- Bayes 决策：假设分类数，各类分布已知。

最小错误率：

$$P(w_i|\vec{x})=\frac{p(\vec{x}|w_i)P(w_i)}{\sum_{i=1}^{n}p(\vec{x}|w_i)P(w_i)}$$

取w_i使得后验概率$P(w_i|\vec{x})$取得最大值

错误率$P(e)=\int_{-\infty}^{+\infty}P(e|\vec{x})p(\vec{x})d\vec{x}$，

$$其中P(e|\vec{x})=\begin{cases}P(w_1|\vec{x}), & P(w_2|\vec{x})>P(w_1|\vec{x})\\ P(w_2|\vec{x}), & P(w_1|\vec{x})>P(w_2|\vec{x})\end{cases}$$

$$P(e)=P(w_2)\int_{R_1}p(\vec{x}|w_2)\,d\vec{x}+P(w_1)\int_{R_2}p(\vec{x}|w_1)\,d\vec{x}$$

$$=P(w_2)P_2(e)+P(w_1)P_1(e)$$

最小风险：决策空间$\{\propto_1,\propto_2,\propto_3,\dots\}$，损失函数$\lambda(\alpha_i,w_j)$

风险$R(\alpha_i|\vec{x})=\sum_{j=1}^{c}\lambda(\alpha_i,w_j)P(w_j|\vec{x})$，求最小。

$\lambda(\alpha_i,w_j)=I_{(i=j)}$时，退化成最小错误率

限定一类错误率：$\min\gamma=P_1(e)+\lambda(P_2(e)-\varepsilon_0)$

求导得$\int_{R_1}p(\vec{x}|w_2)\,d\vec{x}=\varepsilon_0$，似然比$\lambda=\frac{p(t|w_1)}{p(t|w_2)}$

判别$\frac{p(x|w_1)}{p(x|w_2)}>\lambda$，则为$w_1$。

最小最大决策：先验$P(w_i)$未知，损失$\lambda_{ij}=\lambda(\alpha_i,w_j)$已知

则$R=\int R(\alpha(\vec{x})|\vec{x})p(\vec{x})d\vec{x}$　　　　$(P(w_1)+P(w_2)=1)$

$$=\int_{R_1}[\lambda_{11}P(w_1)p(\vec{x}|w_1)+\lambda_{12}P(w_2)p(\vec{x}|w_2)]d\vec{x}+\cdots$$

$$=\lambda_{22}+(\lambda_{12}-\lambda_{22})\int_{R_1}p(\vec{x}|w_2)d\vec{x}+P(w_1)\Big[(\lambda_{11}-\lambda_{22})+(\lambda_{21}-\lambda_{11})\int_{R_2}p(\vec{x}|w_1)d\vec{x}-(\lambda_{12}-\lambda_{22})\int_{R_1}p(\vec{x}|w_2)d\vec{x}\Big]$$

$=a+P(w_1)b$（分界面固定时），扫描分界面$R'\leq R$

取$P(w_1)$使$R'=R$（相切），从而根据此$P(w_1)$设计分类器。

多类分类器：判别函数$g_i(x)=\ln p(\vec{x}|w_i)+\ln P(w_i)$

分类器$i^*=argmax_i\,g_i(x)$

正态分布分类：

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}exp\{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\}$$

高维

$$p(x)=\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

线性变换$y=Ax$，$y\sim N(A\mu,A\Sigma A^T)$

$$g_i(x)=-\frac{d}{2}ln2\pi-\frac{1}{2}ln|\Sigma_i|+lnP(w_i)-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

- 概率密度函数估计：概率密度函数未知。

[参数估计]**最大似然**：$l(\theta)=p(\chi|\theta)=\prod_{k=1}^{N}p(x_k|\theta)$

$max\ H(\theta)=\ln l(\theta)=\sum_{k=1}^{N}\ln p(x_k|\theta)$，$\frac{dH(\theta)}{d\theta}=0\ \to\hat{\theta}$

正态分布：$\hat{\mu}=\frac{1}{N}\sum_{k=1}^{N}x_k$，$\hat{\sigma}^2=\frac{1}{N}\sum_{k=1}^{N}(x_k-\hat{\mu})^2$（有偏）

可识别性$\forall\theta\neq\theta'\ \exists x\ st\ p(x|\theta)\neq p(x|\theta')$，离散往往不可识别

贝叶斯估计：已知先验$p(\theta)$，似然$p(x|\theta)$，则后验

$\max\ p(\theta|x)=\frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$ 数据独立抽取

贝叶斯学习：增量学习

$$p(\theta|\chi^N)=\frac{p(x_N|\theta)p(\theta|\chi^{N-1})}{\int p(x_N|\theta)p(\theta|\chi^{N-1})d\theta}$$

$p(x|\chi^N)=\int p(x|\theta)p(\theta|\chi^N)\,d\theta$，$\lim_{N\to\infty}p(x|\chi^N)=p(x)$

正在分布：$p(x|\mu)\sim N(\mu,\sigma^2)$，$p(\mu)\sim N(\mu_0,\sigma_0^2)$，则有

后验$p(\mu|\chi)\sim N(\mu_N,\sigma_N^2)$，其中$m_N=\frac{1}{N}\sum_{k=1}^{N}x_k$

$$\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0\qquad\sigma_N^2=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

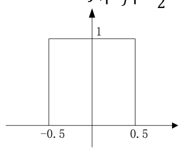
学习得$p(x|\chi)=\int p(x|\mu)p(\mu|\chi)\,d\mu\sim N(\mu_N,\sigma^2+\sigma_N^2)$

非参数估计：N个 iid 样本，估计$p(x)$，构造区域R，面积V，

其中有k个样本落入R中，则平均密度$\hat{p}(x)=\frac{k}{NV}$

构造含x区域序列$\{R_1,R_2,\dots,R_N\}$，满足$\lim_{N\to\infty}V_N=0$，$\lim_{N\to\infty}k_N=\infty$，$\lim_{N\to\infty}k_N/N=0$，则$\hat{p}_N(x)\to p(x)$

Parzen 窗：R_N为d维窗体，$V_N=h_N^d$，$\varphi(u)=I_{\forall j,|u_j|<\frac{1}{2}}$



密度$\hat{p}_N(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{V_N}\varphi\left(\frac{x-x_i}{h_N}\right)$

窗满足$\varphi(u)\geq 0$，$\int\varphi(u)\,du=1$

收敛条件$\sup_u\varphi(u)<\infty$，$\lim_{\|u\|\to\infty}\varphi(u)\prod_{i=1}^{d}u_i=0$

$p(x)$在x连续，$\lim_{N\to\infty}V_N=0$，$\lim_{N\to\infty}NV_N=\infty$

误差产生：分类问题中误差产生的原因

贝叶斯误差$P(e)$：固有的，分类器设计阶段无法消除

模型误差：严重误差，可通过改变模型避免

估计误差：增加样本，改进估计方法

维数问题：维数灾难，1 维需N个样本，d维需N^d个

解决：特征独立性$p(x,y)=p(x)p(y)$，低维流形降维

过拟合/过学习：模型过于复杂→参数$\vec{w}$过多→参数值大

解决：Bayesian 方法可以避免过拟合，正则化$\frac{\lambda}{2}\|\vec{w}\|^2$

样本少→参数少（简单参数化模型，低阶次，对角矩阵）

- 高斯混合模型/EM 算法：多高斯加权求和。

GMM：$p(X|\Theta)=\sum_{i=1}^{M}\alpha_i p_i(X|\theta_i)$，$\Theta=(\alpha_1,.\alpha_M,\theta_1,.\theta_M)$

