Document Retrieval

Goal: Retrieve documents of interest
Task 1: Find Similar Documents

Setup

- **Input:** Query article
- **Output:** Set of k similar articles
k-Nearest Neighbor

Articles

\( X = \{ x^1, \ldots, x^N \}, \quad x^i \in \mathbb{R}^d \)

Query:

\( x \in \mathbb{R}^d \)

k-NN

Goal:

Find \( k \) articles in \( X \) closest to \( x \)

Formulation:

\( X_{\text{NN}} = \{ x_{\text{NN}1}, \ldots, x_{\text{NN}k} \} \subseteq X \)

\( \forall x^i \in X \setminus X_{\text{NN}} \)

\( d(x^i, x) \geq \max_{x_{\text{NN}i} \in X_{\text{NN}}} d(x_{\text{NN}i}, x) \)
Nearest Neighbor with KD Trees

- Traverse the tree looking for the nearest neighbor of the query point.
Task 2: Cluster Documents

Setup

- **Input:** Corpus of documents
- **Output:** Topic assignment per document
A Generative Model

- Documents: \( x^1, \ldots, x^D \)
- Associated topics: \( z^1, \ldots, z^D \)
- Parameters: \( \theta = \{ \pi, \beta \} \)

Generative model:

\[
\begin{align*}
  z^d &\sim \pi \quad d = 1, \ldots, D \\
  w_i^d | z^d &\sim \beta_{z^d} \quad i = 1, \ldots, N
\end{align*}
\]

Bayesian approach:

- \( \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \)
- \( \beta_k \sim \text{Dir}(\chi_1, \ldots, \chi_N) \)

\( \beta_k \) is a V-dim pmf

size of vocab.

word prob. for cluster/topic \( z^d \)

©Emily Fox 2013
Inference

- **Two tasks**
  - **Point estimation:**
    - Expectation-Maximization (EM)
  - **Characterize posterior:**
    - Gibbs sampling
    - Variational methods
    - Stochastic variational inference

\[ \hat{\theta}_{ML}, \text{ or } \hat{\theta}_{MAP} \sim p(\theta) \text{ prior} \]
EM Algorithm

- Initial guess: $\hat{\theta}^{(0)}$
- Estimate at iteration $t$: $\hat{\theta}^{(t)}$

**E-Step**

Compute

\[
U(\theta, \hat{\theta}^{(t)}) = E[\log p(y | \theta) | x, \hat{\theta}^{(t)}]
\]

**M-Step**

Compute

\[
\hat{\theta}^{(t+1)} = \arg \max_\theta U(\theta, \hat{\theta}^{(t)}) + \log p(\theta)
\]
Collapsed Gibbs Sampling

\[ \pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \quad z^i \sim \pi \]

\[ \{\mu_k, \Sigma_k\} \sim F(\phi) \quad x^i \mid z^i \sim \mathcal{N}(x^i; \mu_{z^i}, \Sigma_{z^i}) \]

- Collapsed sampler

For \( i = 1, \ldots, N \)

\[ z^k(t) \sim \pi \left( z^i \mid z^{(t)}, \ldots, z^{i-1(t)}, z^{i+1(t)}, \ldots, z^{N(t)}, X_1: \mathbb{W}, \phi, \alpha, \lambda \right) \]
Task 3: Mixed Membership Model

**Setup:** Document may belong to multiple clusters

- EDUCATION
- FINANCE
- TECHNOLOGY

mixed membership
Latent Dirichlet Allocation (LDA)

Each doc is a mixture of these corpus-wide topics.

Every word is assigned to a topic.

