Matrix-Based Algorithms for Data Mining

Tao Li
School of Computing and Information Sciences
Florida International University

Joint work with Dr. Chris Ding and Dr. Michael I. Jordan
Matrix-Based Algorithms for Data Mining

- **Motivation**
- Matrix-Based Algorithms for Clustering
- Non-Negative Matrix Factorization (NMF) for Clustering
- Non-Negative Matrix Factorization (NMF) for Consensus Clustering and Semi-Supervised Clustering
- Adaptive Dimension Reduction
A recent, fast growing, trend is to use eigenvectors and matrix algorithms to solving challenging problems in data mining.
Why Matrix and Eigenvectors?

Matrix and Linear algebra

• Relatively simple
  – in comparison to probabilistic, information-theoretic, graph-theoretic approaches
• Well-developed branch of mathematics
  – knowledge accumulated since 1700
• Many mature software tools available—developed by scientific computing community
How did I get into this area as a computer scientist?

Motivated by a research project in computer science
Current computing systems:
• Growing Number of Components
• Heterogeneous
• Dynamic
System management: key for high performance and availability
- collects and processes data in real-time for automatic actions
- correlation knowledge needs to be constructed and maintained

-Example
  - rebooting signature: host_down is followed by host_up in five seconds => filtering
  - host_down without host_up => an operator should be paged (availability problem)
Key Challenges for Automatic Management

• Need automatic and efficient approaches
  – IBM Autonomic Computing (AC) initiative
  – System Self-Management

• Automatically acquiring needed knowledge from the log data

• Heterogeneous nature of the system
  – Diversity and disparity in data reporting
    • Different formats (syntax) and content (semantic)
  – Difficult to correlate logs across different components
Log Data Organization

- Different reporting schemes by different components
  - Component A: A has started
  - Component B: B has begun execution

- Difficult to correlate logs across different components
  - Think about to implement the following rule: if any component has started, notify the system operators
  - Need to know all the different schemes for “start”
  - How about add a new component?

- Problem: How to organize the log messages with disparate formats and contents into a canonical form?
Solutions: Common Situations

• Common Situations

  – Encode semantics about the messages
    • Situation assignment requires to understand the message.

  – A set of common categories
    • E.g., start, stop, dependency, create, connection
Computing System Management: a high level conceptual view

- Realtime Analysis
  - Anomaly Detection
  - Fault Diagnosis
  - Problem Determination
  - Real Time Management
  - Correlation/Dependency Knowledge
  - Knowledge Base
  - Summarization/Visualization
  - Rule Construction
  - Temporal Pattern Discovery
  - Knowledge Management
  - Situation Identification and Categorization

- Offline Analysis
  - Planning/Actions
  - Active Data Collection
  - Historical Data Collection
  - Log Data Organization

- Log Data Organization
  - Log Adapter
  - Historical Data Collection
  - Logs
  - Component logs
  - Situati, Identification and Categorization

- Logs
  - Logs
Solutions: Common Situations

• Common Situations
  – Encode semantics about the messages
    • Situation assignment requires to understand the message.
  – A set of common categories
    • E.g., *start*, *stop*, *dependency*, *create*, *connection*

• Two Challenges:
  – Automatically infer the common semantic situations
    • Clustering problem
  – Categorize the log messages into a set of common situations
    • Classification problem
Clustering Problem

K-means:

• Select $k$ centers somehow
• Find a partition of $W$ such that the points within each cluster are “similar” to each other
• Repeat until the centers don’t change (or change very little)
  – Partition the data according to the $k$ centers
  – Use the means of the cluster to find $k$ new centers

- Treat each attribute equally
- Distance Computation
  – Curse of dimensionality
**K-means clustering**

- Computationally Efficient (order-$mn$)
- Most widely used in practice
  - Benchmark to evaluate other algorithms

Given $n$ points in $m$-dim: \( X = (x_1, x_2, \cdots, x_n)^T \)

\[ \text{K-means objective} \quad \min J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - c_k \|^2 \]

- Also called “isodata”, “vector quantization”
- Developed in 1960’s (Lloyd, MacQueen, Hartigan, etc)
Log Data Example

