Modern Information Retrieval

Clustering[1]

Hamid Beigy

Sharif university of technology

December 9, 2022

[1] Some slides have been adapted from slides of Manning, Yannakoudakis, and Schütze.

Table of contents

Introduction

Introduction

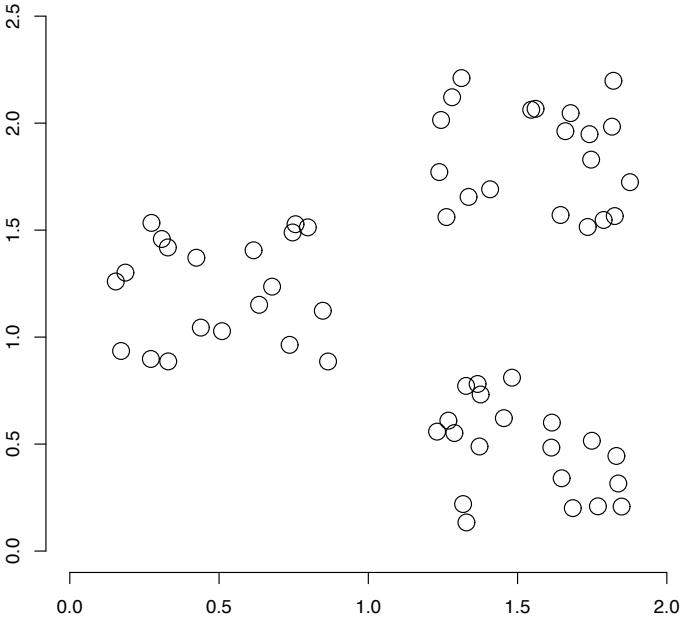
1. (Document) clustering is the process of grouping a set of documents into clusters of similar documents.
2. Documents within a cluster should be similar.
3. Documents from different clusters should be dissimilar.
4. Clustering is the most common form of unsupervised learning.
5. Unsupervised = there are no labeled or annotated data.

Clustering

The cluster hypothesis

1. **Cluster hypothesis.** Documents in the same cluster behave similarly with respect to relevance to information needs.
2. All applications of clustering in IR are based (directly or indirectly) on the cluster hypothesis.
3. Van Rijsbergen's original wording (1979): "closely associated documents tend to be relevant to the same requests".

Data set with clear cluster structure



Goals of clustering

1. General goal: put related docs in the same cluster, put unrelated docs in different clusters.
2. We'll see different ways of formalizing this.
3. The number of clusters should be appropriate for the data set we are clustering.
   - Initially, we will assume the number of clusters K is given.
   - Later: Semi–automatic methods for determining K
4. Secondary goals in clustering
   - Avoid very small and very large clusters
   - Define clusters that are easy to explain to the user

Flat vs. Hierarchical clustering

1. Flat algorithms
   - Usually start with a random (partial) partitioning of docs into groups
   - Refine iteratively
   - Main algorithm: K-means
2. Hierarchical algorithms
   - Create a hierarchy
   - Bottom-up, agglomerative
   - Top-down, divisive

Hard vs. Soft clustering

1. Hard clustering: Each document belongs to exactly one cluster.
   - More common and easier to do
2. Soft clustering: A document can belong to more than one cluster.
   - Makes more sense for applications like creating browsable hierarchies
   - You may want to put *sneakers* in two clusters:
     - sports apparel
     - shoes
   - You can only do that with a soft clustering approach.

Flat algorithms

1. Flat algorithms compute a partition of N documents into a set of K clusters.
2. Given: a set of documents and the number K
3. Find: a partition into K clusters that optimizes the chosen partitioning criterion
4. Global optimization: exhaustively enumerate partitions, pick optimal one
   - Not tractable
5. Effective heuristic method: K-means algorithm

K-means

K-means

- Perhaps the best known clustering algorithm
- Simple, works well in many cases
- Use as default / baseline for clustering documents

Document representations in clustering

1. Vector space model
2. We measure relatedness between vectors by Euclidean distance, which is almost equivalent to cosine similarity.
3. Almost: centroids are not length-normalized.

K-means: Basic idea

1. Each cluster in K-means is defined by a centroid.
2. Objective/partitioning criterion: minimize the average squared difference from the centroid
3. Recall definition of centroid:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

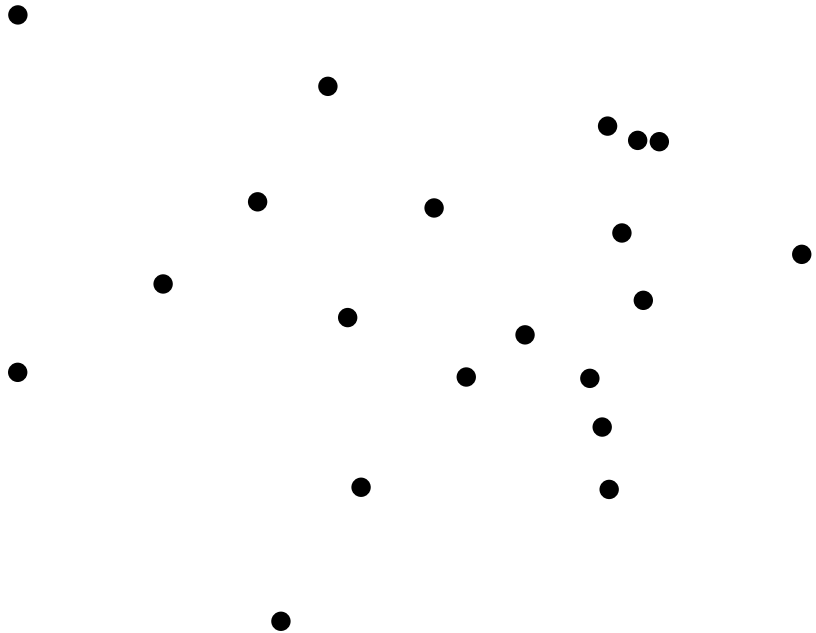
   where we use ω to denote a cluster.
4. We try to find the minimum average squared difference by iterating two steps:
   - reassignment: assign each vector to its closest centroid
   - recomputation: recompute each centroid as the average of the vectors that were assigned to it in reassignment

