Applied Data Science and Machine Intelligence

Clustering: All sessions

Nandan Sudarsanam,
Department of Management Studies,
Robert Bosch Centre for Data Science and AI (RBC-DSAI),
Indian Institute of Technology Madras

Plan for the session

- 6 micro-cases to motivate clustering
- Overview of the different types of clustering/clusters
- Deep dive into two algorithms
  - K-means
  - Hierarchical
  - GMM
- Cluster evaluation through silhouettes and size determination through gap-statistic

Understanding clustering through micro-cases

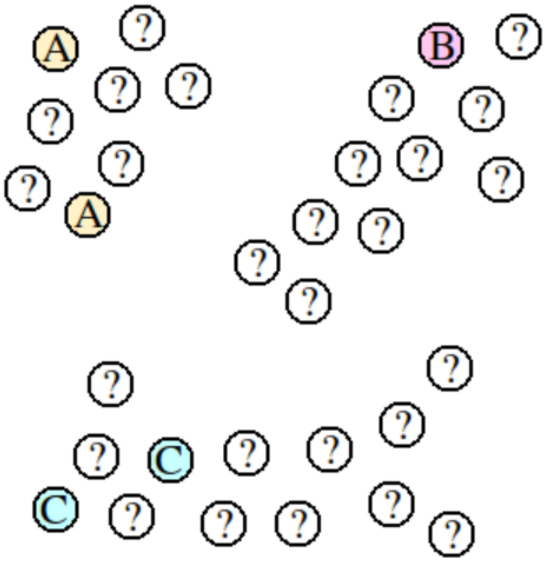
- The Bell computers case
  - Laptops are used for different purposes
  - Laptops are configurable in an online store (RAM, CPU, GPU, CD-Drive, ports, warranty)
  - Can we use historic purchase choices to infer the usage of the product?
- The AirCom case
  - One of the oldest telecom service providers
  - The migration to mobile phones, intro of 3G/4G brought in a lot of competition and niche entrants. Resulting in decline of market share.
  - The organization needs to understand what product plans are desirable (profitable is another question)
  - At a customer level, can we study the current call usage, sms, data? Frequency, quantum, night/day, weekend/weekday

Understanding clustering though micro-cases

- The Gmail case
  - By now we understand that:
    - Divides data into groups (clusters or segments or partitions)
    - Unsupervised learning
  - Why is it different from classification? The Gmail spam example
- More cases:
    - Marketing/Sales. Know your customer. Difference between target and consumption audience. Stories from Ford and the make-up industry.
    - Communicating information – Google News, sentiment analysis.
    - Biology, Climate, Medicine.

Introduction to Clustering

- Why do this?
  - For better understanding. For better processing by subject-matter experts.
  - Could serve as precursor to further Data Analysis. Labelled data is hard to get. Take the following classification example:



Wikipedia contributors. "Transduction (machine learning)." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 5 May. 2016. Web. 19 Oct. 2018.

Types of Clustering and Clusters

- Types of Clustering
  - Exclusive versus Overlapping versus Fuzzy
  - Hierarchical versus Partitional
  - Complete versus Partial
- Types of Clusters
  - Well separated
  - Prototype Based
  - Graph based – types of networks
  - Density based

K-Means Clustering

- What is it?
  - Prototype based approach
  - An iterative procedure that starts with K clusters. How do we know K?
  - Ideal when all variables are quantitative (Should be able to compute distances)
  - If we use the notion that $C_1, \ldots C_K$ denotes K sets of the n observations.
  - Then $C_1 \cup C_2 \cup \cdots C_K = \{1,2..n\}$ and $C_K \cap C_{K'} = \emptyset$
  - We seek to minimize some measure of within cluster variation

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Hastie et al

K-means

- The objective function using Euclidean distance:

$$\underset{C_1,\ldots,C_K}{\text{minimize}}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\right\}.$$