Sample Messages from Log Files

<table>
<thead>
<tr>
<th>Messages/Terms</th>
<th>Start</th>
<th>Application</th>
<th>Version</th>
<th>Create</th>
<th>Temporary</th>
<th>Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 – Start</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Key Observations**
  - Many terms
  - Situations share different terms
  - Situation interpretation via terms
  - Associations between terms and situations

- **Key Ideas**
  - Adaptively measuring data similarities
  - Explore the relationships between data and feature
  - Simultaneous data and feature clustering
Clustering Model

Given W: n data points with m binary features with k groups

- Data-Cluster Coefficients D and Feature-Cluster Coefficients F

\[ D = \begin{bmatrix}
0 & 1 & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{bmatrix} \]  
\[ F = \begin{bmatrix}
0 & 1 & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0
\end{bmatrix} \]

- \( D^T \) (with normalization) is an approximation of original W
- Clustering problem Finding D and F
Talk Outline

• Motivation

• *Matrix-Based Algorithms for Clustering*

• Non-Negative Matrix Factorization (NMF) for Clustering

• Non-Negative Matrix Factorization (NMF) for Consensus Clustering

• Non-Negative Matrix Factorization (NMF) for Semi-Supervised Clustering
Optimization: Matrix Perspective

• The goal of clustering is

$$\arg \min_{D,F} O = \frac{1}{2} \|W - DF^T\|^2_F, \|X\|_F = \sqrt{\sum x_{ij}^2}$$

• The objective function can be minimized by alternatively optimize one of D or F while fixing the other

$$\frac{\partial O}{\partial D} = -WF + DF^TF$$

$$\frac{\partial O}{\partial F} = -W^TD + FD^TD$$

• If F and D are orthogonal, avoiding computation of the inverse
First try: Iterative Feature and Data (IFD) Clustering

• The optimization rule:

\[ D = WF \]

\[ F = W^T D \]

• Basically the optimizing rules show the *mutually reinforcing relationship* between the data and features.

• If a feature \( f \) is shared by many points that have high weights associated with a cluster \( c \), then feature \( f \) has a high weight associated with \( c \). Similar for data point \( d \).

Li & Ma, SDM 2004
Extension I: Adaptive Subspace Clustering

• Explicitly models the subspace structure
  – F specifies the subspace structure
  – IFD is a special case where F is a binary matrix

• WF: the projection of the data points into the subspaces defined by F

• $S = (D^T D)^{-1} D^T WF$ : the projection of the centroids into the subspaces defined by F

• Clustering Objective: Minimizing

$$O(D, F, S) = \frac{1}{2} \|WF - DS\|^2_F$$

Li, Ma & Ogihara, SIGIR 2004
## Extension II: A General Model

A General model for clustering binary data (Li, SIGKDD 2005)

- $W = AXB^T + E$ where $E$ denotes the error component
- $AXB^T$ characterizes the information of $W$ that can be described by the cluster structure
- $A$ and $B$ explicitly designate the cluster membership for data points and feature respectively. $X$ specifies cluster representations.
- An elegant basis for connecting various clustering methods while highlighting their differences
- Different variations of the general model with different constraints and relaxations

<table>
<thead>
<tr>
<th>Methods</th>
<th>$B$</th>
<th>$A$</th>
<th>$X$</th>
<th>Optimization Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>The General Model</td>
<td>$b_{jc} \in {0, 1}$</td>
<td>$a_{ik} \in {0, 1}$</td>
<td></td>
<td>Two-side</td>
</tr>
<tr>
<td></td>
<td>$\sum_{j=c}^C b_{jc} = 1$</td>
<td>$\sum_{i=k}^K a_{ik} = 1$</td>
<td>$x_{kc} = \frac{\sum_{i=1}^n \sum_{j=1}^m a_{ik} b_{jc} w_{ij}}{\sum_{i=1}^n \sum_{j=1}^m a_{ik} b_{jc}}$</td>
<td>K-Means</td>
</tr>
<tr>
<td>Block Diagonal</td>
<td>$b_{jk} \in {0, 1}$</td>
<td>$a_{ik} \in {0, 1}$</td>
<td></td>
<td>Block Diagonal</td>
</tr>
<tr>
<td>Clusterin</td>
<td>$\sum_{j=k}^K b_{jc} = 1$</td>
<td>$\sum_{i=k}^K a_{ik} = 1$</td>
<td>$X = IK \times K$</td>
<td>Optimization</td>
</tr>
<tr>
<td>One-Side K-Means</td>
<td>$B = I$</td>
<td>$a_{ik} \in {0, 1}$</td>
<td></td>
<td>Alternating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\sum_{i=k}^K a_{ik} = 1$</td>
<td>$X = (ATA)^{-1}ATW$</td>
<td>Least Square</td>
</tr>
<tr>
<td>Iterative Feature</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Clustering</td>
<td>Arbitrary</td>
<td>Arbitrary</td>
<td>$X = IK \times K$</td>
<td>Mutually Reinforcing</td>
</tr>
<tr>
<td>Spectral Relaxation</td>
<td>Orthonormal</td>
<td>Orthonormal</td>
<td>$X = ATWB$</td>
<td>Two-Side</td>
</tr>
</tbody>
</table>