K-means Algorithm

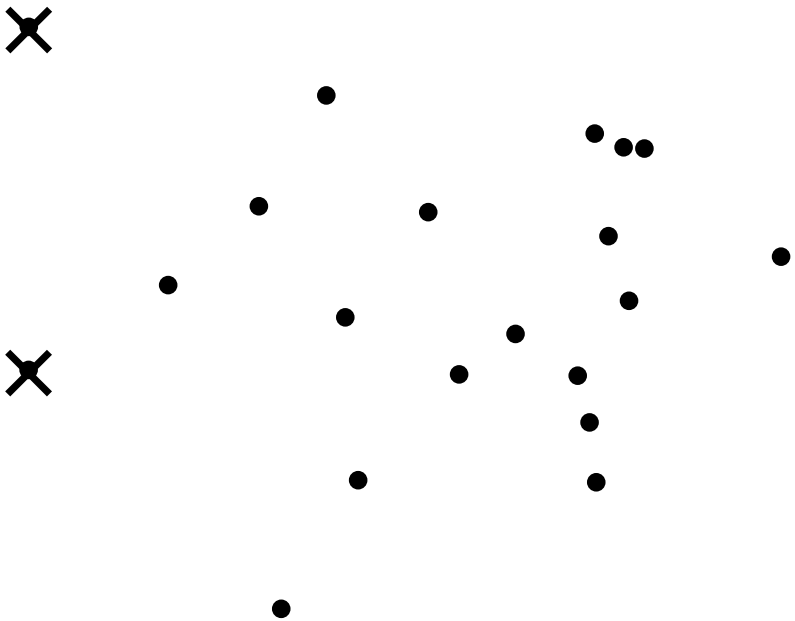
K-MEANS$(\{\vec{x}_1, \ldots, \vec{x}_N\}, K)$

1 $(\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_K) \leftarrow$ SELECTRANDOMSEEDS$(\{\vec{x}_1, \ldots, \vec{x}_N\}, K)$
2 **for** $k \leftarrow 1$ **to** K
3 **do** $\vec{\mu}_k \leftarrow \vec{s}_k$
4 **while** stopping criterion has not been met
5 **do for** $k \leftarrow 1$ **to** K
6 **do** $\omega_k \leftarrow \{\}$
7 **for** $n \leftarrow 1$ **to** N
8 **do** $j \leftarrow \arg\min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$
9 $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$ *(reassignment of vectors)*
10 **for** $k \leftarrow 1$ **to** K
11 **do** $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$ *(recomputation of centroids)*
12 **return** $\{\vec{\mu}_1, \ldots, \vec{\mu}_K\}$

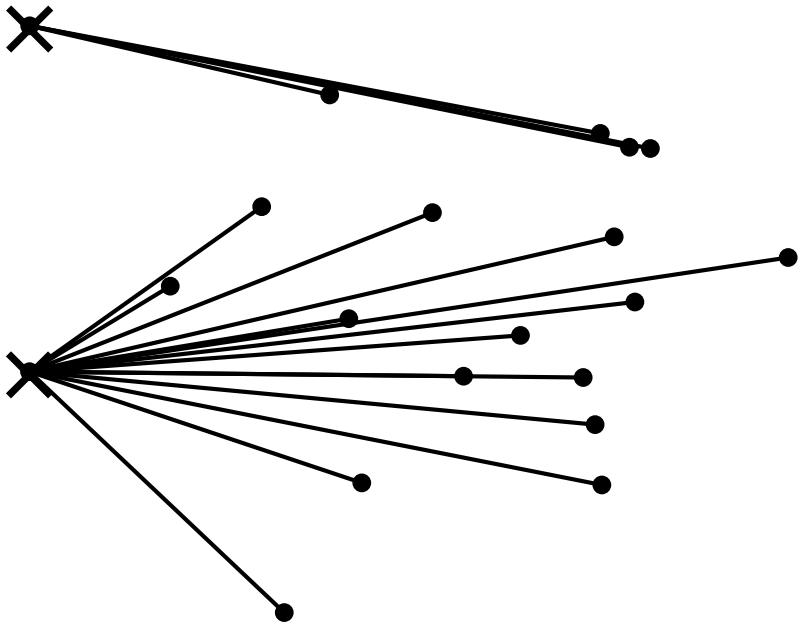
K-means example (dataset for $K = 2$)



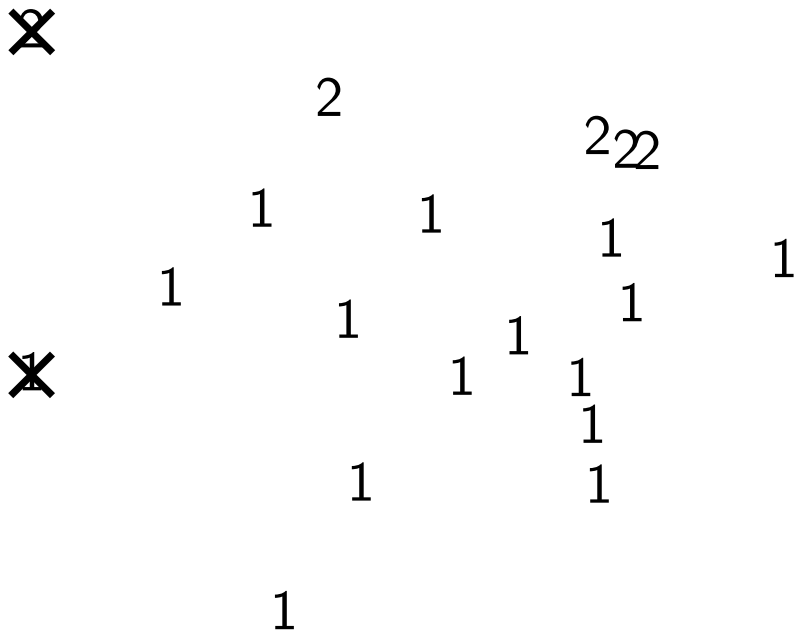
K-means example (initial centroids)