James et al

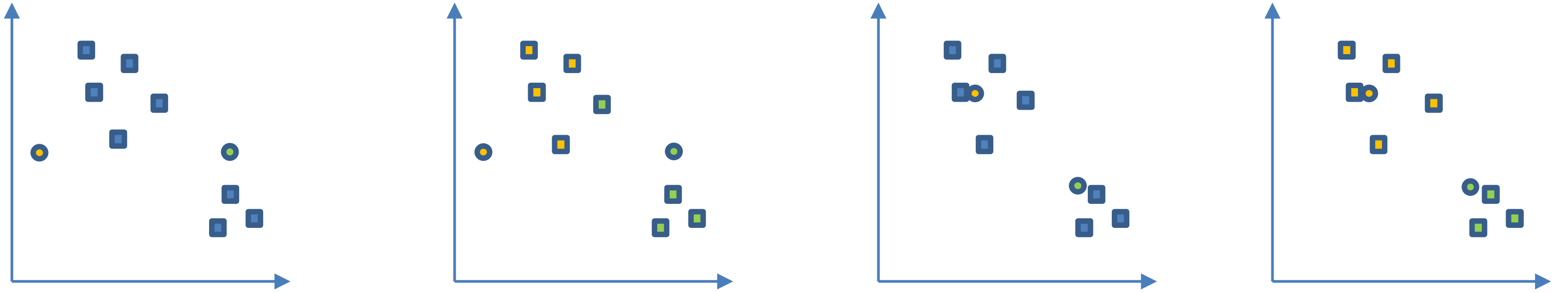
- K is the number of clusters, $|C_K|$denotes the number of elements in cluster C_K, i represents each element and j represents the features.
- This is a hard problem to solve, since there are K^nways to partition the data.
- Fortunately, we have an efficient algorithm:

How does it work?

1. Initialize some cluster centers (K)
2. Assign each point to the closest cluster centre
3. Once all points have been assigned, recompute the cluster centre to be the centroid of the assignment derived from step 2
4. Repeat steps 2 and 3 until no further changes in centroid

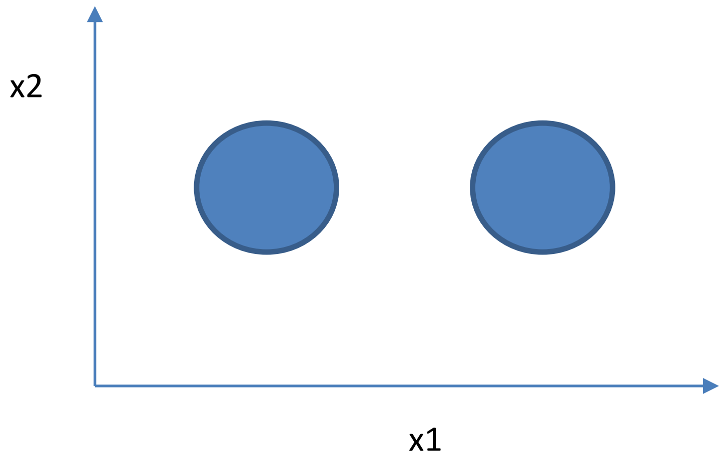
K-means

- K-means Graphically



K-Means clustering

- What would happen at different starting points



- We will explore soft k-means and GMMS in the next session

K - medoids

- Extending K-means to K-medoids
  - Going beyond quantitative variables to “arbitrarily defined dissimilarities” (Hastie et. al)
  - Example:

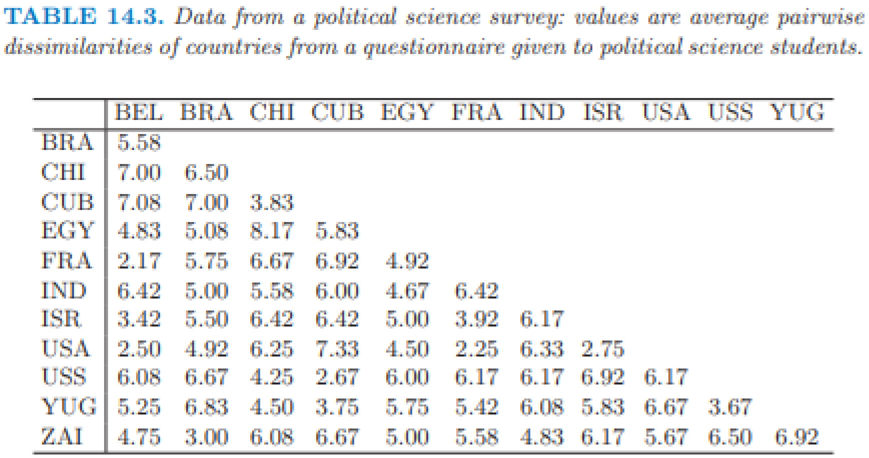


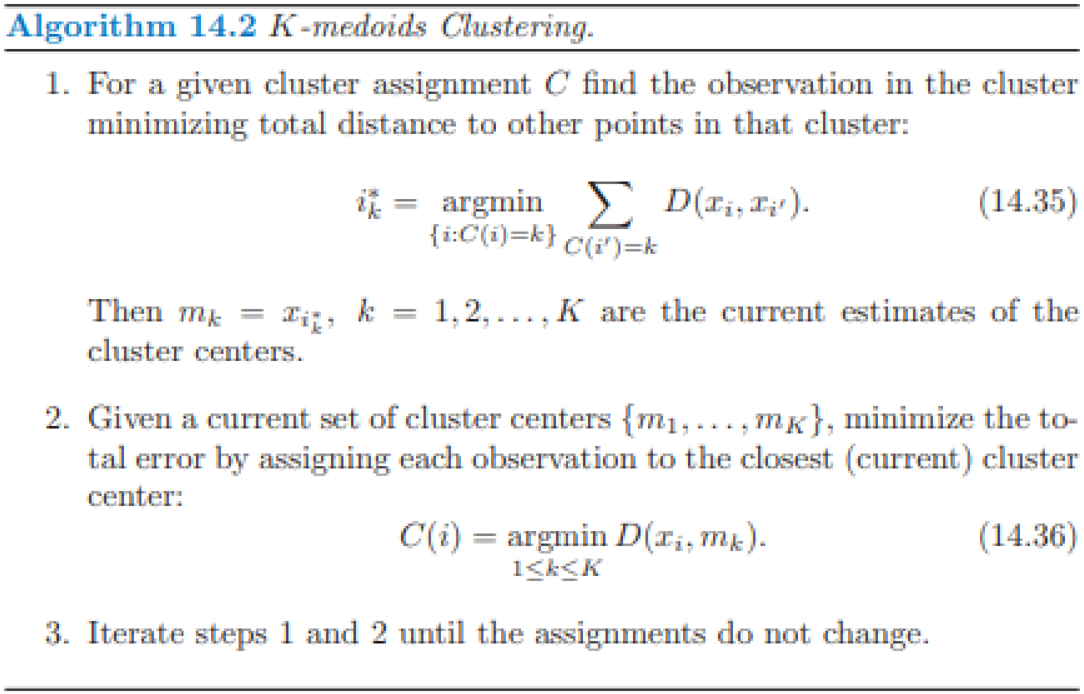
TABLE 14.3. *Data from a political science survey: values are average pairwise dissimilarities of countries from a questionnaire given to political science students.*

	BEL	BRA	CHI	CUB	EGY	FRA	IND	ISR	USA	USS	YUG
BRA	5.58										
CHI	7.00	6.50									
CUB	7.08	7.00	3.83								
EGY	4.83	5.08	8.17	5.83							
FRA	2.17	5.75	6.67	6.92	4.92						
IND	6.42	5.00	5.58	6.00	4.67	6.42					
ISR	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USS	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
YUG	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
ZAI	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. New York: springer, 2009.

K-medoids

- Algorithm:



Algorithm 14.2 *K-medoids Clustering.*

1. For a given cluster assignment C find the observation in the cluster minimizing total distance to other points in that cluster:

$$i_k^* = \underset{\{i:C(i)=k\}}{\operatorname{argmin}} \sum_{C(i')=k} D(x_i, x_{i'}). \qquad (14.35)$$

   Then $m_k = x_{i_k^*}$, $k = 1, 2, \ldots, K$ are the current estimates of the cluster centers.

2. Given a current set of cluster centers $\{m_1, \ldots, m_K\}$, minimize the total error by assigning each observation to the closest (current) cluster center:

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} D(x_i, m_k). \qquad (14.36)$$

3. Iterate steps 1 and 2 until the assignments do not change.

Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. New York: springer, 2009.

