出面向任务解决的算法设计（包括求解策略，计算机求解的方式、路径、质量和效率等）。为促进数据使用，降低不确定性分析不足时的数据滥用，模型误用，生态破坏和价值折损。

首先，需要了解不确定性有几种来源？

数据是一种资源，自带不确定性，对不确定性的解读需要：不确定性感知，稳定结构认知，这需要对不确定性进行测量和描述，分析不确定的影响，进而提出面向任务解决的算法设计（包括求解策略，计算机求解的方式、路径、品质和效率）。为促进数据的使用，抵制不确定性分析不足时的数据滥用，生态破坏，价值折损。

不确定性有几种来源？

可视化可以看到180度的微观数据所勾勒出来的景象。听着语音助手暖暖的声音进入陌生的国度，一点不觉得

1. 被建系统内在随机性：数据带有噪声
2. 有限的或不完整的观测：
   - 2.1 看不见的维度;视角受限
   - 2.2 无法量化的指标;
   - 2.3 数据量不足，极大似然估计的优越性（渐进正态性）无法获得体现
3. 不完全建模：建模中必须舍弃一些信息，舍弃的信息会使得预测出现不确定性；建模的工作流；建模的模型集

基于有限次观测$x_1, \ldots, x_N$ 前提下，对随机变量x 的概率分布$p(x)$ 进行建模，这个问题称为密度估计。严格来说，密度估计的提法并不准确，因为可以产生同一组有限观测数据集的概率分布可以说是不计其数，选择一个合适的分布和模型选择是类似的，这也是模式识别中的一个基本问题。

大纲

- Probability Distribution $p(x)$.二项分布和多项分布
- Parameter Method.特点是整个概率分布只依赖于少量可调节的参数，为了把这种模型应用到密度估计问题中，常常需要一个进程的设计，以便给参数一个比较恰当的值。
  有两种有关估计过程：在频率派给出的是准则最优化的方案，比如似然函数确定参数的具体值。相反，在贝叶斯学派，给定观察数据，引入参数的先验分布，然后用贝叶斯定理来计算对应的后验概率分布，这里，共轭先验（conjugateprior）有着很重要的作用。它保证后验概率分布的函数形式与先验分布一致，极大地简化了贝叶斯分析。
  - Frequentist treatment: MLE
  - Posterior distribution, conjugate priors$P(\theta|x) \propto P(\theta)P(x|\theta)$
    共轭分布例如，多项分布参数的共轭先验被叫做狄利克雷分布（Dirichlet distribution），高斯分布中均值的共轭先验是另一个高斯分布。所有这些分布都是指数分布族（exponential family）的特例。
- NonParametric Method:不假定分布的具体函数形式。

二值变量

- 二值随机变量$x\epsilon\{0,1\}$

$$p(x=1|\mu)=\mu;\ \ E(x)=\mu;\ \ Var(x)=\mu(1-\mu);\ \ 0\leq\mu\leq 1$$

- $Bern(x|\mu)=\mu^x(1-\mu)^{1-x}$;
- 观测集$\mathcal{D}=\{x_1....x_N\}$,如果观测是独立的从$p(x|\mu)$ 抽取，可以构造关于μ 的似然函数如下：

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

- 对数似然函数

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

- $\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n$;$m$ 表示$x=1$ 的观测数量，$\mu_{ML}=\frac{m}{N}$;
- 如果规模集给定，$p(m|N,\mu)=C_N^m\mu^m(1-\mu)^{N-m}$;

$$E(m)=N\mu;\ \ Var(m)=N\mu(1-\mu)$$

Beta分布,Beta-Binomial 共轭

- $Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}, 0 \leq \mu \leq 1;$

$$E(\mu) = \frac{a}{a+b}; \quad Var(\mu) = \frac{ab}{(a+b)^2(a+b+1)}.$$

- 将似然函数和先验连起来

$$p(\mu|m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1};$$

- 预测分布

$$\begin{aligned} p(x=1|\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu \\ &= \int_0^1 \mu p(\mu|\mathcal{D})d\mu = E(\mu|\mathcal{D}) = \frac{m+a}{m+a+l+b}; \end{aligned}$$

- 注意到θ的后验期望在整个数据集上做平均就是θ的先验期望：

$$\mathbb{E}_\theta[\theta] = \mathbb{E}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]]$$

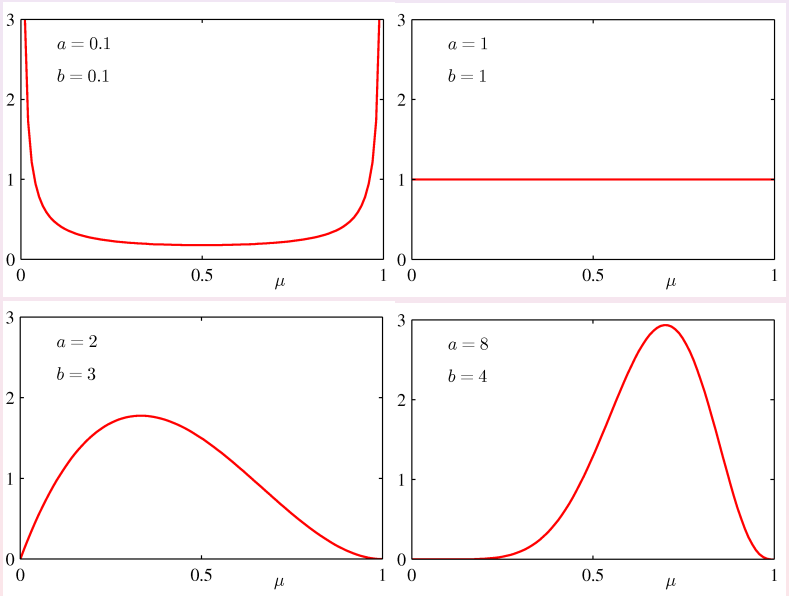
- 同理：

$$\text{var}_\theta[\theta] = \mathbb{E}_\mathcal{D}[\text{var}_\theta[\theta|\mathcal{D}]] + \text{var}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]]$$

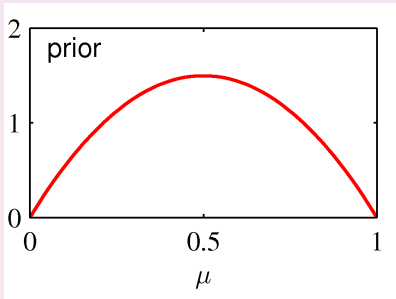
- 第一项是θ的平均后验方差，第二项是θ的后验均值的方差。平均来看，θ的后验方差小于先验方差。后验均值的方差越大，这个方差的减小就越大，需要注意的地方，这个结果只在平均的情况下成立，对于一个特定的观测数据集，有可能后验方差小于先验方差。

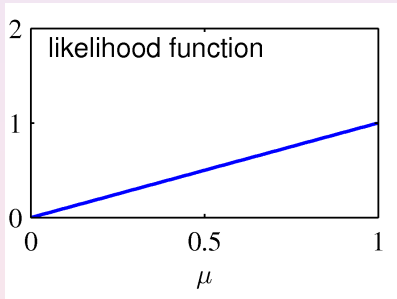
Beta分布不同超参数的函数图像

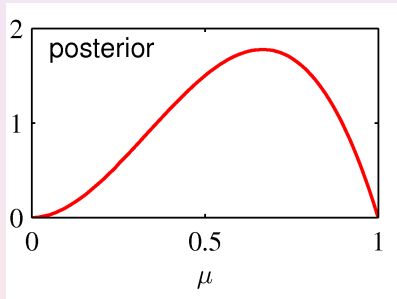
- a和b不一定是整数；样本如何影响到参数估计？
- Example: Prediction: all future tosses will land heads up， 1/1，Overfitting to D
- 怎样控制参数才可以得到更好的预测？
- 随着观测到的数据量越多，m和l会增大，而后验分布的峰变尖了，这样不确定性会下降



- 这里得到的启发是：可以设计一种序列估计过程，这种方法不依赖于先验和似然的选择，只要数据独立性的假设，每次使用一小部分观测数据产生一个后验估计，然后再使用下一组观测值的时候，丢掉前一组继续估计，如此迭代，这一估计方法可以应用于数据流问题，每次不需要将所有的数据一次性地导进数据，节省内存。







多项分布

多项分布是二项分布扩展到多维的情况。多项分布是指单次试验中的随机变量的取值不再是0 和1，而是有多种离散值可能$\{1, 2, ..., k\}$。比如投掷6个面的骰子实验，N 次实验结果服从$K = 6$ 的多项分布。其中：

- K-dimensional vector $x = \underbrace{(0, 1, 0, 0, ..., 0)^T}, \sum_{k=1}^{K} x_k = 1;$
- Probability Distribution:
  $p(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}, \mu_k = p(x_k = 1),$
  $\mu = (\mu_1, ..., \mu_K)^T, \ \mu_k \geq 0, \sum \mu_k = 1;$
- Likelihood Function
  $p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\sum_n x_{nk}} = \prod_{k=1}^{K} \mu_k^{m_k}$
- MLE $\mu_k^{ML} = m_k/N; m_k$是N次观测中$x_k = 1$ 的观测次数。

