

Laboratorio di Apprendimento Automatico

Fabio Aiolli

Università di Padova

What is clustering?

- **Clustering**: the process of grouping a set of objects into classes of similar objects
 - The commonest form of *unsupervised learning*
 - Unsupervised learning = learning from raw data, as opposed to supervised data where a classification of examples is given
 - A common and important task that finds many applications
 - Not only Example Clustering (e.g. feature)

The Clustering Problem

Given:

- A set of documents $D=\{d_1,...d_n\}$
- A similarity measure (or distance metric)
- A partitioning criterion
- A desired number of clusters K

Compute:

- An assignment function $\gamma : D \rightarrow \{1,...,K\}$
 - None of the clusters is empty
 - Satisfies the partitioning criterion w.r.t. the similarity measure

Issues for clustering

- Representation for clustering
 - Document representation
 - Vector space? Normalization?
 - Need a notion of similarity/distance
- How many clusters?
 - Fixed a priori?
 - Completely data driven?
 - Avoid “trivial” clusters - too large or small
 - In an application, if a cluster's too large, then for navigation purposes you've wasted an extra user click without whittling down the set of documents much.

Objective Functions

- Often, the goal of a clustering algorithm is to optimize an objective function
- In this cases, clustering is a search (optimization) problem
- $K^N / K!$ different clustering available
- Most partitioning algorithms start from a guess and then refine the partition
- Many local minima in the objective function implies that different starting point may lead to very different (and unoptimal) final partitions

What Is A Good Clustering?

- Internal criterion: A good clustering will produce high quality clusters in which:
 - the **intra-class** (that is, intra-cluster) similarity is high
 - the **inter-class** similarity is low
 - The measured quality of a clustering depends on both the document representation and the similarity measure used

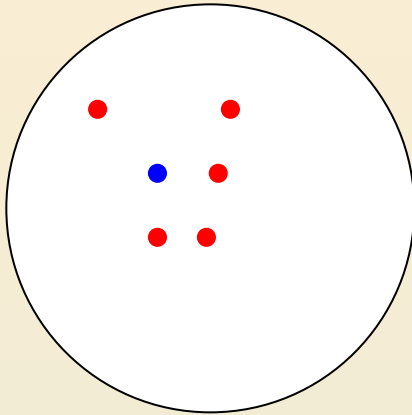
External criteria for clustering quality

- Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data
- Assesses a clustering with respect to **ground truth**
- Assume documents with C gold standard classes, while our clustering algorithms produce K clusters, $\omega_1, \dots, \omega_K$ with n_i members.

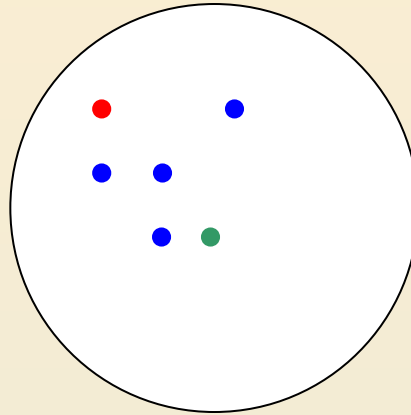
External Evaluation of Cluster Quality

- Simple measure: **purity**, the ratio between the dominant class in the cluster π_i and the size of cluster ω_i
- Others are entropy of classes in clusters (or mutual information between classes and clusters)

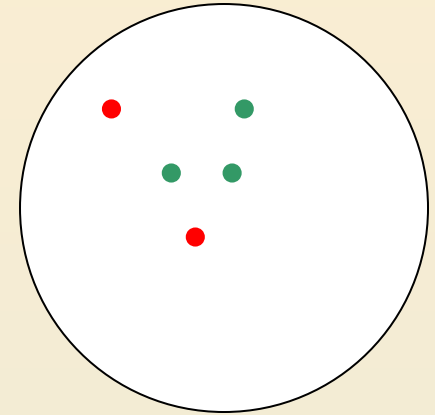
Purity example



Cluster I



Cluster II



Cluster III

Cluster I: Purity = $1/6 (\max(5, 1, 0)) = 5/6$

Cluster II: Purity = $1/6 (\max(1, 4, 1)) = 4/6$

Cluster III: Purity = $1/5 (\max(2, 0, 3)) = 3/5$

Rand Index

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	A (tp)	C (fn)
Different classes in ground truth	B (fp)	D (tn)