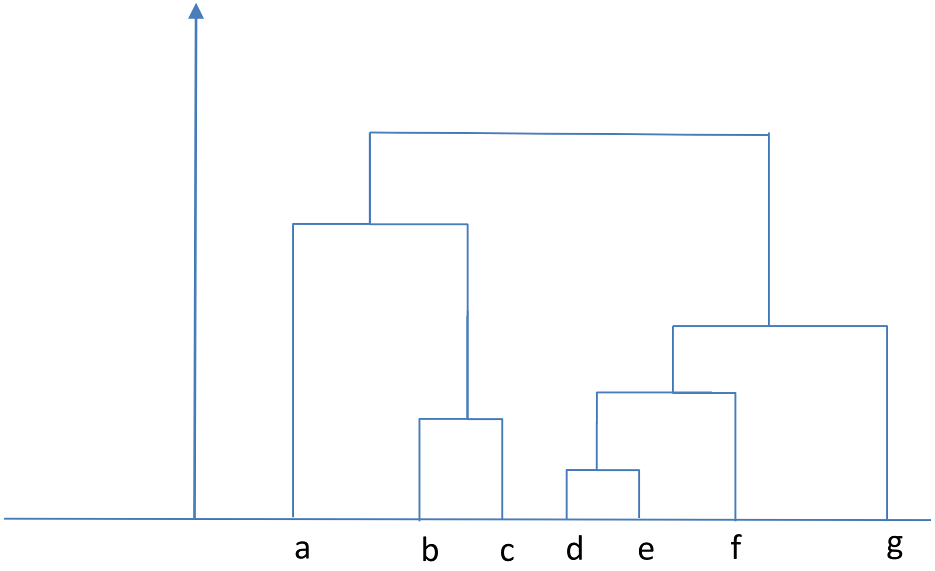
- Computational considerations

Hierarchical Clustering

- Does not require a fixed number of clusters
- Creates nested clusters
- Types:
  - Agglomerative (Bottom-up)
  - Divisive (Top-down)
- Requires defining some measure of dissimilarity between clusters
  - Single Linkage
  - Complete Linkage
  - Group Average
  - Ward’s method, increase in SSE from a merger

Agglomerative Hierarchical Clustering

- Monotonicity
- Graphical representation through Dendrogram

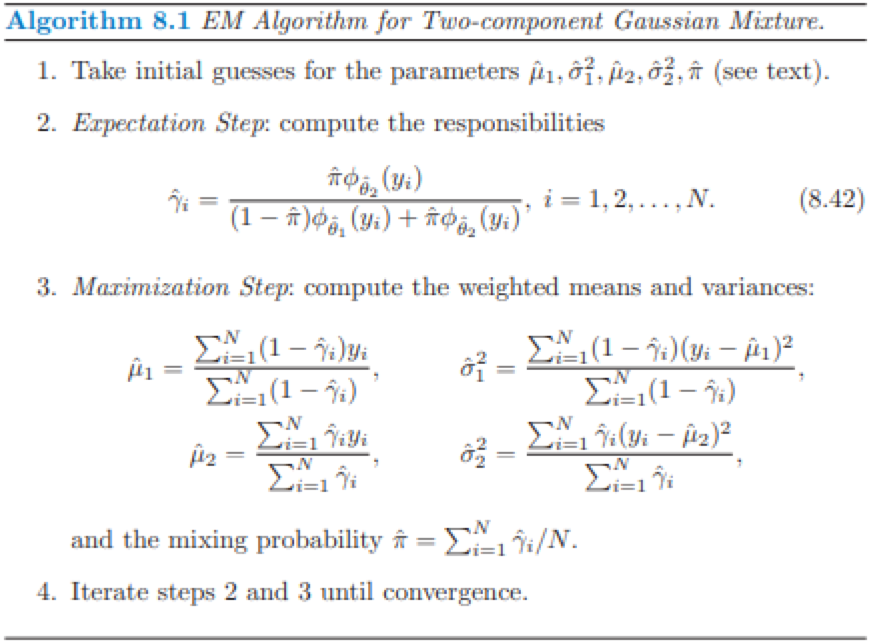


Gaussian Mixture Models (GMMs)

- Can be seen as a soft version of K-means
- Remember Exclusive vs overlapping vs fuzzy?
- A drawback of k-means is that it does not account for variances
- In GMMs we assume that each cluster is a Gaussian, with it’s own mean and variance.
- We use the Expectation- Maximization (EM) algorithm to learn the parameters and assign responsibilities

E-M algorithm

- The steps of the E-M algorithm for a two Gaussians From: Hastie, Trevor, et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. New York: springer, 2009.