权$\sum_{i=1}^{M}\alpha_i=1$，$p_i(x|w_i,\theta_i)\sim N(\mu_i,\Sigma_i)$

伪 EM：似然$ln\,p(X|\theta)=\sum_{i=1}^{N}ln\sum_{j=1}^{c}N(x_i|\mu_j,\Sigma_j)P(w_j)$

求导迭代 1)E-STEP

$$P(w_k|x_i,\mu_k,\Sigma_k)=\frac{N(x_i|\mu_k,\Sigma_k)P(w_k)}{\sum_{j=1}^{c}N(x_i|\mu_j,\Sigma_j)P(w_j)}$$

2)M-STEP

$$\mu_k=\frac{\sum_{i=1}^{N}P(w_k|x_i,\mu_k,\Sigma_k)x_i}{\sum_{i=1}^{N}P(w_k|x_i,\mu_k,\Sigma_k)}$$

3)

$$P(w_k)=\frac{1}{N}\sum_{i=1}^{N}P(w_k|x_i,\mu_k,\Sigma_k)$$

4)

$$\Sigma_k=\frac{\sum_{i=1}^{N}P(w_k|x_i,\mu_k,\Sigma_k)(x_i-\mu_k)^{\mathrm{T}}(x_i-\mu_k)}{\sum_{i=1}^{N}P(w_k|x_i,\mu_k,\Sigma_k)}$$

EM 算法：数据$X=\{x_i\}_{i=1}^N$不完整，隐含变量$Z=(X,Y)$
设$q(y)$为Y分布密度，则似然

$$L(\theta)=ln\sum_y p(x,y|\theta)=ln\sum_y q(y)\frac{p(x,y|\theta)}{q(y)}\geq\sum_y q(y)\,ln\frac{p(x,y|\theta)}{q(y)}$$
$$=\sum_y q(y)\,ln\,p(x,y|\theta)-\sum_y q(y)\,ln\,q(y)=F(q,\theta)$$

找到下界，EM 找最大$F(q,\theta)$
E-SETP：取$q_{[k+1]}(y)=p(y|x,\theta_{[k]})$，此时$F(q,\theta)=L(\theta)$
M-STEP：记$Q(\theta_{[k]},\theta)=E[\ln p(x|y,\theta)|x,\theta_{[k]}]$
$=\sum_y p(y|x,\theta_{[k]})\ln p(x,y|\theta)$
取$\theta_{[k+1]}=\arg\max_\theta Q(\theta_{[k]},\theta)$，迭代到收敛
GEM 算法：满足$Q(\theta_{[k]},\theta_{[k+1]})>Q(\theta_{[k]},\theta)$，依然收敛

EM 参数估计：由于$\sum_{i=1}^N ln(\sum_{j=1}^N\alpha_j,p_j(x_i|\theta_j))$不好计算
选似然$L(\Theta)=ln\,p(X,Y|\Theta)=\sum_{i=1}^N ln(\alpha_{y_i}p_{y_i}(x_i|\theta_{y_i}))$找下界
E-SETP：参数$\Theta^g=(\alpha_1^g,.\alpha_M^g,\theta_1^g,.\theta_M^g)$，求函数$Q(\Theta,\Theta^g)$

$$Q(\Theta,\Theta^g)=\sum_{l=1}^M\sum_{i=1}^N ln(\alpha_l)\,p(l|x_i,\Theta^g)+\sum_{l=1}^M\sum_{i=1}^N ln(p_l(x_i|\theta_l))\,p(l|x_i,\Theta^g)$$

M-SETP：独立优化α_l和θ_l，用拉格朗日$\lambda[\sum_l\alpha_l-1]$得
$\alpha_l=\frac{1}{N}\sum_{i=1}^N P(l|x_i,\Theta^g)$，$\mu_l=..$，$\Sigma_l=..$，**结果同“伪 EM”**。

- 线性判别函数：d维 $g(x)=w^Tx+w_0$。

其中w是超平面$g(x)=0$的法向量，任意样本x有
向超平面投影$x=x_p+r\frac{w}{\|w\|}$，投影距离$r=\frac{g(x)}{\|w\|}$
升维线性化：$g(x)=(x-a)(x-b)\to g'(x)=a^Ty$
Fisher 准则：投影到直线，在直线上最容易分开
N个样本$\chi=\{x_i\}_{i=1}^N$，第一类χ_1，投影$y_n=w^Tx_n$
类内均值$m_i=\frac{1}{N_i}\sum_{x\in\chi_i}x$，总类内$S_w=\sum_i S_i$
类内离散度$S_i=\sum_{x\in\chi_i}(x-m_i)^T(x-m_i)$
类间离散$S_b=(m_1-m_2)(m_1-m_2)^T$
目标函数$\max J_F(w)=\frac{(\tilde{m}_1-\tilde{m}_2)^2}{\tilde{S}_1^2+\tilde{S}_2^2}=\frac{w^TS_bw}{w^TS_ww}$
拉格朗日方法$L=w^TS_bw-\lambda(w^TS_ww-C)$，$\frac{\partial L}{\partial w}=0$
得$S_w^{-1}S_bw=\lambda w$，取特征值$\lambda_{\max}$及其对应特征向量w

$$w=S_w^{-1}(m_1-m_2)$$

Fisher 分类：分界面$y_0=\frac{\tilde{m}_1+\tilde{m}_2}{2}$ 或 $\frac{N_2\tilde{m}_1+N_1\tilde{m}_2}{N_1+N_2}$

Fisher 问题
　适合哪种数据分布：高斯/类高斯分布
　多类：多个 Fisher
　散度矩阵前考虑先验加权.
　可以投影到平面，可以投影到一般的低维空间。
　总的类内散度矩阵不一定可逆，数据中有冗余，降维到可逆。
感知准则函数：需要样本$\{y_i\}_{i=1}^N$线性可分，即$\exists a,s.t.$
$\begin{cases}ay_i>0,\forall y_i\in w_1\\ ay_i<0,\forall y_i\in w_2\end{cases}$，记$y'_i=\begin{cases}y_i,\forall y_i\in w_1\\ -y_i,\forall y_i\in w_2\end{cases}$，$ay'_i>0$
目标函数$\max J_P(a)=\sum_{y\in Y^k}(-a^Ty)$，$Y^k$为分错样本
梯度$\nabla J_P(a)=\frac{\partial J_P(a)}{\partial a}=\sum_{y\in Y^k}(-y)$，迭代公式
$a(k+1)=a(k)-\rho_k\nabla J=a(k)+\rho_k\sum_{y\in Y^k}y$，收敛
快速：逐个遍历$\{y_i\}_N^1$，$a(k+1)=a(k)+y_iI_{(a(k)y_i\leq 0)}$
多类：$c-1$个两类问题（i-非i），或C_2^c个两类问题。