Table 1: Each row lists a variation and its associated constraints
Quick Summary on Clustering

**Subspace Clustering**
(Agrawal et al, 1998
Aggarwarl et al, 1999)

**Binary Matrix Decomposition**
(Kolda et al, 1998
Koyuturk et al, 2003)

**Non-negative Factorization**
(Lee et al, 1999; Lee et al, 2000
Xu et al, 2003)

**Information Bottleneck**
(Tishby et al, 1999
Slonim et al, 2001)

**Additive Clustering**
(Shepard et al, 1979
Desarbo et al, 1982)

**Adaptive Dimension Reduction**
(Ding et al, 2002
Carlotta et al, 2002)

**Traditional Clustering**
(Jain et al, 1998
Han et al, 2001)

**Co-clustering**
(Hartigan, 1975; Govaet, 1985
Dhillon, 2000; Dhillon et al, 2003)

**Non-distance Based**
(Ramkumar et al, 1998)

**CoFD Algorithm**
(Li et al, 2003; Zhu et al, 2002)

**Spectral Clustering**
(Sarkar et al, 1996; Shi et al, 1997
Weiss, 1999; Ng, 2001)

**Image Segmentation**

**Graph Partition**
Related Work Summary

Subspace Clustering
(Agrawal et al, 1998
Aggarwal et al, 1999)

Binary Matrix Decomposition
(Kolda et al, 1998
Koyuturk et al, 2003)

Non-negative Factorization
(Lee et al, 1999;
Lee et al, 2000
Xu et al, 2003)

Information Bottleneck
(Tishby et al, 1999
Slonim et al, 2001)

Additive Clustering
(Shepard et al, 1979
Desarbo et al, 1982)

Spectral Clustering
(Sarkar et al, 1996; Shi et al, 1997
Weiss, 1999; Ng, 2001)

Image Segmentation

Graph Partition

Adaptive Dimension Reduction
(Ding et al, 2002
Carlotta et al, 2002)

Traditional Clustering
(Jain et al, 1998
Han et al, 2001)

Partitional

Hierarchical

Density/Grid based

Co-clustering
(Hartigan, 1975; Govaet, 1985
Dhillon, 2000; Dhillon et al, 2003)

Non-Distance

Non-distance Based
(Ramkumar et al, 1998)

CoFD Algorithm
(Li et al, 2003; Zhu et al, 2002)
Talk Outline

• Motivation
• Matrix-Based Algorithms for Clustering
• *Non-Negative Matrix Factorization (NMF) for Clustering*
• Non-Negative Matrix Factorization (NMF) for Consensus Clustering and Semi-Supervised Clustering
• Adaptive Dimension Reduction
Binary Data Matrix in our work is a special case of Non-negative Matrix
Nonnegative Matrix Factorization
(NMF)

Data Matrix: $n$ points in $p$-dimensional space:

$$X = (x_1, x_2, \cdots, x_n)$$

$x_i$ is an image, document, webpage, etc.

