Foundation of Data Science
Lecture 7, Module 1
Fall 2022

Rumi Chunara, PhD

Machine Learning

- There are two main categories of machine learning:
  - supervised learning already discussed
  - unsupervised learning focus today!

- **Unsupervised learning:**
  - Extracting structure from data
    - Example: segment grocery store shoppers into “clusters” that exhibit similar behaviors
  - Goal is “representation”

Unsupervised Learning

- Unsupervised learning has some clear differences from supervised learning. With **unsupervised learning:**
  - There is no clear objective
  - There is no *right answer* (hard to tell how well you are doing)
  - There is no response variable, just observations with features
  - Labeled data is not required

Unsupervised Learning: Example

Unsupervised learning example: Image clustering

- Input data: Images from Google
- Features: Numerical representations of the images
- Response: There isn't one (no hand-labeling required!)

Perform unsupervised learning

- Cluster the images based on "similarity"
- Might find a "dog cluster", might not
- You're done!

Sometimes, unsupervised learning is used as a "preprocessing" step for supervised learning!

Clustering

- Clustering is a technique for finding similarity groups in data, called **clusters**.
  - It groups data instances that are similar (near) to each other in one cluster and data instances that are very different (far away) from each other into different clusters.
- Clustering is often modeled as an unsupervised learning task as no classes denoting an *a priori* grouping of the data instances are given, which is the case in supervised learning.
- Historically, clustering is often considered synonymous with unsupervised learning.
  - However, association rule mining is also unsupervised, for example!
  - For example, the rule {onlions, potato} => {burger} found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat.

What is clustering for?

- Example 1: groups people of similar sizes together to make “small”, “medium” and “large” t-shirts.
  - Tailor-made for each person: too expensive
  - One-size-fits-all: does not fit all
- Example 2: In marketing, segment customers according to their similarities
  - To do targeted marketing
- Example 3: Given a collection of text documents, we want to organize them according to their content similarities
  - To produce a topic hierarchy

What is clustering for?

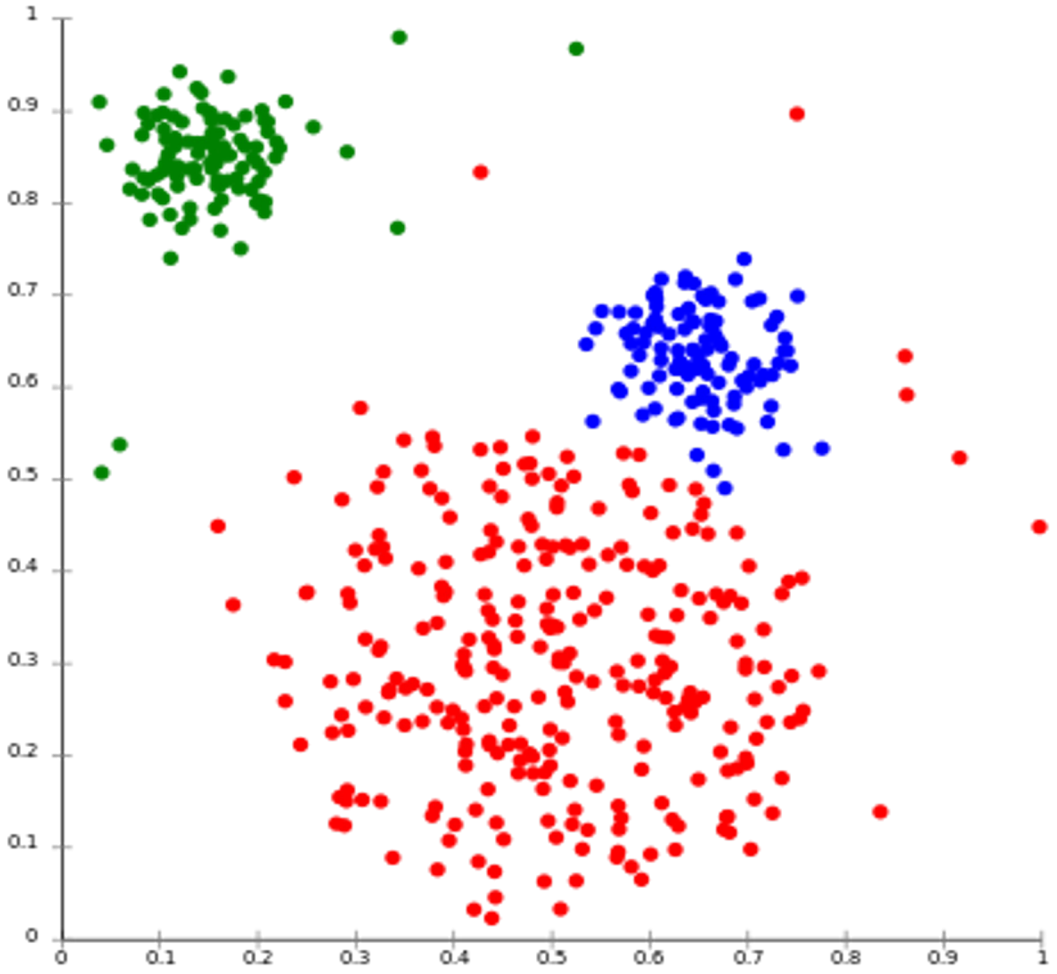
- In fact, clustering is one of the most utilized data mining techniques
  - It has a long history, and used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc
  - In recent years, due to the rapid increase of online documents, text clustering becomes important

Aspects of clustering

- A clustering algorithm
  - Partitional clustering
  - Hierarchical clustering
- A distance function (similarity, or dissimilarity)
- Clustering quality
  - Inter-cluster distance ⇒ maximized
  - Intra-cluster distance ⇒ minimized
- The quality of a clustering result depends on the algorithm, the distance function, and the application

Stereotypical Clustering

Note: Points are samples plotted in feature space



Cluster Bias

- Human beings conceptualize the world through categories represented as *exemplars* (Rosch 1973, Estes 1994).