Dirichlet Distribution, Dirichlet-Multinomial 共轭

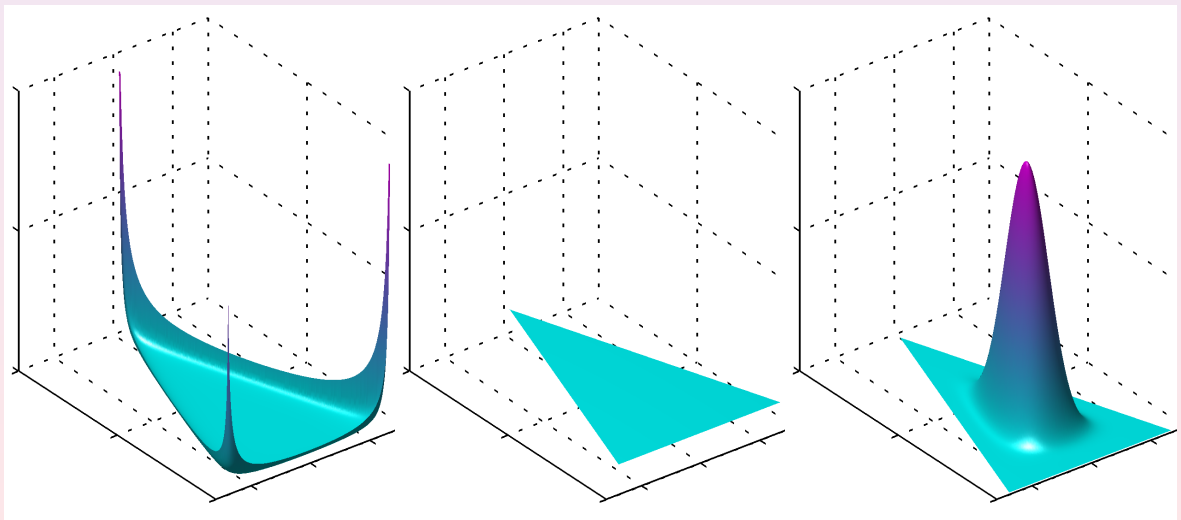
Dirichlet分布是beta分布在高维度上的推广。Dirichlet 分布的密度函数形式跟beta 分布的密度函数类似：$p(\mathcal{D}|\mu) = \prod_{k=1}^{K} \mu_k^{m_k}$

- Conjugate prior:
  $p(\mu|\alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}; \alpha_0 = \sum_{k=1}^{K} \alpha_k$
- Normalized Form
  $Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1),\ldots,\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1};$
- Posterior
  $p(\mu|\mathcal{D},\alpha) \propto p(\mathcal{D}|\mu)p(\mu|\alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1};$
  $p(\mu|\mathcal{D},\alpha) = Dir(\mu|\alpha + m) =$
  $\frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1),\ldots,\Gamma(\alpha_k + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}. m = (m_1, \ldots, m_K)^T, \alpha_k$可以理解为Dirichlet 分布$x_k = 1$ 的有效观测数。

极大似然估计

- $$\ln p(\mathcal{D}|\mu) = \sum_{k=1}^{K} m_k \ln \mu_k$$
- $\mu_k = \frac{m_k}{N}$

三个变量的Dirichilet分布图像，左图$\alpha_k = 0.1$, 中图$\alpha_k = 1$, 右图$\alpha_k = 10$,



Gaussian Distribution

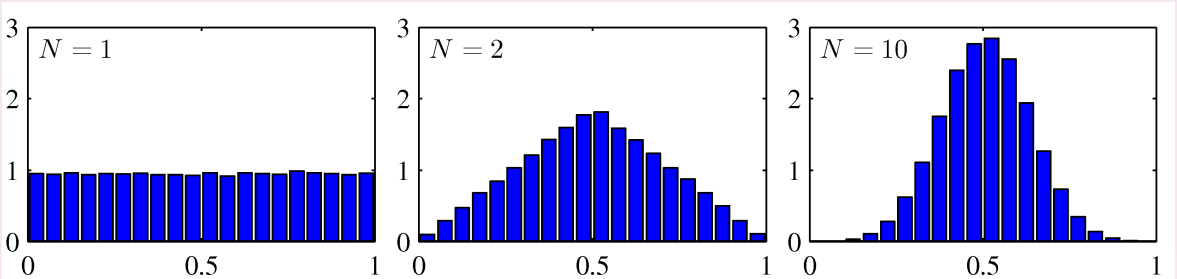
- 2.3.1 ML for Gaussian
- 2.3.2 Bayesian inference for Gaussian
- 2.3.3 Sequential estimation
- 2.3.4 * Mixture of Gaussian
- 2.3.5 * Periodic variables

Gaussian Distribution

- Widely used model for the distribution of continuous variables:

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

- The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows. Example: N uniform $[0, 1]$ random variables.



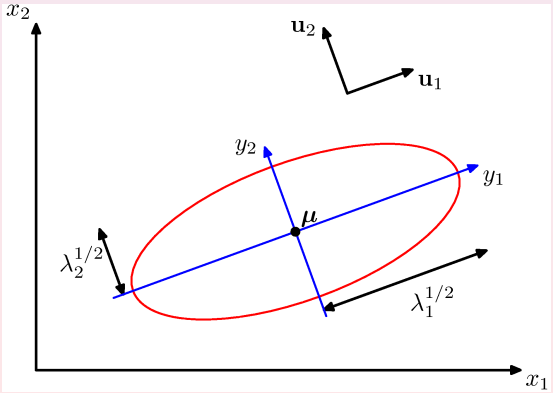
高斯分布的几何形式

- 高斯对x的依赖是通过二次型表达出来的

$$\Delta^2 = (x-\mu)^T\Sigma^{-1}(x-\mu); \quad \Sigma = \sum_{i=1}^{D}\lambda_i u_i u_i^T;$$

Δ称为μ和x之间的马氏距离（Mahalanobis Distance），当Σ是单位矩阵的时候，就是欧式距离。

- 协方差矩阵特征向量方程$\Sigma u_i = \lambda_i u_i; i = 1, ..., D, u_i^T u_j = I_{ij}, I_{ij} = 1, i = j, I_{ij} = 0, others$



- 协方差矩阵可以表示成特征向量的展开形式：

$$\Sigma = \sum_{i=1}^{D} \lambda_i u_i u_i^T;$$

- 类似地

$$\Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} u_i u_i^T$$

- 二次型的另外一种形式

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}; y_i = u_i^T (x - \mu)$$

- $\{y_i\}$表示成单位正交向量u_i关于原始的x_i 坐标经过平移和旋转后形成的新坐标系$\mathbf{y} = \mathbf{U}(\mathbf{x} - \mu)$.
- U是一个矩阵，它的行是向量u_i^T,它满足$UU^T = I$

- 由协方差矩阵的特征向量定义的在由y_i定义的新坐标下高斯分布的形式是：

$$p(y) = p(x)|J| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\{-\frac{y_j^2}{2\lambda_j}\}$$

  在这里联合分布可以表达成为独立分布的乘积
- 高斯分布的特征是：

$$\mathbb{E}x = \Sigma, \quad \text{var}x = \Sigma$$

- 高斯分布的优点:一个对称的协方差矩阵有$\frac{D(D+1)}{2}$个独立参数，适应性很强，μ有D个独立参数，总计有$\frac{D(D+3)}{2}$个参数.
- 高斯分布的局限：对于比较大的D参数的总数以D^2的速度增长，这会导致求逆无法计算，可以通过限定只在对角矩阵上进行计算，但是这样将丧失了对相关性进行分析的能力。另一个局限是：单峰性，不能很好地表示多峰分布。

条件高斯分布

高斯分布的性质：两组变量的联合分布是高斯分布，那么以一组变量为条件，另一组变量同样是高斯分布。假设$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$

- 相应的均值$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$
- 协方差矩阵$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$
- 协方差矩阵的逆，记为$\Lambda = \Sigma^{-1}$, 称为精度矩阵(precision matrix)；精度矩阵$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$
- 协方差矩阵是对称矩阵，$\Sigma_{ab}^T = \Sigma_{ba}$; 精度矩阵也是对称矩阵，$\Lambda_{ab}^T = \Lambda_{ba}$;

