Cluster Analysis

(1) Similar Coefficient:

① Transf of Data:

If we have data $X = (x_{ij})_{n\times m}$. $(x_{ij})_{j=1}^{m}$ is from π_i. Then we obtain:

mean: $\bar{x}_j = \sum_i x_{ij}/n$. SD: $s_j = \sqrt{\frac{1}{n-1}\sum_i (x_{ij}-\bar{x}_j)^2}$

Range: $R_j = \max_i x_{ij} - \min_i x_{ij}$.

i) Centralization:

$x^*_{ij} = x_{ij} - \bar{x}_j$. i.e. $X^* = X - \frac{1}{n} J_n^n X$.

$S^* = \sum_{t=1}^{n} x^*_{ti} x^*_{tj}/n-1 = S$ (invariant)

ii) Standardilization:

$x^*_{ij} = (x_{ij} - \bar{x}_j)/s_j$, indept with unit.

iii) Range Standardilization:

$x^*_{ij} = (x_{ij} - \bar{x}_j)/R_j$. indept with unit.

Rmk: Normalization: $x^*_{ij} = (x_{ij} - \min_i x_{ij})/R_j$

iv) logarithm:

$x^*_{ij} = \log x_{ij}$. where $x_{ij} > 0$. For linearity.

Def: i) Minkowski Dist: $d_{ij}(p) = \left(\sum_{t=1}^{m} |x_{it} - x_{jt}|^p\right)^{\frac{1}{p}}$

ii) Lance Dist: $d_{ij}(L) = \frac{1}{m}\sum_{t=1}^{m} |x_{it} - x_{jt}| / (x_{it} + x_{jt})$

iii) Bias Dist: $d_{ij} = \frac{1}{m}\left(\sum_{k}\sum_{t}^{m} (x_{ik} - x_{jk})(x_{it} - x_{jt}) r_{kt}\right)^{\frac{1}{2}}$

Rmk: i)、ii) are often used in indept r.v.'s but iii) isn't. ii) isn't sensitive with large eigenvalue.

Def: 在数量化理论，特定性变量称为项目，而项目中不同的值称为类目.

eg. 若有 n 个样品 X_i $(1 \le i \le n)$，m 个项目.

k^{th} 项目有 r_k 个类目，则定义 X_i 的取值：

$X_i = (\delta_i(k,1) \cdots \delta_i(k,r_k))$，$1 \le k \le m$.

$$\delta_i(k,l) = \begin{cases} 1, & i^{th} \text{ sample 属于 } k^{th} \text{ 项目 } l^{th} \text{ 类目.} \\ 0, & \text{otherwise.} \end{cases}$$

Def: X_i couples with X_j on k^{th} 项目 l^{th} 类目 if $\delta_i(k,l) = \delta_j(k,l)$. 若都为 1 则称为 1-1 配对. 若都为 0，则称为 0-0 配对.

Def: In qualitative case, Distance between X_i and X_j is defined by:

i) m_1 为 X_i、X_j 的 1-1 配对数；m_0 为 0-0 配对数，m_- 为不配对数. $(m_1 + m_0 + m_- = \sum_{k=1}^{m} r_k)$

$d_{ij} = m_2 / m_1 + m_-$

ii) $d_{ij}^2 = \sum_{k}^{m}\sum_{l}^{r_k} (\delta_i(k,l) - \delta_j(k,l))^2$. i.e. 不配对的次数.

Rmk: 存在情况：sample x_i 属于 k^{th} 项目，不同的类目，若不能同时兼取时。

Define distance in ij：m_{ij}^* 为不同比对项目总数。$d_{ij} = m_{ij}^*/m$.

Def: Similar coefficient between x_i, x_j is c_{ij}.

satisfies i) $|c_{ij}| \le 1$. $\forall i,j$ ii) $c_{ij} = \pm 1 \Leftrightarrow x_i = a x_j$. $a \neq 0$.

iii) $c_{ij} = c_{ji}$ iv) $|c_{ij}| \uparrow 1$. x_i, x_j are more similar.

