Week17-notes

Monday, 10 May 2021 12:15

Exercise.

Show that

$$\boxed{\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2}$$

where $\bar{x}_{kj}=\frac{1}{|C_k|}\sum_{i\in C_k}x_{ij}$.

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}\underbrace{-\bar{x}_{kj}+\bar{x}_{kj}}_{0}-x_{i'j})^2$$

$$=\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left((x_{ij}-\bar{x}_{kj})-(x_{i'j}-\bar{x}_{kj})\right)^2$$

$$=\text{---}||\text{---}\left[(x_{ij}-\bar{x}_{kj})^2-2(x_{ij}-\bar{x}_{kj})(x_{i'j}-\bar{x}_{kj}) + (x_{i'j}-\bar{x}_{kj})^2\right]$$

$$\sum_{i\in C_k}\boxed{\sum_{i'\in C_k}}$$

$$=\frac{1}{|C_k|}\sum_{i,i'}\sum_{j=1}^{p}\underbrace{(x_{ij}-\bar{x}_{kj})^2}$$

Do not depend on i', only on i and j

$$-\frac{2}{|C_k|}\sum\sum(x_{ij}-\bar{x}_{kj})(x_{i'j}-\bar{x}_{kj})$$

$$+\boxed{\frac{1}{|C_k|}\sum_{i,i'\in C_k}}\sum_{j=1}^{p}(x_{i'j}-\bar{x}_{kj})^2$$

$$=\frac{\boxed{|C_k|}}{\cancel{|C_k|}}\underbrace{\sum_{\boxed{i\in C_k}}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2}-\frac{2}{|C_k|}\sum\sum(x_{ij}-\bar{x}_{kj})(x_{i'j}-\bar{x}_{kj})^2$$

$$+\frac{\cancel{|C_k|}}{\cancel{|C_k|}}\underbrace{\sum_{i'\in C_k}\sum_{j=1}^{p}(x_{i'j}-\bar{x}_{kj})^2}_{\text{Replace } i' \text{ with } i}$$

Mellomregning.

For example, say that we want to sum

$C_k=\{1,2,3\}$

$$\underbrace{\sum_{i,i'\in C_k} i}=\sum_{i\in C_k}\sum_{i'\in C_k} i$$

$$=\sum_{i=1}^{3}\sum_{i'=1}^{3} i$$

$$=\sum_{i=1}^{3} i\cdot\underbrace{\sum_{i'=1}^{3}(1)}_{=3}$$

$$=\sum_{i=1}^{3} i\cdot 3$$

$$=3\sum_{i=1}^{3} i$$

$$=|C_k|\sum_{i=1}^{3} i$$

$$=|C_k|\sum_{i\in C_k} i$$

Cardinality of a set.

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

subtracting the mean

$$-\frac{2}{|C_k|}\underbrace{\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})(x_{i'j}-\bar{x}_{kj})}_{0}$$

Mellomregning #2

$$\sum_{i\in C_k}\sum_{i'\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})(x_{i'j}-\bar{x}_{kj})$$

$$=\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})\overbrace{\sum_{i'\in C_k}(x_{i'j}-\bar{x}_{kj})}^{0}$$

$$=\boxed{\text{— || —}}\left[\sum_{i'\in C_k}x_{i'j}-\sum_{i'\in C_k}\bar{x}_{kj}\right]$$

$$=\text{— || —}\left[\sum_{i'\in C_k}x_{i'j}-\bar{x}_{kj}\cdot\cancel{|C_k|}\right]$$

$$\bar{x}_{kj}\rightarrow\frac{1}{\cancel{|C_k|}}\sum_{i'\in C_k}x_{i'j}$$

$$=\text{— || —}\left[\sum_{i'\in C_k}x_{i'j}-\sum_{i'\in C_k}x_{i'j}\right]$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{0}$$

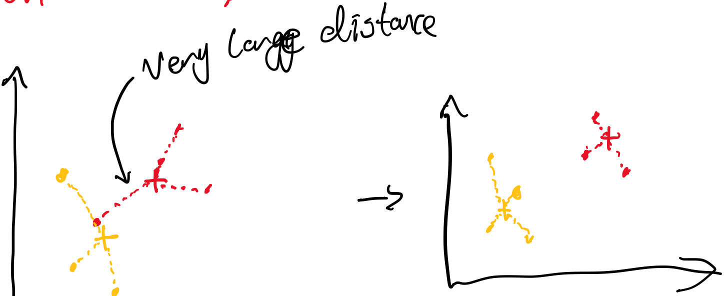
$$= 0.$$

So we are done, shown that

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$$

$$=2\sum_{i\in C_k}\sum_{j=1}^{p}(\overbrace{x_{ij}-\underset{\substack{\uparrow\\ \text{centroids}}}{\bar{x}_{kj}}})^2$$

After each iteration, the right side is decreasing, and therefore the k-means algorithm decrease the objective function (10.11)/the LHS.

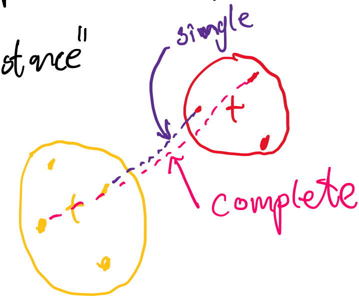


Exercise 4

"smallest"

Hierarchical clustering with single and complete linkage.

"distance"



a) {1,2,3} and {4,5}.

Not enough information.

2 = d(1,4) = 2
2 = d(1,5) = 3
2 = d(2,4) = 1
2 = d(2,5) = 3
2 = d(3,4) = 4
2 = d(3,5) = 1

Single: 1

Complete: 4

↑ Fuse at a higher value.

Same single and complete distance
→ fusing at same height.

b) Two clusters {5}, {6}, then single and complete is the same, so fuse at same height.

''co

in

6

s

60
election)

en

L

i