CSE6242 / CX4242: Data & Visual Analytics

Clustering

Duen Horng (Polo) Chau
Assistant Professor
Associate Director, MS Analytics
Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos, Parishit Ram (GT PhD alum; SkyTree), Alex Gray
Clustering in Google Image Search

Video: http://youtu.be/WosBs0382SE
http://googlesystem.blogspot.com/2011/05/google-image-search-clustering.html
Clustering

The most common type of unsupervised learning

High-level idea: group similar things together

“Unsupervised” because clustering model is learned without any labeled examples
Applications of Clustering

- google news
- IMDB (movie sites)
- anomaly detection
- detecting population subgroups (community detection)
  - as in healthcare
- Twitter hashtags
  - text-based clustering
- (Age detection)
Clustering techniques you’ve got to know

K-means
Hierarchical Clustering
DBSCAN
K-means (the “simplest” technique)

Summary

• We tell K-means the value of $k$ (#clusters we want)
• Randomly initialize the k cluster “means” (“centroids”)
• Assign each item to the cluster whose mean the item is closest to (so, we need a similarity function)
• Update the new “means” of all k clusters.
• If all items’ assignments do not change, stop.

Java demo: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
YouTube video demo: https://youtu.be/IuRb3y8qKX4?t=3m4s
K-means

What’s the catch?


Need to **decide k ourselves**.

- How to find the optimal k?

Only locally optimal (vs global)

- Different initialization gives different clusters
  - How to “fix” this?
- “Bad” starting points can cause algorithm to converge slowly

- Can work for relatively large dataset
  - Time complexity \(O(d \cdot n \log n)\) per iteration
    (assumptions: \(n \gg k\), dimension d is small)

http://www.cs.cmu.edu/~./dpelleg/download/kmeans.ps
Hierarchical clustering

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

High-level idea: build a tree (hierarchy) of clusters

Agglomerative (bottom-up)
- Start with individual items
- Then iteratively group into larger clusters

Divisive (top-down)
- Start with all items as one cluster
- Then iteratively divide into smaller clusters
Ways to calculate distances between two clusters

**Single linkage**
- minimum of distance between clusters
- similarity of two clusters = similarity of the clusters’ **most similar** members

**Complete linkage**
- maximum of distance between clusters
- similarity of two clusters = similarity of the clusters’ **most dissimilar** members

**Average linkage**
- distance between cluster centers
Example from Wikipedia

Raw data

Dendrogram
Clustered Iris data set
(the labels give the true flower species)
Hierarchical clustering for large datasets?

- OK for small datasets (e.g., <10K items)
- Time complexity between $O(n^2)$ to $O(n^3)$ where $n$ is the number of data items
- Not good for millions of items or more
- But great for understanding concept of clustering
DBSCAN

“Density-based spatial clustering with noise”
https://en.wikipedia.org/wiki/DBSCAN

Received “test-of-time award” at KDD’14 — an extremely prestigious award.

Only need two parameters:
1. “radius”
2. minimum number of points (e.g., 4) required to form a dense region

Yellow “border points” are density-reachable from red “core points”, but not vice-versa.
Visualizing Clusters
D3 has some built-in techniques

Visualizing Graph Communities (using colors)
Visualizing Graph Communities
(using colors and convex hulls)

Visualizing Graph Communities as Matrix

https://bost.ocks.org/mike/miserables/

Require good node ordering!

Les Misérables Co-occurrence
Visualizing Graph Communities as Matrix

Require good node ordering!

Fully-automated way: “Cross-associations”
http://www.cs.cmu.edu/~christos/PUBLICATIONS/kdd04-cross-assoc.pdf
Graph Partitioning

If you know, or want to, specify #communities, use METIS, the most popular graph partitioning tools
http://glaros.dtc.umn.edu/gkhome/views/metis
Visualizing **Topics** as Matrix

Termite: Visualization Techniques for Assessing Textual Topic Models
Jason Chuang, Christopher D. Manning, Jeffrey Heer. AVI 2012.
http://vis.stanford.edu/papers/termite
Visualizing **Topics** as Matrix

**Termite: Visualization Techniques for Assessing Textual Topic Models**
Jason Chuang, Christopher D. Manning, Jeffrey Heer. AVI 2012.
http://vis.stanford.edu/papers/termite