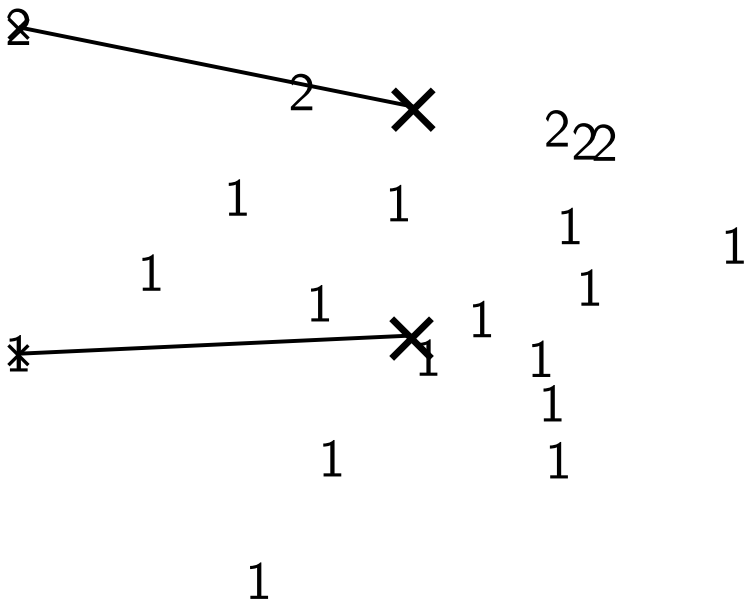
K-means example (Assign points to closest center)



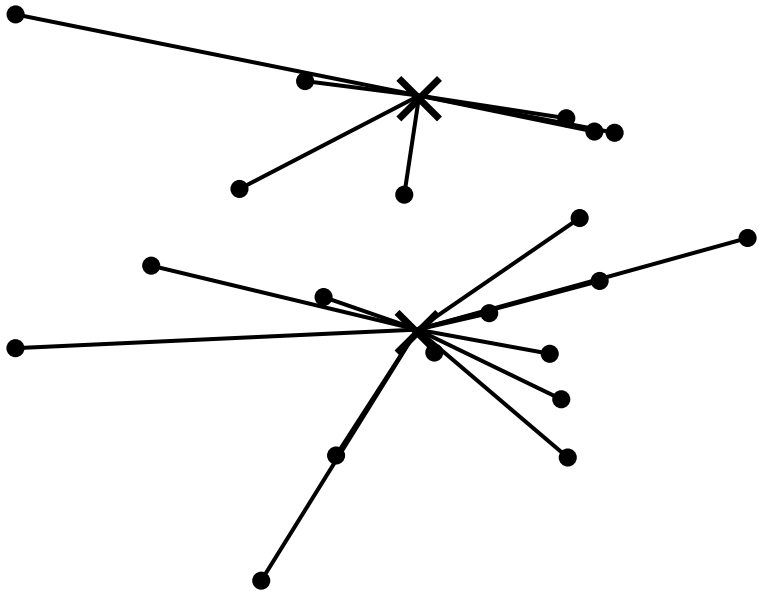
K-means example (Cluster assignment)



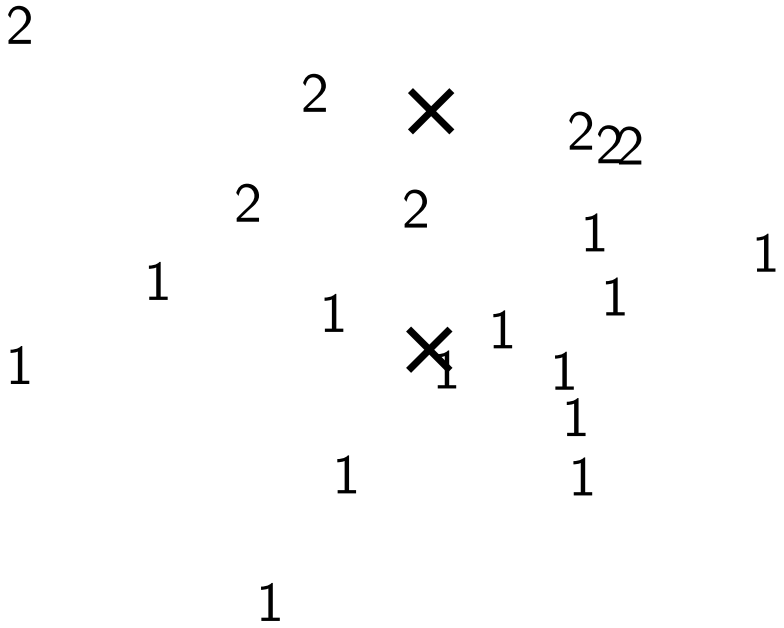
K-means example (Recompute cluster centroids)



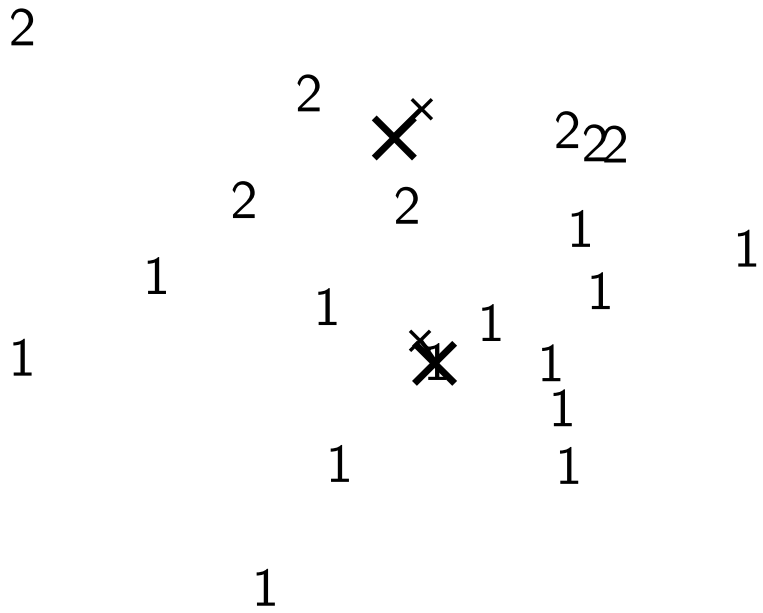
K-means example (Assign points to closest centroid)



