Clustering

- Document clustering
  - Motivations
  - Document representations
  - Success criteria
- Clustering algorithms
  - K-means
  - Model-based clustering (EM clustering)
What is clustering?

- **Clustering** is the process of grouping a set of physical or abstract objects into classes of similar objects
  - It is the commonest form of unsupervised learning
    - Unsupervised learning = learning from raw data, as opposed to supervised data where the correct classification of examples is given
  - It is a common and important task that finds many applications in IR and other places

Why cluster documents?

- Whole corpus analysis/navigation
  - Better user interface
- For improving recall in search applications
  - Better search results
- For better navigation of search results
- For speeding up vector space retrieval
  - Faster search
Navigating document collections

• Standard IR is like a book index
• Document clusters are like a table of contents
• People find having a table of contents useful

Index
Aardvark, 15
Blueberry, 200
Capricorn, 1, 45-55
Dog, 79-99
Egypt, 65
Falafel, 78-90
Giraffes, 45-59
…

Table of Contents
1. Science of Cognition
   1.a. Motivations
      1.a.i. Intellectual Curiosity
      1.a.ii. Practical Applications
   1.b. History of Cognitive Psychology
2. The Neural Basis of Cognition
   2.a. The Nervous System
   2.b. Organization of the Brain
   2.c. The Visual System
3. Perception and Attention
   3.a. Sensory Memory
   3.b. Attention and Sensory Information Processing

Corpus analysis/navigation

• Given a corpus, partition it into groups of related docs
  – Recursively, can induce a tree of topics
  – Allows user to browse through corpus to find information
  – Crucial need: meaningful labels for topic nodes.
• Yahoo!: manual hierarchy
  – Often not available for new document collection
For improving search recall

- Cluster hypothesis - Documents with similar text are related
- Therefore, to improve search recall:
  - Cluster docs in corpus a priori
  - When a query matches a doc $D$, also return other docs in the cluster containing $D$
- Hope if we do this: The query “car” will also return docs containing *automobile*
  - Because clustering grouped together docs containing *car* with those containing *automobile*.

Why might this happen?
For better navigation of search results

- For grouping search results thematically
  - clusty.com / Vivisimo

- And more visually: Kartoo.com
Navigating search results (2)

• One can also view grouping documents with the same sense of a word as clustering.
• Given the results of a search (e.g., jaguar, *NLP*), partition into groups of related docs.
• Can be viewed as a form of word sense disambiguation.
• E.g., *jaguar* may have senses:
  – The car company
  – The animal
  – The football team
  – The video game
• Recall query reformulation/expansion discussion.
For speeding up vector space retrieval

• In vector space retrieval, we must find nearest doc vectors to query vector
• This entails finding the similarity of the query to every doc
  – slow (for some applications)
• By clustering docs in corpus a priori
  – find nearest docs in cluster(s) close to query
  – inexact but avoids exhaustive similarity computation

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 criterion: The quality of a clustering is also measured by its ability to discover some or all of the hidden patterns or latent classes
  – Assessable with gold standard data
External Evaluation of Cluster Quality

- Assesses clustering with respect to ground truth
- Assume that there are C gold standard classes, while our clustering algorithms produce \( k \) clusters, \( \pi_1, \pi_2, \ldots, \pi_k \) with \( n_i \) members.
- Simple measure: purity, the ratio between the dominant class in the cluster \( \pi_i \) and the size of cluster \( \pi_i \)
  \[
Purity(\pi_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad j \in C
  \]
- Others are entropy of classes in clusters (or mutual information between classes and clusters)

\[
Purity(\pi_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad j \in C
\]

Purity

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
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.

What makes docs “related”?

- Ideal: semantic similarity.
- Practical: statistical similarity
  - We will use cosine similarity.
  - Docs as vectors.
  - For many algorithms, easier to think in terms of a distance (rather than similarity) between docs.
  - We will describe algorithms in terms of cosine similarity.

\[
\text{Cosine similarity of normalized } D_j, D_k : \\
\text{sim}(D_j, D_k) = \sum_{i=1}^{m} w_{ij} \times w_{ik} \\
\text{Aka normalized inner product.}
\]
Recall doc as vector

- Each doc $j$ is a vector of $tf \times idf$ values, one component for each term.
- Can normalize to unit length.
- So we have a vector space
  - terms are axis - aka features
  - $n$ docs live in this space
  - even with stemming, may have 20,000+ dimensions
  - do we really want to use all terms?
    - Different from using vector space for search. Why?

Intuition

Postulate: Documents that are “close together” in vector space talk about the same things.
Clustering Algorithms

- Partitioning “flat” algorithms
  - Usually start with a random (partial) partitioning
  - Refine it iteratively
    - $k$ means/medoids 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
How hard is clustering?

• One idea is to consider all possible clusterings, and pick the one that has best inter and intra cluster distance properties.

• Suppose we are given $n$ points, and would like to cluster them into $k$-clusters.
  – How many possible clusterings?

• Too hard to do it brute force or optimally.

• Solution: Iterative optimization algorithms.
  – Start with a clustering, iteratively improve it (e.g., K-means).

K-Means

• Assumes documents are real-valued vectors.