Each doc has its own prevalence of topics in doc.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Documents</th>
<th>Topic proportions and assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 0.04</td>
<td>Seeking Life's Bare (Genetic) Necessities</td>
<td></td>
</tr>
<tr>
<td>dna 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>genetic 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>life 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>evolve 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>organism 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>brain 0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neuron 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nerve 0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>data 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>number 0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>computer 0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- ©Emily Fox 2013
Variational Methods

- Recall task: Characterize the posterior
- Turn posterior inference into an optimization task
- Introduce a “tractable” family of distributions over parameters and latent variables
  - Family is indexed by a set of “free parameters”
  - Find member of the family closest to: $p(\theta, z | x)$
    - Call the family $Q$ and want $q \in Q$ that is closest to $p(\theta, z | x)$
- Questions:
  - How do we measure “closeness”?
  - If the posterior is intractable, how can we approximate something we do not have to begin with?
Variational Methods

Similarity measure:

\[ D(q(z, \theta) \| p(z, \theta \mid x)) = E_q[\log q(z, \theta)] - E_q[\log p(z, \theta \mid x)] \]

\[ = E_q[\log q(z, \theta)] - E_q[\log p(z, \theta, x)] - \log p(x) \]

Evidence lower bound (ELBO)

\[ \log p(x) = D(q(z|\theta) \| p(z|\theta \mid x)) + \mathcal{L}(q) \geq \mathcal{L}(q) \]

Therefore, minimizing KL is equivalent to maximizing a lower bound on the marginal likelihood:

- Max \( \mathcal{L}(q) = \min D(q \parallel p) = \max \text{lower bound of } \log p(x) \)

\[ \mathcal{L}(q) = E_q[\log p(\theta, z, x)] - E_q[\log q(\theta, z)\]
Task 2: Cluster Documents

**Setup**

- **Input:** Corpus of documents
- **Output:** Topic assignment per document
New Approach: Spectral Clustering

- **Goal:** Cluster observations
- **Method:**
  - Use similarity metric between observations
  - Form a similarity graph
  - Use standard linear algebra and optimization techniques to cut graph into connected components (clusters)
Setup

- Data: \(x^1, \ldots, x^N\)
- Similarity metric:
  \[ S_{ij} = \text{similarity between } x^i \text{ and } x^j \]
- Similarity graph
  - Nodes
  - Edge weights

Problem: Want to partition graph such that edges between groups have low weights
Types of Graphs

- **ε-neighborhood:**
  - Only include edges with distances < \( \varepsilon \)
  - Treat as unweighted \( w_{i,j} = \varepsilon \).

- **k-NN:**
  - Connect \( v_i \) and \( v_j \) if \( v_j \) is a k-NN of \( v_i \)
  - Weighted by similarity \( s_{ij} = w_{i,j} \)
  - Directed \( \rightarrow \) undirected

- **Mutual k-NN:**
  - Same as k-NN, but only include mutual k-NN
Issues with Choosing Graph

Choosing graph construction techniques and parameters is non-trivial

Choice matters

From von Luxburg 2007
Graph Terminology I

- Weighted adjacency matrix

\[ W = (w_{ij})_{i,j=1,\ldots,N} \]

- \( w_{ij} = 0 \) \( \Rightarrow \) no edge between \( v_i \) and \( v_j \)
- \( w_{ij} \geq 0 \)

Sparse matrix

Large \( w_{ij} \)

Low weight \( w_{ij} \)
Graph Cuts

- **Problem:** Partition graph such that edges between groups have low weights
- Define: $W(A, B) = \sum_{i \in A, j \in B} w_{i,j}$
- MinCut problem:
  \[
  \text{Cut}(A_1, \ldots, A_k) = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \overline{A_i})
  \]
  Choose $A_1, \ldots, A_k = \arg\min \text{Cut}(A_1, \ldots, A_k)$
  $A_1, \ldots, A_k$ disjoint $\subseteq V$
- Trivial to solve for $k=2$
Issues with MinCut

- MinCut favors isolated clusters

nothing working against this
Cuts Accounting for Size

- Ratio cuts (RatioCut)
- Normalized cuts (Ncut)
- Lead to “balanced” clusters

- First need more graph terminology…
  to measure “size” of the clusters
Graph Terminology II

- Two measures of size of a subset
  - Cardinality:
    \[ |A| = \text{# of vertices in } A \]
  - Volume:
    \[ \text{vol}(A) = \sum\limits_{i \in A} \sum\limits_{j=1}^{N} w_{ij} \]
Cuts Accounting for Size