条件概率分布

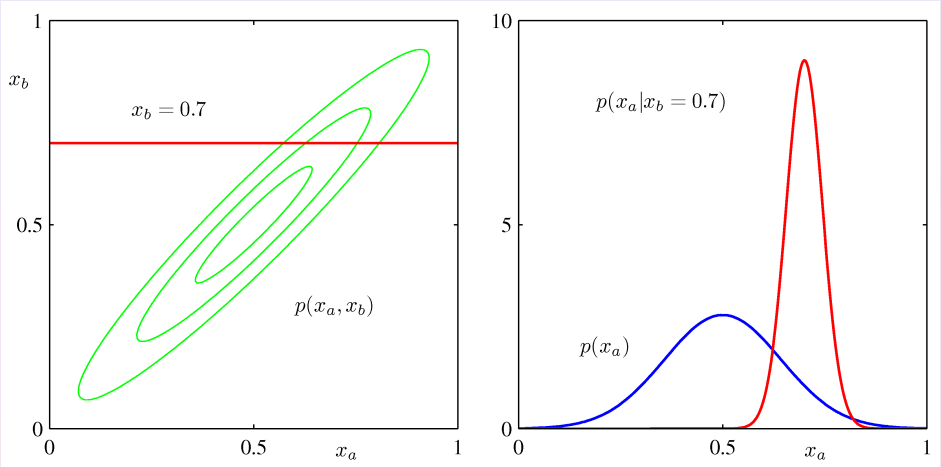
$p(x_a|x_b) = \mathcal{N}(x_a|\mu_{a|b}, \Lambda_{aa}^{-1}), \mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$;

用分块协方差矩阵表达的条件高斯分布的期望和协方差矩阵

- In matrix form, we have that
$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}$$
- 这里$M = (A - BD^{-1}C)^{-1}$.
- $\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$ 于是
$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1};$$
$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1};$$
- 条件分布$p(x_a|x_b)$的均值和协方差如下：
$$\mu_{a|b} = \mu_a - \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b);$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba};$$

  条件分布用分块精度矩阵表达比用分块协方差矩阵有更简洁的表达形式。

Gaussian Distribution边缘分布



边缘概率分布

$$p(x_a) = \mathcal{N}(x_a|\mu_a, \Sigma_{aa});$$

Gaussian Distribution高斯变量的贝叶斯定理

令x的边缘分布和条件分布的形式如下：

$$p(x) = N(x|\mu, \Lambda^{-1});$$

$$p(y|x) = N(y|Ax + b, L^{-1})$$

μ,A, b是控制均值的参数，Λ和L 是精度矩阵，如果x 的维度是D, y 的维度是M, 矩阵A 的大小为$D \times M$.

- $E[y] = A\mu + b; Cov[y] = L^{-1} + A\Lambda^{-1}A^T$
- y的边缘分布是

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

- x在给定y的条件下的条件分布是

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

其中$\Sigma = (\Lambda + A^T LA)^{-1}$.

ML for Gaussian Distribution

- Log Likelihood function

$$\ln p(x|\mu.\Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$$

- MLE

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

$$\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})(x_n-\mu_{ML})^T;$$

- Expectation

$$E[\mu_{ML}] = \mu; \quad E[\Sigma_{ML}] = \frac{N-1}{N}\Sigma.$$

Sequential Estimation

- For ML

$$\begin{aligned}\mu_{ML} =& \frac{1}{N}\sum_{n=1}^{N} x_n; \\ =& \frac{1}{N}x_N + \frac{1}{N}\sum_{n=1}^{N-1} x_n; \\ =& \frac{1}{N}x_N + \frac{N-1}{N}\mu_{ML}^{(N-1)} \\ =& \mu_{ML}^{(N-1)} + \frac{1}{N}(x_N - \mu_{ML}^{(N-1)})\end{aligned}$$

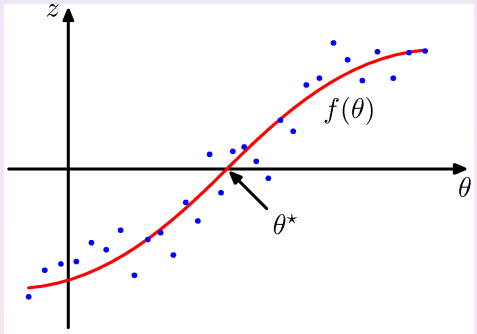
- For Bayes

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

序列估计Robins-Monro Algorithm

Consider μ and z governed by $p(z, \mu)$,当θ 给定时，z 的条件期望定义了一个回归函数

$$f(\theta) = E(z|\theta) = \int zp(z|\theta)dz$$



目标是寻找θ^* 使得$f(\theta^*) = 0$.

- Successive estimates of θ^* are then given by

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1}z(\theta^{(N-1)})$$

a_N满足收敛条件：$\lim_{N\to\infty} a_N = 0, \sum_{N=1}^{\infty} a_N = \infty, \sum_{N=1}^{\infty} a_N^2 < \infty.$

极大似然估计的Robins-Monro Algorithm

- 对数似然函数

$$\frac{\partial}{\partial\theta}\{\frac{1}{N}\sum_{n=1}^{N}-\ln p(x_n|\theta)\}|_{\theta_{ML}}=0$$

- 交换导数与求和，取极限$N\to\infty$

$$-\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_n|\theta)=E_x[-\frac{\partial}{\partial\theta}\ln p(x|\theta)]$$

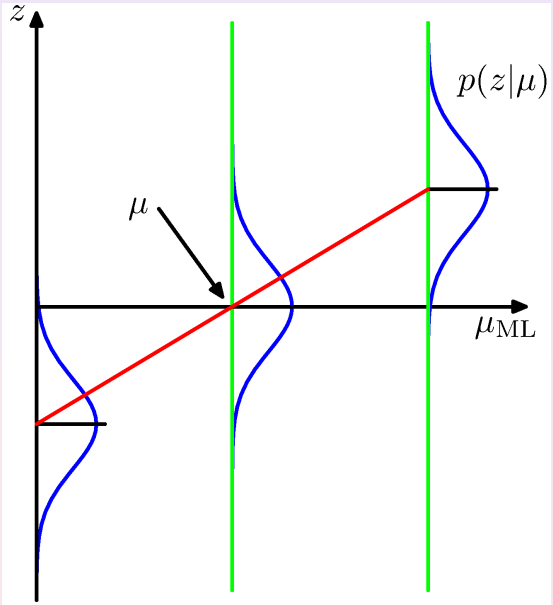
- Robbins-Monro法：

$$\theta^{(N)}=\theta^{(N-1)}-\alpha_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}[-\ln p(x_N|\theta^{(N-1)})]$$

正态分布均值的序列估计算法

- 参数$\theta^{(N)}$是正态分布均值$\mu_{ML}^{(N)}$的估计，随机变量z的形式为

$$z = -\frac{\partial}{\partial \mu_{ML}} \ln p(x|\mu_{ML}, \sigma^2) = -\frac{1}{\sigma^2}(x - \mu_{ML})$$



z的分布对应于对数似然函数的导数，由$-(x - \mu_{ML}/\sigma^2)$给出，定义回归函数的期望是一条直线，由$-(\mu - \mu_{ML})/\sigma^2$给出，真值是$\mu, a_N = \frac{\sigma^2}{N}$.

Bayes Inference for Gaussian▶ Unknown mean known Variance

- Known variance, unknown mean

$$p(x|\mu) = -\prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{2\pi\sigma^2}\exp\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\}$$

- Prior

$$p(\mu) = N(\mu|\mu_0, \sigma_0^2);$$

$$p(\mu|x) \propto p(x|\mu)p(\mu)$$

- Posterior

$$p(\mu|x) = N(\mu|\mu_N, \sigma_N^2)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML} \qquad (2.141)$$

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n; \quad \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

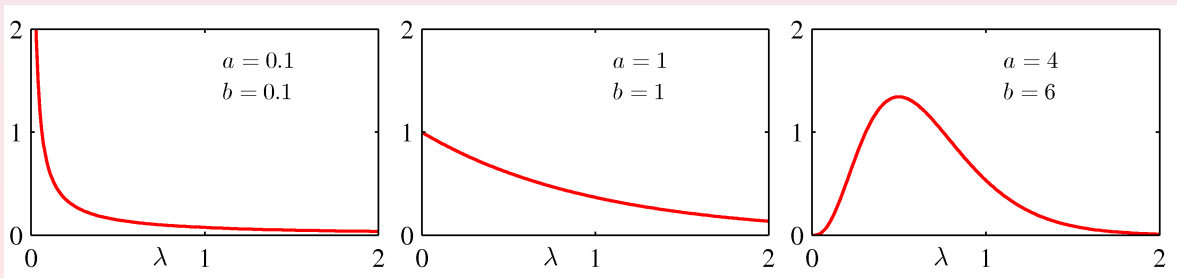
- Likelihood

$$p(x|\lambda) = -\prod_{n=1}^{N} N(x_n|\mu, \lambda^{-1})$$

$$\propto \lambda^{N/2} \exp\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\};$$

- Prior

$$Gam(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda); \quad \lambda = \frac{1}{\sigma^2} \qquad (2.145)$$



- 考虑一个先验$Gam(\lambda|a_0,b_0)$根据公式(2.145)给出的似然函数,有
- Posterior

$$p(\lambda|x) \propto \lambda^{a_0-1}\lambda^{N/2}\exp\{-b_0\lambda - \frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\};$$