- 支持向量机：分界面$g(x)=w^Tx+b$。

满足$\begin{cases}w_1: d_i=+1\to w^Tx_i+b\geq+1\\ w_2: d_i=-1\to w^Tx_i+b\leq-1\end{cases}$，支持向量$x^{(s)}$
投影距$r(x^{(s)})=\frac{|g(x^{(s)})|}{\|w\|}=\frac{1}{\|w\|}$，间距$margin=2r=\frac{2}{\|w\|}$
目标$max\frac{2}{\|w\|}\to min\,f(w)=\frac{w^Tw}{2}$，凸优化问题
约束$g_i(w)=d_i(w^Tx_i+b)-1\geq 0$，$i=1,2,\dots,N$
拉格朗日$J(w,b,\alpha)=\frac{w^Tw}{2}-\sum_{i=1}^N\alpha_i[d_i(w^Tx_i+b)-1]$
求$min_w\,max_\alpha\,J(w,b,\alpha)$，为$\Phi(w,\alpha)=J(w,b,\alpha)$的鞍点
鞍点(w',α')满足$\Phi(w',\alpha)\leq\Phi(w',\alpha')\leq\Phi(w,\alpha')$
证明：$\Phi(w,\alpha)=f(w)-\sum_{i=1}^N\alpha_ig_i(w)$，$w'$是$min\,f(w)$解

$$f(w')-\sum_{i=1}^N\alpha_ig_i(w')\leq f(w')-\sum_{i=1}^N\alpha'_ig_i(w')\leq f(w)-\sum_{i=1}^N\alpha'_ig_i(w)$$

上式为鞍点定义：1)左$\Rightarrow\sum_{i=1}^N(\alpha_i-\alpha'_i)g_i(w')\geq 0$
取$\alpha_1=\alpha'_1+1,\alpha_i=\alpha'_i\Rightarrow g_i(w')\geq 0$，$\Rightarrow w'$可行解
2) 取$\alpha=0$，左$\Rightarrow\sum_{i=1}^N\alpha'_ig_i(w')\leq 0$，$\sum_{i=1}^N\alpha'_ig_i(w')=0$
则右$\Rightarrow f(w')\leq f(w)-\sum_{i=1}^N\alpha'_ig_i(w)\leq f(w)$，$\Rightarrow w'$最优解[#]
求解对偶问题：$Q(\alpha)=min_{w,b}\,J(w,b,\alpha)$，$max_{\alpha\geq 0}\,Q(\alpha)$
由

$$\begin{cases}\frac{\partial J(w,b,\alpha)}{\partial w}=0\to w=\sum_{i=1}^N\alpha_id_ix_i\\ \frac{\partial J(w,b,\alpha)}{\partial b}=0\to\sum_{i=1}^N\alpha_id_i=0\end{cases}$$

代入$J(w,b,\alpha)$得
- **目标**

$$max\,Q(\alpha)=\sum_{i=1}^N\alpha_i-\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N\alpha_i\alpha_j\,d_id_jx_i^Tx_j$$

约束$\sum_{i=1}^N\alpha_id_i=0$，$\alpha_i\geq 0$，$i=1,2,\dots,N$，求得$\alpha$
则**分类器**（支持向量$\alpha_i\neq 0$，非支持向量$\alpha_i=0$）

$$g(x)=w^Tx+b=\sum_{i=1}^{N^{(s)}}\alpha_id_is_i^Tx+b$$

少数支持向量（$1\ to\ N^{(s)}$）决定了分类超平面
- **线性不可分时**：引入罚向量$\{\xi_i\}_{i=1}^N$

目标：C为罚值

$$min\,f(w,\xi)=\frac{w^Tw}{2}+C\sum_{i=1}^N\xi_i$$

约束$g_i(w)=d_i(w^Tx_i+b)-1+\xi_i\geq 0$，$i=1,2,\dots,N$
对偶问题：几乎完全相同，约束部分改为$0\leq\alpha_i\leq C$

■ **非线性情况**：将$Q(\alpha)$中的$x_i^T x_j$替换为$\Phi(x_i)\cdot\Phi(x_j)$
核函数$K(x_i,x_j)=\Phi(x_i)\cdot\Phi(x_j)$，($\Phi$函数形式可以不知)
目标：

$$max\ Q(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\ d_i d_j K(x_i,x_j)$$

约束$\sum_{i=1}^{N}\alpha_i d_i=0$，$\alpha_i\ge 0$，$i=1,2,\dots,N$，求得$\alpha$
则**分类器**：

$$g(x)=w^T x+b=\sum_{i=1}^{N^{(s)}}\alpha_i d_i K(s_i,x)+b$$

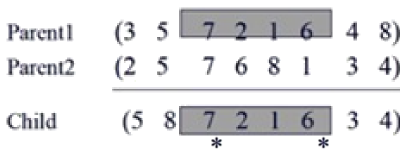
核函数需满足：$\forall g(x),\int g(x)^2dx<\infty,s.t.$
$\int K(x,y)g(x)g(y)dxdy>0$

Polynomial 多项式核：$K(x,y)=(x\cdot y+\xi)^p$
RBF 高斯核(无穷阶)：$K(x,y)=exp\{-\|x-y\|^2/2\sigma^2\}$
Sigmoid 核：$K(x,y)=tanh(\kappa x\cdot y-\delta)$

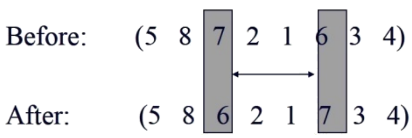
SVM 问题：
多类问题：多个两类问题
执行效率：较高。快速算法
是否可把各种不同参数情况下的解都找到：不可以
一类问题的 SVM：构造非本类样本

产生式模型：计算$p(\vec{x}|w_i)$，[优]可得$p(\vec{x})$，结合领域知识，[缺]计算量大。如 GMM,隐马尔科夫 HMM
判别式模型：直接算$P(w_i|\vec{x})$或判别函数$f(\vec{x})$，省略的条件概率模型的细节，准确率往往高。
计算后验的好处：修正最小化风险决策准则。如果仅仅有一个判别函数，那么损失矩阵的任何改变都要求用训练数据并重新解决分类问题。选择拒绝策略。

● 遗传算法：繁殖，竞争，选择，生存。
Crossover 交叉：
旅行家问题
最短路

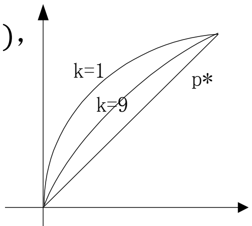


Mutation 变异：

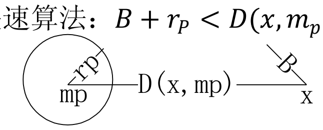


算法核心：问题编码（小数二进制），适应度，后代

● 近邻法：k近邻，类数c，错误率ρ，贝叶斯ρ^*
1 近邻$g_i(x)=min_k\|x-x_i^k\|$，k样本号，i所属类号
错误率：$\rho^*\le\rho\le\rho^*(2-\frac{c}{c-1}\rho^*)$，
$k\to\infty,\rho\to\rho^*$



快速算法：$B+r_P<D(x,m_p)$



距离度量：马氏距离$\delta_M(x,y)=\sqrt[s]{\sum_{j=1}^{d}|x_j-y_j|^s}$
棋盘$S=1$，欧式$S=2$，切比雪夫$S=\infty$ (max$|x_j-y_j|$)
相似性度量：向量夹角