- **Ratio cuts (RatioCut)**
  - $k=2$
    \[ \text{RatioCut}(A, \tilde{A}) = \text{cut}(A, \tilde{A}) \left( \frac{1}{|A|} + \frac{1}{|\tilde{A}|} \right) \]
  - General $k$
    \[ \text{RatioCut}(A_1, \ldots, A_k) = \frac{1}{2} \sum_i \frac{W(A_i, \tilde{A}_i)}{|A_i|} \]

- **Normalized cuts (Ncut)**
  - $k=2$
    \[ \text{Ncut}(A, \tilde{A}) = \text{cut}(A, \tilde{A}) \left( \frac{1}{\text{vol}(A)} + \frac{1}{\text{vol}(\tilde{A})} \right) \]
  - General $k$
    \[ \text{Ncut}(A_1, \ldots, A_k) = \frac{1}{2} \sum_i \frac{W(A_i, \tilde{A}_i)}{\text{vol}(A_i)} \]

- **Problem is NP-hard!** Look at relaxation.

©Emily Fox 2013
Graph Terminology III

- **Degree**
  \[ d_i = \sum_{j=1}^{N} w_{ij} \]  
  *only counts neighbors*

- **Degree matrix**
  \[
  D = \begin{bmatrix}
  d_1 & 0 \\
  0 & d_2 \\
  \vdots & \vdots \\
  0 & d_N
  \end{bmatrix}
  \]
Restating Cut Metric

\[ \text{Volume}(A) - \text{"association}(A)" = \text{cut cost} \]
Restating Cut Metric

\[ x^T W x \]

binary vector that's an indicator on set \( A \)

\[ 1_A \]
Restating Cut Metric

\[ x^T D x = \text{Volume} \]

Sum of weights in row 2

\[ x^T \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & d_N \end{bmatrix} x = \text{Vol}(A) \]
Restating Cut Metric

\[ x^T D x - x^T W x = x^T (D - W) x \]

\( \mathbf{1}_A \) key

\( \text{cut} (A, \bar{A}) \)
Graph Laplacian

Definition: \[ L = D - W \]

Facts:

- Symmetric, positive semi-definite
- Eigenvalues: \[ 0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N \]
- Invariance to self-edges: \[ L_{ii} = d_i - w_{ii} \quad \text{don't depend on } w_{ii} \]
- Inner product in $L$ space:
  \[ \forall f \in \mathbb{R}^N, \quad f'Lf' = \frac{1}{2} \sum_{ij} w_{ij} (f_i - f_j)^2 \quad \text{useful later} \]
Proposition:

- The multiplicity $k$ of eigenvalue 0 of $L$ is equal to the number of connected components $A_1, \ldots, A_k$.

Furthermore, $U_1 \cdots U_k = \mathbb{I}_{A_1} \cdots \mathbb{I}_{A_k}$.

Proof: Assume graph is connected ($k=1$)

$$0 = U_1^T L U_1 = \sum_{i,j} w_{ij} (U_{1i} - U_{1j})^2$$

If $w_{ij} > 0 \Rightarrow U_{1i} = U_{1j}$.

Since $\exists$ a path bt all $i,j$, then $U_1 = \text{constant} = \mathbb{1}$.
Relationship to Identifying Connected Components

- Proposition:
  - The multiplicity $k$ of eigenvalue 0 of $L$ is equal to the number of connected components

- Proof: Assume $k$ connected components

Assume WLOG that they're ordered $A_1, \ldots, A_k$

$L = \begin{pmatrix} L_1 & 0 \\ 0 & L_k \end{pmatrix}$

graph Laplacian for subgraph $A_i$

Eigenvalues $L = \bigcup \text{Eigval} (L_i) + \text{corr, eigvec } 1_{A_i}$

$\Rightarrow$ eigvecs are indicators on the partition
Example – Mixture of Gaussians

Before we dive into the theory of spectral clustering, we would like to illustrate its principle on a very simple toy example. This example will be used at several places in this tutorial, and we chose it because it is so simple that the relevant quantities can easily be plotted.

This toy data set consists of a random mixture of Gaussians. As similarity function on this data set we choose the Gaussian similarity function drawn according to a mixture of four Gaussians. The first row of Figure 1 shows the histogram of a sample drawn from this distribution (the histogram of the sample).

