Learning with Local Consistency

Deng Cai (蔡登)
College of Computer Science
Zhejiang University
dengcai@gmail.com
What is Local Consistency?

- Nearby points (neighbors) share similar properties.
- Traditional machine learning algorithms:
  - *k*-nearest neighbor classifier
Local Consistency Assumption

- A lot of \textit{unlabeled} data
- \textbf{Local} consistency
  - $k$-nearest neighbors
  - $\epsilon$-neighbors
  - ...

...
Local Consistency Assumption

- Put edges between neighbors (nearby data points)
- Two nodes in the graph connected by an edge share similar properties.
Local Consistency Assumption

- Similar properties
  - Labels
  - Representations
  - \( x: f(x) \)

- \( W \in \mathcal{R}^{n \times n} \): weight matrix of the graph

\[
\min \frac{1}{2} \sum_{i,j} W_{ij} \left(f(x_i) - f(x_j)\right)^2 \\
\min y^T (D - W)y \\
\min y^T Ly \\
s.t. \quad y^T Dy = 1
\]

\( y_i = f(x_i) \)
\( y = [y_1, \ldots, y_n]^T \)

Local Consistency and Manifold Learning

- Manifold learning
- We only need **local consistency**

$$\min \sum_{i,j} W_{ij} (f(x_i) - f(x_j))^2$$
How to use the local consistency idea?
Local Consistency in Semi-Supervised Learning

- Supervised learning

\[ f^* = \arg\min_f \frac{1}{m} \sum_{i=1}^{m} l(x_i, y_i, f) + \lambda \|f\|^2 \]

- Squared loss: ridge regression (regularized least squares)

- Hinge loss: SVM

- Semi-Supervised learning (with local consistency)

\[ f^* = \arg\min_f \frac{1}{m} \sum_{i=1}^{m} l(x_i, y_i, f) + \lambda_1 \|f\|^2 + \lambda_2 \sum_{i,j=1}^{n} W_{ij} (f(x_i) - f(x_j))^2 \]

- Laplacian least squares and Laplacian SVM.

Manifold Regularization

- Semi-Supervised learning (with local consistency)

\[ f^* = \underset{f}{\text{argmin}} \frac{1}{m} \sum_{i=1}^{m} l(x_i, y_i, f) + \lambda_1 \|f\|^2 + \lambda_2 \sum_{i,j=1}^{n} W_{ij} \left( f(x_i) - f(x_j) \right)^2 \]

- Laplacian least squares

\[ a^* = \left( XX^T + \lambda_1 I + \lambda_2 XLX^T \right)^{-1} Xy \]

- Ridge regression (regularized least squares)

\[ a^* = \left( XX^T + \lambda I \right)^{-1} Xy \]

How to use the local consistency idea?

- Matrix factorization
  - Non-negative matrix factorization

- Topic modeling
  - Probabilistic latent semantic analysis

- Clustering
  - Gaussian mixture model
Matrix Factorization (Decomposition)

\[ X = [x_1, \cdots, x_n] \in \mathbb{R}^{p \times n} \Rightarrow X \approx UV^T \]

\[ X \approx \hat{X} = UV^T \]

- Approximation
- Left factor
- Right factor
Matrix Factorization (Decomposition)

\[ X \approx UV^T \]

\[
\begin{bmatrix}
    x_{11} & x_{12} & \cdots & x_{1n} \\
    x_{12} & x_{22} & \cdots & x_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{1m} & x_{2m} & \cdots & x_{nm}
\end{bmatrix}
\approx
\begin{bmatrix}
    u_{11} & \cdots & u_{k1} \\
    u_{12} & \cdots & u_{k2} \\
    \vdots & \ddots & \vdots \\
    u_{1m} & \cdots & u_{km}
\end{bmatrix}
\times
\begin{bmatrix}
    v_{11} & v_{12} & \cdots & v_{1n} \\
    \vdots & \vdots & \ddots & \vdots \\
    v_{k1} & v_{k2} & \cdots & v_{kn}
\end{bmatrix}
\]

\[
\begin{bmatrix}
    x_i
\end{bmatrix}
\approx v_{1i} \cdot \begin{bmatrix}
    u_1
\end{bmatrix}
+ v_{2i} \cdot \begin{bmatrix}
    u_2
\end{bmatrix}
+ \cdots
+ v_{ki} \cdot \begin{bmatrix}
    u_k
\end{bmatrix}
\]
Singular Value Decomposition

- For an arbitrary matrix $X$ there exists a factorization (Singular Value Decomposition = SVD) as follows:

$$X = U \Sigma V^T \in \mathbb{R}^{n \times m}$$

- Where
  - (i) $U \in \mathbb{R}^{n \times k}$, $\Sigma \in \mathbb{R}^{k \times k}$, $V \in \mathbb{R}^{m \times k}$
  - (ii) $U'U = I$ \hspace{1cm} $V'V = I$ Orthonormal columns
  - (iii) $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_k)$, $\sigma_i \geq \sigma_{i+1}$ Singular values (ordered)
  - (iv) $k = \text{rank}(X)$

The LSA via SVD can be summarized as follows:

