Advanced Text Analysis for Business (IDS-566)

Lecture 7

Mar 9, 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

Singular Value Decomposition

- LSA

Topic Modeling:

- LDA

Clustering:

- K-means

Final exam

- Review

Unsupervised Learning

- Supervised learning
  - With labeled data
  - Training data: X, y
- Unsupervised learning
  - Without labeled data
  - Training data: 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

Singular Value Decomposition

- $X = U\Sigma V^T$
- X: m × n
- U: m × m → Unitary matrix
- Σ: m × n → Rectangular diagonal matrix
- V: n × n → Unitary matrix

Singular Value Decomposition

- $X = U\Sigma V^T$
- $UU^T = I$
- $\Sigma_{ii} \geq 0$
- $VV^T = I$

SVD Example

X =
[[0 1 2 3]
[4 5 6 7]
[8 9 10 11]]

U=
[[-0.14733887 0.90087891 0.40820312]
[-0.50048828 0.28808594 -0.81640625]
[-0.85302734 -0.32446289 0.40820312]]

Eigenvalues:
22.4, 1.95, 8e-16

S =
[[22.40625   0.   0.   0.   ]
[0.   1.95507812 0.   0.   ]
[0.   0.   8.17e-16   0.   ]]

Vt=
[[-0.39379883 -0.4609375 -0.52783203 -0.59472656]
[-0.73828125 -0.29589844 0.14624023 0.58837891]
[-0.47875977 0.83642578 -0.23730469 -0.1206665]
[0.26635742 0.00177097 -0.80224609 0.53417969]]

Truncated SVD

X =
[[0 1 2 3]
[4 5 6 7]
[8 9 10 11]]

U=
[[-0.14733887 0.90087891 0.40820312]
[-0.50048828 0.28808594 -0.81640625]
[-0.85302734 -0.32446289 0.40820312]]

S =
[[22.40625   0.   0.   0.   ]
[0.   1.95507812   0.   0.   ]
[0.   0.   8.17e-16   0.   ]]

Vt=
[[-0.39379883 -0.4609375 -0.52783203 -0.59472656]
[-0.73828125 -0.29589844 0.14624023 0.58837891]
[-0.47875977 0.83642578 -0.23730469 -0.1206665]
[0.26635742 0.00177097 -0.80224609 0.53417969]]

Truncated SVD

X =
[[0 1 2 3]
[4 5 6 7]
[8 9 10 11]]

U_2=
[[-0.14733887 0.90087891]
[-0.50048828 0.28808594]
[-0.85302734 -0.32446289]]

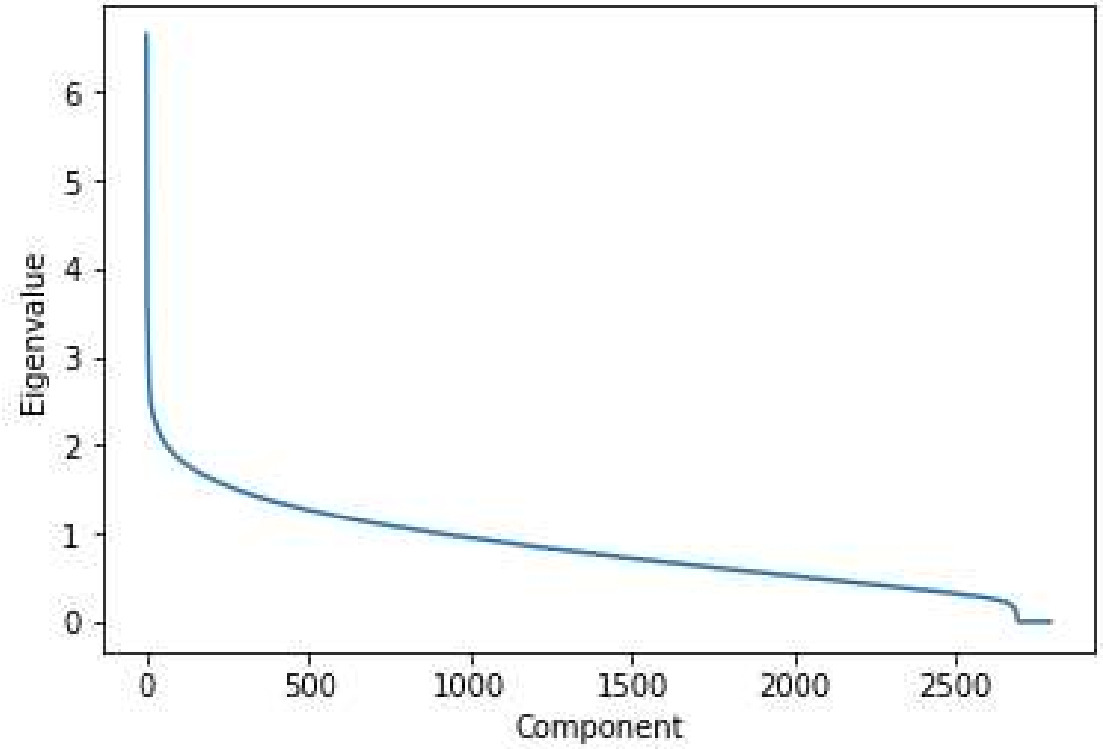
S_2 =
[[22.40625   0   ]
[0.   1.95507812]]