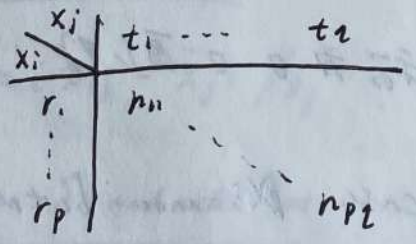
e.g. i) $c_{ij}(1) = \dfrac{x_i^T x_j}{\sqrt{(x_i^T x_i)(x_j^T x_j)}}$ (cosine)

ii) $c_{ij}(2) = r_{ij} = \Sigma (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j) \Big/ \sqrt{\Sigma (x_{ti} - \bar{x}_i)^2 \Sigma (x_{tj} - \bar{x}_j)^2}$

Rmk: Define distance by c_{ij}:

$d_{ij} = 1 - |c_{ij}|$. or $d_{ij}^2 = 1 - c_{ij}^2$.

iii) In qualitative case: 对 x_i 有 p 个项目，取值为 $(r_1 \cdots r_p)$，x_j 有 q 个项目取值为 $(t_1 \cdots t_q)$.



n_{kl} = number of (x_i, x_j) $= (r_k, t_l)$

$$\chi_{ij}^2 = n\left(\sum_k \sum_l \left(\frac{n_{kl}^2}{n_{k\cdot} n_{\cdot l}} - 1\right)\right)$$

Def: $c_{ij} = \sqrt{\chi_{ij}^2 / \chi_{ij}^2 + n}$

(2) Hierarchical Method:

If we have samples $X_i = (X_{i1} \cdots X_{im})$. $1 \le i \le n$.

The ideal is seeing n samples as n classes initally. Then define a distance. Classify two sample as a class which has minimal distance. Repeat the procedure...

Process: i) Data Transform ii) Calculate the matrix of distance $(d_{ij})_{n\times n} \triangleq D_0$

iii) If $(k,l) = \arg\min_{i,j} d_{ij}$. Then merge X_k, X_l as a class $\{X_k, X_l\}$.

iv) Calculate the distance w.r.t remaining classes to obtain $(d'_{ij})_{m\times m} \triangleq D_1$

v) Repeat the procedure until remain one class.

e.g. $X = (X_1 \cdots X_5) = (1.2, 4.5, 6, 8)$. $d_{ij} = |X_i - X_j|$.

i) $D^0 = \begin{pmatrix} 0 & \boxed{1} & 3.5 & 5 & 7 \\ & 0 & 2.5 & 4 & 6 \\ & & 0 & 1.5 & 3.5 \\ & & & 0 & 2 \\ & & & & 0 \end{pmatrix}$ $G_i = \{X_i\}$. $1 \le i \le 5$, $k = 5$

ii) d_{12} is min $\Rightarrow$ Merge X_1, X_2 as $CL4 = \{X_1, X_2\}$

iii) $D' =$

	X_3	X_4	X_5	$CL4$
X_3	0	$\boxed{1.5}$	3.5	2.5
X_4		0	2	4
X_5			0	6
$CL4$				0

$(d(X_i, CL4) = \min_{X_j \in CL4} d(X_i, X_j))$

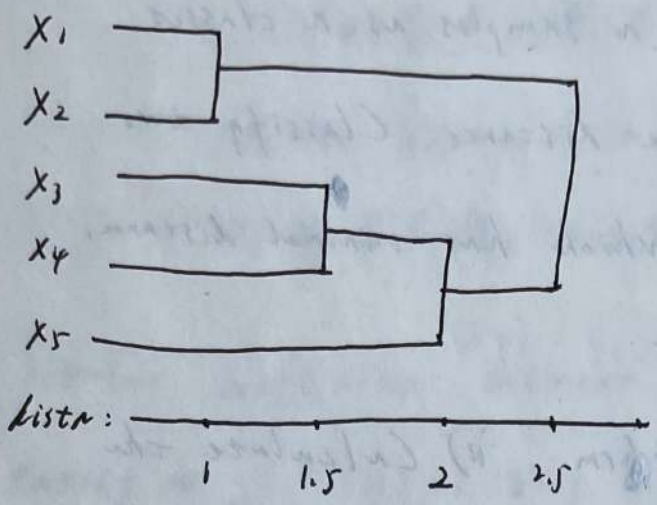
d'_{12} is min $\Rightarrow$ $CL3 = \{X_3, X_4\}$

iv) $D^2 =$

	X_5	$CL4$	$CL3$
X_5	0	6	$\boxed{2}$
$CL4$		0	2.5
$CL3$			0

$\Rightarrow$ $CL2 = \{X_5, X_3, X_4\}$.

v) Finally. $CL1 = \{CL2, CL4\}$.



