Advanced Text Analysis for Business (IDS-566)

Lecture 5

Feb 23, 2018

Course Overview

- Instructor
  - Ehsan M. Ardehaly PhD, ehsan@uic.edu
  - Office hours: 4:45 - 5:45 pm F, BLC L270
  - Teacher assistant: 4:00 - 5:00 pm W, BLC L270
- Objectives:
  - Text mining
  - Applications for business decisions
  - Study of machine learning concepts
  - Design and implementation of text mining approaches

Assignments-3

- Grade: 20%
- Sentiment analysis
- Due date: 3/13/2018

- Submission:
  - Notebook (code + analysis) → PDF
  - Word document with code as an appendix → PDF

Agenda

Unsupervised learning

- Motivation

Dimensionality reduction:

- LSA

Clustering:

- K-means

Text analysis:

- Document clustering

Unsupervised Learning

- Supervised learning
  - With labeled data
  - Training data: X, y

- Supervised learning
  - Without labeled data
  - Training data: X

Supervised learning

- Pros:
  - Easier modeling
- Cons:
  - Requires human annotation:
    - Expensive
    - Noisy
    - Bias
    - Temporal dynamic

Unsupervised learning

- Pros:
  - Lost of data (BigData)
- Cons:
  - Hard to model

Unsupervised models

- Dimensionality reduction
- Clustering
- Topic Modeling
- Anomaly detection

Dimensionality reduction

- The curst of dimensionality
- Textual data has a high dimension
- Solution:
  - Reduce the dimension
  - Analysis the data in low dimension
  - Reducing noise

Motivation

- Discover hidden concepts
  - Words that occur often together (co-occurrence).
- Remove noise
  - Not all words are useful
- Visualization
  - 2D or 3D plot

Problem statement

- High dimension data: X (e.g. 50000x20000)
- Low dimension data: Z (e.g. 50000x100)
- Transformation function: f
  - $Z = f(X)$

Inverse transform

- Transformation function: f
  - $Z = f(X)$
- Inverse transformation function: f^{-1}
  - $X' = f^{-1}(Z)$
  - $X \approx X'$

Latent Semantic Analysis (LSA)

- Based on Singular Value Decomposition (SVD)
  - Also know as truncate SVD
- Suitable for sparse data (e.g. text)
- Fast training

LSA applications

- Document clustering
- Word clustering
- Word relations (synonymy and polysemy)
- Information retrieval

Application in NLP

- Synonymy
  - Different words describe the same idea.
  - These words may be close to each other in the lower dimension.
- Polysemy
  - Same word has multiple meanings (e.g. chair).
  - Hard to capture with LSA.

LSA components

- Component:
  - Each dimension in lower dimension
  - Each component is a vector with size of number of features.
- Component matrix (W)
- $Z = XW$

Document similarity

- Reduce dimension
- Measure similarity between documents in lower dimension
- Advantage:
  - Lower dimension is more dense

Topic model

- Abstract topics in a collection of documents.
- Each component could be consider as a topic.
- Words inside a topic often co-occur together.