Factorization (low-rank approximation)

$$X \approx FG^T$$

Nonnegative Matrices

$$X_{ij} \geq 0, F_{ij} \geq 0, G_{ij} \geq 0$$

$$F = (f_1, f_2, \cdots, f_k) \quad G = (g_1, g_2, \cdots, g_k)$$
Pixel vector
Lee and Seung (1999): Parts-of-whole Perspective

\[ X = (x_1, x_2, \cdots, x_n) \]

\[ = \text{Matrix of images} \]

\[ F = (f_1, f_2, \cdots, f_k) \quad G = (g_1, g_2, \cdots, g_k) \]
Meanwhile ……

Several studies *empirically* show the usefulness of NMF for *pattern discovery/clustering*

Research shows

NMF factors give *holistic* pictures of the data

i.e., NMF is doing data clustering
NMF Gives Holistic Pictures

F factors
NMF Gives Holistic Pictures II

F factors

Original data
NMF is doing “Data Clustering”

NMF => K-means Clustering
Reformulate K-means Clustering

Cluster membership

\[
\begin{bmatrix}
C_1 & C_2 & C_3 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{bmatrix}
= (h_1, h_2, h_3) = H
\]

K-means Clustering

\[
J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \| x_i - c_k \|^2
\]

\[
J = \| X - CH^T \|^2
\]

(Zha, Ding, Gu, He, Simon, NIPS 2001)
(Ding & He, ICML 2004)
K-means Clustering Theorem

\( G \)-orthogonal NMF is equivalent to relaxed K-means clustering.

\[
\min_{G^T G = I, G \geq 0} \| X \pm - F \pm G^T \|^2
\]

Requires only \( G \)-orthogonality and nonnegativity.

\[
F = (f_1, f_2, \cdots, f_k) \implies \text{cluster centroids}
\]

\[
G = (g_1, g_2, \cdots, g_k) \implies \text{cluster indicators}
\]

(Ding, Li, Jordan, 2006)
NMF Generalizations

(Li & Ding, ICDM 2006)

SVD: \[ X_\pm = F_\pm G_\pm^T = U\Sigma V^T \]

Semi-NMF: \[ X_\pm = F_\pm G_+^T \]

Convex-NMF: \[ X_\pm = X_\pm W + G_+^T \]

Kernel-NMF: \[ \phi(X_\pm) = \phi(X_\pm)W + G_+^T \]

Tri-NMF: \[ X_\pm = F_+ S_\pm G_+^T \]

Symmetric NMF: \[ X = QSQ^T \]

(Ding, Li, Jordan, 2006)

(Ding, Li, Peng, Park, KDD 2006)
NMF $\Leftrightarrow$ PLSI

NMF objective functions

- Frobenius norm
  \[ J_{KL} = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \log \frac{x_{ij}}{(FG^T)_{ij}} - x_{ij} + (FG^T)_{ij} \]

- KL-divergence:

Probabilistic LSI (Hoffman, 1999) is a latent variable model for clustering:

\[ J_{PLSI} = \sum_{i=1}^{m} \sum_{j=1}^{n} x(w_i, d_j) \log p(w_i, d_j) \]

\[ p(w_i, d_j) = \sum_k p(w_i \mid z_k) p(z_k) p(d_j \mid z_k) \]

We can show \[ J_{PLSI} = -J_{NMF-KL} + \text{constant} \]

(Ding, Li & Peng, AAAI 2006)
Orthogonal Nonnegative Tri-Factorization

3-factor NMF with explicit orthogonality constraints

$$\min_{F^TF=I, F \geq 0} \| X_\pm - F_\pm S_\pm G^T \|_2^2$$

1. Solution is unique
2. Can’t reduce to NMF

Simultaneous K-means clustering of rows and columns

$$F = (f_1, f_2, \ldots, f_k) \implies \text{Row cluster indicators}$$

$$G = (g_1, g_2, \ldots, g_k) \implies \text{Column cluster indicators}$$

(Ding, Li, Peng, Park, KDD 2006)
Symmetric NMF: \( W \approx QSQ^T \)

Symmetric NMF is a special case of Tri-factorization

Symmetric NMF

\[
\min_{Q^TQ=I, Q \geq 0, S \geq 0} \| W - QSQ^T \|^2
\]

Update Q:

\[
Q_{jk} \leftarrow Q_{jk} \sqrt{\frac{(WQS)_{jk}}{(QQ^TWQS)_{jk}}}
\]