Rand index: symmetric version

$$RI = \frac{A + D}{A + B + C + D}$$

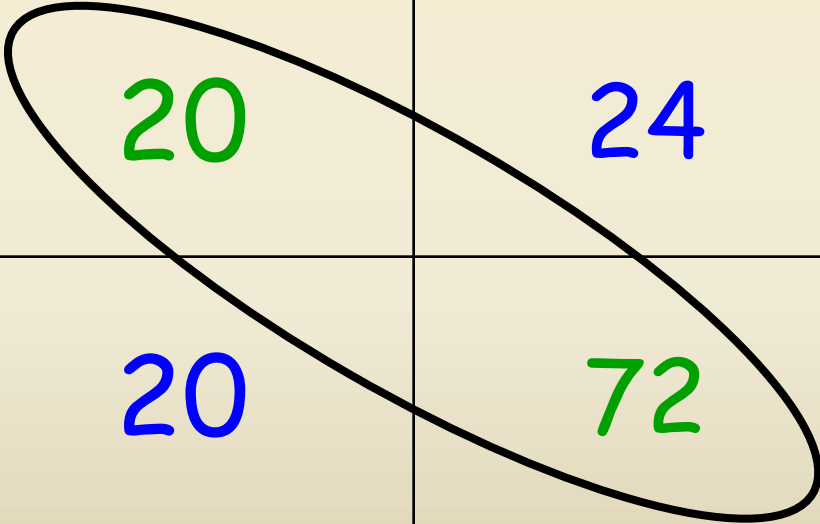
Compare with standard Precision and Recall.

$$P = \frac{A}{A + B}$$

$$R = \frac{A}{A + C}$$

Rand Index example: 0.68

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	20	24
Different classes in ground truth	20	72



Clustering Algorithms

- Partitional algorithms
 - Usually start with a random (partial) partitioning
 - Refine it iteratively
 - K means clustering
 - Model based clustering
- Hierarchical algorithms
 - Bottom-up, agglomerative
 - Top-down, divisive

Partitioning Algorithms

- Partitioning method: Construct a partition of n documents into a set of K clusters
- Given: a set of documents and the number K
- Find: a partition of K clusters that optimizes the chosen partitioning criterion
 - Globally optimal: exhaustively enumerate all partitions
 - Effective heuristic methods: K -means and K -medoids algorithms

K-Means

- Assumes documents are real-valued vectors.
- Clusters based on *centroids* (aka the *center of gravity* or mean) of points in a cluster, c :

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

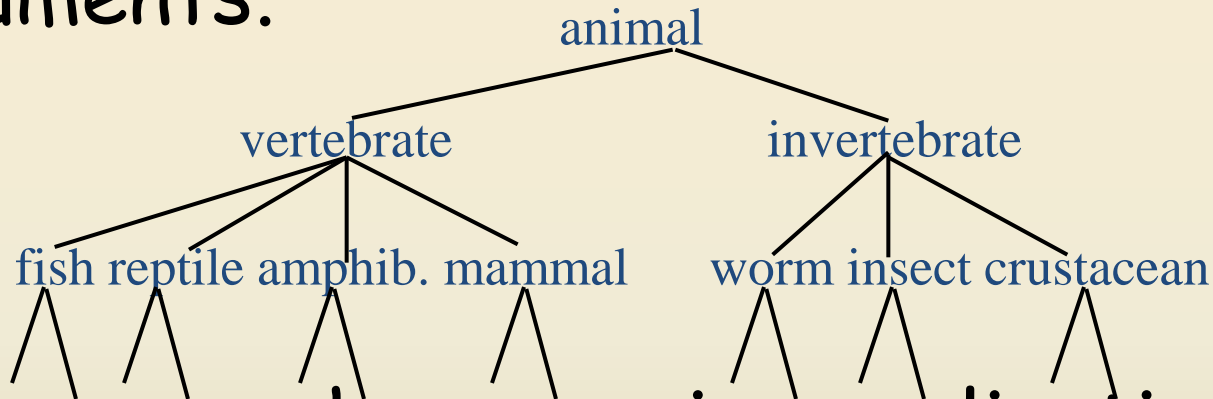
- Reassignment of instances to clusters is based on distance to the current cluster centroids.
 - (Or one can equivalently phrase it in terms of similarities)

How Many Clusters?

- Number of clusters K is given
 - Partition n docs into predetermined number of clusters
- Finding the “right” number of clusters is part of the problem
 - Given docs, partition into an “appropriate” number of subsets.
 - E.g., for query results - ideal value of K not known up front - though UI may impose limits.