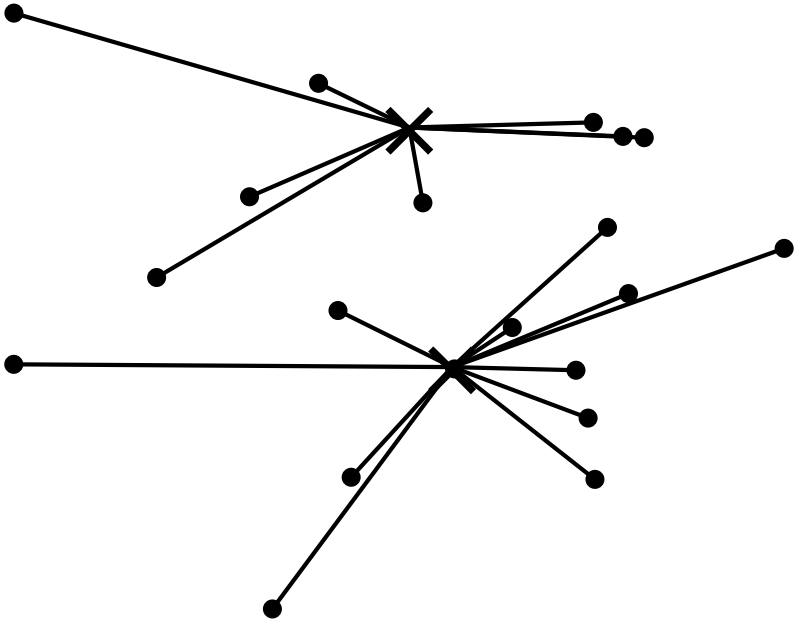
K-means example (Cluster assignment)



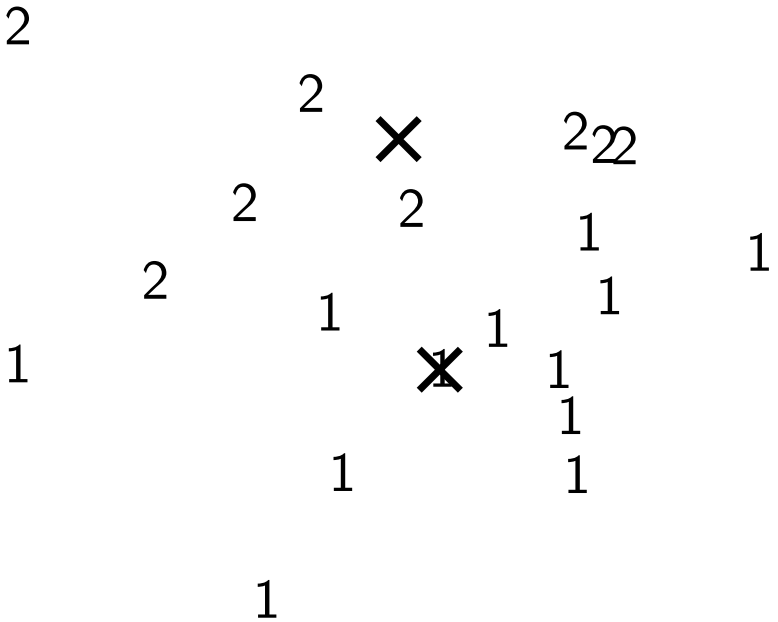
K-means example (Cluster assignment)



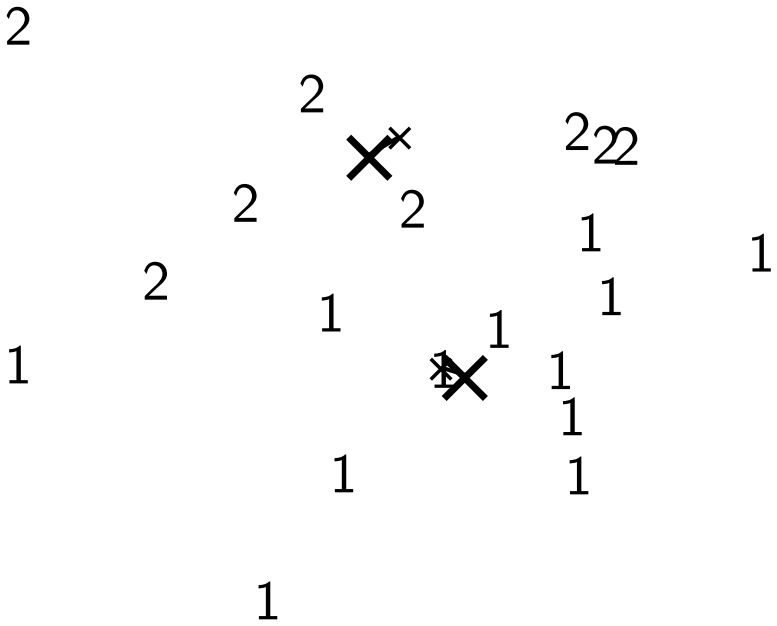
K-means example (Recompute cluster centroids)



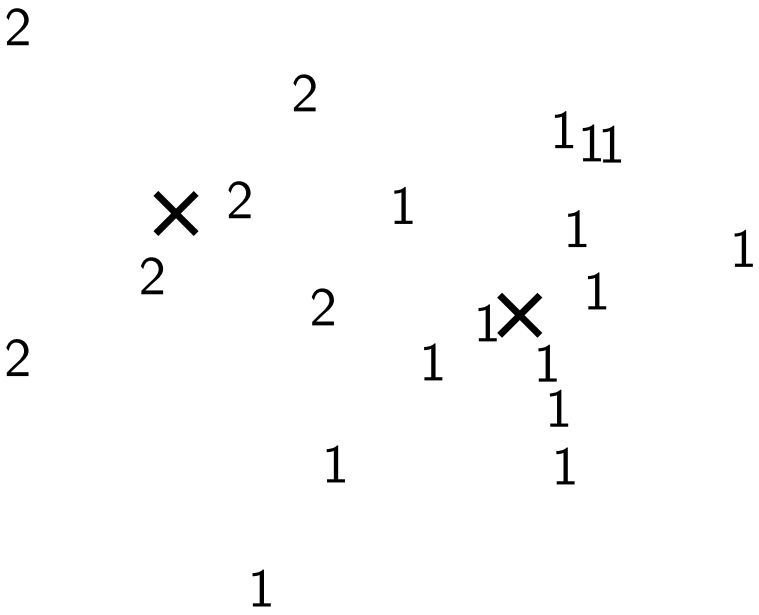
K-means example (Assign points to closest centroid)



