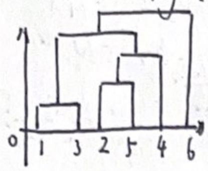
多统期末cheat sheet.

清華大學

L12 聚类分析. 1. Hierachical

Agglomerative 自下而上.

距离
- 欧式
- single (MIN): 对两类中最近的点
- complete (MAX): 不易受异常值影响
- Group Average
- Ward's Linkage (SSE)

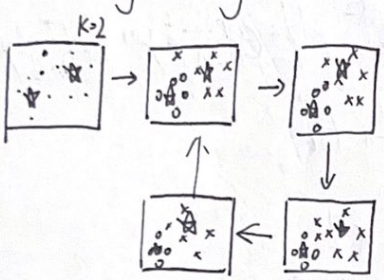


x方向上坐标无需有序

y方向上高低一定严格不同.

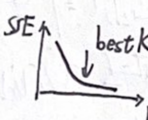
不足: 计算复杂度高, 不能纠错.

2. Partitioning Clustering (K-Means)



Picking K:

① $SSE=\sum_{i=1}^{k}\sum_{x\in C_i} dist^2(m_i,x)$



ELbow

② Empirical $k\approx\sqrt{\frac{n}{2}}$

③ Cross validation.

L11 LDA判别分析.

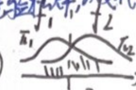
Fisher Approach: Assumption $\Sigma_1=\Sigma_2$

$$\max_{\vec{a}} \frac{(\vec{a}^T(\bar{x}_1-\bar{x}_2))^2}{\vec{a}^T S_{pooled}\vec{a}} \qquad \hat{a}=S_{pooled}^{-1}\vec{d} \qquad \vec{d}=\bar{x}_1-\bar{x}_2$$

$$S^2_{pooled}=\frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2} \qquad y=\hat{a}^T\vec{x}$$

maximum $D^2=(\bar{x}_1-\bar{x}_2)^T S_{pooled}^{-1}(\bar{x}_1-\bar{x}_2)$

ECM方法: p_1,p_2为π_1,π_2先验概率 需知f_1,f_2,先验概率,以及代价.



$$ECM\triangleq C(2|1)P(x\in\pi_1,x\in R_2)+C(1|2)P(x\in\pi_2,x\in R_1)$$

$$=C(2|1)p_1+\int_{R_1}[C(1|2)p_2f_2(x)-C(2|1)p_1f_1(x)]dx \text{ 求 } \min.$$

$$\Rightarrow R_1=\left\{x: \frac{C(1|2)p_2}{C(2|1)p_1}<\frac{f_1(x)}{f_2(x)}\right\} \qquad R_2=R_1^C$$

正态分布下的结论

$\Sigma_1=\Sigma_2$ $f_i(x)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-u_i)^T\Sigma^{-1}(x-u_i)\right\}$

$$R_1: (u_1-u_2)^T\Sigma^{-1}x-\frac{1}{2}(u_1-u_2)^T\Sigma^{-1}(u_1+u_2)\geq \ln\left[\left(\frac{C(1|2)}{C(2|1)}\right)\frac{p_2}{p_1}\right]$$

$$\text{sample}\Rightarrow (u_1-u_2)^T S_{pooled}^{-1}x-\frac{1}{2}(\bar{x}_1-\bar{x}_2)S_{pooled}^{-1}(\bar{x}_1+\bar{x}_2)\geq \ln\left[\frac{C(1|2)}{C(2|1)}\left(\frac{p_2}{p_1}\right)\right]$$

if $\frac{C(1|2)}{C(2|1)}\left(\frac{p_2}{p_1}\right)=1 \Rightarrow \hat{y}\geq\frac{1}{2}(\bar{y}_1+\bar{y}_2)\Rightarrow LDA$

$\Sigma_1\neq\Sigma_2$ QDA.

$$R_1: -\frac{1}{2}x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x+(u_1^T\Sigma_1^{-1}-u_2^T\Sigma_2^{-1})x-k\geq \ln\left[\left(\frac{C(1|2)}{C(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

APER: Confusion matrix

		predict membership π_1	π_2	
Actual membership	π_1	n_{1C}	n_{1M}	n_1
	π_2	n_{2M}	n_{2C}	n_2

$APER=\frac{n_{1M}+n_{2M}}{n_1+n_2}$

Cross validation: 样本均分为n份,每份中用其他样本作为训练集,该份验证.