● 特征的提取与选择：不是特征越多越好。
考虑特征的分类价值和提取代价，可分离性判据J_P
$J_P\ge 0$，完全可分$\max J_P$，不可分$J_P=0$
特征选择：D个特征选d个，$J(x_1)\ge J(x_2)\ge..\ge J(x_D)$
寻优算法：最优搜索-分支定界(NP 难)。次优搜索：
1)单独最优组合：$J(X)=\sum_{i=1}^{D}J(x_i)$，$J(X)=\prod_{i=1}^{D}J(x_i)$
2)顺序前进：单最优x_{i1}，两个最优(x_{i1},x_{i2})…到d个
3)顺序后退：D个，每次减 1 个，到d个
4)前进后退：0 个，加 2 减 1，到d个。5)增 l 减 r
Relief 算法：训练集$X=\{x_i\in R^D\}_{i=1}^{N}$，选$d$个，循环$n$次
1)初始化：权向量$w=[w_1,w_2,w_3,\dots,w_D]^T=0$
2)开始：$for\ i=1\ to\ n$
$x=RandGet(X)$ //随机获取
$h=SameClassNear(X,x)$ //同类最近
$m=DiffClassNear(X,x)$ //异类最近
$for\ j=1\ to\ d$ //更新权
$w_j=w_j-\frac{diff(j,x,h)}{n}+\frac{diff(j,x,m)}{n}$ //差异
$next\ j$
$next\ i$
3)返回：$return\ w,(d\ of\ max\ w)$ // w中d个最大的
差异函数$diff(j,x,y)$：离散$diff(j,x,y)=I(x_j\ne y_j)$。
连续$diff(j,x,y)=|x_j-y_j|/|max(j)-min(j)|$。
Relief-f 多类特征选择：
1)初始化权向量$w=[w_1,w_2,w_3,\dots,w_D]^T=0$
2) $for\ i=1\ to\ n$
$x=RandGet(X)$ //随机获取
$\{h_l\}_{l=1}^{k}=SameClassNear(X,x,\mathrm{k})$ //同类 k 近邻
$for\ each\ class\ C\ne Class(x)$ //对于每异类
$\{m_l(C)\}_{l=1}^{k}=DiffClassNear(X,C,x,k)$ //k 近邻
$for\ j=1\ to\ d$ //更新权

$$w_j=w_j-\frac{\sum_{l=1}^{k}diff(j,x,h_l)}{n\times k}+\sum_{C\ne Class(x)}\frac{P(C)}{1-P(Class(x))}\frac{\sum_{l=1}^{k}diff(j,x,m_l(C))}{n\times k}$$

$next\ j$
$next\ i$
3) $return\ w,(d\ of\ max\ w)$ // w中d个最大的

● 主成分分析/K-L 变换。
K-L 变换：$x=[x_1,x_2,x_3,\dots,x_n]^T$，找完备$\forall x=\sum_{i=1}^{\infty}c_iu_i$，正交归一$u_i^Tu_j=I(i=j)$的基向量$u=[u_1,u_2,\dots,x_\infty]^T$。
降维近似有$\hat{x}=\sum_{i=1}^{d}c_iu_i$，误差$\varepsilon=E[(x-\hat{x})^T(x-\hat{x})]$
$\varepsilon=E[\sum_{i=d+1}^{\infty}c_i^2]=\sum_{i=d+1}^{\infty}u_i^TE[x^Tx]u_i$，$\psi=E[x^Tx]$
拉格朗日$g(u_i)=\sum_{i=d+1}^{\infty}(u_i^T\psi u_i-\lambda_i[u_i^Tu_j-1])$，求导得$\psi u_i=\lambda_iu_i$，取特征值$\lambda_i$大的(代表数据在$i$上方差)。
PCA 主成分分析：样本集$X=[\vec{x}_1,\vec{x}_2,\vec{x}_3,\dots,\vec{x}_n]$
零均值化$\hat{X}$，$\psi=\hat{X}\hat{X}^T$，$\psi u_i=\lambda_iu_i$，$Q=[\vec{u}_1,\vec{u}_2,\dots,\vec{u}_d]$
则$X_{PCA}=Q^T\hat{X}$。

降维前需要零均值化！

- 聚类（非监督）：基于相似度的分类。

C均值：N个样本Y，c类$\{\Gamma_i\}_{i=1}^{c}$，类均值$m_i=\frac{1}{N_i}\sum_{y\in\Gamma_i}y$

误差$J(e)=\sum_{i=1}^{c}\sum_{y\in\Gamma_i}\|y-m_i\|^2$，

1)初始化c类，计算$J(e)$

2)$while(true)\ for\ each\ y\in Y$，

$y\in\Gamma_i$，$if\ N_i=1, contiune$

$\forall j\neq i,\ if\ y\in\Gamma_j\to J'(e)<J(e)$，$set\ y\in\Gamma_j$

$next\ y$，$if\ loop\ N\ time\ J(e)\ not\ change,\ break$

3) $return$

ISODATA：最近的两类合并。分裂：类内方差大/样本多

核函数：修改准则$J(e)=\sum_{i=1}^{c}\sum_{y\in\Gamma_i}\Delta(y,K_i)$

其中：核K_i代表集合Γ_i，度量$\Delta(y,K_i)$。

原始核：$K_i=m_i$，$\Delta(y,K_i)=\|y-m_i\|^2$ （C 均值）

正态核：$K_i=\{m_i,\Sigma_i\}$，$\Delta(y,K_i)=ln\,N(y|m_i,\Sigma_i)$

主轴核：$\Delta(y,K_i)$：$(y-m_i)$向主轴投影的长度平方

多级聚类：初始化 N 类，最近的两类合并。

模糊聚类：$J(e)=\sum_{i=1}^{c}\sum_{y\in Y}[\mu_i(y)]^b\|y-m_i\|^2$

隶属度$\sum_{y\in Y}[\mu_i(y)]=1$，$m_i=\frac{\sum_{y\in Y}[\mu_i(y)]^b y}{\sum_{y\in Y}[\mu_i(y)]^b}$，$i=1,.,c$

- 多分类器：提升单分类器性能。

Bagging Predictors 自举预测：用于好却**对噪声不稳定**的单个文类器。每次用子集训练，训练L个分类器，投票

AdaBoost：每次建立弱规则的分类器，通过加权（易分错的样本加权）求和生成强文类器。

输入：样本集$S=\{(\vec{x}_i,y_i)\}_{i=1}^{m}$，标签$y_i\in Y=\{1,\dots,K\}$，

带权弱分类器$y=Learn(\vec{x},\vec{p})$，弱分类器数量L

1)初始化：样本权$\vec{w}^{[1]}=\{1/m\}_{i=1}^{m}$

2)$for\ l=1\ to\ L$ //迭代变量$[l]$

$\vec{p}^{[l]}=\{p_i^{[l]}=w_i^{[l]}/\sum_i w_i^{[l]}\}_{i=1}^{m}$ //$\vec{p}$为归一化的$\vec{w}$

$\boldsymbol{h}^{[l]}(\vec{\boldsymbol{x}})=Learn(\vec{x},\vec{p}^{[l]})$ //训练带权弱分类