Algorithm 8.1 *EM Algorithm for Two-component Gaussian Mixture.*

1. Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$ (see text).
2. *Expectation Step*: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \ldots, N. \qquad (8.42)$$

3. *Maximization Step*: compute the weighted means and variances:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. Iterate steps 2 and 3 until convergence.

- $\hat{\pi}$ is the proportion of the second Gaussian
- $\hat{\gamma}_i$is the expectation of the second Gaussian (the likelihood that a point belongs to the second cluster)
- The Maximization step provides the updated mean and variance of the two Gaussians.
- Upon convergence the cluster membership of each point is captured by $\hat{\gamma}_i$ and (1-$\hat{\gamma}_i$)

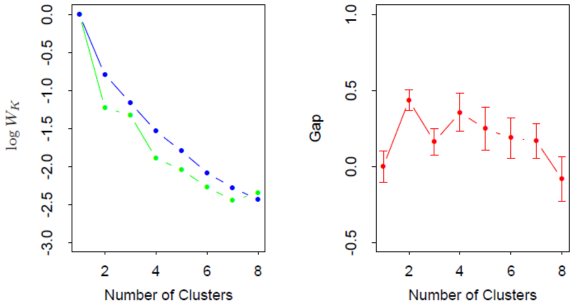
Cluster Validation

- Cluster Evaluation. Why do we need to evaluate?
- Internal versus External (unsupervised versus supervised). Would Cross validation work?
- Internal: cohesion versus separation
- Silhouettes:
  - For each data point i define $a(i)$ as the average dissimilarity of i with the other data points in its cluster. SSE should work.
  - For the same data point i define the vector $\mathrm{c}(i)$ as the average dissimilarity of i with the other data points in each cluster. Define $\mathrm{b(i)} = \min(\mathrm{c}(i))$
  - $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$
  - Average s (across all datapoints) tells us something about the clustering. What is the range of s? what are desirable values?

Empty slide for drawings

How many clusters? Gap Statistic

- The problem with Silhouettes for a wide range of k (natural bias towards high k)
- Say you had some measure of within cluster dissimilarity W_K, which is a function of k (number of clusters), for a given repeatable clustering algorithm.
- As K goes from 1,2…to…N , the values of $W_1, W_2, \ldots, W_N$ tend to generally decrease.
- What if you assumed that the data was uniformly distributed?



Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. New York: Springer series in statistics, 2001.

Principal Components

- Unsupervised motivation: Let us say we want to visualize the data. We can at most use two dimensions/features. How many plots would we need if we had 10 features?
- Supervised motivation: We may be overfitting the model due to high multicollinearity. This could be because some of the correlated variables are capturing the same underlying phenomena. Could we represent this phenomena as a single variable?
- Principal components are a lower dimensional representation of data that captures as much information as possible
  - So do we just knock off the useless features (how do we know what is useless)?
  - We create these dimensions in such a way as to maximize the variance that the points have on those dimensions.
  - http://setosa.io/ev/principal-component-analysis/

Principal components

- So the first PC is $Z_1 = \emptyset_{11}X_1 + \emptyset_{21}X_2 \ldots.$ Where $\sum_{j=1}^{p} {\emptyset_{j1}}^2 = 1$
- The first data point in $Z_1$is $Z_{1i} = \emptyset_{11}X_{1i} + \emptyset_{21}X_{2i}..$
- Optimization problem:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \right\} \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1.$$

Gareth, James, et al. *An introduction to statistical learning: with applications in R*. Spinger, 2013..

- Breakdown of Variances

$$\underbrace{\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2}_{\text{Var. of data}} = \underbrace{\sum_{m=1}^{M}\frac{1}{n}\sum_{i=1}^{n}z_{im}^2}_{\text{Var. of first } M \text{ PCs}} + \underbrace{\frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - \sum_{m=1}^{M}z_{im}\phi_{jm}\right)^2}_{\text{MSE of } M\text{-dimensional approximation}}$$

Gareth, James, et al. *An introduction to statistical learning: with applications in R*. Spinger, 2013..

- Proportion explained: 1-$\frac{RSS}{TSS}$