Vt_2=
[[-0.39379883 -0.4609375 -0.52783203 -0.59472656]
[-0.73828125 -0.29589844 0.14624023 0.58837891]]

LSA vs. SVD

- $X \approx U_k \Sigma_k V_k^T$
- $Z = U_k \Sigma_k$ ← Lower dimension
- $W = V_k^T$ ← Components
- $X = ZW$

Eigenvalues



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.

Latent Dirichlet Allocation

- Generative model
- Find similarity in some parts of data
- A topic model
- A graphical model

LDA

- Doc 1 and 3: 100% topic A
- Doc 2, 5, 6: 100% topic B
- Doc 3: 60% topic A, 40 % topic B

- Topic A: 50% politic, 30% economy, 20% finance
- Topic B: 20% politic, 60% economy, 20% finance

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

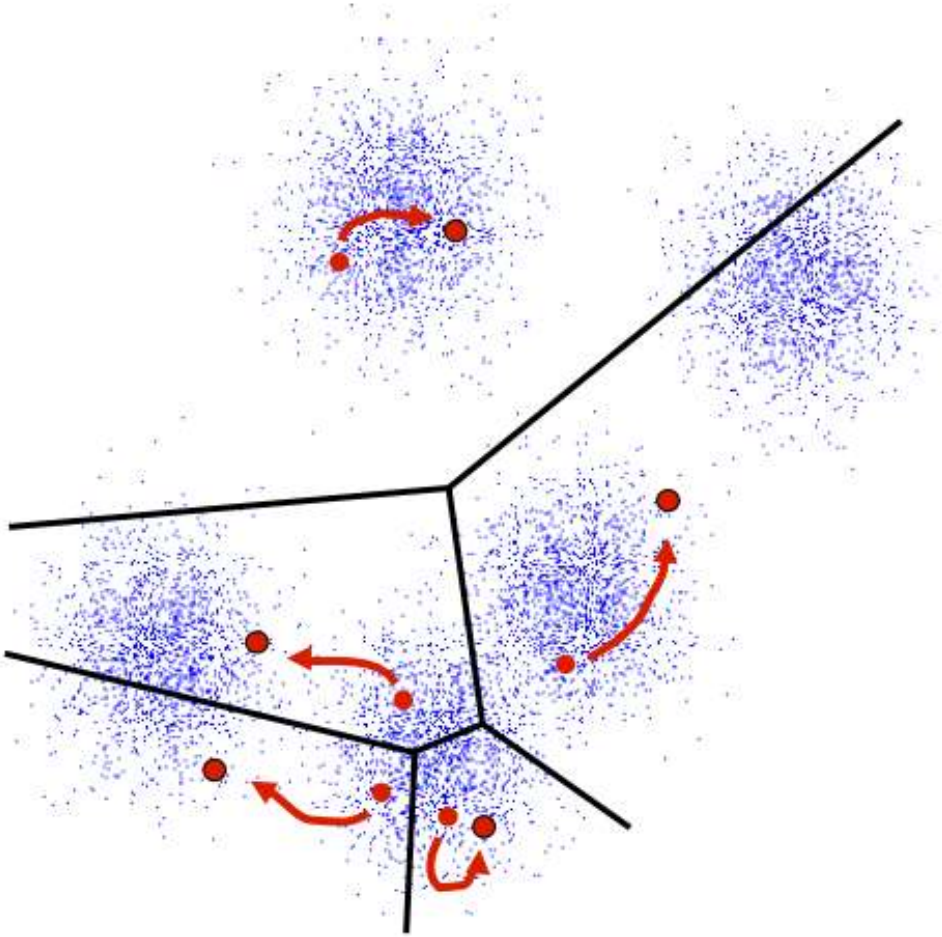
Initialization:

- Select random K centroids

Until converges:

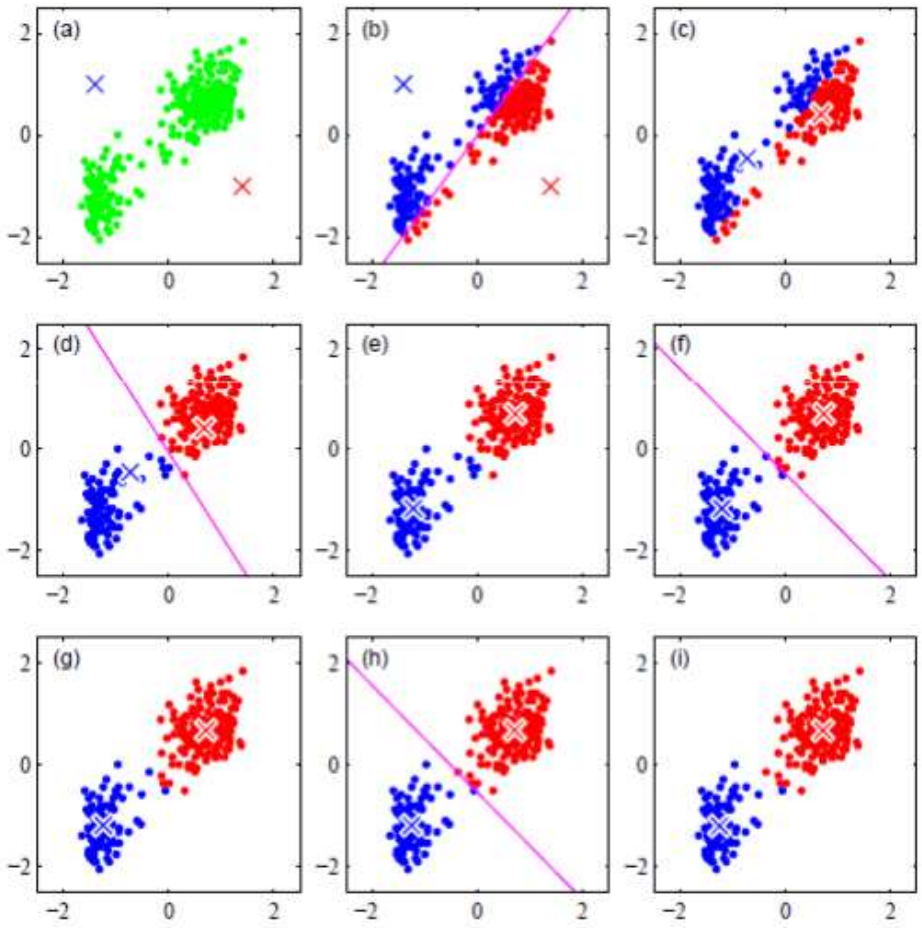
- Create clusters by assigning samples to the closest centroid.
- Change the centroid to the mean of samples in the cluster.

