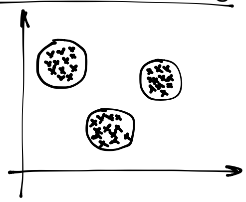
K-means clustering



Assumption: k clusters centred at unknown points (centroids) $\underset{\sim}{\mu}_1, \cdots, \underset{\sim}{\mu}_k$

- find disjoint sets $G_1, \cdots, G_k$ of observations to minimize

$$\sum_{j=1}^{k} \sum_{i \in G_j} \| \underset{\sim}{x}_i - \underset{\sim}{\mu}_j \|^2$$

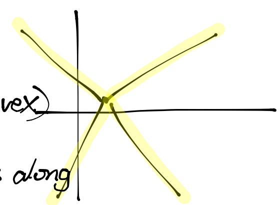
$$= \sum_{i=1}^{n} \min(\|\underset{\sim}{x}_i - \underset{\sim}{\mu}_1\|^2, \cdots, \|\underset{\sim}{x}_k - \underset{\sim}{\mu}_k\|^2) \text{ total within group sum of squares}$$

R function: kmeans ← several different options for algorithm
← can allow the algorithm to use more than 1 starting value.

- works best if have spherical clusters.
  - transform variables?

Alternative approach:

- divide $\mathbb{R}^p$ into k disjoint (convex) regions.
- Assume density of observations along boundaries is low.
- how to express this mathematically?



Model-based clustering

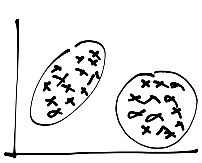
Model: Mixture model

$\underset{\sim}{X}$ has $f(\underset{\sim}{x}) = \lambda_1 f_1(\underset{\sim}{x}, \underset{\sim}{\theta}_1) + \lambda_2 f_2(\underset{\sim}{x}, \underset{\sim}{\theta}_2) + \cdots + \lambda_k f_k(\underset{\sim}{x}, \underset{\sim}{\theta}_k)$

- $\underbrace{\lambda_1, \lambda_2, \cdots, \lambda_k}_{\text{unknown}} \geq 0$. (proportions) with $\lambda_1 + \cdots + \lambda_k = 1$
- $\underset{\sim}{\theta}_1, \cdots, \underset{\sim}{\theta}_k$ are unknown parameter vectors
- Assumption: For each $\underset{\sim}{x}$, $f_i(\underset{\sim}{x}, \underset{\sim}{\theta}_i) f_j(\underset{\sim}{x}, \underset{\sim}{\theta}_j) \doteq 0$ if $i \neq j$

Example: $f_j(\underset{\sim}{x}, \underset{\sim}{\theta}_j) = N_p(\underset{\sim}{\mu}_j, C_j)$ ← describe shape of cluster

↑ cluster centre



Problem: Given data $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$, estimate $\lambda_1, \cdots, \lambda_k$ and $\underset{\sim}{\theta}_1, \cdots, \underset{\sim}{\theta}_k$

Maximum Likelihood estimation

Maximize $\ln L(\lambda_1, \cdots, \lambda_k, \underset{\sim}{\theta}_1, \cdots, \underset{\sim}{\theta}_k) = \sum_{i=1}^{n} \ln(\lambda_1 f_1(\underset{\sim}{x}_i, \underset{\sim}{\theta}_1) + \cdots + \lambda_k f_k(\underset{\sim}{x}_i, \underset{\sim}{\theta}_k))$

EM algorithm

E: expectation M: maximization

Idea: Assume first that we know which cluster $\underset{\sim}{x}_i$ belongs to.

i.e. we observe $\underset{\substack{\uparrow\\ \text{cluster}}}{(C_1}, \underset{\sim}{x_1}), \cdots, (\underset{\substack{\uparrow\\ \text{indicator}}}{C_n}, \underset{\sim}{x_n})$ where C_i takes values in $\{1, \cdots, k\}$

Then the likelihood function becomes

$$L_c(\lambda_1, \cdots, \lambda_k, \underset{\sim}{\theta_1}, \cdots, \underset{\sim}{\theta_k}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{k} [\lambda_j f_j(\underset{\sim}{x_i}, \underset{\sim}{\theta_j})]^{I(C_i=j)} \right\} \qquad I(C_i=j) = \begin{cases} 1 & \text{if } C_i=j \\ 0 & \text{ow} \end{cases}$$

$$\ln L_c(\lambda_1, \cdots, \lambda_k, \underset{\sim}{\theta_1}, \cdots, \underset{\sim}{\theta_k}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \{ I(C_i=j) \ln(\lambda_j) + I(C_i=j) \ln f_j(\underset{\sim}{x_i}, \underset{\sim}{\theta_j}) \}$$

$$= \ln L_c^{(1)}(\lambda_1, \cdots, \lambda_k) + \ln L_c^{(2)}(\underset{\sim}{\theta_1}, \cdots, \underset{\sim}{\theta_k})$$

MLEs of $\lambda_1, \cdots, \lambda_k$:

$$\hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} I(C_i = j) = \text{proportion of cluster } j \text{ in sample}$$

But, we don't observe $C_1, \cdots, C_n$!

EM algorithm: estimate $\Delta_{ij} = I(C_i = j)$. (for $i = 1, \cdots, n$; $j = 1, \cdots, k$)

E-step: Estimate Δ_{ij} by $E[\Delta_{ij} \mid \underset{\sim}{x_i} ; \underbrace{\hat{\lambda}_1, \cdots \hat{\lambda}_k, \hat{\underset{\sim}{\theta}}_1, \cdots, \hat{\underset{\sim}{\theta}}_k}_{\text{correct estimates}}]$

$$= \frac{\hat{\lambda}_j f_j(\underset{\sim}{x_i} ; \hat{\underset{\sim}{\theta}}_j)}{\sum_{l=1}^{k} \hat{\lambda}_l f_l(\underset{\sim}{x_i} ; \hat{\underset{\sim}{\theta}}_l)} \quad \text{(Bayes rule)}$$

<u>M-step</u>: Update estimates of $\hat{\lambda}_1, \cdots, \hat{\lambda}_k$ and $\hat{\underset{\sim}{\theta}}_1, \cdots, \hat{\underset{\sim}{\theta}}_k$

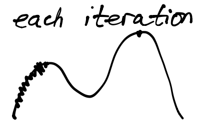
$$\hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{\Delta}_{ij}$$

and $\hat{\underset{\sim}{\theta}}_1, \cdots, \hat{\underset{\sim}{\theta}}_k$ maximize $\ln L_c^{(2)}(\underset{\sim}{\theta_1}, \cdots, \underset{\sim}{\theta_k}) = \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{\Delta}_{ij} \ln f_j(\underset{\sim}{x_i}, \underset{\sim}{\theta_j})$

Now, iterate E- and M-step <u>until convergence</u>!

<u>Notes</u>:

① Don't necessarily have convergence to MLEs i.e. maximizers of $\ln L(\lambda_1, \cdots, \lambda_k, \underset{\sim}{\theta_1}, \cdots, \underset{\sim}{\theta_k})$
But at each iteration of E- and M-steps maximized $\ln L$ will increase.



② Performance of EM algorithm depends very strongly on initial estimates of $\underset{\sim}{\theta_1}, \cdots, \underset{\sim}{\theta_k}$
– can usually take initial estimates of $\lambda_1, \cdots, \lambda_k$ equal to $\frac{1}{k}$.