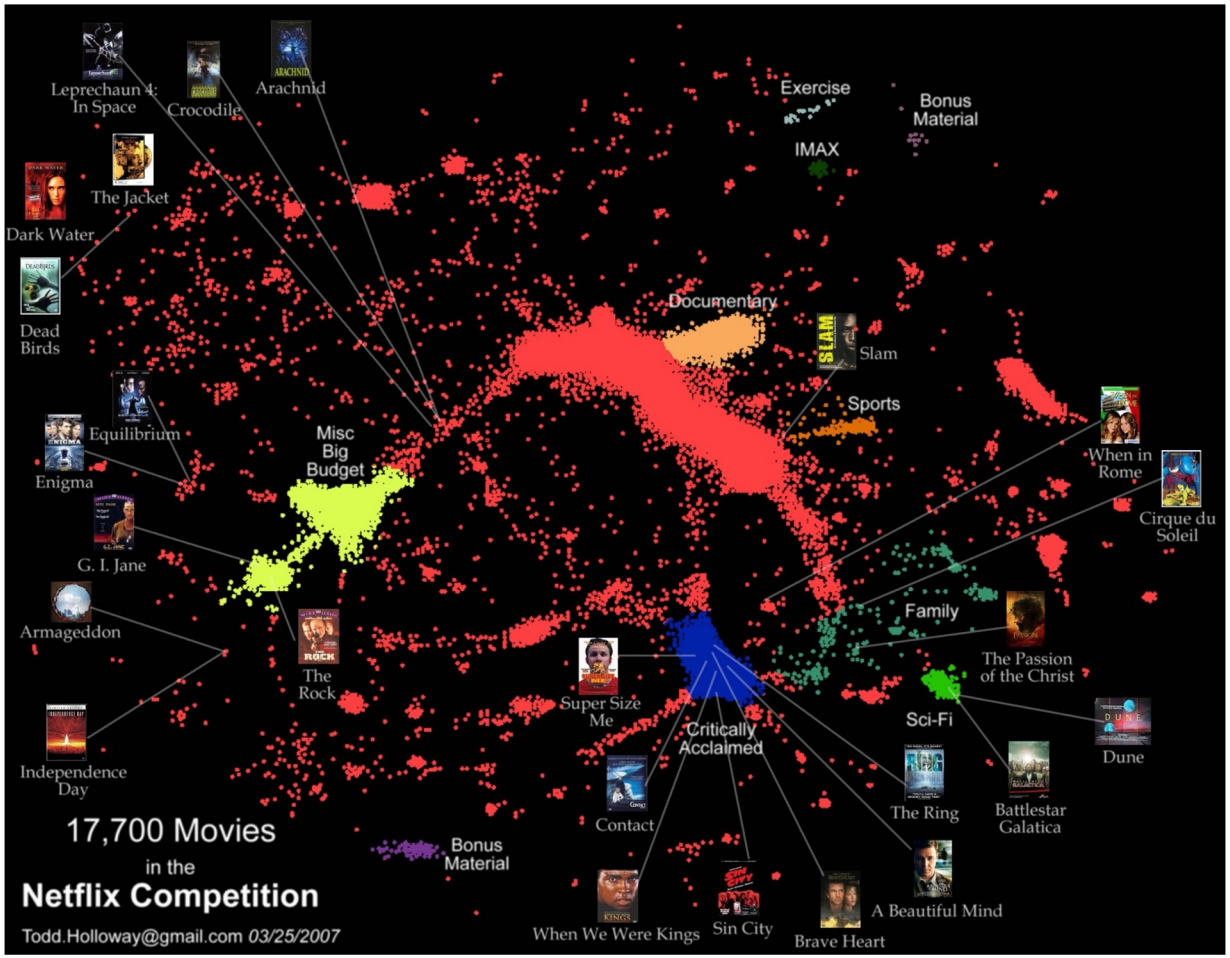


- We tend to see cluster structure whether it is there or not.
- Works well for dogs, but…

Clustering Illusion: https://en.wikipedia.org/wiki/Clustering_illusion

Netflix

- Taste is more of a continuum than discrete clusters
- Factor models and kNN do much better than discrete cluster models



Cluster Bias

Upshot:

- **Clustering is used more than it should be** because people assume an underlying domain has discrete classes in it.

- In reality the underlying data is usually **continuous**.

- Just as with taste preferences on Netflix, continuous models (dimensionality reduction, kNN) tend to perform better.

Terminology

- **Hierarchical clustering:** clusters form a hierarchy. Can be computed bottom-up or top-down.
- **Flat clustering:** no inter-cluster structure.

- **Hard clustering:** items assigned to a unique cluster.
- **Soft clustering:** cluster membership is a real-valued function, distributed across several clusters.