$\varepsilon^{[l]}=\sum_i p_i^{[l]}I(h^{[l]}(\vec{x}_i)\neq y_i)$ //计算分类误差

$if\ (\varepsilon^{[l]}>0.5)\ break\ with\ error$ //无可救药

$\boldsymbol{\beta}^{[l]}=\varepsilon^{[l]}/(1-\varepsilon^{[l]})$ //计算加权因子

$\vec{w}^{[l+1]}=\left\{w_i^{[l+1]}=w_i^{[l]}\left[\beta^{[l]}\right]^{1-I(h^{[l]}(\vec{x}_i)\neq y_i)}\right\}_{i=1}^{m}$ //更新

$next\ l$

3)$return\ \ h(\vec{x})=argmax_y\sum_{l=1}^{L}\left[ln\frac{1}{\beta^{[l]}}\right]I(h^{[l]}(\vec{x})=y)$

性质：1)$lim_{l\to\infty}\varepsilon^{[l]}=0$，2)泛化能力取决样本 margin

问题：弱问分类如何带样本权训练：数字带权，带权投票

AdaBoost 快速人脸检测：简单矩形特征，分级分类

Random subspace：每次从D随机选d特征，训练L个分类器，投票$h(\vec{x})=argmax_y\sum_{l=1}^{L}I(h^{[l]}(\vec{x})=y)$。

多分类器融合方法：1)决策层输出（C1: 3; C2: 1）：投票

2)排序层输出（C1: 3,2,1; C2: 1,3,2）：Borda 计数（No2=2）

3)度两侧输出（$C1: 1-0.6, 2-0.3, 3-0.1$）：和/积/均值

投票的弊病：排序票-最低票淘汰，其票数归第二名。

Vote1； 35%: A, B, C

33%: B, C, A

32%: C, A, B →C 淘汰 A wins

Vote2； 37%: A, B, C （A 努力，从 B 获取 2%的票）

31%: B, C, A

32%: C, A, B →B 淘汰 C wins

- 决策树：特征离散取值，非实数，无序，无距离。

特征用多元组4 $tuple$: $\{red, round, sweet, small\}$表示。

二叉树：多叉树可用多个二叉树代替。实数：分区间。

+ CART 算法：

$P(\omega_j)$是节点N处属于ω_j类样本占总样本数的比例.

熵不纯度：

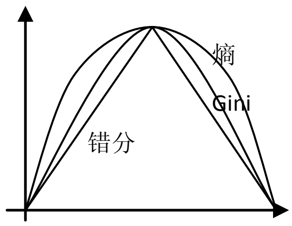
$$i(N)=-\sum_j P(\omega_j)log_2(P(\omega_j))$$

Gini 不纯度：

$$i(N)=1-\sum_j P^2(\omega_j)$$

错分不纯度：

$$i(N)=1-max_j P(\omega_j)$$



不纯度的变化：$\Delta i(N)=i(N)-\sum_{j=1}^{B}P(N_j)i(N_j)$

$i(N)$为上层节点的不纯度,

$P(N_j)$为分到j节点的元素占总元素的比例

$i(N_j)$为在下层某节点的不纯度

寻找查询使得$\Delta i(N)$最大。即<u>**不纯度函数取值最小**</u>.

+ **剪枝**：先让决策树充分生长，然后依次检测每相邻两个叶节点，如果将这两个叶节点合并造成的不纯度的增加小于某个阈值，则合并这两个叶节点。

+ **复杂度**: 假如给定n个d维训练样本，建树的平均时间复杂度为$O(dn(logn)^2)$。识别的平均时间复杂度为$O(logn)$。

+ ID3 算法

1. 可以考虑实数变量, 将其按区间划分.
2. 多叉树
3. 问题深度与样本数有关
4. 不考虑剪枝

+ C4.5 算法

1. 考虑剪枝: 建议一组规则, 对其排序.
2. 当高优先级满足时, 即退出

- 没有免费的午餐：没有一种完美的分类器。

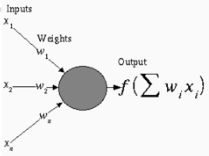
- 神经网络

训练网络：

随机选择初始权重

当误差较大时：对每一次训练

1.输入添加到网络；2.计算从输入层到隐藏层的输出；3.计算输出误差；4.反向传播调整



反向传播：单层

$e_j(n) = d_j(n) - y_n(n)$

$\mathrm{E}(n) = \frac{1}{2}\sum_{j\in C} e_j^2(n)$

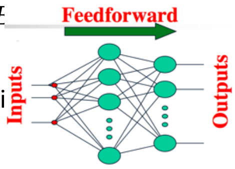
$v_j(n) = \sum_{i=0}^{p} w_{ij}(n)\, y_i(n) \quad y_j(n) = \varphi_j(v_j(n))$

$\frac{\partial E(n)}{\partial w_{ij}(n)} = \frac{\partial E(n)}{\partial e_{j(n)}}\frac{\partial e_j(n)}{\partial y_j(n)}\frac{\partial y_j(n)}{\partial v_j(n)}\frac{\partial v_j(n)}{\partial w_{ij}(n)} = -e_j(n)\varphi_j'(v_j(n))y_j(n)$

$w_i(n+1) = w_i(n) - \eta\nabla E$

多隐含层：

k：output unit　j:inner unit



$\mathrm{E}(n) = \frac{1}{2}\sum_{j\in C} e_j^2(n)$

$\frac{\partial E(n)}{\partial y_j(n)} = \sum_k e_k(n)\frac{\partial e_j(n)}{\partial y_j(n)} = \sum_k e_k(n)\frac{\partial e_k(n)}{\partial v_{k(n)}}\frac{\partial v_k(n)}{\partial y_j(n)}$

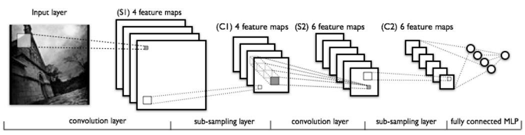
$\frac{\partial E(n)}{\partial y_j(n)} = -\sum_k -e_k(n)\varphi_k'(v_k(n))w_{kj}(n) = -\delta_k(n)w_{kj}(n)$

$v_j - y_j - v_k - y_k - e_k$：正向传播/反向传播

经验：使用 ReLU 函数；

SGD；训练集乱序；降 lr 训练；正则化防过学习

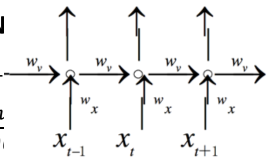
CNN: Convolutional Neural Networks　2D ConvNets 最好



RNN: Recurrent Neural N

$y_t = \varphi(t) \quad v_t = w_v y_{t-1}$

$\frac{\partial E_t}{\partial w_v} = \sum_{k=1}^{t}\frac{\partial E_t}{\partial y_t}\frac{\partial y_t}{\partial v_t}\frac{\partial v_t}{\partial v_k}\frac{\partial v_k}{\partial w_v}$



$\frac{\partial E_t}{\partial y_t} = -1$

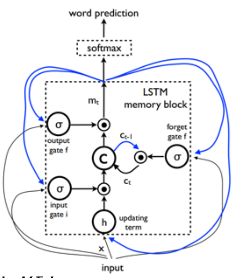
$\frac{\partial v_t}{\partial v_k} = \prod_{i=k+1}^{t}\frac{\partial v_t}{\partial y_{i-1}}\frac{\partial y_{i-1}}{\partial v_{i-1}} = \prod_{i=k+1}^{t} w_v\varphi'(t)$