$a_N = a_0 + \frac{N}{2}$;
$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma_{ML}^2$ (2.151)
其中σ_{ML}^2是方差的极大似然估计。

- 从这里可以看出，观察数据点的效果是把先验分布中a_0 这个值提高了$\frac{N}{2}$, 如果将先验分布中的参数a_0 看成是$2a_0$ 个先验“高效”观测，那么后验分布则是合成两种情景下的观测数；同样的道理，根据公式(2.151)N个数据点对参数b 贡献了$\frac{N\sigma_{ML}^2}{2}$, 先验分布看成方差为$\frac{b_0}{a_0}$ 个高效先验观测

- Likelihood

$$p(x|\mu,\lambda)=\prod_{n=1}^{N}(\frac{\lambda}{2\pi})^{1/2}\exp\{-\frac{\lambda}{2}(x_n-\mu)^2\};$$

$$\propto[\lambda^{1/2}\exp(-\frac{\lambda\mu^2}{2})]^N\exp\{\lambda\mu\sum_{n=1}^{N}x_n-\frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\}$$

- Prior

$$p(\mu,\lambda)\propto[\lambda^{1/2}\exp(-\frac{\lambda\mu^2}{2})]^\beta\exp\{c\lambda\mu-d\lambda\}$$

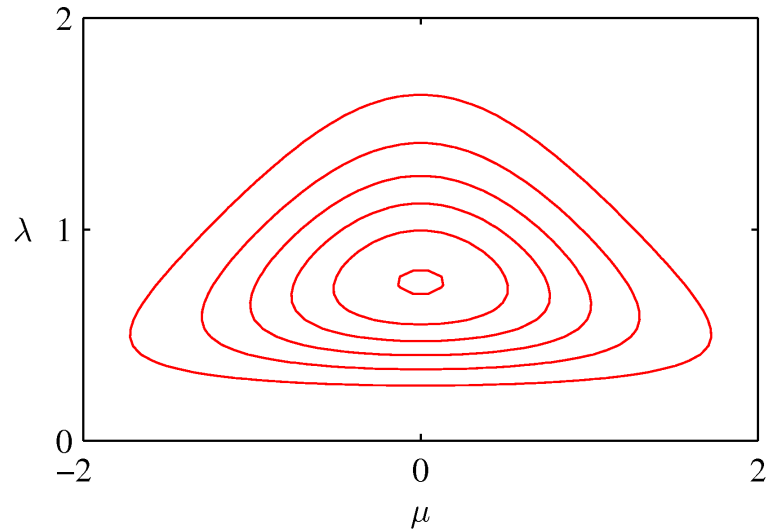
$$=\exp\{-\frac{\beta\lambda}{2}(\mu-c/\beta)^2\}\lambda^{\beta/2}\exp\{-(d-\frac{c^2}{2\beta})\lambda\}$$

其中c,d和β都是常数

- Normal-gamma or Gaussian-gamma

$$p(\mu,\lambda)=N(\mu|\mu_0,(\beta\lambda)^{-1})Gam(\lambda|a,b)$$

正态分布-Gamma分布在参数$\mu_0 = 0, \beta = 2, a = 5, b = 6$条件下的分布轮廓线



Multivariate conjugate priors

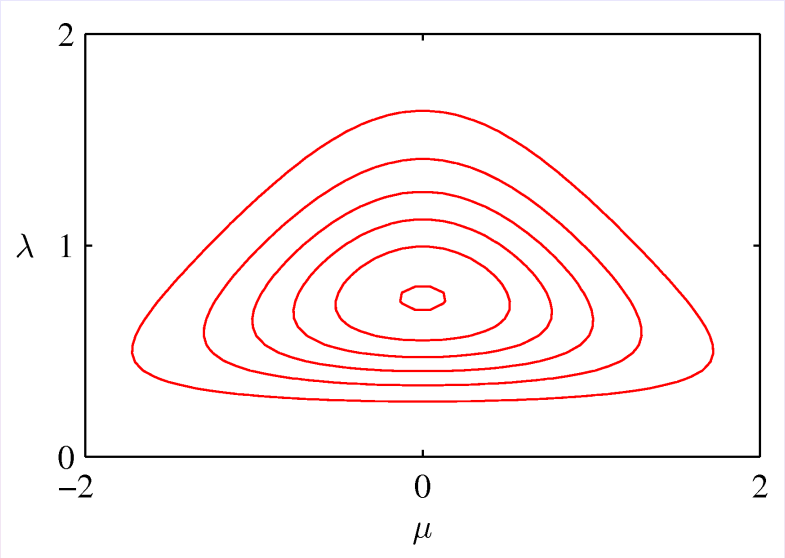
- μ未知，精度Λ已知，$p(\mu)$高斯；
- μ已知，精度Λ未知，$p(\Lambda)$ 共轭先验是Wishart 分布

$$\mathcal{W}(\Lambda|W,\nu) = B|\Lambda|^{\frac{(\nu-D-1)}{2}}\exp(-\frac{1}{2}\mathrm{Tr}(W^{-1}\Lambda)) \qquad (2.155)$$

Tr(.)表示矩阵的迹。B是归一化系数。

- Λ和μ未知，$p(\mu,\Lambda)$Gaussian-Wishart,

$$p(\mu,\Lambda|\mu_0,\beta,W,\nu) = N(\mu|\mu_0,(\beta\Lambda)^{-1})\mathcal{W}(\Lambda|W,\nu)$$



- The decision problem:
  - given x, predict t according to a probabilistic model $p(x,t)$
- binary classification:$t \in \{0,1\} \leftrightarrow \{C_1, C_2\}$;
- Important quantity: $p(C_k|x)$

$$p(C_k|x) = \frac{p(x,C_k)}{p(x)} = \frac{p(x,C_k)}{\sum_{k=1}^{2} p(x,C_k)}$$

$\rightarrow$ getting $p(x,C_i)$ is the (central!) inference problem

$p(x|C_k)p(C_k)$ $\propto$ likelihood $\times$ prior

决策论

- 决策是对不确定的方案做决定，它的基本特点是通过计算决策的后果来影响选择。（包括状态的选择、参数的选择到模型的选择）；
- **参数集：**参数的所有自然可能的不同的状态$\theta = \{\theta\}$，用于表示影响到决策的控制因素，比如出门带口罩的影响因素是是否收到空气质量警报短信，一个小区的照明用电的大小取决于这个小区的入住率，常驻率等因素的影响。
- **决策集：**所有可能的决策结果$\Delta = \{\delta\}$;
- **损失函数：**对于一个决策行动δ和参数θ, 评价不同决策优劣的函数$l(\theta, \delta)$.

决策方法

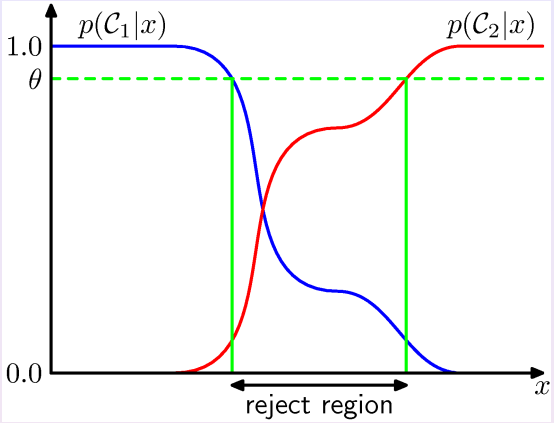
有很多不同的决策方式，通常而言有：

- 极小极大决策方法，在每一种参数状态下，取损失最大的决策，再从已选出的决策里选择损失最小的决策：特点是这是一种相对保守的选择，而且假设每种参数发生的可能性是相等的；
- 期望损失决策法，当参数出现的可能性不同时，体现了一个好的决策应该在参数空间上一致地好的基本原则；$R = \int_\theta l(\delta, \theta) dF(\theta)$;
- 例题（练习题1）：假设一种新产品即将投放市场，有两种方案，自己生产或转让专利，损失根据销售前景有所不同。

可能的市场前景	自己生产δ_1	转让专利δ_2
销售好θ_1	-80	-40
销售一般θ_2	-20	-10
销售较差θ_3	10	-5

- 极小极大原则下，最好的决策是转让专利δ_2。
- 最小期望损失下的决策是$\pi(\theta_1) = 0.2, \pi(\theta_2) = 0.5, \pi(\theta_3) = 0.3$, 决策的期望损失$R(\delta_j) = \sum_{i=1}^{3} \pi(\theta_i) l(\theta_i, \delta_j), j = 1, 2, R(\delta_1) < R(\delta_1)$.