K-means



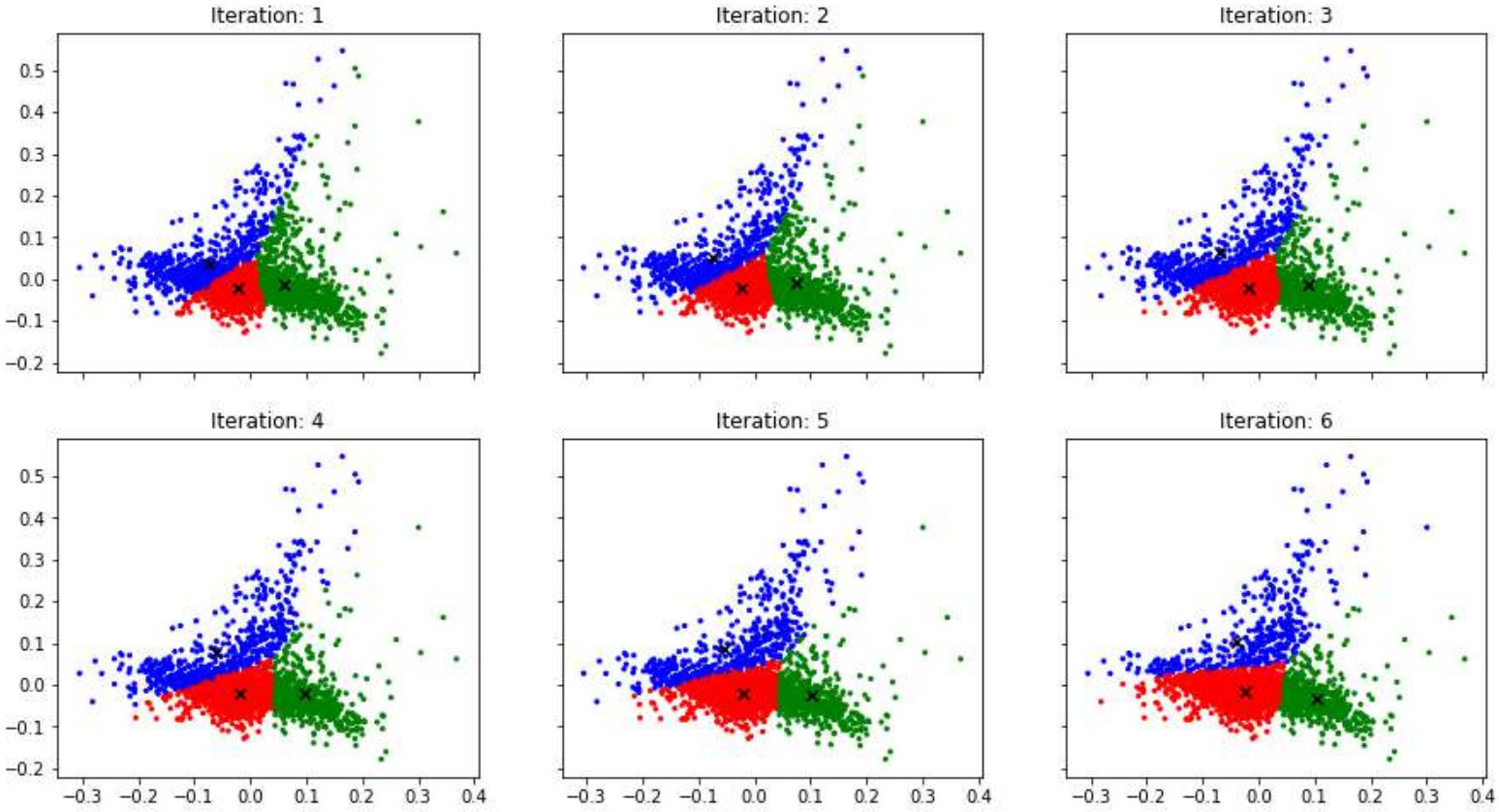
http://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf
David Sontag, New York University

Example 1



http://www.cs.haifa.ac.il/~rita/uml_course/lectures/kmeans.pdf
Rita Osadchy, K-Means

Example 2


Iteration: 1
Iteration: 2
Iteration: 3
Iteration: 4
Iteration: 5
Iteration: 6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4

Final exam

- 3/16 6:00 PM – 7:30 PM
- Burnham Hall 208
- Written exam
- Close book
- No electronical devices (mobile, calculator, ...)