In Figure 1 we show the first eigenvalues and eigenvectors of the unnormalized Laplacian of a fully connected graph and the 10-nearest neighbor graph. In Figure 1 we show the first eigenvalues and eigenvectors of an eigenvector.

The reason is that the clusters are 0, and the corresponding eigenvectors are cluster indicator vectors. The reason is that the clusters are 0, and the corresponding eigenvectors are cluster indicator vectors. The reason is that the clusters are 0, and the corresponding eigenvectors are cluster indicator vectors.

From von Luxburg 2007

still info here
Graph Laplacians and Ratio Cuts

- Ratio cuts for $k=2$
- Define cluster indicator variables:
  \[ f_{A_i} = \begin{cases} \sqrt{\frac{1}{|A|}} & v_i \in A \\ -\sqrt{\frac{1}{|\overline{A}|}} & v_i \in \overline{A} \end{cases} \]
- Properties:
  \[ \sum f_{A_i} = |A| \sqrt{\frac{1}{|A|}} \sqrt{\frac{1}{|\overline{A}|}} - |\overline{A}| \sqrt{\frac{1}{|A|}} \sqrt{\frac{1}{|\overline{A}|}} = 0 \]
  \[ \|f_A\|^2 = N \]
- RatioCut
  \[ \text{RatioCut}(A, \overline{A}) = \frac{f_A^T L f_A}{\|f_A\|^2} \quad \text{for } f_A \text{ as} \]
- Reformulating RatioCut problem

\[ \min_{A \subset V} f_A^T L f_A \quad \text{s.t. } f_A \text{ defined as above, } f_A^T 1 = 1, \|f_A\|^2 = N \quad \text{for } f_{A_i} \text{ are in a discrete set} \]
Relaxation to Formulation

- Let \( f \) be arbitrary continuous vector

\[
\min_{f \in \mathbb{R}^N} \langle f', L f \rangle \quad \text{s.t.} \quad \| f \| = 1
\]

- Rayleigh-Ritz Theorem

\[
f = U_2(L) = \text{eigvec assoc. w/ 2nd smallest eigenval}
\]

- Which vector maximizes objective subject to constraint that the vector is orthogonal to the first eigenvector and has bounded norm?
Mapping Back to Partition

- To obtain partition, transform continuous $f$ to a discrete indicator

- Cluster coordinates

$$f_i \in \mathbb{R} \text{ into } C, \bar{C}$$

using k-means

- Return

$$\begin{cases} v_i \in A \text{ if } f_i \in C \\ v_i \in \bar{A} \text{ if } f_i \in \bar{C} \end{cases}$$
Ratio Cuts for General $k$

- Define cluster indicator variables:
  \[
  F_{i,j} = \begin{cases} 
  1 / \sqrt{|A_j|} & \text{if } i \in A_j \\
  0 & \text{otherwise}
  \end{cases}, \quad \forall A_j \in \mathbb{R}^{N \times k}, \quad F_A F_A = I
  \]

- RatioCut
  \[
  \text{RatioCut}(A_1, \ldots, A_k) = \sum_{i=1}^{k} f_{A_i}^T L f_{A_i} = \text{Tr}(F_A^T L F_A)
  \]

- Reformulating RatioCut problem
  \[
  \min_{A_1, \ldots, A_k} \text{Tr}(F_A^T L F_A)
  \]
  \[
  \quad \text{and } F_A \text{ s.t. } F_A^T F_A = I
  \]

- Relaxation
  \[
  \min_{F \in \mathbb{R}^{N \times k}} \text{Tr}(F^T L F)
  \]
  \[
  \text{ s.t. } F^T F = I
  \]
Ratio Cuts for General k

- Relaxation:
  \[
  \min_{F \in \mathbb{R}^{N \times k}} \text{Tr}(F'LF) \quad \text{s.t.} \quad F'F = I
  \]

- Solution: standard trace min problem
  \[
  \Rightarrow \text{choose } F \text{ containing first } k \text{ eigenvectors of } L
  \]
  \[
  \begin{bmatrix}
  \mathbf{u}_1 & \cdots & \mathbf{u}_k
  \end{bmatrix}
  \]