Hierarchical Graph

iv) We can determine the number of groups by requiring the distance of group $< d_0$

Eg. If require $d < d_0 = 2$. Then classify:

$\{X_1, X_2\}$. $\{X_3, X_4\}$. $[X_5]$. $k = 3$.

$d < 3$: $[X_1, X_2]$, $[X_3, X_4, X_5]$, $k = 2$.

① Distance of Groups:

i) $D_{pq}^{L} = d(G_p, G_q) =: \min\{d_{ij} \mid X_i \in G_p, X_j \in G_q\}$

or $D_{pq}^{S} = \max\{d_{ij} \mid X_i \in G_p, X_j \in G_q\}$

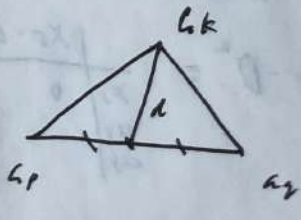
Rmk: If merge G_p, G_q to G_r. Then $D_{rk} =$

$\min\{D_{pk}, D_{qk}\}$. or $= \max\{D_{pk}, D_{qk}\}$.

ii) By Formula: $D_{rk}^2 = \frac{1}{2}(D_{pk}^2 + D_{qk}^2) + \beta D_{pq}^2$. $-\frac{1}{4} \le \beta \le 0$.

where $G_r = G_p \cup G_q$

Rmk: When $\beta = -\frac{1}{4}$. $D_{rk}^2 = d^2$



iii) Average linkage: $D_{pq} = \frac{1}{n_p n_q} \sum_{i \in G_p} \sum_{j \in G_q} d_{ij}^2$.

Rmk: If $G_r = G_p \cup G_q$. $n_r = n_p + n_q$. Then we have:

$$D_{rk}^2 = \frac{n_p}{n_r} D_{pk}^2 + \frac{n_q}{n_r} D_{qk}^2, \quad k \neq p, q.$$

iv) Centroid Method: $D_{pq}^2 = (\bar{X}_p - \bar{X}_q)^T (\bar{X}_p - \bar{X}_q)$

Rmk: If $G_r = G_p \cup G_q$. $n_r = n_p + n_q$. Then $\bar{X}_r =$

$\frac{1}{n_r}(n_p \bar{X}_p + n_q \bar{X}_q)$. It's easy to check:

$$D_{rk}^2 = \frac{n_p}{n_r} D_{pk}^2 + \frac{n_q}{n_r} D_{qk}^2 - \frac{n_p n_q}{n_r^2} D_{pq}^2, \quad k \neq p, q.$$

v) Ward Method: It based on ANOVA. $D_{pq}^2 = W_r - W_p - W_q$.

where $G_r = G_p \cup G_q$. $W_p = \sum_i^{n_p} (X_{(i)}^{(p)} - \bar{X}^{(p)})^T (X_{(i)}^{(p)} - \bar{X}^{(p)})$

Rmk: It merges two group whose distance is min. i.e.

$\widehat{Var}$ of $G_r = G_p \cup G_q$ is min.

Note $\bar{X}_r = \frac{n_p}{n_r} \bar{X}_p + \frac{n_q}{n_r} \bar{X}_q \Rightarrow D_{pq}^2 = \frac{n_p n_q}{n_r} (\bar{X}_p - \bar{X}_q)^T (\bar{X}_p - \bar{X}_q)$

It differs from iv) a constant.

If $G_r = G_p \cup G_q$. Then $D_{rk}^2 = \frac{n_k + n_p}{n_r + n_k} D_{pk}^2 + \frac{n_k + n_q}{n_r + n_k} D_{qk}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2$

② Properties:

i) Monotone:

Def: If in Hierarchical cluster process. D_i is min distance in i^{th} stage. s.t. $D_1 \leq D_2 \cdots \leq D_k$. Then we say the distance we used is monotone.

prop. The distance ① i). iii). v) are monotone

but not ① ii) (when $\beta=\frac{1}{4}$). iv)

Pf: 1') It's easy to check.

2') $d_{AB}^2 = d_{AC}^2 = 1.1$. $d_{BC}^2 = 1$. for points A, B, C.

is counterexample for the latter.

ii) At each k^{th} stage. $D_{ij}^L \geq D_{ij}^S$. $\forall i, j$.

③ Classes:

i) Def (Rao): $T>0$ is fixed. If G satisfies $d_{ij} \leq T$

for $\forall i, j \in G$. Then G is class w.r.t T.

Rmk: Or we can refine: $\frac{1}{|G|-1} \sum_{j \in G} d_{ij} \leq T$. $\forall i \in G$.