Update S:

\[
S_{kl} \leftarrow S_{kl} \sqrt{\frac{(QTWQ)_{kl}}{(Q^TQSQ^TQ)_{kl}}}
\]
Semi-NMF: \[ X_\pm = F_\pm G^T_+ \]

- For any mixed-sign input data (centered data)
- Clustering and Low-rank approximation

\[
\min \| X - FG^T \|
\]

Update F: \[ F = XG(G^T G)^{-1} \]

Update G: \[ G_{ik} \leftarrow G_{ik} \sqrt{\frac{(X^TF)^+_{ik} + [G(FF)^-]_{ik}}{(X^TF)^-_{ik} + [G(FF)^+]_{ik}}} \]  

(Ding, Li, Jordan, 2006)
Can we extend NMF for other data mining problems?

Open up additional applications for NMF
Talk Outline

• Motivation
• Matrix-Based Algorithms for Clustering
• Non-Negative Matrix Factorization (NMF) for Clustering
  • *Non-Negative Matrix Factorization (NMF) for Consensus Clustering and Semi-Supervised Clustering*
• Adaptive Dimension Reduction
An Illustrating Example
K-means Clustering
Combining Multiple Clusterings
Consensus Clustering Applications
(Li, Ma & Ogihara, CIKM 2004)

• Aggregation of Clusterings
  – Improve clustering performance and stability
• Clustering heterogeneous data sources
• Distributed clustering
• Clustering with multiple criteria
• And more
Consensus Clustering

T Clusterings \( P = (P_1, P_2, \ldots, P_T) \)

Clustering \( P^t \) \( P^t = (C_1^t, C_2^t, \ldots, C_k^t) \)

Connectivity Matrix \( M_{ij}(P^t) = \begin{cases} 1 & (i,j) \in C_k(P^t) \\ 0 & \text{Otherwise} \end{cases} \)

Consensus Clustering

\[
\min_{P^*} J = \frac{1}{T} \sum_{t=1}^{T} d(P^t, P^*) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i,j=1}^{n} \left[ M_{ij}(P^t) - M_{ij}(P^*) \right]^2
\]
Consensus Clustering is Equivalent to Clustering Consensus Association
(Li, Ding & Jordan, ICDM 2007)

**Consensus Association**

\[ M_{ij} = \frac{1}{T} \sum_{t=1}^{T} M_{ij}(P^t) \]

Let \( U_{ij} = M_{ij}(P^*) \)

**Consensus Objective**

\[ J = \frac{1}{T} \sum_{t} \sum_{ij} \left( M_{ij}(P^t) - \tilde{M}_{ij} + \tilde{M}_{ij} - U_{ij} \right)^2 \]

\[ = \frac{1}{T} \sum_{t} \sum_{ij} \left( M_{ij}(P^t) - \tilde{M}_{ij} \right)^2 + \sum_{ij} \left( \tilde{M}_{ij} - U_{ij} \right)^2 \]

Constant

**Solving U**

\[ \text{Min } J = \text{Min } \sum_{i,j=1}^{n} \left( \tilde{M}_{ij} - U_{ij} \right)^2 = \left\| \tilde{M} - U \right\|^2 \]
NMF Formulations

Cluster indicator

\[
\begin{bmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{bmatrix} = H
\]

\[
\begin{align*}
\text{Min}_{U} & \left\| \tilde{M} - U \right\|^2 \\
\text{Min}_{H \geq 0} & \left\| \tilde{M} - HH^T \right\|^2 \\
H & \text{ is cluster indicator} \\
D & = \text{diag}(H^T H) = \text{diag}(n_1, \ldots, n_k) \\
\text{Min}_{H^T H = D, H \geq 0} & \left\| \tilde{M} - HH^T \right\|^2 \\
\text{Symmetric NMF}
\end{align*}
\]
Weighted Consensus Clustering

• In consensus clustering, no all clusterings are useful
  – Subset of clusterings might be highly correlated
  – Some clusterings are not “good” in quality
  – Redundancy

• Weighted Consensus Clustering (Li & Ding, SDM 2008)
  – Based on NMF framework
  – Each input clustering is weighted and the weights are automatically determined via optimization
Semi-Supervised Clustering

• We may have some idea about the data
  – Before Clustering: Prior Knowledge
  – After Clustering: User Feedback