L8 & 9 因子分析.

$$\psi=\begin{bmatrix}\psi_1 & & \\ & \ddots & \\ & & \psi_p\end{bmatrix}$$

$$\underset{p\times 1}{X-u}=\underset{p\times m}{L}\underset{m\times 1}{F}+\underset{p\times 1}{\varepsilon} \qquad E(F)=0 \quad E(\varepsilon)=0 \quad Cov F=I_{m\times m} \quad Cov(\varepsilon)=\underset{p\times p}{\psi}$$

$$Cov(\varepsilon,F)=\underset{p\times m}{0}.$$

PCA法

$$S=\underset{p\times m}{\hat{L}}\hat{L}^T+\underset{p\times p}{\hat{\psi}} \qquad \hat{L}=[\sqrt{\hat{\lambda}_1}e_1,\cdots\sqrt{\hat{\lambda}_m}e_m]$$

special variance: $S-\hat{L}\hat{L}^T$

$$l_{1i}^2+\cdots+l_{pi}^2=\hat{\lambda}_i$$

proportion $\frac{\hat{\lambda}_1+\cdots\hat{\lambda}_m}{tr(S)}$

MLE法：

$$L(\mu,\Sigma)=(2\pi)^{-\frac{(n-1)p}{2}}|\Sigma|^{-\frac{(n-1)}{2}}\exp\{-\frac{1}{2}tr[\Sigma^{-1}(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^T)]\}$$

$$\times(2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\{-\frac{n}{2}(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)\}$$

假设：common factors 和 specific factors 都是正态的

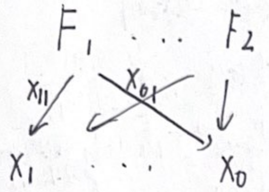
$F\sim N_m(0,I)$ $\varepsilon\sim N_p(0,\psi)$ $F\perp\varepsilon$.

communalities:

$\hat{h}_i^2=\hat{l}_{i1}^2+\cdots+\hat{l}_{im}^2$ for $i=1,2,\ldots p$ 共性方差

$\psi_i=\sigma_{ii}-\hat{h}_i^2$ 特殊方差.

$$\Sigma=\begin{bmatrix}\sigma_{ii}&\\&\end{bmatrix}=\begin{bmatrix}l_i^T\\\end{bmatrix}\begin{bmatrix}l_i&\end{bmatrix}+\begin{bmatrix}\psi_i&\\&\end{bmatrix}$$



$\left(\begin{matrix}\text{Proportion of total sample}\\\text{variance due to } j\text{th factor}\end{matrix}\right)=\dfrac{\hat{l}_{1j}^2+\hat{l}_{2j}^2+\cdots+\hat{l}_{pj}^2}{s_{11}+\cdots+s_{pp}}$ 因子j的解释能力.

$\hat{h}_i^2=\hat{l}_{i1}^2+\hat{l}_{i2}^2+\cdots+\hat{l}_{im}^2$

$$\Sigma=E((X-\mu)(X-\mu)^T)=E((LF+\varepsilon)(LF+\varepsilon)^T)$$
$$=E(LFF^TL^T+\varepsilon F^TL^T+LF\varepsilon^T+\varepsilon\varepsilon^T)$$
$$=LE(FF^T)L^T+E(\varepsilon\varepsilon^T)=LIL^T+\Psi=LL^T+\Psi$$

$\hat{L}^*=V^{\frac{1}{2}}\hat{L}_z$ $\hat{\psi}^*=V^{\frac{1}{2}}\hat{\Psi}_zV^{\frac{1}{2}}$ $V=diag(\Sigma)=diag(\sigma_{ii})$

V即正则化项.

两种做法优劣：PC：随m增大计算方便

MLE：相关数与协方差之间L和Ψ可直接转化.

选择更合适的：看可解释性.