LSTM: Long Short-Term M

Input Gate

Output Gate

Forget Gate



$o_t = \sigma(W_o x_t + U_o m_{t-1})$

$f_t = \sigma(W_f x_t + U_f m_{t-1})$

$i_t = \sigma(W_i x_t + U_i m_{t-1})$

$c_t = f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1})$

$m_t = o_t \odot c_t$

$p_{t+1} = Softmax(m_t)$

- 非线性降维：流型

ISOMAP

1.寻找每个点的 k 近邻（或一定距离内的点）$O(DN^2)$

2.计算与邻居的距离，定义图 G，$d_G(i,j) = d_x(i,j)$,不是近邻 $d_G(i,j) = \infty$；按照如下规则调整图 G:　$O(DN^3)$

$d_G(i,j) = \min\{d_G(i,j), d_G(i,k) + d_G(k,j)\}$

3.使用 MDS　$O(pN^2)$

a）计算$A = [-\frac{1}{2}d_G^2(\mathrm{i},\mathrm{j})\ ]$;

b）计算$B = HAH, H = I - n^{-1}ll^T, l = (1,1,\cdots,1)^T$;

c）计算 B 的特征值$\lambda_1,\lambda_2,\cdots,\lambda_{n-1}$和对应的特征向量$v_1,v_2,\cdots,v_{n-1}$,规范化使$v_i^T v_j = \delta_{ij}$;

d）p 维空间中，$\mathrm{X} = \mathrm{V}\Lambda^{\frac{1}{2}}, \Lambda^{\frac{1}{2}} = diag\left(\lambda_1^{\frac{1}{2}},\lambda_2^{\frac{1}{2}},\cdots,\lambda_p^{\frac{1}{2}},\right), \mathrm{V} = v_1,v_2,\cdots,v_P,$

LLE

1.为每个点找到 K 个近邻

2.计算权重矩阵 W，损失函数$\varepsilon = \sum_{i=1}^{N}|X_i - \sum_j W_{ij}X_j|^2$

3. $\phi(\mathrm{Y}) = \sum_{i=1}^{N}|\bar{Y}_i - \sum_j W_{ij}\bar{Y}_j|^2$　$\bar{Y}$为 d 维

$= \sum M_{ij}\bar{Y}_i\bar{Y}_j \quad M_{ij} = \delta_{ij} - w_{ij} - w_{ji} - \sum_k w_{ki}w_{kj}$

计算 M 的特征值由大到小$\lambda_1,\lambda_2,\cdots,\lambda_N$和对应的特征向量$v_1,v_2,\cdots,v_N$, $\mathrm{Y} = [v_1,v_2,\cdots,v_P]$

- 谱聚类

相似性图：G=（V,E）,其中顶点v_i对应于一个样本点x_i;

边权重$\mathrm{w}_{ij} \geq 0$;无连接$\mathrm{w}_{ij} = 0$。

v_i的度：$d_i = \sum_j^n w_{ij}$;度矩阵$\mathrm{D} = \mathrm{diag}(d_1,\cdots,d_n)$;

L=D-W;归一化：$L_{\mathrm{rw}} = D^{-1}L = I - D^{-1}W$

未归一化谱聚类：

1.输入：相似性矩阵$S \in R^{n\times n}$,类别数 k

2.构造相似性图，设加权邻接矩阵为 W

3.计算未归一化 Graph Laplacian L

4.计算 L 的前 k 个特征向量$u_1,\cdots,u_k$,并令$\mathrm{U} = [u_1,\cdots,u_k]$

5.$y_i \in R^k$为 U 的第 i 行构成的向量

6.使用 C-均值聚类方法将$y_i, \mathrm{i} = 1,\ldots,\mathrm{n}$,聚为 k 类$C_1,\cdots,C_k$

7.输出：最终聚类$A_1,\cdots,A_k$,其中$\mathrm{A}_i = \{j|y_j \in C_i\}$

归一化谱聚类:将步骤 4 中 L 改为L_{rw}，其他同上

[作业]===

- 贝叶斯决策作业

1.证明马氏距离$\mathrm{r}(\mathrm{a},\mathrm{b}) = \sqrt{(a-b)^T\Sigma^{-1}(a-b)}$满足距离定义

先证方差矩阵Σ正定：$\forall a \in R^n$,

$a^T\Sigma a = \sum_{i=1}^{N}\sum_{j=1}^{N}c_i c_j Cov(x_i,x_j) = E[\sum_{i=1}^{N}c_i(x_i - Ex_i)]^2 \geq 0$

（等号成立$\Leftrightarrow x_i = \mathrm{E}x_i \Leftrightarrow Rank(\Sigma) = 1 \Rightarrow \Sigma$不可逆，矛盾）

$\Rightarrow \Sigma$正定$\Rightarrow \Sigma^{-1}$正定$\Rightarrow \exists P$可逆$, st\ \Sigma^{-1} = P^T P$

则$r(a,b) = \sqrt{(P(a-b))^T P(a-b)}$

6.正态分布$p(x)\sim N(\mu,\Sigma)$，$q(x)\sim N(m,L)$，**其 K-L 距离**

$$KL(p,q)=-\int p(x)\,ln\frac{q(x)}{p(x)}dx$$

$$=-\frac{1}{2}\{ln\frac{|\Sigma|}{|L|}+\frac{d}{2}-tr(L^{-1}\Sigma)-(\mu-m)^TL^{-1}(\mu-m)\}$$

7.高维正态$\chi=\{x_k\}_{k=1}^N$，$p(x|\mu)\sim N(\mu,\Sigma^2)$，$p(\mu)\sim N(\mu_0,\Sigma_0^2)$，求

$$p(\mu|\chi)=\frac{p(\chi|\mu)p(\mu)}{\int p(\chi|\mu)p(\mu)du}=\alpha\prod_{i=1}^{N}p(x_k|\mu)p(\mu)$$

$$=\alpha^{'}\,exp\{-\frac{1}{2}[\sum_{k=1}^{N}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)$$

$$+(\mu-\mu_0)^T\Sigma_0^{-1}(\mu-\mu_0)]\}$$

$p(\mu|\chi)$也是高斯分布，设$p(\mu|\chi)\sim N(\mu_N,\Sigma_N)$，则有

$$\mu_N=(N\Sigma^{-1}+\Sigma_0^{-1})^{-1}(N\Sigma^{-1}\bar{x}+\Sigma_0^{-1}\mu_0)$$

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

- 参数估计作业

1.样本$\chi=\{x_k\}_{k=1}^N$，估计参数$p(x|P)=P^x(1-P)^{1-x}$，则似然

$$H(p)=ln\,l(p)=\sum_{k=1}^{N}ln\,p^{x_k}(1-P)^{1-x_k}=\sum_{k=1}^{N}[x_k\,ln\,p+(1-x_k)\,ln(1-p)]$$

$$\frac{\partial H(p)}{\partial p}=0\Rightarrow\hat{p}=\frac{1}{N}\sum_{k=1}^{N}x_k$$