• Instance Level Constraints
  – Pairwise Constraints
  – Are x and y in the same cluster?

\[
\begin{array}{c}
\text{must-link} \\
\begin{array}{cc}
\circ x & \circ y \\
\end{array}
\end{array}
\quad \begin{array}{c}
\text{cannot-link} \\
\begin{array}{cc}
\circ x & \cdots & \circ y \\
\end{array}
\end{array}
\]
Constraints

**Must-link Constraints**

\[ A = \{(i_1, j_1), \ldots, (i_a, j_a)\}, \quad a = |A| \]

**Cannot-link Constraints**

\[ B = \{(i_1, j_1), \ldots, (i_b, j_b)\}, \quad b = |B| \]

A, B can be viewed as symmetric matrices containing \{0,1\}
NMF Formulations
(Li, Ding & Jordan, ICDM 2007)

Must-link Condition

\[ \text{Max}_H \sum_{(i,j) \in A} (HH^T)_{ij} = \sum_{ij} A_{ij} (HH^T)_{ij} = \text{Tr} H^T A H \]

Cannot-link Condition

\[ \sum_{(i,j) \in B} (HH^T)_{ij} = \text{Tr} H^T B H = 0 \text{ or } \text{Min}_H \text{Tr} H^T B H \]

Semi-Supervised Clustering

\[ \text{Max}_{H^T H = I, H \geq 0} \text{Tr} \left[ H^T WH + \alpha H^T A H - \beta H^T B H \right] \]

\[ W = XX^T \]
NMF-based Algorithm

\[
\begin{align*}
\text{Max}_{H^T H = I, H \geq 0} & \quad Tr[ H^T WH + \alpha H^T AH - \beta H^T BH ] \\
\text{Max}_{H^T H = I, H \geq 0} & \quad Tr[ H^T (W^+ - W^-) H ] \\
\text{Min}_{H \geq 0} & \quad \left\| (W^+ - W^-) - HH^T \right\|^2
\end{align*}
\]

Update H:

\[
H_{ik} \leftarrow H_{ik} \sqrt{ \frac{(W^+ H)_{ik}}{(W^- H)_{ik} + [HH^T H]_{ik}}} 
\]
Talk Outline

• Motivation
• Matrix-Based Algorithms for Clustering
• Non-Negative Matrix Factorization (NMF) for Clustering
• Non-Negative Matrix Factorization (NMF) for Consensus Clustering and Semi-Supervised Clustering
• Adaptive Dimension Reduction
Dimension Reduction and Clustering

- Consider skewed distributions

PCA

Standard PCA fails to provide the most discriminant subspace!
Dimension Reduction and Clustering

- Skewed data distributions

We wish to find the most discriminant subspace in an unsupervised way!
Relationship between LDA and K-means Clustering

Data: \( X = (x_1, x_2, \cdots, x_n) \) \( \sum_i x_i / n = 0 \)

Scatter matrix:
\[
S_t = \sum_{i=1}^{n} x_i x_i^T, \quad S_b = \sum_k n_k m_k m_k^T \quad S_w = \sum_k \sum_{i \in C_k} (x_i - m_k)(x_i - m_k)^T
\]

K-means clustering:
\[
\min_{J_K}, \quad J_K = \sum_k \sum_{i \in C_k} ||x_i - m_k||^2 \quad J_K = \text{Tr} S_w = \text{Tr} (S_t - S_b)
\]

K-means clustering equivalent to \( \min S_w \) or \( \max S_b \)
LDA and K-means Clustering

LDA directions $U$ are determined by

$$\max_U \operatorname{Tr} \frac{U^T S_b U}{U^T S_w U}$$

$$\Rightarrow \min_U \operatorname{Tr}(U^T S_w U) \quad \text{and} \quad \max_U \operatorname{Tr}(U^T S_b U)$$

K-means and LDA optimize same objectives!