K-means example (Cluster assignment)



K-means example (After convergence)



K-means is guaranteed to converge: Proof

1. Residual sum of squares (RSS) decreases during each reassignment step, because each vector is moved to a closer centroid

$$RSS = \sum_{k=1}^{K} \sum_{x \in C_k} |x - \mu_k|^2$$

2. There is only a finite number of clusterings.
3. Thus, we must reach a fixed point.
4. Finite set & monotonically decreasing evaluation function implies convergence

Initialization of K-means

1. Random seed selection is just one of many ways K-means can be initialized.
2. Random seed selection is not very robust: It's easy to get a suboptimal clustering.
3. Better ways of computing initial centroids:
   - Select seeds using some heuristic.
   - Use hierarchical clustering to find good seeds
   - Select i (e.g., $i = 10$) different random sets of seeds, do a K-means clustering for each, select the clustering with lowest RSS.
   - Use the following optimization function.

$$K = \arg\min_k [RSS(k) + \lambda k]$$

   - How do you select λ?
   - Using other objective functions.

Model-based clustering

- k-means is closely related to a probabilistic model known as the Gaussian mixture model.

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

- π_k, μ_k, Σ_k are parameters. π_k are called mixing proportions and each Gaussian is called a mixture component.
- The model is simply a weighted sum of Gaussian. But it is much more powerful than a single Gaussian, because it can model multi-modal distributions.

- Note that for $p(x)$ to be a probability distribution, we require that $\sum_k \pi_k = 1$ and that for all k we have $\pi_k > 0$. Thus, we may interpret the π_k as probabilities themselves.
- Set of parameters $\theta = \{\{\pi_k\}, \{\mu_k\}, \{\Sigma_k\}\}$

Model-based clustering (cont.)

- Let use a K-dimensional binary random variable z in which a particular element z_k equals to 1 and other elements are 0.
- The values of z_k therefore satisfy $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$
- The marginal distribution of x equals to

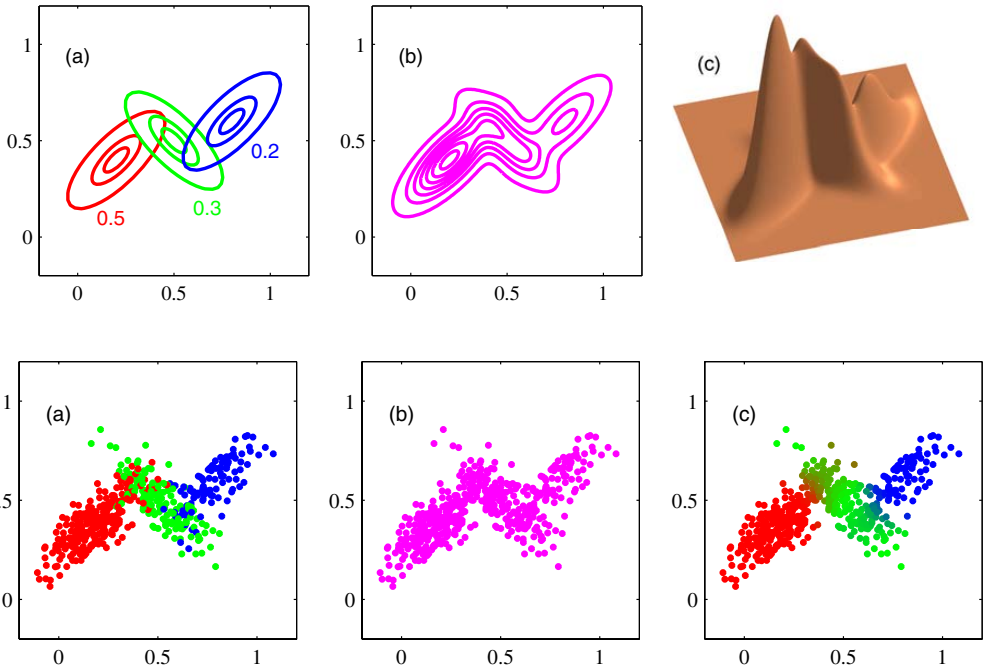
$$p(x) = \sum_z p(z)p(x|z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

- We can write $p(z_k = 1|x)$ as

$$\gamma(z_k) = p(z_k = 1|x)$$

- We shall view π_k as the prior probability of $z_k = 1$, and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed x.

Gaussian mixture model (example)



1. Initialize μ_k, Σ_k, and π_k, and evaluate the initial log likelihood.
2. E step Evaluate $\gamma(z_{nk})$ using the current parameter values

$$p(z_k = 1|x) = \gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

3. M step Re-estimate the parameters using the current value of $\gamma(z_{n_k})$

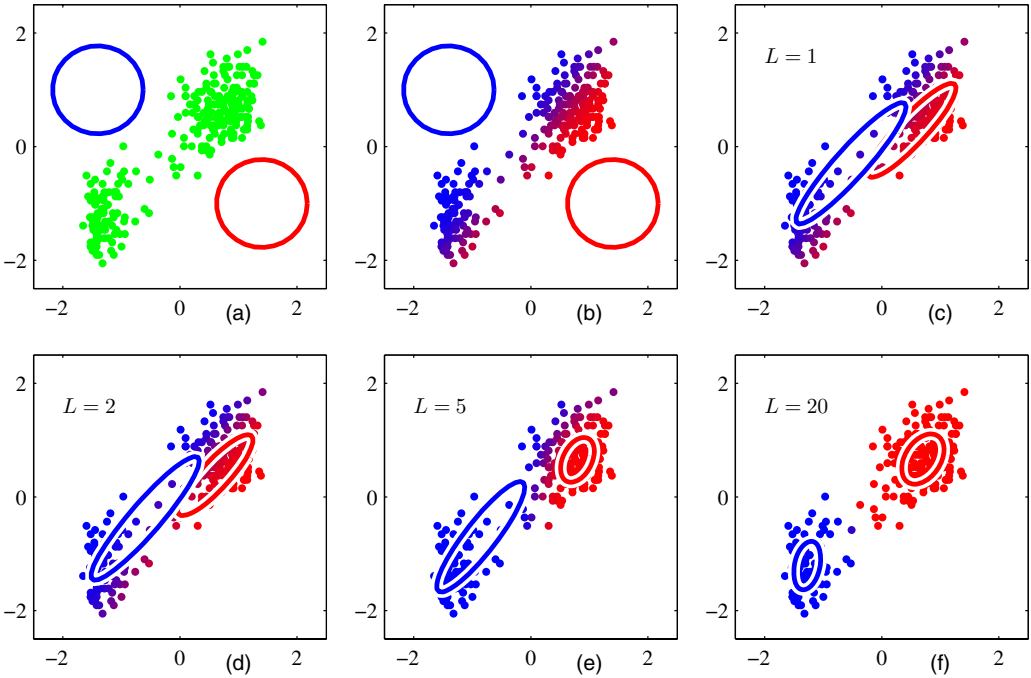
$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{n_k}) x_n$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{n_k})(x_n - \mu_k)(x_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.