2.参数估计$x_i\in\{1,\dots,m\}$，$\sum_{i=1}^d x_i=m$，$\theta_i\epsilon(0,1)$，$\sum_{i=1}^d\theta_i=1$

$$p(x|\theta)=\frac{m!\prod_{i=1}^d\theta_i^{x_i}}{\prod_{i=1}^d(x_i!)}$$

似然

$$H(\theta)=ln\,l(\theta)=\sum_{k=1}^{N}\sum_{i=1}^{d}x_i^{(k)}ln\theta_i+ln\,m!-\sum_{i=1}^{d}ln(x_i^{(k)}!)+\lambda(\sum_{i=1}^{d}\theta_i-1)$$

令$\frac{\partial H(\theta)}{\partial\theta_i}=\frac{\sum_{k=1}^N x_i^{(k)}}{\theta_i}+\ \lambda=0\ \Rightarrow\ \theta_i=-\frac{\sum_{k=1}^N x_i^{(k)}}{\lambda}$，又$\sum_{i=1}^d\sum_{k=1}^N x_i^{(k)}=$

$\sum_{i=1}^d(-\lambda\theta_i)=mN$得$\lambda=-m$，则$\theta_i=\frac{\frac{1}{N}\sum_{k=1}^N x_i^{(k)}}{m}$

3.**错误模型的 MLE 估计**：第 1 类$p(x|w_1)\sim N(0,1)$，第 2 类实际 $p(x|w_2)\sim N(1,10^6)$却假设为$\hat{p}(x|w_2)\sim N(\mu,1)$。则通过 MLE 估计$\hat{\mu}=\frac{1}{N}\sum_{k=1}^N x_k\xrightarrow{N\to\infty}1$。此时贝叶斯决策面$t=\frac{1+\hat{\mu}}{2}=0.5$，而实际决策面$t=\pm3.72$。

4.**EM 参数估计**：均匀分布x_1,x_2独立，设丢失数据$Y=(y_1,y_2)$ 则 E-STEP：初始$\theta_{[0]}=(0,0,10,10)^T$

$$Q(\theta_{[0]},\theta)=\int p(y|x,\theta_{[0]})\,ln\,p(x,y|\theta)\,dy$$

其中 ????? $-5ln[|\theta_3-\theta_1||\theta_4-\theta_2|]$; $\theta_{1,2}\le0,\theta_{3,4}\ge10$何来

$$p(y|x,\theta_{[0]})=\frac{p(y,x|\theta_{[0]})}{p(x|\theta_{[0]})}=\frac{\prod_{i=1}^5|10-0|^{-2}}{\prod_{i=1}^4|10-0|^{-2}}=10^{-2}$$

$$ln\,p(x,y|\theta)=\begin{cases}-5ln[|\theta_3-\theta_1||\theta_4-\theta_2|];\ \theta_{1,2}\le\cdots,\theta_{3,4}\ge\cdots\\-\infty\end{cases}$$

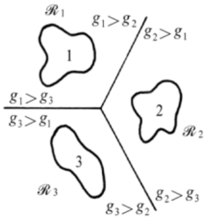
M-STEP：$\theta_{[1]}=arg\,max_\theta\,Q(\theta_{[0]},\theta)$

$$=arg\,min_\theta|\theta_3-\theta_1||\theta_4-\theta_2|,\theta_{1,2}\le0,\theta_{3,4}\ge10$$

$$=(0,0,10,10)^T=\theta_{[0]}$$，收敛。

此题目实际上不适用于 EM 算法，该结果与初始直接相关。

- 线性分类器作业



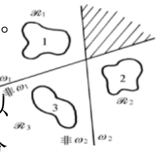
1.线性可分：对c类构建c个线性函数$g_i(x)$

样本可通过$i^*=argmax_i\,g_i(x)$正确分类

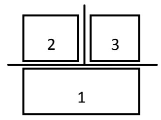
完全线性可分：构建$c-1$个（i-非 i）型

线性判别面，每个类都能和非本类线性可分。

完全线性可分⟹线性可分



2.成对线性可分：存在$c(c-1)/2$个超平面，可以将每两个类线性分开。成对线性可分⇏线性可分



3.两个同方差的正态分布，对数似然判别等价于 Fisher 判别（取特定阈值）。

证明：正态分布对数似然

$$g_i(x)=-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)+ln\,P(w_i)$$

Fisher 准则：$S_w=\sum S_i=2\Sigma$，$y=w^Tx$，$w=S_w^{-1}(m_1-m_2)$

则$y=h(x)=w^Tx=\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}x$

分类器$g(x)=g_1(x)-g_2(x)$

$$=(\mu_1-\mu_2)^T\Sigma^{-1}x-\frac{1}{2}[\mu_1^T\Sigma^{-1}\mu_1-\mu_2^T\Sigma^{-1}\mu_2]+\ln\frac{P(w_1)}{P(w_2)}$$

$=2h(x)+c$，选取 Fisher 阈值$-c/2$，等价 ???

6.马氏距离$\delta_M(x,y)=\sqrt[S]{\sum_{j=1}^d|x_j-y_j|^S}$，证明$S\ge1$时满足距离定义；证明$0<S<1$前 3 条满足，而当$(x_i-y_i)(y_i-z_i)\ge0$时，$\delta_M(x,z)\ge\delta_M(x,y)+\delta_M(y,z)$。

经典证明，复杂。

- 分界面/特征选择作业

1.**最近邻方法的分界面**一定是**分段线性的**。

K 近邻方法的分界面：样本有限时是分段线性的。待证明

2.证明欧式距离经过线性变换还是距离。即证明$\delta(x,y)=\sqrt{\sum_{j=1}^d|x_j-y_j|^2\alpha_j^2}$是距离：1) $\delta(x,y)\ge0$，等号$\Leftrightarrow x=y$。

2) $\delta(x,y)=\delta(y,x)$。3) $\delta(x,z)=\sqrt{\sum_{j=1}^d|x_j-z_j|^2\alpha_j^2}=$

$$\sqrt{\sum_{j=1}^d|x_j-y_j+y_j-z_j|^2\alpha_j^2}\le\sqrt{\delta^2(x,y)+\delta^2(y,z)}$$

4.从$p(x|w_1)$中抽取x_1，从从$p(x|w_2)$中抽取x_2建立最近邻分类器，从$p(x|w_1)$中抽取x测试，计算错误率$P_1(e)$。

建立最近邻分类器，分界面$t=(x_1+x_2)/2$，错分情况有

(1) $x<t,x_1>x_2$

(2) $x>t,x_1<x_2$，$P_1(e)=P_{1(1)}(e)+P_{1(2)}(e)$

其中

$$P_{1(1)}(e)=\int_{-\infty}^{+\infty}p(x_1|w_1)\left[\int_{-\infty}^{x_1}p(x_2|w_2)\left(\int_{-\infty}^{(x_1+x_2)/2}p(x|w_1)dx\right)dx_2\right]dx_1$$

$$P_{1(2)}(e)=\int_{-\infty}^{+\infty}p(x_2|w_2)\left[\int_{-\infty}^{x_2}p(x_1|w_1)\left(\int_{(x_1+x_2)/2}^{+\infty}p(x|w_1)dx\right)dx_1\right]dx_2$$

