Text Mining

Emanuele Guidotti

2019/2020

Clustering

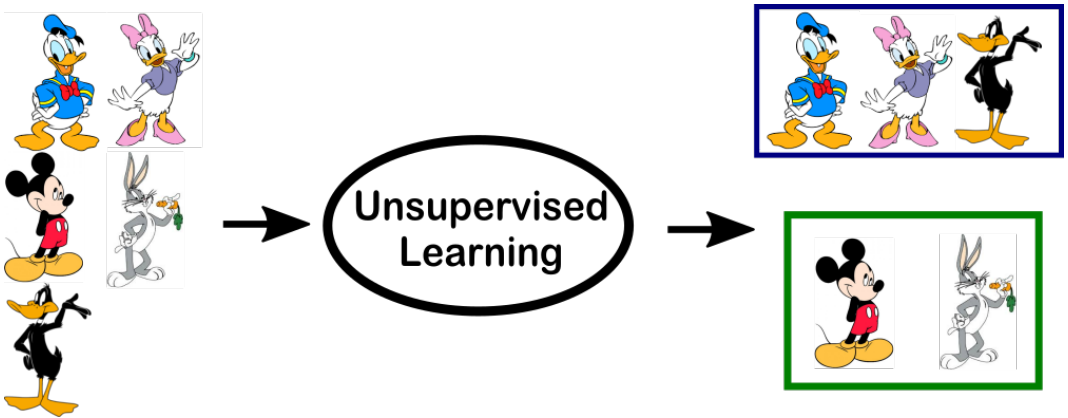
Clustering algorithms are **unsupervised learning algorithms** aimed at **grouping similar items together**.

Clustering analysis is broadly used in applications such as...

- exploratory data analysis
- market research
- customer segmentation
- topics discovery

... and in every task of grouping a set of objects in such a way that objects in the same group (called *cluster*) are more **similar** to each other than to those in other groups.

What does similarity mean?



Unsupervised grouping of similar (?) items... We need to explicitly define a **similarity function** to measure similarity between two objects.

Similarity-Based Clustering

Explicitly define a function to measure (dis)similarity between two objects. This is usually a **metric**, thus satisfying:

- symmetry: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
- separation: $d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$
- triangular inequality: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$

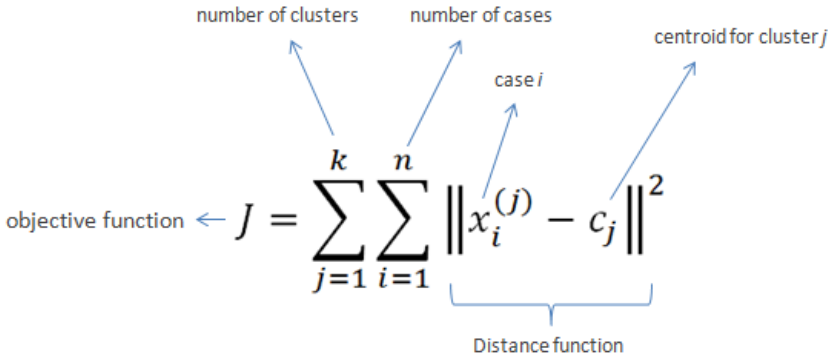
The choice of the similarity function is crucial.

While the cosine similarity is quite popular, the Euclidean distance is usually not well-suited for text clustering, unless term vectors are properly **normalized**.

Clustering Algorithms | K-Means Clustering

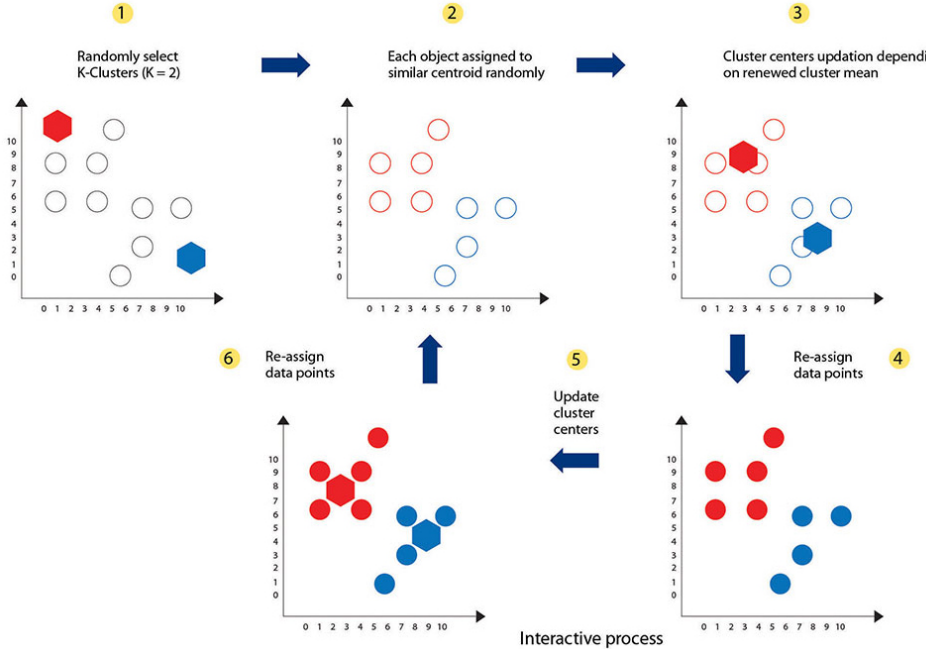
K-Means Clustering

K-Means is **based on Euclidean distances** between data points, partitioning the observations into k sets so as to minimize the objective function:



Algorithm

- Represent text objects as term vectors
- Start with k randomly selected points and assume they are the centroids of k clusters (initialization points)
- Assign every point to a cluster whose centroid is the closest to the point
- Re-compute the centroid for each cluster based on the newly assigned points in the cluster
- Repeat this process until the algorithm converges, i.e. when the assignments no longer change


1
Randomly select
K-Clusters (K = 2)
2
Each object assigned to
similar centroid randomly
3
Cluster centers updation depending
on renewed cluster mean
4
Re-assign
data points
5
Update
cluster
centers
6
Re-assign
data points
Interactive process

Clustering Algorithms | Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering

Hierarchical clustering has the distinct advantage that **any valid measure of distance can be used**. In fact, the observations themselves are not required: all that is used is a matrix of distances.

The Agglomerative Hierarchical Clustering is a bottom-up approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

In general, the results of hierarchical clustering are usually presented in a **dendrogram**.

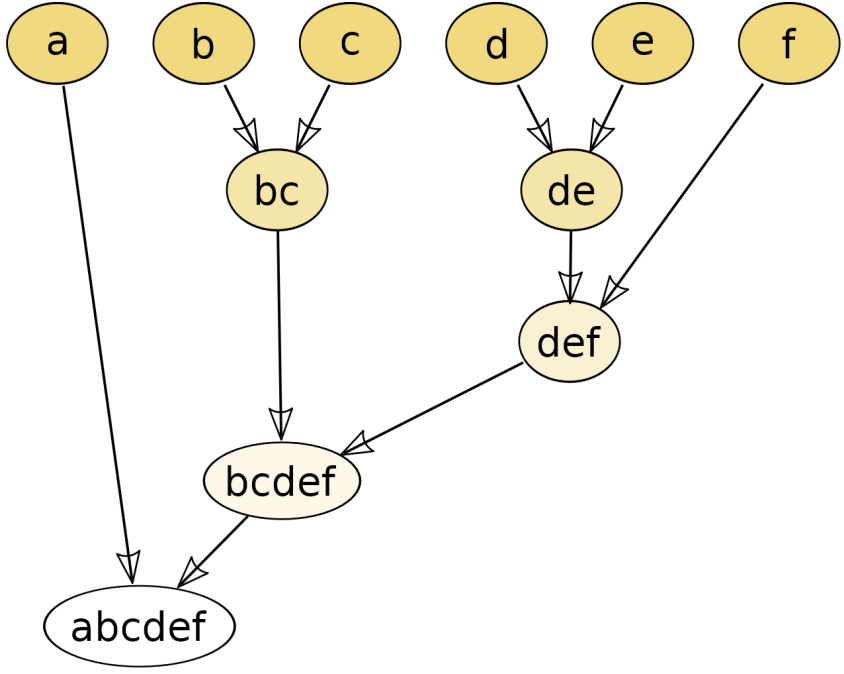
Algorithm

- Represent text objects as term vectors and define a similarity function (e.g. Euclidean distance, cosine similarity, etc.)
- Treat each point as a separate cluster
- Identify the two **clusters that are closest together**
- Merge the two most similar clusters
- Iterate until some stopping criterion is met (number of clusters, similarity threshold, etc.).

Linkage criteria

How to identify the two clusters that are closest together? We need a measure of group similarity, i.e. distance between sets of observations (*linkage criterion*). Some common choices:

- **Single Linkage**: the similarity between two clusters is the greatest similarity (minimum distance) between any pairs from these two clusters.
- **Complete Linkage**: the similarity between two clusters is the smallest similarity (maximum distance) between any pairs from these two clusters.
- **Average Linkage**: the similarity between two clusters is the average similarity/distance between the pairs in the two clusters.


a
b
c
d
e
f
bc
de
def
bcdef
abcdef

Evaluation

Direct Evaluation

How close are the system-generated clusters to the expected, manually labelled, clusters?

- Given a test set, manually label the dataset and create the ideal cluster result; e.g. assign each cluster to the class which is most frequent in the cluster
- The problem becomes a supervised learning task, i.e. a classification problem
- Apply the standard evaluation metrics used for classification tasks

Example: Purity

- Assign each document to the class which is most frequent in its cluster.
- Count the number of correctly assigned documents and divide by the total number of documents.

Example: Rand Index

- Create a list of all possible pairs of documents.
- Label each pair as (1) the two documents are similar or (0) the two documents are not similar.
- Count the number of times two similar documents are in the same cluster or two dissimilar documents are in different clusters. Divide by the total number of pairs.

Indirect Evaluation

How useful are the clustering results for the intended application?

- Create a test set for the intended application to quantify the performance of any (machine learning) system for this application
- Choose a benchmark system to compare with
- Add a clustering algorithm to the benchmark system
- Compare the performance of the clustering enhanced system with the benchamrk, in terms of any performance measure for the application

Take Home Concepts

Take Home Concepts

- Clustering is the process of grouping similar entities together
- Clustering is an unsupervised machine learning technique
- The choice of the similarity function is crucial
- How K-Means works
- How Agglomerative Hierarchical Clustering works
- Evaluation of clustering can be done both directly and indirectly

References

References

- https://towardsdatascience.com/similarity-measures-e3dbd4e58660
  - read (6 mins)
- http://charuaggarwal.net/text-content.pdf
  - 4.1, 4.2, 4.2.1, 4.2.1.1, 4.3, 4.3.1, 4.3.2