Model-based clustering (example)



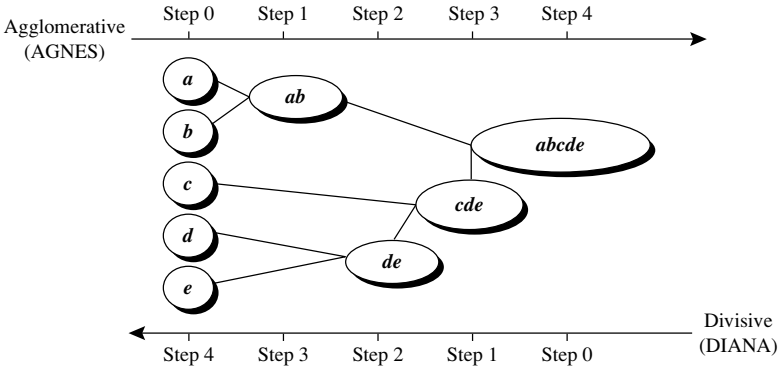
Hierarchical clustering

Hierarchical clustering

- Imagine we now want to create a hierarchy in the form of a binary tree.
- Assumes a similarity measure for determining the similarity of two clusters.
- Up to now, our similarity measures were for documents.
- We will look at different cluster similarity measures.

Hierarchical methods

- A hierarchical clustering method works by grouping data objects into a hierarchy or "tree" of clusters.

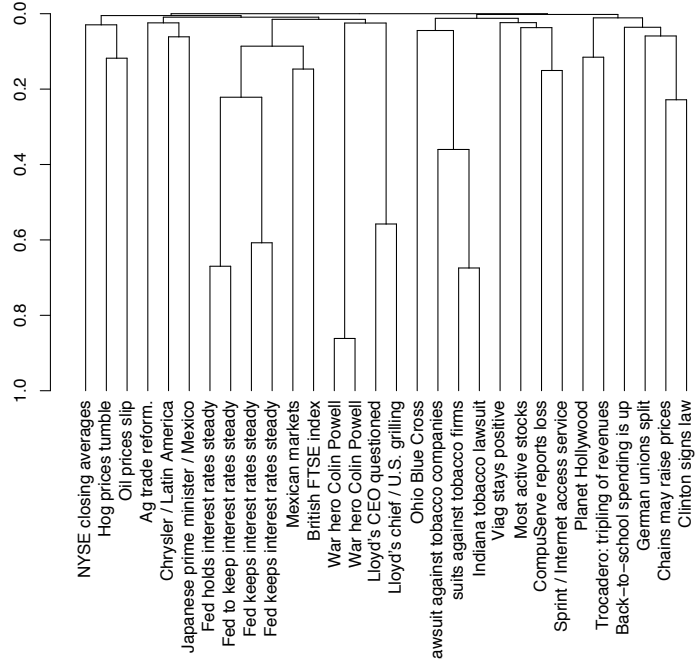


- Hierarchical clustering methods
  - Agglomerative hierarchical clustering
  - Divisive hierarchical clustering

Distance measures in hierarchical methods

1. Whether using an agglomerative method or a divisive method, a core need is to measure the distance between two clusters, where each cluster is generally a set of objects.
2. Four widely used measures for distance between clusters are as follows, where $|p - q|$ is the distance between two objects or points, p and q; μ_i is the mean for cluster, C_i; and n_i is the number of objects in C_i. They are also known as linkage measures.
   - Minimum distance
   $$d_{min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} |p - q|$$
   - Maximum distance
   $$d_{max}(C_i, C_j) = \max_{p \in C_i, q \in C_j} |p - q|$$
   - Mean distance $d_{mean}(C_i, C_j) = |\mu_i - \mu_j|$
   - Average distance $d_{min}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, q \in C_j} |p - q|$

Dendrogram

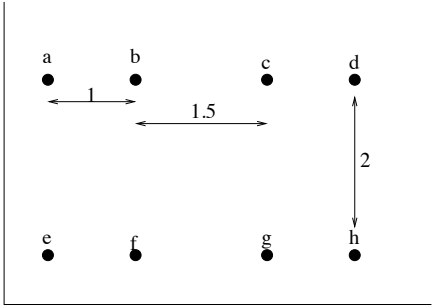

1.0
0.8
0.6
0.4
0.2
0.0
NYSE closing averages
Hog prices tumble
Oil prices slip
Ag trade reform.
Chrysler / Latin America
Japanese prime minister / Mexico
Fed holds interest rates steady
Fed to keep interest rates steady
Fed keeps interest rates steady
Fed keeps interest rates steady
Mexican markets
British FTSE index
War hero Colin Powell
War hero Colin Powell
Lloyd's CEO questioned
Lloyd's chief / U.S. grilling
Ohio Blue Cross
Lawsuit against tobacco companies
suits against tobacco firms
Indiana tobacco lawsuit
Viag stays positive
Most active stocks
CompuServe reports loss
Sprint / Internet access service
Planet Hollywood
Trocadero: tripling of revenues
Back−to−school spending is up
German unions split
Chains may raise prices
Clinton signs law

Time complexity of hierarchical methods

1. We must compute the similarity of all $N \times N$ pairs of documents.
2. In each of N iterations:.
   - We scan the $O(N^2)$ similarities to find the maximum similarity.
   - We merge the two clusters with maximum similarity.
   - We compute the similarity of the new cluster with all other clusters.
3. There are $O(N)$ iterations, each performing a $O(N^2)$ operation.
4. Overall complexity is $O(N^3)$.
5. Depending on the similarity function, a more efficient algorithm is possible.