• Clusters based on centroids (aka the center of gravity or mean) of points in a cluster, $c$:

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

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

Let $d$ be the distance measure between instances. Select $k$ random instances \{$s_1, s_2, \ldots, s_k$\} as seeds. Until clustering converges or other stopping criterion:

For each instance $x_i$:

Assign $x_i$ to the cluster $c_j$ such that $d(x_i, s_j)$ is minimal.

(*Update the seeds to the centroid of each cluster*)

For each cluster $c_j$

\[s_j = \mu(c_j)\]
Termination conditions

• Several possibilities, e.g.,
  – A fixed number of iterations.
  – Doc partition unchanged.
  – Centroid positions don’t change.

Does this mean that the docs in a cluster are unchanged?

Time Complexity

• Assume computing distance between two instances is $O(m)$ where $m$ is the dimensionality of the vectors.
• Reassigning clusters: $O(kn)$ distance computations, or $O(knm)$.
• Computing centroids: Each instance vector gets added once to some centroid: $O(nm)$.
• Assume these two steps are each done once for $i$ iterations: $O(i kmn)$.
• Linear in all relevant factors, assuming a fixed number of iterations, more efficient than hierarchical agglomerative methods
Seed Choice

- Results can vary based on random seed selection.
- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.
  - Select good seeds using a heuristic (e.g., doc least similar to any existing mean)
  - Try out multiple starting points
  - Initialize with the results of another method.

Recap

- Why cluster documents?
  - For improving recall in search applications
  - For speeding up vector space retrieval
  - Navigation
  - Presentation of search results
- \( k \)-means basic iteration
  - At the start of the iteration, we have \( k \) centroids.
  - Each doc assigned to the nearest centroid.
  - All docs assigned to the same centroid are averaged to compute a new centroid;
    - thus have \( k \) new centroids.
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.
• Can usually take an algorithm for one flavor and convert to the other.

\( k \) not specified in advance

• Say, the results of a query.
• Solve an optimization problem: penalize having lots of clusters
  – application dependent, e.g., compressed summary of search results list.
• Tradeoff between having more clusters (better focus within each cluster) and having too many clusters
Given a clustering, define the Benefit for a doc to be the cosine similarity to its centroid.

Define the Total Benefit to be the sum of the individual doc Benefits.

**Why is there always a clustering of Total Benefit** $n$?

Penalize lots of clusters

For each cluster, we have a Cost $C$.
Thus for a clustering with $k$ clusters, the Total Cost is $kC$.
Define the Value of a clustering to be $\text{Value} = \text{Total Benefit} - \text{Total Cost}$.
Find the clustering of highest value, over all choices of $k$.
- Total benefit increases with increasing $K$. But can stop when it doesn’t increase by “much”. The Cost term enforces this.
Convergence

• Why should the K-means algorithm ever reach a *fixed point*?
  – A state in which clusters don’t change.

• K-means is a special case of a general procedure known as the *Expectation Maximization (EM) algorithm*.
  – EM is known to converge.
  – Number of iterations could be large.

Convergence of K-Means

• Define goodness measure of cluster $k$ as sum of squared distances from cluster centroid:
  – $G_k = \sum_i (v_i - c_k)^2$ (sum all $v_i$ in cluster $k$)

• $G = \sum_k G_k$

• Reassignment monotonically reduces $G$ since each vector is assigned to the closest centroid.

• Recomputation monotonically decreases each $G_k$ since: ($m_k$ is number of members in cluster)
  – $\sum (v_{in} - a)^2$ reaches minimum for:
  – $\sum -2(v_{in} - a) = 0$
K-means issues, variations, etc.

- Recomputing the centroid after every assignment (rather than after all points are re-assigned) can improve speed of convergence of K-means
- Assumes clusters are spherical in vector space
  - Sensitive to coordinate changes, weighting etc.
- Disjoint and exhaustive
  - Doesn’t have a notion of “outliers”

Soft Clustering

- Clustering typically assumes that each instance is given a “hard” assignment to exactly one cluster.
- Does not allow uncertainty in class membership or for an instance to belong to more than one cluster.
- Soft clustering gives probabilities that an instance belongs to each of a set of clusters.
- Each instance is assigned a probability distribution across a set of discovered categories (probabilities of all categories must sum to 1).
Hierarchical Clustering

- Build a tree-based hierarchical taxonomy \((\text{dendrogram})\) from a set of unlabeled examples.

- One option to produce a hierarchical clustering is recursive application of a partitional clustering algorithm to produce a hierarchical clustering.

---

“The Curse of Dimensionality”

- Why document clustering is difficult
  - While clustering looks intuitive in 2 dimensions, many of our applications involve 10,000 or more dimensions…
  - High-dimensional spaces look different: the probability of random points being close drops quickly as the dimensionality grows.
  - One way to look at it: in large-dimension spaces, random vectors are almost all almost perpendicular. Why?
- Solution: Dimensionality reduction … important for text
Related Tasks

- **TDT**
  - Topic Detection: “Dynamic” Clustering
  - Topic Tracking: on-line categorization
  - Story Segmentation
  - First Story Detection
  - New Information Detection
  - Story Link Detection
- **TIDES**
  - All of the above in multilingual and multimedia
- Word cloud
- And others…