Or $\frac{1}{n(n-1)} \sum_{i,j \in G} d_{ij} \leq T$. $d_{ij} \leq M$ $M > T$. $n = |G|$. $\forall i, j \in G$.

Or any partition of $G = G_1 + G_2$. $D(G_1, G_2) \leq T$.

ii) To determine the number of classes:

- Set a threshold value. e.g. $T=2$. in the example

Then $k=3$.

- Observe the data plot

- By statistics: e.g. If $G = \sum_1^k G_i$. $T_k = \sum_1^k \| (X_{ij} - \bar{X}) \|^2$

$= \sum_1^k W_i + \sum_1^k n_i (\bar{X}_i - \bar{X})^T (\bar{X}_i - \bar{X}) = P_k + B_k$

Denote: $R_k^2 = B_k / T_k$. When R_k^2 ↑, effect ↑.

Rmk: Refine: $\widetilde{R}_k^2 = R_{k+1}^2 - R_k^2$. $\widetilde{R}_k^2$ ↑ means set $G_k = \sum_1^{k+1} G_i$ has more effect.

(3) Progressive Cluster:

Idea: Employ a coarse classification

⇒ make adjustment on it basing on some rule

① Initial Classification:

i) 选取凝聚点:

· Base on information or experience. · Randomly.

· 人为分为大类. 计算每一类的重心.

· 确定某个 d 为半径. 以每个样品为中心做一个球 $r=d$. 确定 D. 取第一个凝聚点为 X_i s.t. $B(X_i, r)$ 内的样品数最多. 再依次选点. s.t. 它与其它凝聚点的距离 $>D$ 且球内样品数最多.

ii) 依据凝聚点初始分类:

· 将每个样品分进距其最近的凝聚点的类中.

· 对数据做标准化处理. $X_{i\cdot} = \sum_j X_{ij}$. $k = \max_{i,j} X_{i\cdot} - X_{j\cdot}$.

若需分为 k 类. 对样品 $X_i = (X_{i1} \cdots X_{in})$ 计算:

$$\frac{(k-1)(X_{i\cdot} - \min_i X_{i\cdot})}{K} + 1$$

取 $l \in \mathbb{Z}^+$ 距其最近. 分进 Π_l 类.

② Stepwise Adjustment:

1') 初始：选凝聚点 ⇒ 按距离最近原则分类.

重新计算每一类的重心，以之作为新的凝聚点.

重复这个过程。若某一次选出的凝聚点与上一次的完全一致，则停止。

2') 分类函数：若有样本 $\{X_i\}_1^n$，分为 k 类 $\{G_i\}_1^k$.

标记重心为 $\{\bar{X}_k\}$. $|G_i| = n_i$. t_i 为 X_i 的类别

Def: $d(X_i, G_j) = (X_i - \bar{X}_j)^T (X_i - \bar{X}_j)$

$$L(G_1 \cdots G_k) = \sum_{i=1}^{n} d(X_i, G_{t_i}) \quad (SSD).$$

Rmk: The process above is to make $L(G_1 \cdots G_k)$ to minimal.

Ex. $\{X_i\}_1^5 = \{1, 4, 5, 7, 11\}$. $d_{ij} = |X_i - X_j|$

取 $d = 2$. $D = 2d = 4$.

	X_1	X_2	X_3	X_4	X_5
密度	0	1	2	1	0

1') 第一取 X_3 为凝聚点，然后再取 X_1、X_5.

2') 计算重心：$G_1^0 = \{X_2, X_3, X_4\}$. $G_2^0 = \{X_1\}$.

$G_3^0 = \{X_5\}$. $\bar{X}_1^0 = \frac{11}{3}$, $\bar{X}_2^0 = 1$. $\bar{X}_3 = 11$

3') 重新按凝聚点 $\{\frac{11}{3}, 1, 11\}$ 进行分类，新的凝聚点仍为 $\frac{11}{3}, 1, 11$，过程终止.