可以考虑先去顶测试样本 x，再去定其最近邻（两种情况），最后扫描其他样本。此方法适用于取 N 个样本。

- Assignments for PCA and K-Means

1.两类 K-L 变换降维。类均值$\vec{m}_1, \vec{m}_2$，类内总离散$S_w = S_1 + S_1$
特征分解$U^T S_w U = \Lambda$，$\Lambda = diag(\lambda_1, \dots, \lambda_n)$，$U = (\vec{u}_1, \dots, \vec{u}_n)$
由于对S_w分解，希望类内离散度小，因此取$\lambda_{\min}$对应的$\vec{u}$。
降维主成分$Y_i = (\vec{y}_1, \dots, \vec{y}_m) = Q(X_i - \vec{m}_i)$

2. K-L 变换的坐标轴不变性。由$S_w = P_1\Sigma_1 + P_2\Sigma_2$且$B^T S_w B = I$
代入$P_1 B^T \Sigma_1 B + P_2 B^T \Sigma_2 B = I \Rightarrow P_1 B^T \Sigma_1 B = I - P_2 B^T \Sigma_2 B$ (*)
设U_1是$P_1 B^T \Sigma_1 B$产生的 K-L 坐标轴，即$[P_1 B^T \Sigma_1 B]U_1 = \lambda_1 U_1$
代入(*)式$[P_2 B^T \Sigma_2 B]U_1 = (1 - \lambda_1)U_1 \triangleq \lambda_2 U_1$。
即 1)坐标轴相同，2)特征矩阵满足$\Lambda_1 = I - \Lambda_2$

3. 证明$f(x) = \sum_{k=1}^{N}(x_k - x)^T \Sigma^{-1}(x_k - x)$在$\{x_k\}_{k=1}^{N}$的均值处取得最小值。求导

$$\frac{df(x)}{dx} = \sum_{k=1}^{N}[\Sigma^{-1} + (\Sigma^{-1})^T](x - x_k) = N[\Sigma^{-1} + (\Sigma^{-1})^T]\left(x - \frac{1}{N}\sum_{k=1}^{N}x_k\right) = 0$$

由Σ非奇异,Σ^{-1}可逆,$trace(\Sigma^{-1}) \neq 0$,$[\Sigma^{-1} + (\Sigma^{-1})^T] \neq \mathbf{0}_{N\times N}$

则$x = \frac{1}{N}\sum_{k=1}^{N}x_k$时$f(x)$取得极值,又$f(x)$为二次函数,最小值。

4.证明将一个聚类H（元素数$N \geq 2$）分为两个非空聚类，会使得目标函数$J_e = \sum_i\left[\sum_{x_k \in H_i}\|x_k - m\|^2\right]$减小（空聚类$H_i$不计入）

聚类H：$m = \frac{1}{N}\sum_{k=1}^{N}x_k$，$J_e = \sum_{k=1}^{N}\|x_k - m\|^2$。

拆分H_1：$m_1 = \frac{1}{N-1}\sum_{k=1}^{N-1}x_k$，$J_{e1} = \sum_{k=1}^{N-1}\|x_k - m_1\|^2$

H_2：$m_2 = x_N$，$J_{e2} = \|x_N - m_2\|^2 = 0$

有$J_e = \sum_{k=1}^{N-1}\|x_k - m\|^2 + \|x_N - m\|^2$

$$= \sum_{k=1}^{N-1}\|x_k - m\|^2 + \frac{1}{N-1}\sum_{k=1}^{N-1}\|x_N - m\|^2$$

$$\geq \sum_{k=1}^{N-1}\left\|x_k - m + \frac{1}{N-1}(x_N - m)\right\|^2$$

$$= \sum_{k=1}^{N-1}\|x_k - m_1\|^2 + 0 = J_{e1} + J_{e2}$$

5.PCA 特征值：记$\lambda, \vec{u}$为$S = XX^T$的特征值和向量，即$Su = \lambda u$
两边左乘X^T，$(X^T X)X^T = \lambda X^T u$。记$G = X^T X, \vec{v} = X^T \vec{u}$，则有$Gv = \lambda v$。$\Rightarrow S$特征值是$G$特征值。同理：$G$也是$S$特征值。

优缺点==

贝叶斯：

优点：传统分类方法，在理论上满足分类错误率最小，对于服从特定模型的样本有较好的分类结果，是其他分类算法分析的理论基础。

缺点：依赖模型（类先验概率，类条件概率分布的形式和具体参数），因此模型可能选错，模型的参数也可能过拟合，从而导致分类效果下降。

SVM：

优点：将低位空间线性不可分问题变换到高维空间，使其线性可分，由于只需要进内积计算，并没有增加多少计算复杂度，推广能力与变换空间维数无关，具有较好的推广能力，相对于传统方法，对模型具有一定的不敏感性。

缺点：缺少理论证明，VC 维一般情况下如何计算和估计的问题还没有得到解决。

近邻法：

优点：错误率在贝叶斯错误率 P*和两倍贝叶斯错误率 2P*之间，算法直观容易理解，易于实现，可以使用任何分布的样本，算法适用性强。

缺点：需将所有样本存入计算机中，每次决策都要计算待识别样本 x 与全部训练样本的距离并进行比较，存储和计算开销大；当错误的代价很大时，会产生较大风险。错误率的分析是渐进的，这要求样本为无穷，实际中这一条件很难达到。

分级聚类：

最近距离：

优点：能正确处理带状分布的样本

缺点：两类样本中间又几个隔得比较近的样本是，两类样本会被错误地聚为一类。

最远距离：优缺点正好与最近距离聚类方法相反

均值距离：效果介于以上两者之间。

2.正态分布多类分类器

判别函数$g_i(x) = \ln p(\vec{x}|w_i) + \ln P(w_i)$

$$= -\frac{d}{2}ln2\pi - \frac{1}{2}ln|\Sigma_i| + lnP(w_i) - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$$

分类$c = argmax_i\, g_i(x)$

3.第 1 类错误$P_1(e) = \int_{R_2} p(\vec{x}|w_1)\, d\vec{x}$(实际 1 类却分为 2 类)

4.正态$N(\mu_i, \sigma^2 I)$两类分类器，先验相等，证$P(e) = \frac{\int_a^\infty e^{-u^2/2}du}{2\pi}$

其中$a = \|\mu_2 - \mu_1\|/2\sigma$。证明：存在正交单位阵$Q$，$\vec{y} = Q\vec{x}$

且$Q(\mu_2 - \mu_1) = (\|\mu_2 - \mu_1\|, 0,0,\dots,0)$，（$Q$距离不变性）

5.0/1 变量$x \in R^d$, $p_{ij} = P(x_i = 1|w_j)$，求其贝叶斯决策$g_j(x)$。

$$p(\vec{x}|w_i) = \left[\prod_{i=1,x_i=1}^{d} p_{ij}\right]\left[\prod_{i=1,x_i=0}^{d} (1-p_{ij})\right] = \left[\prod_{i=1}^{d} x_i p_{ij}\right]\left[\prod_{i=1}^{d}(1-x_i)(1-p_{ij})\right]$$

$g_j(x) = ln\, p(\vec{x}|w_j) + ln\, P(w_j)$

$$= \sum_{i=1}^{d} x_i\, ln\frac{p_{ij}}{1-p_{ij}} + \sum_{i=1}^{d} ln(1-p_{ij}) + ln\, P(w_j)$$