Hierarchical clustering (example)

1. Consider 8 points in 2-D plan and their Euclidean distances.



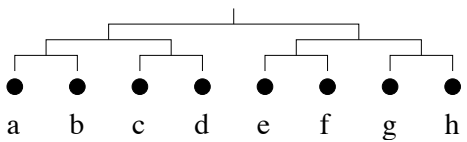
b	1						
c	2.5	1.5					
d	3.5	2.5	1				
e	2	$\sqrt{5}$	$\sqrt{10.25}$	$\sqrt{16.25}$			
f	$\sqrt{5}$	2	$\sqrt{6.25}$	$\sqrt{10.25}$	1		
g	$\sqrt{10.25}$	$\sqrt{6.25}$	2	$\sqrt{5}$	2.5	1.5	
h	$\sqrt{16.25}$	$\sqrt{10.25}$	$\sqrt{5}$	2	3.5	2.5	1
	a	b	c	d	e	f	g

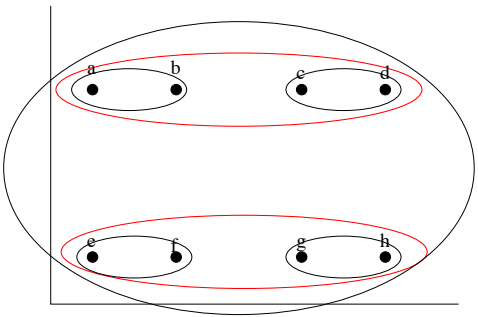
Hierarchical clustering (single link example)

1. In the first step, a-b, c-d, e-f, g-h merged.

c–d	1.5		
e–f	2	$\sqrt{6.25}$	
g–h	$\sqrt{6.25}$	2	1.5
	a–b	c–d	e–f

2. The final clustering is



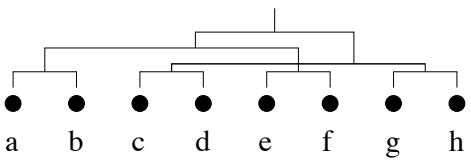


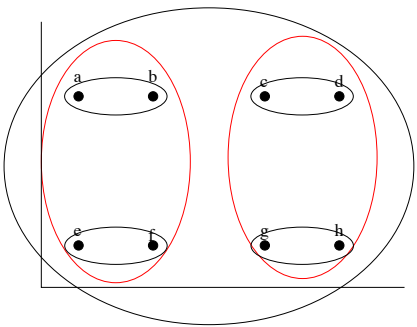
Hierarchical clustering (complete link example)

1. In the first step, a-b, c-d, e-f, g-h merged.

c–d	2.5	1.5				
	3.5	2.5				
e–f	2	$\sqrt{5}$	$\sqrt{10.25}$	**$\sqrt{16.25}$**		
	$\sqrt{5}$	2	$\sqrt{6.25}$	$\sqrt{10.25}$		
g–h	$\sqrt{10.25}$	$\sqrt{6.25}$	2	**$\sqrt{5}$**	2.5	1.5
	$\sqrt{16.25}$	$\sqrt{10.25}$	$\sqrt{5}$	2	**3.5**	2.5
	a–b		c–d		e–f	

2. In this step, we merge a-b/e-f and c-d/g-h.





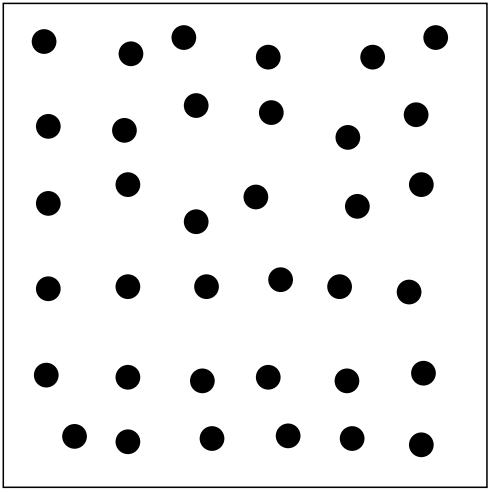
Flat or Hierarchical clustering?

1. When a hierarchical structure is desired: we should use hierarchical algorithm
2. Humans are bad at interpreting hierarchical clusterings (unless cleverly visualized)
3. For high efficiency, use flat clustering.
4. For deterministic results, use hierarchical agglomerative clustering.
5. Hierarchical agglomerative clustering also can be applied if K cannot be predetermined (can start without knowing K)

Evaluation of clustering

- Cluster evaluation assesses the feasibility of clustering analysis on a data set and the quality of the results generated by a clustering method. The major tasks of clustering evaluation include the following:
  1. Assessing clustering tendency : In this task, for a given data set, we assess whether a nonrandom structure exists in the data. Clustering analysis on a data set is meaningful only when there is a nonrandom structure in the data.
  2. Determining the number of clusters in a data set : Algorithms such as k-means, require the number of clusters in a data set as the parameter. A simple method is to set the number of clusters to about $\sqrt{n/2}$ for a data set of n points.
  3. Measuring clustering quality : After applying a clustering method on a data set, we want to assess how good the resulting clusters are. There are also measures that score clusterings and thus can compare two sets of clustering results on the same data set.

Assessing clustering tendency



Cluster validation and assessment

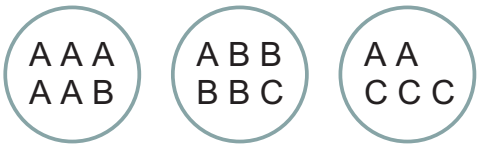
- How good is the clustering generated by a method?
- How can we compare the clusterings generated by different methods?
- Clustering is an unupervised learning technique and it is hard to evaluate the quality of the output of any given method.
- If we use probabilistic models, we can always evaluate the likelihood of a test set, but this has two drawbacks:
  1. It does not directly assess any clustering that is discovered by the model.
  2. It does not apply to non-probabilistic methods.
- We discuss some performance measures not based on likelihood.