③ Pointwise Adjustment: (K-Method)

一个自然的想法是不用一次性修改一批凝聚点
而是对每个样品归类同时改变凝聚点。

1') 确定初始凝聚点数 k，取前 k 个样品为
k 个凝聚点。

2') 确定距离函数，计算类间最小距离 c
与类内最大距 R。

3') 计算凝聚点两两距离，若 $< c$，则合并
取重心为新的凝聚点，直到所有凝聚
点间距 $\geq c$，此时数量 $k_0 \leq k$。

4') 对剩下 $n-k$ 个样品逐个归类：若该样品
与凝聚点最小距 $> R$，则单独作为凝聚
点，否则 $\leq R$，则与最近的凝聚点合并，重
新计算重心，作为新凝聚点，$\Rightarrow$ 重复 3')

5') 再将样品 (n个) 以此列层按 4') 分类，若各
样点二次归类结果一致，则不必计算重心。

e.g. $\{X_i\}_1^5 = \{1, 4, 5, 7, 11\}$，$k=3$，$c=2$，$R=3$，欧式距离。

1') $d(X_1, X_2) = 3$，$d(X_1, X_3) = 4$，$d(X_2, X_3) = 1 < c = 2$，
合并 X_2, X_3，凝聚点为 $1, \frac{4+5}{2}$，且已有
两类 $G_1 = \{X_1\}$，$G_2 = \{X_2, X_3\}$，$d(X_1, G_2) = 3.5 > 2$。

2') 对剩下的 X_4, X_5 分类，$d(G_1, X_4) = 6$，$d(G_2, X_4)$
$= 2.5 < 3$，X_4 分入 G_2，重心为 $\frac{16}{3}$；而 $d(G_1, X_5)$

$=10$. $d(G_2, X_5) = \frac{17}{3} > 3$. 因此 X_5 单独作为凝聚点成为一类。

3) 重新分类，结果仍为 $G_1 = \{X_1\}$, $G_2 = \{X_2, X_3, X_4\}$, $G_3 = \{X_5\}$.

Rmk: 最终分类与样品考虑顺序有关，若按 $X_5 \to X_1$, k-c-k不变，则结果为：

$G_1 = \{X_5\}$, $G_2 = \{X_4\}$, $G_3 = \{X_2, X_3\}$, $G_4 = \{X_1\}$.

因此前k个凝聚点最好取有代表性的点。

(4) Fisher's Algorithm:

在某些样品分类中，我们要求对 $X_1 \cdots X_n$ 分类后的次序不变，i.e. 每一类为 $\{X_i, \cdots X_{i+k}\}$ 的形式，此时若分为k类则共有 $\binom{n-1}{k-1}$ 种.

process:

i) 类的直径：

设 $G = \{X_k\}_i^j$. 则记 $G = \{i, i+1, \cdots j\}$. $\bar{X}_G = \frac{1}{j-i+1} \sum_{t=i}^{j} X_t$.

$D(i,j) = \sum_i^j (X_t - \bar{X}_G)^T (X_t - \bar{X}_G)$. 为 G 的直径.

ii) Loss Function:

记分法 $b(n,k)$ 为：$G_1, G_2 \cdots G_k$. $G_i = \{j_i \cdots j_{i+1} - 1\}$.

loss function: $L(b(n,k)) = \sum_{t=1}^{k} D(i_t, i_{t+1}-1)$, i.e.

当 n,k 固定时. 我们要使 $L(b(n,k))$ 达到极小. 此时. 各类内的离差最小. 记 $P(n,k)$ 为 n 条件分类法。

iii) 递推公式:

$$L(P(n,2)) = \min_{2\le j\le n} \{ D(1,j-1) + D(j,n) \}.$$

$$L(P(n,k)) = \min_{k\le j\le n} \{ L(P(j-1,k-1)) + D(j,n) \}.$$

iv) 求法:

从 $k=2$ 开始. 计算 $L(P(j,2))$, $j\in[2,n]$

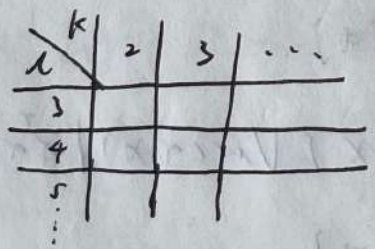
若目前已知 $L(P(i,k-1))$. $i\in[k-1,n]$.

则可寻找 i_{k-1}. 使. $L(P(j,k)) = L(P(i_{k-1},k-1))$

$+ D(i_{k-1},j)$. 得到 $\{L(P(i,j)).\ 1\le i\le n.\ i\le j\le n\}$.

e.g. 1) 计算 $\{D(i,j)\} \Rightarrow L(P(j,2))$.

2) 递推出 $\{L(P(i,j))\}_{i,j}$. 作图:



$L(P(1,k))$

v) 类数不确定:

作图. 取 k 为拐点:

(取 k=3 或 4).