Lecture 2 review

- Tokenization
  - Sequence of words.
- Regular expression
  - A sequence of characters that define a search pattern.
- Unigram, bigram, n-gram
- Document to Term Matrix (DTM)
- Tf-idf transformation

Lecture 2 review

- Zipf's law
- Sparse matrix:
  - LIL matrix
  - CSR matrix
  - CSC matrix
- Sentiment analysis
  - Lexicon

Lecture 3 review

- Supervised learning vs unsupervised learning
- Document classification pipeline
- K-Nearest Neighbor
- Impact of K
- Bias vs. variance
- Generative learning

Lecture 3 review

- Naïve Bayes
  - Bernoulli
  - Multinomial
  - Gaussian
- Discriminative Learning
- Logistic regression
  - Logistic function (sigmoid)
  - Decision boundary

Lecture 4 overview

- Likelihood
- Maximum Likelihood Estimate (MLE)
- Negative log-likelihood
- Regularization
  - L1
  - L2
  - Elastic net

Lecture 4 overview

- Gradient descent algorithm
- Learning rate
- Multi-class logistic regression
  - Softmax function
- Metrics
  - Confusion matrix
  - Accuracy, precision, recall, F1
- Learning curve
- K-folds cross-validation

Lecture 5 review

- Binary vs. categorical cross-entropy
- Hidden layer
- Activation functions
  - Relu
  - Tanh
- MLP
- Stochastic Gradient Decent (SGD)
- Creating batches

Lecture 5 review

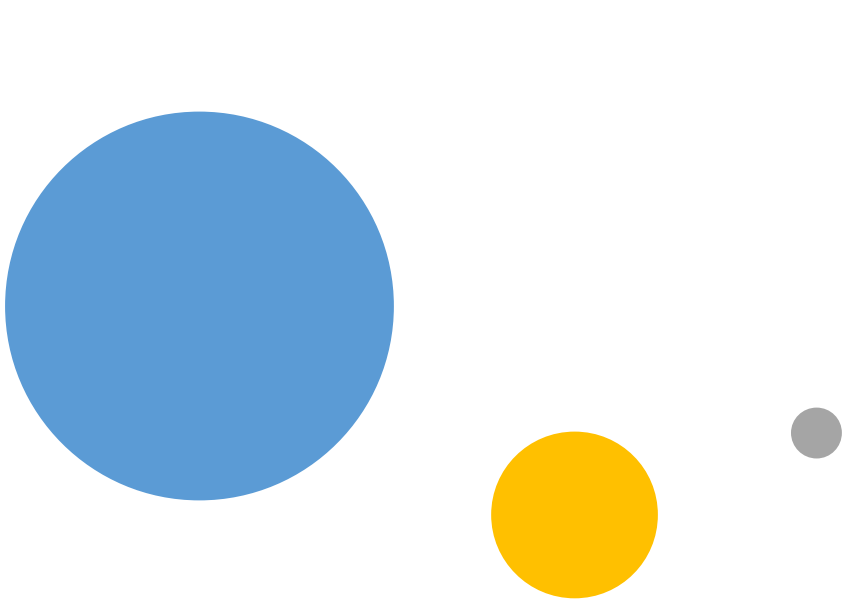
- Training a neural network
  - Feed forward
  - Backpropagation
  - Validation (optional)
- Regularization layers
  - Dropout
  - Batch Normalization
- Model output size and weights
- Word embedding (word2vect)

Lecture 6 and 7 review

- Dimensionality reduction
- SVD
  - Eigenvalues
- LSA
  - Truncate SVD
  - Components
- K-means
  - Inertia
  - Training process
  - Elbow method

Lecture 6 and 7 review

- Document clustering
- Word clustering
- Clustering in low dimension
- Topic modeling
  - LSA
  - LDA



Thank you!

Any questions?