Hierarchical Clustering

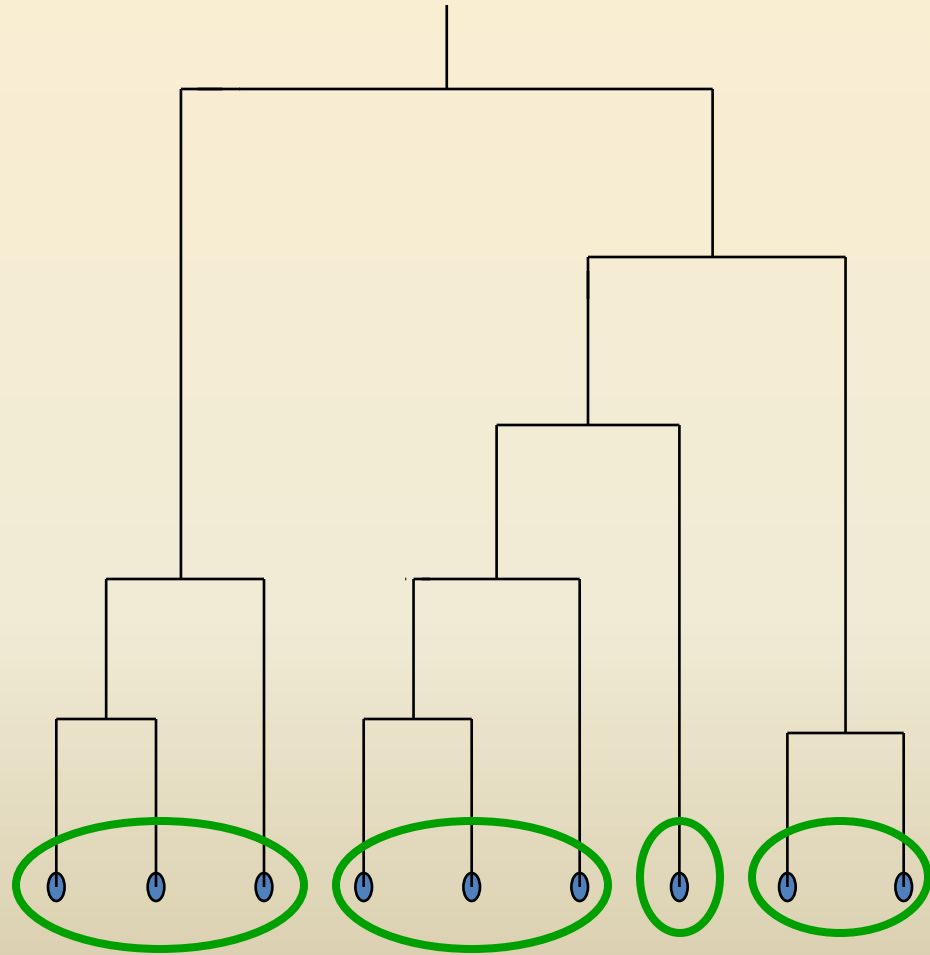
- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of documents.



- One approach: recursive application of a partitioning clustering algorithm

Dendrogram: Hierarchical Clustering

Clustering obtained by cutting the dendrogram at a desired level: each **connected** component forms a cluster.



The dendrogram

- The y-axis of the dendrogram represents the **combination similarities**, i.e. the similarities of the clusters merged by a the horizontal lines for a particular y
- Assumption: The merge operation is **monotonic**, i.e. if s_1, \dots, s_{k-1} are successive combination similarities, then $s_1 \leq s_2 \leq \dots \leq s_{k-1}$ must hold

Hierarchical Agglomerative Clustering (HAC)

- Starts with each doc in a separate cluster
 - then repeatedly joins the **closest pair** of clusters, until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

Closest pair of clusters

- Many variants to defining closest pair of clusters
- *Single-link*
 - Similarity of the *most* cosine-similar (single-link)
- *Complete-link*
 - Similarity of the “furthest” points, the *least* cosine-similar
- *Centroid*
 - Clusters whose centroids (centers of gravity) are the most cosine-similar
- *Average-link*
 - Average cosine between pairs of elements

Summarizing

Single-link	Max sim of any two points	$O(N^2)$	Chaining effect
Complete-link	Min sim of any two points	$O(N^2 \log N)$	Sensitive to outliers
Centroid	Similarity of centroids	$O(N^2 \log N)$	Non monotonic
Group-average	Avg sim of any two points	$O(N^2 \log N)$	OK