后验分布的拒绝决策功能



在实际应用中，后验分布除了选出最优决策以外，还有一种功能是给出不宜决策的区域，如图1.26所示，用于决策的好的后验分布就像对两组数据作了很好的凝练，这样选定一个域值就可以比较清晰的将两组数据分开；但是如果后验分布有重叠，那么通过后验分布的比较，可以给出分辨两类最优的域值所在，当设定域值过小，则对应着后验分辨失效的区域，从而可以根据失效区域和域值的高低来决定后验分布对分类的有效性。从以上分析来看，一个决策方案的4项步骤：

- （要素集）决策基本要素：参数空间，决策集，损失函数；
- （数据道）分辨参数的数据依赖特征：测量性与估计性；
- （决策轩）期望损失最小产生最优决策；
- （失效性）设定分界域值，给出决策失效的区域。

- 假如根据市场调查数据显示，生产厂家的品牌知名度($x = 0$表示知名度低，$x = 1$表示知名度高)与参数θ 发生的可能性有关系，根据这些分析结果如何决策？

	$x = 0$	$x = 1$
销售好θ_1	0.3	0.7
销售一般θ_2	0.6	0.4
销售较差θ_3	0.7	0.3

两类分类问题-loss-sensitive decision

- 两个状态$\{\omega_1, \omega_2\}$;
- 决策集：$\{\delta_1, \delta_2\}$;
- 损失矩阵Cost/Loss：$l_{ij} = l(\delta_i, \omega_j)$表示当真实的状态是$\omega_j$ 判定为第i类的损失。
- 条件风险：

$$
\begin{aligned}
R(\delta_1|x) =& l_{11}p(\omega_1|x) + l_{12}p(\omega_2|x) \\
R(\delta_2|x) =& l_{21}p(\omega_1|x) + l_{22}p(\omega_2|x)
\end{aligned}
$$

$$
\delta = \begin{cases} \delta_1(\omega_1), & R(\delta_1|x) < R(\delta_2|x); \\ \delta_2(\omega_2), & R(\delta_1|x) > R(\delta_2|x). \end{cases} \tag{2}
$$

$$\delta = \begin{cases} \delta_1(\omega_1), & \frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{l_{12}-l_{22}}{l_{21}-l_{11}} \frac{P(\omega_2)}{P(\omega_1)}; \\ \delta_2(\omega_2), & others. \end{cases} \tag{3}$$

作业1

从经典案例数据iris的Petal.Length选择versicolor(记为ω_1)和virginica(记为ω_2) 两类，损失矩阵如下：

$$L = \begin{bmatrix} 0 & 1 \\ 3 & 0 \end{bmatrix} \tag{4}$$

1. 假设两类数据是正态分布，$p_{\omega_1}(x) \sim N(\mu_1, \sigma_1^2)$,$p_{\omega_2}(x) \sim N(\mu_2, \sigma_2^2)$, 给出$\omega_1$和$\omega_2$期望和方差$(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)$ 的极大似然估计，根据分布密度最大原则给出$x = 5.0$ 的判别类别；
2. 如果$P(\omega_1) = 0.3, P(\omega_2) = 0.7$,请给出决策规则，设计程序探索两类的分界点，求解分界点估计；
3. 假设μ_1有先验分布$p(\mu_1) \sim N(4.0, 0.3^2)$, μ_2 有先验分布$p_{\mu_2}(x) \sim N(5.8, 0.5^2)$, 请给出后验分布条件风险的判别规则，用模拟数据给出ω_1 和ω_2的判决区域。
4. 根据先验分布给出决策规则，再次给出$x = 5.0$的判别类别，和问题1的结果比较，有怎样的不同；

作业2

从经典案例数据iris中取出Sepal.Length和Petal.Width两个变量，选择setosa(记为ω_1) 和virginica(记为ω_2) 两类制成训练集，

- 假设两类数据服从二元正态分布$p_{\omega_1}(x) \sim N(\mu_1, \Sigma^2 = \sigma^2 I), p_{\omega_2}(x) \sim N(\mu_2, \Sigma^2)$, 给出期望$\mu_1, \mu_2$和方差$\sigma^2$) 的极大似然估计，根据分布密度最大原则给出$(x = 5.0, y = 1.7)$ 的判别类别；
- 如果$P(\omega_1) = P(\omega_2)$请给出判别函数，求解判别函数$g(x) = w^T(x - x_0)$ 中的(x_0, w)；
- 如果$P(\omega_1) = 0.3, P(\omega_2) = 0.7$,请给出判别函数，求解判别函数$g(x) = w^T(x - x_0)$中的$(x_0, w)$比较和问题2的判别函数之间的差别；
- 记$\sigma^2 = \lambda^{-1}$，假设λ_1有先验分布$p(\lambda_1) \sim Gamma(0.1, 0.1)$, λ_2 有先验分布$p(\lambda_1) \sim Gamma(1, 1)$,(参数表示如PRML(2.145)) 请给出后验分布条件风险的判别规则，用模拟数据给出ω_1 和ω_2的判决区域。

作业要求

- 无程序的作业交doc格式或tex源程序作业，最好用markdown交有程序的作业:包括源程序和试验输出的html,pdf，R markedown 格式（或Python markdown）以及本次作业所涉及的数据请本次课堂的助教协助提供；
- 本次作业3月24日（下周六）前交给第二次课堂助教，由助教登记后交给老师，作业名称和规范以第一次助教规定为准；

Bayesian Classification

王 星

中国人民大学统计学院
E-mail:wangxingwisdom@126.com

October 15, 2020

大纲

- 分类器，判别函数和决策面
- 二次判别
- Fisher 判别

- The decision problem:
  - given x, predict t according to a probabilistic model $p(x,t)$
- binary classification:$t \in \{0,1\} \leftrightarrow \{C_1, C_2\}$;
- Important quantity: $p(C_k|x)$.

$$p(C_k|x) = \frac{p(x,C_k)}{p(x)} = \frac{p(x,C_k)}{\sum_{k=1}^{2} p(x,C_k)}$$

$$\rightarrow \text{getting } p(x,C_i) \text{ is the (central!) inference problem}$$

$$= \frac{p(x|C_k)p(C_k)}{p(x)} \propto \text{likelihood} \times \text{prior}$$

- Intuition: choose k that maximizes $p(C_k|x)$.

- 决策域: $\mathcal{R}_i = \{x : pred(x) = C_i\}$.
- 判错率:

$$\begin{aligned} p(misclassification) &= p(x \in \mathcal{R}_1, C_2) + p(x \in \mathcal{R}_2, C_1) \\ &= \int_{\mathcal{R}_1} p(x, C_2)dx + \int_{\mathcal{R}_2} p(x, C_1)dx \end{aligned}$$

- $\Leftrightarrow p(C_1|x) > p(C_2|x)$.
- 对于k类而言: 极小化$\sum_j = \int_{\mathcal{R}_j}(\sum_{k\neq j} p(x, C_k))dx$
  $\Leftrightarrow pred(x) = \text{argmax}_k p(C_k|x)$.

分类问题–推断和决策

- 假设$\{C_i\}, i = 1, ..., K$个不同的类，如果$g_i(x) > g_j(x), i \neq j$, $g_i(x), i = 1, ..., K$ 称为判别函数;
- 例如：$g_i(x) = -R(\delta_i|x)$(这对应于后验风险最小准则)，那么根据错误最小化原理，$g_i(x) = p(C_i|x)$, 于是

$$g_i(x) = p(x|C_i)p(C_i).$$

- 判别函数可以替换为任意一个单调增函数$f(g(.))$,

$$g_i(x) = \ln p(x|C_i) + \ln p(C_i).$$

分类问题–推断和决策

有3种构造判别函数的方法

- 分布依赖型：分类问题可以分为两个阶段：推断和决策，在推断的阶段训练分布，在决策的阶段结合损失函数给出最优的分类。通常又可以分为两种方法：
  - 生成法（generative）
    - 用一个生成模型去推断$p(x|C_k)$;
    - 将先验分布$p(C_k)$和$p(x|C_k)$ 联合得到$p(C_k|x)$;
  - 判别法(discriminant):直接推断$p(C_k|x)$;
- 非分布依赖型:直接学习一个判别函数$g(x)$.
  - 直接解一个有分类变量的映射函数;
  - 而分类问题而言，相当于一个$\{+1,-1\}$的函数

分布依赖型：正态分布的二次判别

- 似然函数

$$p(x|C_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)];$$

- 最小错误率由判别函数决定$g_i(x) = \ln p(x|C_i) + \ln p(C_i)$;
- 多类的情形下（QDA）

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln p(C_i).$$

情形1：$\Sigma_i = \sigma^2 I, I$ stands for the identity matrix

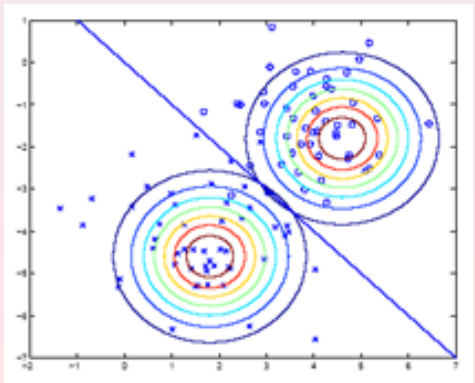
- 判决函数

$$g_i(x) = -\frac{||x-\mu_i||^2}{2\sigma^2} + \ln p(C_i)$$

$$||x-\mu_i||^2 = (x-\mu_i)^T(x-\mu_i) = x^T x - 2\mu_i^T x + \mu_i^T \mu_i$$

- $g_i(x) = w_i^T x + w_{i0}$ Linear discriminant function where

$$w_i = \frac{\mu_i}{\sigma^2}; \quad w_{i0} = -\frac{1}{2\sigma^2}\mu_i^T \mu_i + \ln p(C_i);$$