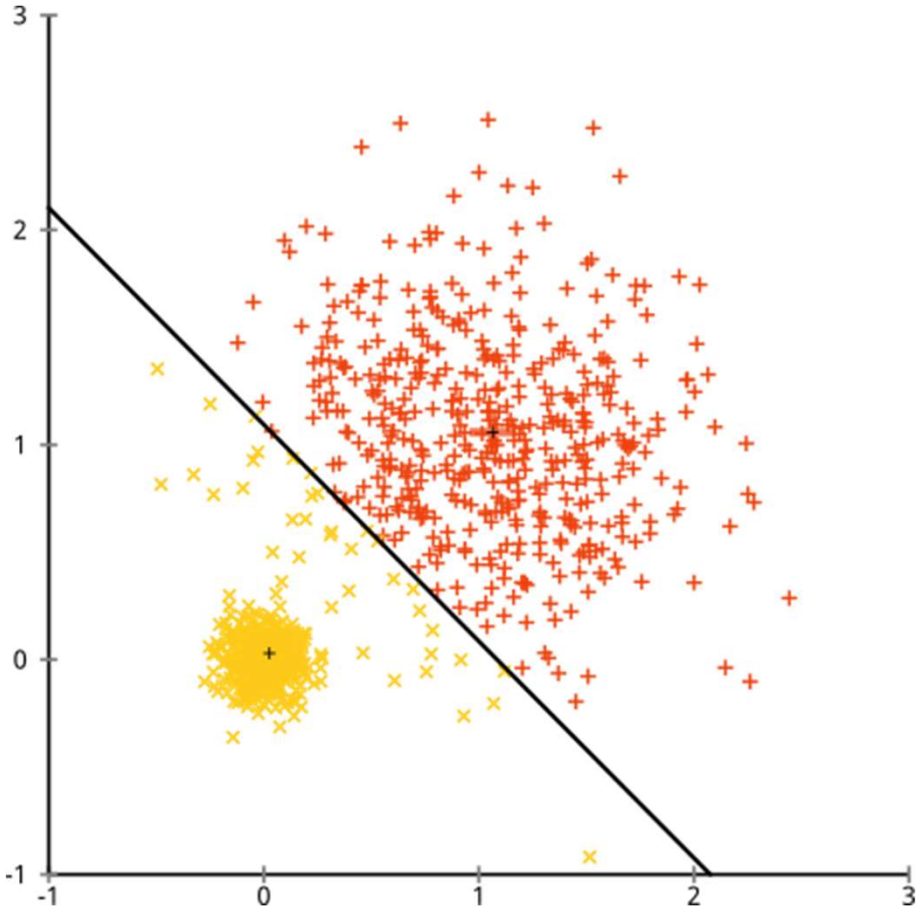
Clustering

- Grouping a set of instances into a cluster.
  - Samples in a cluster are more similar to each other.
- Similarity
  - Distance function
    - Euclidean
    - Cosine

K-means

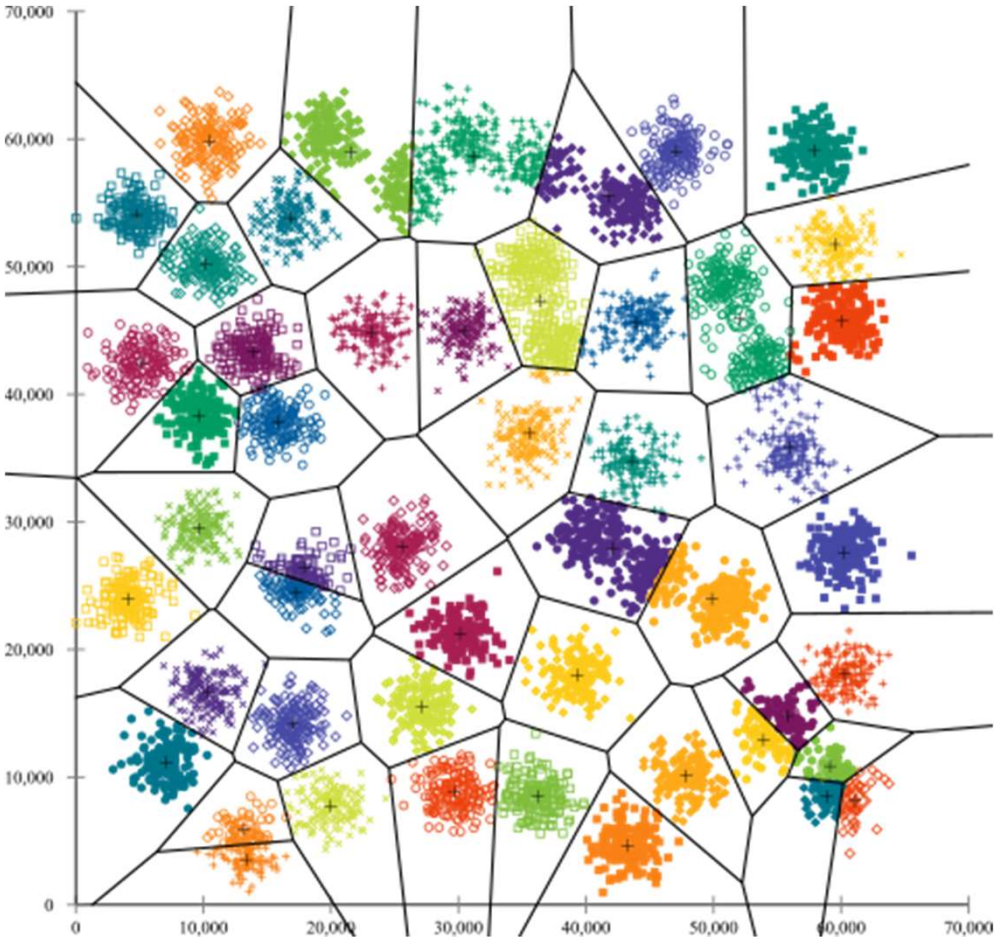
- Finding centroid
- Based on Euclidean distance
- Local optimum
- Sensitive to initialization