因子个数限制 $m<p$.

Fisher Approach

目标：$\dfrac{\bar{y}_1-\bar{y}_2}{s_y}$ $s_y^2=\dfrac{\sum(y_{1j}-\bar{y}_1)^2+\sum(y_{2j}-\bar{y}_2)^2}{n_1+n_2-2}$

优化：$\max\limits_a\dfrac{(\bar{y}_1-\bar{y}_2)^2}{s_y^2}$

$\hat{y}=\hat{a}^Tx=(\bar{x}_1-\bar{x}_2)^TS_{pooled}^{-1}X$ s.t. $\max\limits_a\dfrac{(a^T(\bar{x}_1-\bar{x}_2))^2}{a^TS_{pooled}a}$

典型相关分析：

$X^{(1)}=\begin{bmatrix}X_1^{(1)}\\\vdots\\X_p^{(1)}\end{bmatrix}$ $X^{(2)}=\begin{bmatrix}X_1^{(2)}\\\vdots\\X_q^{(2)}\end{bmatrix}$ $\Sigma=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix}$ $X^{(1)}_{p\times p}$ $X^{(12)}$ $X^{(2)}_{q\times q}$

目标，找到最佳 $\vec{a},\vec{b}$ st. $corr(U,V)$ 最大 $U=a^TX^{(1)}$ $V=b^TX^{(2)}$

$\Rightarrow\max\limits_{a,b\neq0}\dfrac{a^T\Sigma_{12}b}{\sqrt{a^T\Sigma_{11}a}\sqrt{b^T\Sigma_{22}b}}=\max\dfrac{\tilde{a}^T\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}\tilde{b}}{\sqrt{\tilde{a}^T\tilde{a}}\sqrt{\tilde{b}^T\tilde{b}}}=\max\limits_{|\tilde{a}|=|\tilde{b}|=1}\tilde{a}^T\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}\tilde{b}$

$\tilde{a}=\Sigma_{11}^{\frac{1}{2}}a$ $\tilde{b}=\Sigma_{22}^{\frac{1}{2}}b$

不要忘记开方.

对 $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ 作奇异值分解 ⇒ 求 $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$ 特征值

第ith特征值为 ρ_i^{*2} $U=e_i^T\Sigma_{11}^{-\frac{1}{2}}X^{(1)}$ $V=f_i^T\Sigma_{22}^{-\frac{1}{2}}X^{(2)}$

$f_i=\dfrac{1}{\rho_i^*}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}\vec{e}_i$

$U=AX^{(1)}$ $V=BX^{(2)}$

$Cov(U,V)=A\Sigma_{12}B^T=\begin{bmatrix}\rho_1^*&&\\&\ddots&\\&&\rho_p^*\end{bmatrix}$

$Cov(U,U)=A\Sigma_{11}A^T=I$ $Cov(V,V)=B\Sigma_{22}B^T=I$

$Var(U_k)=Var(V_k)=1$ $cov(U_k,U_l)=corr(U_k,U_l)=0$ $k\neq l$

$cov(V_k,V_l)=corr(V_k,V_l)=0$ $k\neq l$

$corr(U,X^{(1)})=A\Sigma_{11}V_{11}^{-\frac{1}{2}}$

$Corr(U,X^{(2)})=A\Sigma_{12}V_{22}^{-\frac{1}{2}}$ $Cov(U,X^{(1)})=Cov(AX^{(1)},X^{(1)})=A\Sigma_{11}$

$Corr(V,X^{(2)})=B\Sigma_{22}V_{22}^{-\frac{1}{2}}$ ⇐ $Corr(U,X^{(1)})=Cov(U,V_{11}^{-\frac{1}{2}}X^{(1)})$

$=A\Sigma_{11}V_{11}^{-\frac{1}{2}}$

$Corr(V,X^{(1)})=B\Sigma_{21}V_{11}^{-\frac{1}{2}}$ $V_{11}=diag(\Sigma_{11})$

标准化不会影响结果

从几何上来看，标准化后 $X^{(1)}$ $X^{(2)}$ 张成空间不变，∴ U_1,V_1 也不变.