线性判别面

- A classifier that uses linear discriminant functions is called “a linear machine”使用线性判别函数的分类模型,称为线性机
- The decision surfaces for a linear machine are pieces of hyperplanes defined by: 线性判别机是超平面

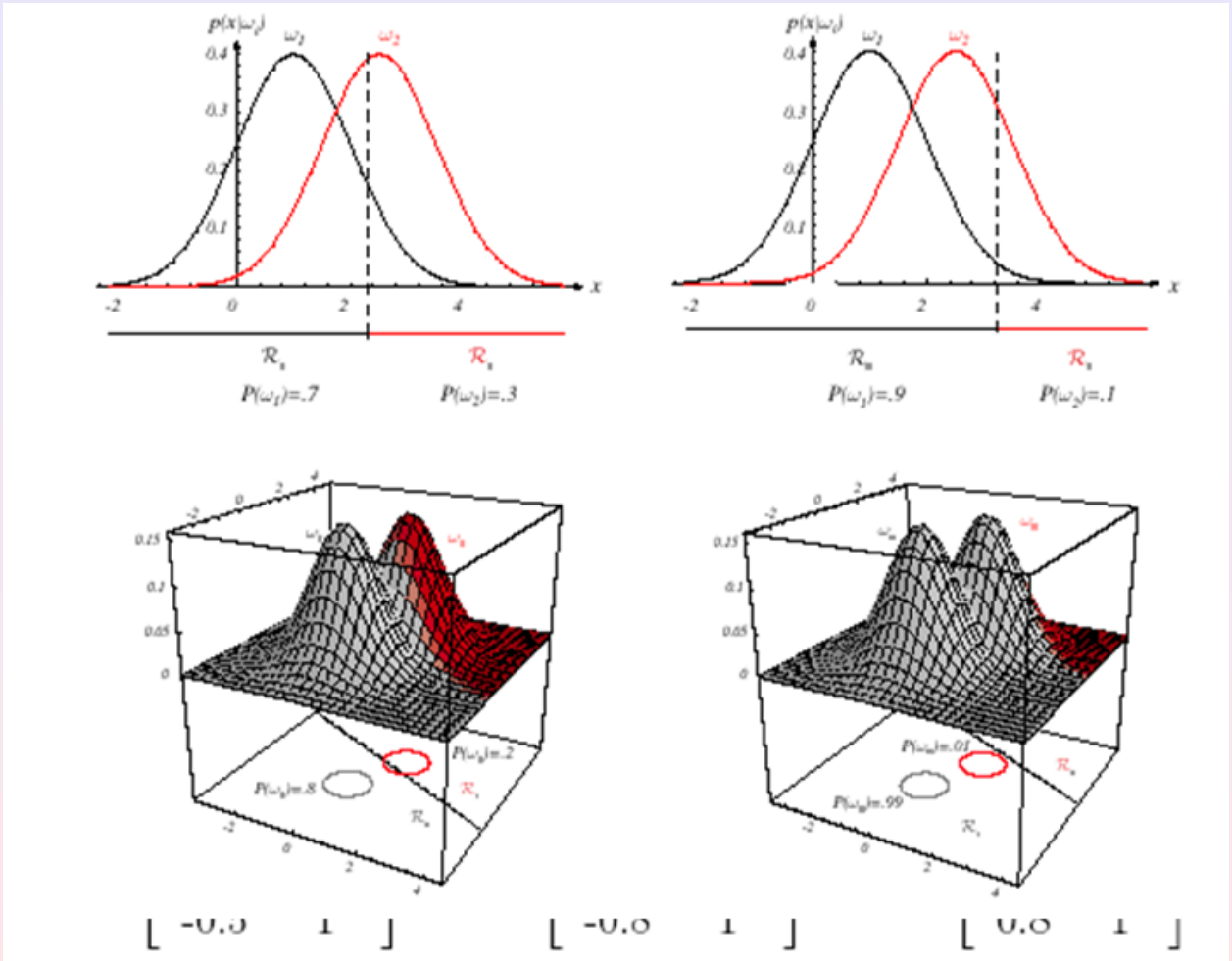
$$g_(x) = g_j(x).$$

$$w^T(x - x_0) = 0;$$

$$w = \mu_i - \mu_j;$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{||\mu_i - \mu_j||} \ln \frac{p(C_i)}{p(C_j)}(\mu_i - \mu_j)$$

先验概率相等和不等的分界面



情形2：$\Sigma_i = \Sigma$线性判别函数

- 判决函数

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln p(C_i).$$

- $g_i(x) = w_i^T + w_{i0}$ Linear discriminant function where

$$w_i = \Sigma^{-1}\mu_i;$$

$$w_{i0} = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \ln p(C_i);$$

w称为权向量，决定判别面的方向，w_0称为阈值，决定了决策面的位置

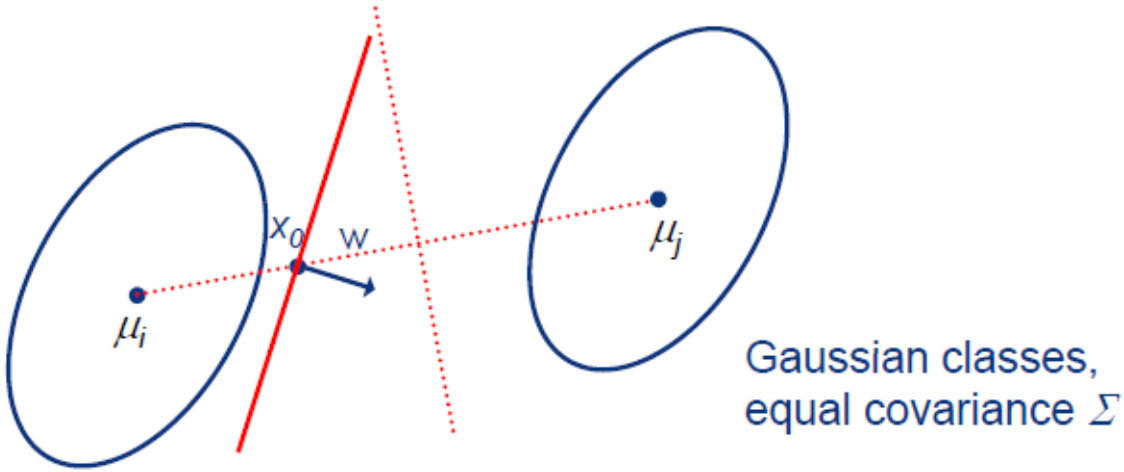
- $w^T(x - x_0) = 0$ where $w = \Sigma^{-1}(\mu_i - \mu_j)$
- (the hyperplane separating R_i and R_j is generally not orthogonal to the line between the means!)

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln p(C_i)/p(C_j)}{(\mu_i - \mu_j)^T\Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

- for a classification problem with Gaussian classes of equal covariance $\Sigma_i = \Sigma$, the BDR boundary is the plane of normal

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

- if $\Sigma_1 = \Sigma_0$, this is also the LDA solution



This gives two different interpretations of LDA

- if and only if the classes are Gaussian and have equal Covariance
- geometric interpretation： plane of normal w,that passes through x_0.

情形3 Σ_i任意

The covariance matrices are different for each category

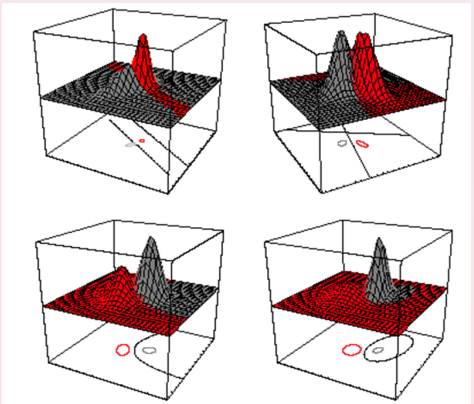
- 判决函数$g_i(x) = x^T W_i x + w_i^T x + w_{i0}$ where

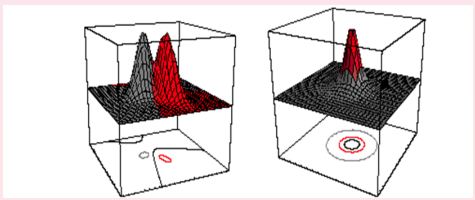
$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i;$$

$$w_{i0} = -\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln p(C_i);$$

- 超二次曲面Hyperquadrics which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids





分布依赖型生成式模型的判别函数的似然解(情形2)Σ相同

- 假设数据集(x_n, t_n),其中$n = 1, ..., N, t_n = 1$表示类别$C_1, t_n = 0$表示类别C_2,
- $p(C_1) = \pi, \ \ p(C_2) = 1 - \pi;$

$$p(x_n, C_1) = p(C_1)p(x_n|C_1) = \pi N(x_n|\mu_1, \Sigma);$$

$$p(x_n, C_2) = p(C_2)p(x_n|C_2) = (1 - \pi)N(x_n|\mu_2, \Sigma);$$

- 似然函数

$$p(t, X|\pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} [\pi N(x_n|\mu_1, \Sigma)]^{t_n} [(1-\pi)N(x_n|\mu_2, \Sigma)]^{(1-t_n)}$$

$t = (t_1, ..., t_N)$.