K-means clusters



https://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Number of clusters



https://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

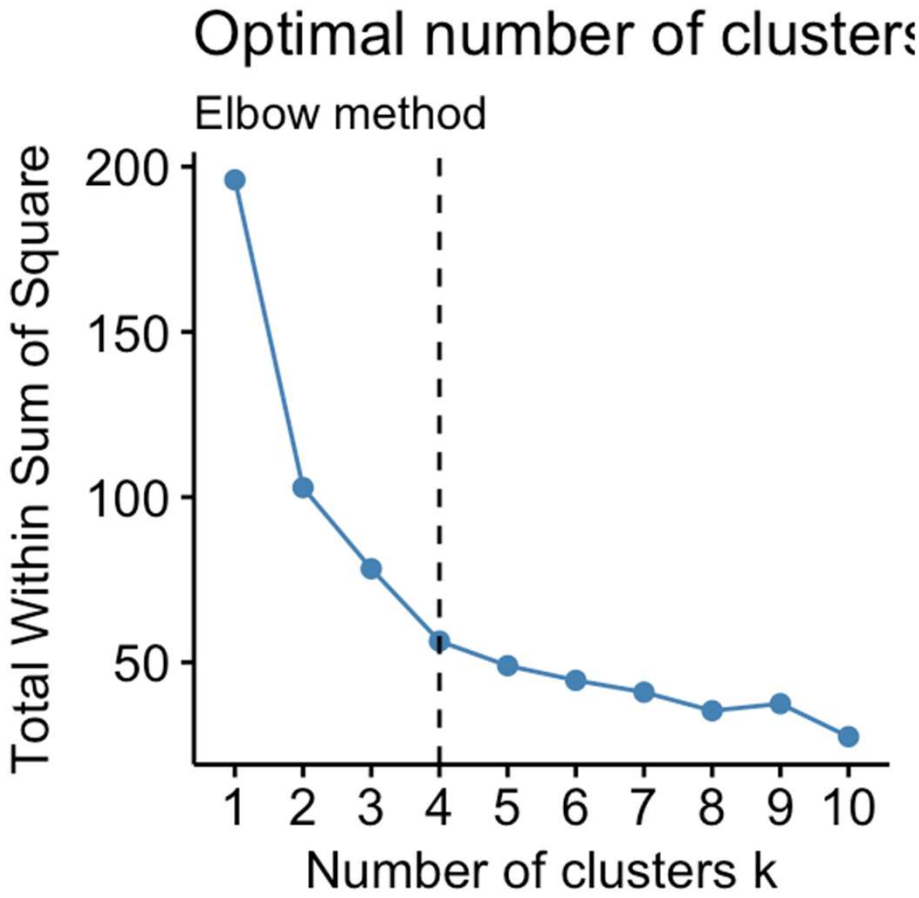
Inertia

- Metric to measure clustering accuracy
- Within sum of square (WSS) distance to the nearest centroid
- $\sum_i \min_{\mu_j \in C} \left\| x_i - \mu_j \right\|^2$

Number of clusters vs. inertia

- More clusters → Lower inertia
- Number of clusters = number of samples
  - Zero inertia

Best number of clusters



http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimal-number-of-clusters-3-must-know-methods/#average-silhouette-method

Document clustering

- Creating the feature matrix:
  - Tf-idf is recommended
- Fit k-means with different number of clusters (e.g 2 – 10)
- Plot inertia
- Select the best one with elbow method (finding knee)

Challenge

- Text is sparse
- Distance metrics may not reflect the similarity

- Solution
  - Reduce the dimension before clustering

Clustering in low dimension (LD)

- Reduce the dimension
  - Using LSA
- Fit K-means
- Tune number of clusters
  - Elbow method

LD vs HD clustering

- Low dimension
  - Pros: Data is more dense.
  - Cons: number of components need to tune too.
- High dimension
  - Pros: Only one tuning parameter (number of clusters)
  - Cons: Sparsity

Clustering accuracy

- We often don’t have labels
  - Hard to measure the cluster accuracy
- If we have the labels
  - We can measure the accuracy
  - Predict all samples in one cluster to one class base on majority vote

Clustering accuracy

- Cluster 1:
  - 100 samples → class A
  - 50 samples → class B
  - 50 samples → class C
  - Predict all samples as class A

Clustering applications

- Cluster analysis
- Feature learning
- Word clustering

Soft clustering

- Each sample may belongs to multiple clusters.
- Each sample has a weight for multiple clusters.
- Higher weight → More similar to the given cluster.

Soft clustering vs. topic modeling

- Similarity
  - Documents in a soft cluster have similar topics
  - A document can belong to multiple topics
- Difference:
  - Clustering is based on distance
  - Topic models are based on likelihood