K-means clustering:

$$\min_{\mathcal{H}} J_K, \quad J_K = \sum_k \sum_{i \in C_k} \|x_i - m_k\|^2$$

$$J_K = \operatorname{Tr} S_w = \operatorname{Tr} (S_t - S_b)$$

K-means clustering equivalent to $\min S_w$ or $\max S_b$
LDA and K-means

- LDA and K-means optimize same objectives
  - LDA is supervised
  - K-means is unsupervised
- Can we do LDA in unsupervised way? Find the most discriminative subspace. (Ding & Li, ICML 2007)

\[
\max_{U,H} \text{Tr} \frac{U^T S_b U}{U^T S_w U}
\]

- $U$ is LDA subspace;
- $H$ is cluster indicator from K-means clustering
LDA and K-means

Fix $U$ and solve for $H$:

$$\max_H \frac{\text{Tr} \ U^T S_b U}{\text{Tr} \ U^T S_w U} = \frac{\text{Tr} \ U^T (S_t - S_w) U}{\text{Tr} \ U^T S_w U} = \frac{\text{Tr} \ U^T S_t U}{\text{Tr} \ U^T S_w U} - 1$$

$\text{Tr} \ U^T S_t U$ is independent of $H$

$$\min_H \text{Tr} \ U^T S_w U = \sum_k \sum_{i \in C_k} ||U^T x_i - U^T m_k||^2$$

This is K-means clustering in subspace $U$
LDA and K-means

Fix $H$ and solve for $U$:

Using standard LDA procedure
Combine **LDA** and **K-means** into a single algorithm (LDA-Km)

**Unsupervised LDA** to find most discriminative subspace.

- **Iterate**
  - K-means clustering in the current subspace
  - Do LDA based on current cluster labels.
LDA-Km Algorithm

Start with PCA
Do unsupervised LDA
Iterate until convergence
Algorithm correctly discovered the LDA subspace
Relation to Earlier Approaches

(LDA-Km deals with full LDA. Earlier approaches do partial LDA.)

Adaptive Dimension Reduction (Ding, et al, 2002)
– Deal with between-class scatter only

\[
\max_{U, H} \text{Tr}(U^T S_b U)
\]

Adaptive Subspace Iteration (Li et al, 2004)
– Deal within-class scatter only

\[
\min_{C, H, U} \|U^T X - CH^T\|^2 = \text{Tr} U^T S_w U
\]

\[
Y = U^T X \implies \min_{C, H} \|Y - CH^T\|^2, \quad \text{s.t. } C \geq 0, H \succeq 0, H^T H = I,
\]

This is equivalent to Relaxed K-means clustering (Ding, et al, 2005)
Relation to Earlier Approaches (II)

(LDA-Km deals with full LDA. Earlier approaches do partial LDA.)

Discriminative Cluster Analysis (De la Torre & Kanade, 2006)
– Deal with between-class scatter only

\[
\min_{H,V,U} \left\| \left( HH^T \right)^{-1/2} \left( H^T - VU^T X \right) \right\|^2
\]

using our notation \( U, H, \) where \( V \) is a new matrix factor.

\[
\Rightarrow \quad \max_{H,U} \text{Tr} \left( \frac{U^T S_b U}{U^T S_t U} \right) \quad \text{(full LDA is } \max_{U,H} \text{Tr} \left( \frac{U^T S_b U}{U^T S_w U} \right) \text{)}
\]

(replace \( U^T S_w U \) by \( U^T S_t U \), which is most important part of LDA)
Current Work

• Data Graph
• Tensor Data
  – document, term, author
  – User, query, advertisements
  – Document, author, year
• Scalability
  – Randomize algorithms
• Multi-way relational data
• Information and knowledge transformation
  – Label propagation on multi-partite graphs
Summary

Matrix-based Algorithms: A new/rich paradigm for unsupervised learning
Acknowledgements

• Funding
  – NSF Career Award
  – NIH
  – Faculty Research Awards from IBM, Xerox and NEC
  – Equipment Donation from IBM

• Collaborators
  – Dr. Chris Ding (University of Texas at Arlington)
  – Dr. Michael I. Jordan (UC Berkeley)
  – Dr. Charles Perng (IBM Research)
  – Dr. Tong Sun (Xerox Research)
  – Dr. Shenghuo Zhu (NEC Research)

• Ph.D. Students
  – Dr. David Kaiser
  – Dr. Wei Peng
  – Bo Shao
  – Dingding Wang
  – Haifeng Wang
  – Yi Zhang
Question?

- Email: taoli@cs.fiu.edu
- http://www.cs.fiu.edu/~taoli

Thank You!