- 求π最优，对数似然函数$\sum_{n=1}^{N} [t_n \ln \pi + (1 - t_n) \ln(1 - \pi)]$;

MLE估计续

- 对π求导数，$\pi^{MLE} = \frac{1}{N}\sum t_n = \frac{N_1}{N} = \frac{N_1}{N_1+N_2}$,其中$N_1$表示第一类$C_1$中数据点的总和；$N_2$表示第二类$C_2$中数据点的总和；
- 对μ求极值，

$$\sum_{n=1}^{N} t_n \ln N(x_n|\mu,\Sigma) = -\frac{1}{2}\sum t_n(x_n-\mu_1)^T(x_n-\mu_1) + Const;$$

$$\mu_1^{MLE} = \frac{1}{N_1}\sum_{n=1}^{N} t_n x_n;\ \mu_2^{MLE} = \frac{1}{N_2}\sum_{n=1}^{N}(1-t_n)x_n;$$

- 考虑协方差矩阵Σ的极大似然解：

$$-\frac{1}{2}\sum_{n=1}^{N} t_n \ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N} t_n(x_n-\mu_1)^T\Sigma^{-1}(x_n-\mu_1)$$

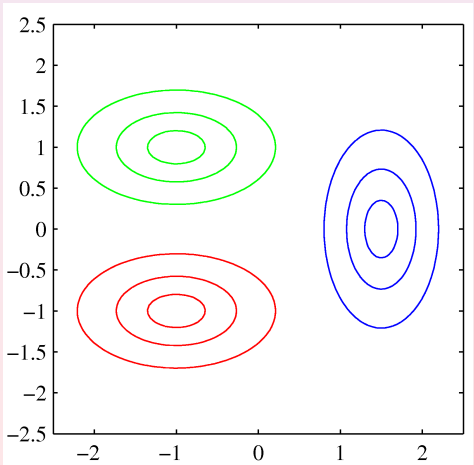
$$-\frac{1}{2}\sum_{n=1}^{N}(1-t_n)\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(1-t_n)(x_n-\mu_2)^T\Sigma^{-1}(x_n-\mu_2)$$

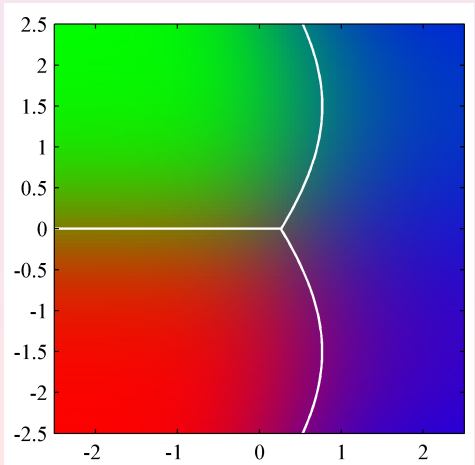
$$= -\frac{N}{2}\ln|\Sigma| - \frac{N}{2}Tr(\Sigma^{-1}S)$$

$$S = \frac{N_1}{N_1 + N_2} S_1 + \frac{N_2}{N_1 + N_2} S_2;$$

$$S_1 = \frac{1}{N_1} \sum_{n \epsilon C_1} (x_n - \mu_1)^T (x_n - \mu_1)^T;$$

$$S_2 = \frac{1}{N_2} \sum_{n \epsilon C_2} (x_n - \mu_2)^T (x_n - \mu_2)^T;$$





指数型分布

- x的分布

$$p(x|\lambda_k) = \ln(x)g(\lambda_k)\exp(\lambda_k^T)\mu(x);$$

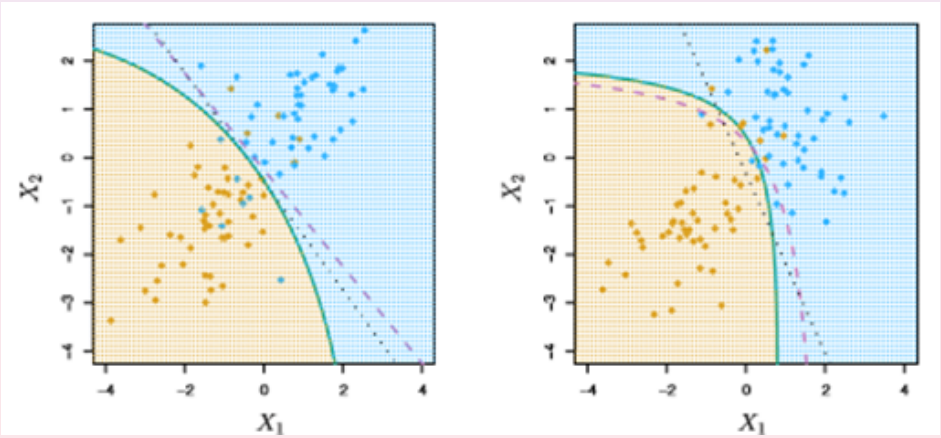
- $a(x)$的形式

$$a(x) = \frac{1}{s}(\lambda_1 - \lambda_2)^T x + \ln g(\lambda_1) + \ln g(\lambda_2) + \ln p(C_1) - \ln p(C_2).$$

- $a_k(x) = \frac{1}{s}\lambda_k^T x + \ln g(\lambda_k) + \ln p(C_k).$

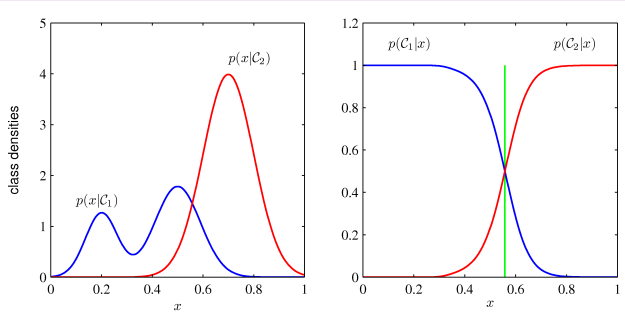
LDA和QDA的区别

- QDA will work best when the variances are very different between classes and we have enough observations to accurately estimate the variances.
- LDA will work best when the variances are similar among classes or we don’t have enough data to accurately estimate the variances.



Decision theory–inference and decision

- 分布依赖型生成式模型:
  - pros: access to $p(x) \rightarrow$ easy detection of outliers I i.e., low-confidence predictions I;
  - cons: estimating the joint probability $p(x, C_k)$ can be computational and data demanding.
- 分布依赖型判别模型:
  - pros: less demanding than the generative approach;

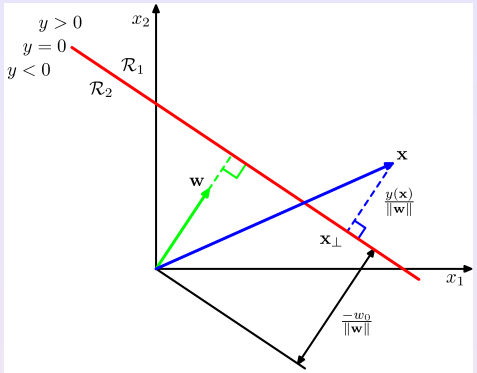


- 判别函数:
  - pros: a single learning problem (vs inference + decision);
  - cons: no access to $p(C_k|x)$ which can have many advantages in practice for (e.g.) rejection and model combination – see page 45.

Reminder:Three different spaces that are easy to confuse

- **Weight-space:**
  - Each axis corresponds to a weight;
  - A point is a weight vector;
  - Dimensionality:=♯weights +1 extra dimension for the loss.
- **Data-space:**
  - Each axis corresponds to an input value;
  - A point is a data vector;
  - A decision surface is a plane;
  - Dimensionality=dimensionality of a data vector;
- **“Case-space”:**
  - Each axis corresponds to a training case;
  - A point assigns a scalar value to every training case.
  - So it can represent the 1-D targets or it can represent the value of one input component over all the training data.
  - Dimensionality =♯training cases.