- To obtain partition:
  cluster rows of \( F \) using \( k \)-means
  \[
  F = \begin{bmatrix}
  \mathbf{f}_1 & \cdots & \mathbf{f}_k
  \end{bmatrix}
  \]
  if \( \mathbf{f}_i \) is in cluster \( w_i \) with \( \mathbf{f}_j \), then the rows are the same
Normalized cuts for $k=2$

Define cluster indicator variables:

$$f_{A_i} = \begin{cases} \frac{\text{vol}(A_i)}{\text{vol}(A)} & v_i \in A \\ -\frac{\text{vol}(A_i)}{\text{vol}(\overline{A})} & v_i \in \overline{A} \end{cases}$$

Properties:

$$(Df_A)'\mathbb{1} = 0 \quad \text{and} \quad f_A' Df_A = \text{vol}(V)$$

Ncut

$$\text{Ncut}(A, \overline{A}) = \frac{f_A' L f_A}{\text{vol}(V)}$$

Reformulating Ncut problem

$$\min_{f_A} f_A' L f_A \quad \text{s.t.} \quad Df_A' \mathbb{1} = \mathbb{1} \quad \text{and} \quad f_A' Df_A = \text{vol}(V)$$
Relaxation to Formulation

- Let $f$ be arbitrary continuous vector

$$\min_{f \in \mathbb{R}^n} \mathcal{L}(f) \quad \text{s.t.} \quad \mathcal{L}(f) = \frac{1}{2} f^T D f = \text{vol}(\mathcal{V})$$

- Rayleigh-Ritz Theorem

$$\min_{g \in \mathbb{R}^n} \mathcal{L}(D^{1/2} g) \quad \text{s.t.} \quad g^T D^{1/2} g = 1 \quad \|g\|^2 = \text{vol}(\mathcal{V})$$

$g = U_2(\mathcal{L}_{\text{sym}})$

$$\Rightarrow f = D^{-1/2} U_2(\mathcal{L}_{\text{sym}}) = U_2(\mathcal{L}_{\text{rw}})$$

equiv to $f$ soln of $Lu = \frac{1}{2}DU$
Normalized Cuts for General k

- Define cluster indicator variables:
  \[ F_{ij} = \begin{cases} 
  1 / \sqrt{\text{vol}(A_j)} & v_i \in A_j \\
  0 & \text{otherwise}
  \end{cases} \]

- Reformulating RatioCut problem
  \[
  \min_{A_1, \ldots, A_k} \text{Tr}(F_A' LF_A) \quad \text{s.t.} \quad F_A' DF_A = I
  \]

- Relaxation
  \[
  \min_{H \in \mathbb{R}^{N \times k}} \text{Tr}(H' D^{-1/2} LD^{-1/2} H) \quad \text{s.t.} \quad H' H = I
  \]

- Solution:
  - H is matrix of first k eigenvectors of \( L_{\text{sym}} \), which is equivalent to the approximate F being the first k eigenvectors of \( L_{\text{rw}} \)
Random Walks on Graphs

- Stochastic process with random jumps from $v_i$ to $v_j$ wp:

- Transition matrix:

- Connection to graph Laplacian:

- Intuitively, want to partition graph s.t. random walk stays in cluster for a while and rarely jumps between clusters
Assume that stationary distribution exists and is unique. Then,

**Proposition:** \( \text{Ncut}(A, \bar{A}) = P(A | \bar{A}) + P(\bar{A} | A) \)

**Proof:**

Minimizing normalized cuts is equivalent to minimizing the probability of transitioning between clusters.
Notes

- No guarantee to quality of approximation

- Sensitive to choice of similarity graph (see earlier)

- Which graph Laplacian to use?
  - If degrees in graph vary significantly, then Laplacians are quite different
  - In general, $L_{rw}$ behaves the best
  - Volume gives better measure of within-cluster similarity than cardinality
  - Normalized cuts has consistency results, Ratio cuts does not
Choosing the number of clusters $k$ can be hard

- Easy when clusters are well-separated

$k$-means to return partition from solution to relaxation is an approach, but not the only

From von Luxburg 2007