Cluster validation and assessment

- The goal of clustering is to assign points that are similar to the same cluster, and to ensure that points that are dissimilar are in different clusters.
- There are several ways of measuring these quantities
  1. Internal criterion : Typical objective functions in clustering formalize the goal of attaining high intra-cluster similarity and low inter-cluster similarity. But good scores on an internal criterion do not necessarily translate into good effectiveness in an application. An alternative to internal criteria is direct evaluation in the application of interest.
  2. External criterion : Suppose we have labels for each object. Then we can compare the clustering with the labels using various metrics. We will use some of these metrics later, when we compare clustering methods.

Purity

- Purity is a simple and transparent evaluation measure. Consider the following clustering.

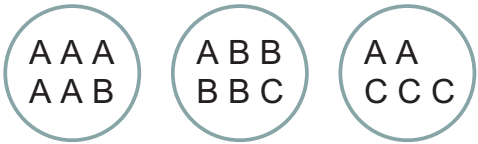


- Let N_{ij} be the number of objects in cluster i that belongs to class j and $N_i = \sum_{j=1}^{C} N_{ij}$ be the total number of objects in cluster i.
- We define purity of cluster i as $p_i \triangleq \max_j \left(\frac{N_{ij}}{N_i}\right)$, and the overall purity of a clustering as

$$purity \triangleq \sum_i \frac{N_i}{N} p_i.$$

Purity

- For the following figure, the purity is



$$\frac{6}{17}\frac{5}{6} + \frac{6}{17}\frac{4}{6} + \frac{5}{17}\frac{3}{5} = \frac{5+4+3}{17} = 0.71$$

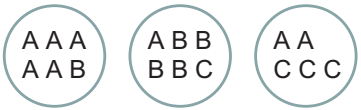
- Bad clusterings have purity values close to 0, a perfect clustering has a purity of 1.
- High purity is easy to achieve when the number of clusters is large. In particular, purity is 1 if each point gets its own cluster. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters.

Rand index

- Let $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_C\}$ be two different clustering of N data points.
- For example, U might be the estimated clustering and V is reference clustering derived from the class labels.
- Define a 2×2 contingency table, containing the following numbers:
  1. TP is the number of pairs that are in the same cluster in both U and V (true positives);
  2. TN is the number of pairs that are in different clusters in both U and V (true negatives);
  3. FN is the number of pairs that are in different clusters in U but the same cluster in V (false negatives);
  4. FP is the number of pairs that are in the same cluster in U but different clusters in V (false positives).
- Rand index is defined as $RI \triangleq \frac{TP+TN}{TP+FP+FN+TN}$
- Rand index can be interpreted as the fraction of clustering decisions that are correct. Clearly $RI \in [0, 1]$.

Rand index (example)

- Consider the following clustering



- The three clusters contain 6, 6 and 5 points, so we have

$$TP + FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40.$$

- The number of true positives

$$TP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20.$$

- Then $FP = 40 - 20 = 20$. Similarly, $FN = 24$ and $TN = 72$.
- Hence Rand index

$$RI = \frac{20 + 72}{20 + 20 + 24 + 72} = 0.68.$$

Adjusted rand index (example)

- Consider the following clustering

- Hence Rand index

$$RI = \frac{20 + 72}{20 + 20 + 24 + 72} = 0.68.$$

- Rand index only achieves its lower bound of 0 if $TP = TN = 0$, which is a rare event. We can define an adjusted Rand index

$$ARI \triangleq \frac{index - \mathbb{E}[index]}{\max index - \mathbb{E}[index]}.$$

Adjusted rand index (example)

- For computing adjusted Rand index, we build a contingency matrix, where columns are gold clusters and rows are obtained clusters.

$$
\begin{aligned}
ARI &\triangleq \frac{index - \mathbb{E}[index]}{\max index - \mathbb{E}[index]} \\
&= \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}
\end{aligned}
$$

- n_{ij} is the count in cell of (i, j) of contingency matrix.
- a_i is the sum of row i of contingency matrix.
- b_j is the sum of column j of contingency matrix.
- Exercise: Assume that the gold clustering is $\{\{A, D\}, \{B, C\}, \{E, F\}\}$ and obtained clustering is $\{\{A, B\}, \{E, F\}, \{C, D\}\}$, calculate ARI.

Mutual information

- We can measure cluster quality is computing mutual information between U and V.
- Let $P_{UV}(i,j) = \frac{|u_i \cap v_j|}{N}$ be the probability that a randomly chosen object belongs to cluster u_i in U and v_j in V.
- Let $P_U(i) = \frac{|u_i|}{N}$ be the be the probability that a randomly chosen object belongs to cluster u_i in U.
- Let $P_V(j) = \frac{|v_j|}{N}$ be the be the probability that a randomly chosen object belongs to cluster v_j in V.
- Then mutual information is defined

$$\mathbb{I}(U,V) \triangleq \sum_{i=1}^{R}\sum_{j=1}^{C} P_{UV}(i,j) \log \frac{P_{UV}(i,j)}{P_U(i)P_V(j)}.$$

- This lies between 0 and $\min\{\mathbb{H}(U), \mathbb{H}(V)\}$.

- Then mutual information is defined

$$\mathbb{I}(U,V) \triangleq \sum_{i=1}^{R}\sum_{j=1}^{C} P_{UV}(i,j)\log\frac{P_{UV}(i,j)}{P_U(i)P_V(j)}.$$

- This lies between 0 and $\min\{\mathbb{H}(U),\mathbb{H}(V)\}$.
- The maximum value can be achieved by using a lots of small clusters, which have low entropy.
- To compensate this, we can use normalized mutual information (NMI)

$$NMI(U,V) \triangleq \frac{\mathbb{I}(U,V)}{\frac{1}{2}[\mathbb{H}(U)+\mathbb{H}(V)]}.$$

- This lies between 0 and 1.

References

Reading

1. Chapter 16 of Information Retrieval Book[2]

[2] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

References

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Questions?