判别函数



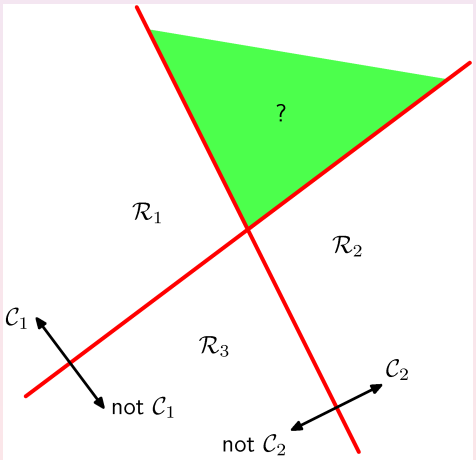
- The planar decision surface in data-space for the simple linear discriminant function:$w^T x + w_0$
- 线性判别函数$y(x) = w^T x + w_0$; 考虑两个决策面上的点x_A, x_B,应该有$w^T(x_A - x_B) = 0$.权值向量与决策面上的任意向量都正交。
- 考虑$y(x) = 0, x\epsilon$决策面上，于是从原点出发到决策面的垂直距离是

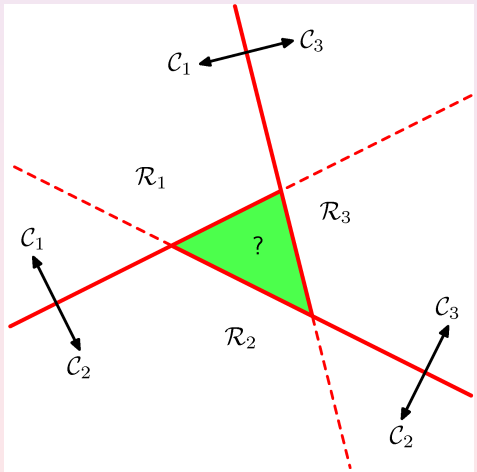
$$\frac{w^T x}{\|w\|} = -\frac{w_0}{\|w\|}.$$

- 任意一点和它在决策面的投影$x_\perp$有$x = x_\perp + r\frac{w}{\|w\|}$, $r = \frac{y(x)}{\|w\|}$.

Discriminant functions for $N > 2$ classes

- One possibility is to use N two-way discriminant functions. Each function discriminates one class from the rest.
- Another possibility is to use $N(N-1)/2$ two-way discriminant functions,Each function discriminates between two particular classes.
- Both these methods have problems





A simple solution

- 使用K个分类函数$g_1(x), ..., g_k(x)$,每个线性判别函数有下面的形式

$$g_k(x) = w_k^T x + w_{k0}$$

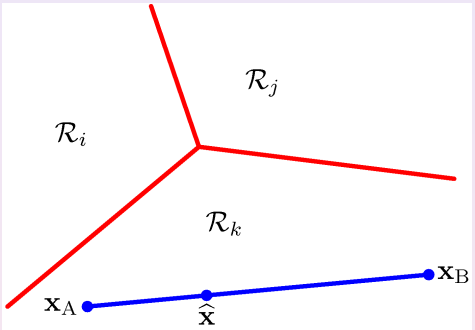


Figure: 如果两个点在同一个决策区域，决策区域满足单连通和凸性质

- 如果$g_k(x_A) > g_l(x_A), g_k(x_B) > g_l(x_B), \forall \hat{x} = \lambda x_A + (1-\lambda)x_B, 0 \leq \lambda \leq 1,$

$$g_k(\hat{x}) > g_l(\hat{x})$$

在分类中运用最小二乘

- This is not the right thing to do and it doesn’t work as well as better methods, but it is easy:
  - It reduces classification to least squares regression.
  - We already know how to do regression. We can just solve for the optimal weights with some matrix algebra
- We use targets that are equal to the conditional probability of the class given the input.
  - When there are more than two classes, we treat each class as a separate problem (we cannot get away with this if we use the “max” decision function).

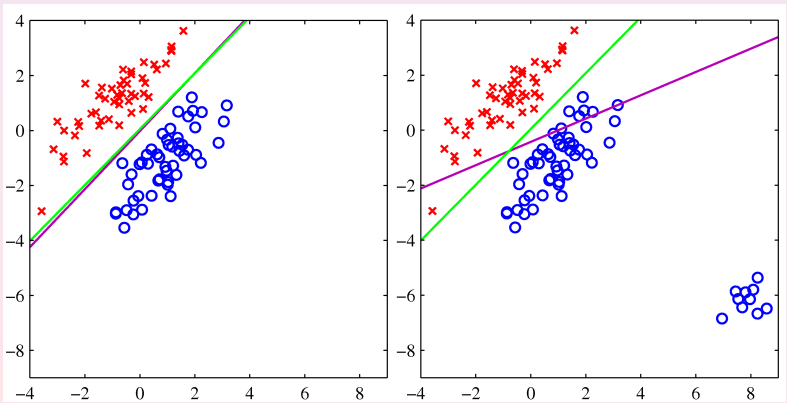


Figure: 最小二乘对于异常数据点敏感

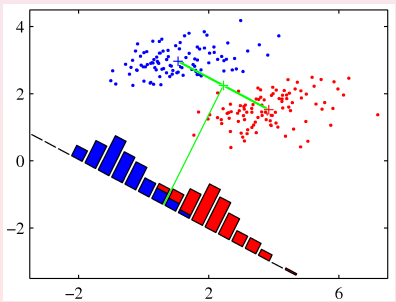
Fisher's线性判别

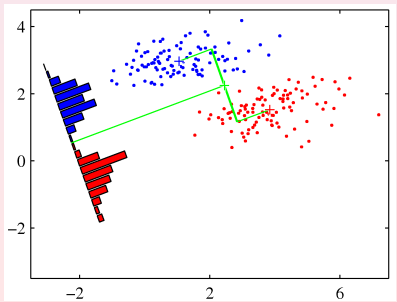
- 从降维的角度重新考虑线性分类模型;
- 对于标准的二分类：假设有d维向量x,通过$y = w^T x$投影到一维：

$$\delta = \begin{cases} C_1, & y \geq -w_0; \\ C_2, & y < -w_0. \end{cases}$$

- 向一维投影会造成信息损失,在d维空间中能够被拆分，但投影到一维却可能出现重叠;
- 想法：调整权向量w,选择能够将两个类分开到最大的投影方向;
- 具体而言，假设C_1有N_1个数据点，C_2有N_2个数据点：

$$\bar{x}_1 = \frac{1}{N_1} \sum_{n \epsilon C_1} x_n; \bar{x}_2 = \frac{1}{N_2} \sum_{n \epsilon C_2} x_n$$





Fisher判别

- What linear transformation is best for discrimination? $y = w^T x$;
- The projection onto the vector separating the class means seems sensible: $w = \bar{x}_2 - \bar{x}_1$.
- But we also want small variance within each class:

$$s_1^2 = \sum_{n \epsilon C_1} (y_n - \bar{x}_1);$$

$$s_2^2 = \sum_{n \epsilon C_2} (y_n - \bar{x}_2);$$

- Fisher’ s objective function is:

$$J(w) = \frac{m_2 - m_1}{s_1^2 + s_2^2}$$

与最小平方的关系

- 最小平方法确定线性判别函数的目标是使得模型的预测与目标值接近；
- 平方误差函数

$$E = \frac{1}{2}\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)^2;$$

- 令E关于w_0和w求导为零

$$\sum_{n=1}^{N}(w^T x_n + w_0 - t_n) = 0 \Rightarrow w_0 = -w^T m;$$

$$\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)x_n = 0, w \propto S_w^{-1}(x_2 - x_1);$$

Fisher判别的推广——-多分类Fisher 判别

- 假设空间维度D大于类别数量K,引入D'个线性特征$y_k = w_k^T x, k = 1, ..., D'$
- 类内协方差矩阵

$$S_W = \sum_{k=1}^{K} S_K$$

- 一种选择$J(W) = Tr\{S_W^{-1} S_B\}$

与最小平方的关系

- 最小平方法确定线性判别函数的目标是使得模型的预测与目标值接近；
- 平方误差函数

$$E = \frac{1}{2}\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)^2;$$

- 令E关于w_0和w求导为零

$$\sum_{n=1}^{N}(w^T x_n + w_0 - t_n) = 0 \Rightarrow w_0 = -w^T m;$$

$$\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)x_n = 0, w \propto S_w^{-1}(x_2 - x_1);$$

与最小平方的关系

- 最小平方法确定线性判别函数的目标是使得模型的预测与目标值接近；
- 平方误差函数

$$E = \frac{1}{2}\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)^2;$$

- 令E关于w_0和w求导为零

$$\sum_{n=1}^{N}(w^T x_n + w_0 - t_n) = 0 \Rightarrow w_0 = -w^T m;$$

$$\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)x_n = 0, w \propto S_w^{-1}(x_2 - x_1);$$