- **Document similarity**
  \[
  \langle \hat{X}, \hat{X} \rangle
  \]

- **Folding-in queries**
  \[
  \hat{q} = \sum_{k}^{-1} V_k q
  \]

---

Non-negative Matrix Factorization

\[ X \approx \tilde{X} = UV^T, \min \| X - UV^T \|_2^2 \]
\[ u_{ij} \geq 0, v_{ij} \geq 0 \]

- The Euclidean distance \( \| X - UV^T \|_2^2 \) is nonincreasing under the update rules

\[
    u_{ik} \leftarrow \frac{(XV)_{ik}}{(UV^T V)_{ik}} u_{ik} \quad v_{jk} \leftarrow \frac{(X^T U)_{jk}}{(V U^T U)_{jk}} v_{jk}
\]

- Can we incorporate the local consistency idea?

Locally Consistent NMF

\[ X \approx UV^T \]

If \( x_i \) and \( x_j \) are neighbors

\[ x_i = v_{1i} \cdot u_1 + v_{2i} \cdot u_2 + \cdots + v_{ki} \cdot u_k \]

\[ x_j = v_{1j} \cdot u_1 + v_{2j} \cdot u_2 + \cdots + v_{kj} \cdot u_k \]

- Neighbor: prior knowledge, label information, \( p \)-nearest neighbors ...

D. Cai, X. He, J. Han, and T. Huang, Graph regularized Non-negative Matrix Factorization for Data Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear.
Locally Consistent NMF

\[
\mathbf{x}_i = v_{1i} \cdot \mathbf{u}_1 + v_{2i} \cdot \mathbf{u}_2 + \cdots + v_{ki} \cdot \mathbf{u}_k
\]

\[
\mathbf{x}_j = v_{1j} \cdot \mathbf{u}_1 + v_{2j} \cdot \mathbf{u}_2 + \cdots + v_{kj} \cdot \mathbf{u}_k
\]

\[
\min \sum_{i,j} W_{ij} (f(x_i) - f(x_j))^2 \quad \min \sum_k \sum_{i,j} W_{ij} (v_{ki} - v_{kj})^2
\]

\[
\min \text{Tr}(V^T LV)
\]

D. Cai, X. He, J. Han, and T. Huang, Graph regularized Non-negative Matrix Factorization for Data Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear.
Objective Function

**NMF:** \[
\begin{align*}
\min \| X - UV^T \|^2 \\
\quad u_{ik} &\leftarrow \frac{(XV)_{ik}}{(UVT^T)_{ik}} u_{ik} \\
\quad v_{jk} &\leftarrow \frac{(X^T U)_{jk}}{(VUT^T)_{jk}} v_{jk}
\end{align*}
\]

**GNMF:** \[
\begin{align*}
\min \| X - UV^T \|^2 + \lambda \text{Tr}(V^T LV) \\
\quad u_{ik} &\leftarrow \frac{(XV)_{ik}}{(UVT^T)_{ik}} u_{ik} \\
\quad v_{jk} &\leftarrow \frac{(X^T U + \lambda WV)_{jk}}{(VUT^T + \lambda DV)_{jk}} v_{jk}
\end{align*}
\]

Graph regularized NMF

D. Cai, X. He, J. Han, and T. Huang, Graph regularized Non-negative Matrix Factorization for Data Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear.
# Clustering Results

<table>
<thead>
<tr>
<th>$K$</th>
<th>NMF</th>
<th>GNMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>81.0±14.2</td>
<td>93.5±10.1</td>
</tr>
<tr>
<td>6</td>
<td>74.3±10.1</td>
<td>92.4±6.1</td>
</tr>
<tr>
<td>8</td>
<td>69.3±8.6</td>
<td>84.0±9.6</td>
</tr>
<tr>
<td>10</td>
<td>69.4±7.6</td>
<td>84.4±4.9</td>
</tr>
<tr>
<td>12</td>
<td>69.0±6.3</td>
<td>81.0±8.3</td>
</tr>
<tr>
<td>14</td>
<td>67.6±5.6</td>
<td>79.2±5.2</td>
</tr>
<tr>
<td>16</td>
<td>66.0±6.0</td>
<td>76.8±4.1</td>
</tr>
<tr>
<td>18</td>
<td>62.8±3.7</td>
<td>76.0±3.0</td>
</tr>
<tr>
<td>20</td>
<td>60.5</td>
<td>75.3</td>
</tr>
<tr>
<td>Avg.</td>
<td>68.9</td>
<td>82.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$K$</th>
<th>NMF</th>
<th>GNMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>95.5±10.2</td>
<td>98.5±2.8</td>
</tr>
<tr>
<td>10</td>
<td>83.6±12.2</td>
<td>91.4±7.6</td>
</tr>
<tr>
<td>15</td>
<td>79.9±11.7</td>
<td>93.4±2.7</td>
</tr>
<tr>
<td>20</td>
<td>76.3±5.6</td>
<td>91.2±2.6</td>
</tr>
<tr>
<td>25</td>
<td>75.0±4.5</td>
<td>88.6±2.1</td>
</tr>
<tr>
<td>30</td>
<td>71.9</td>
<td>88.6</td>
</tr>
<tr>
<td>Avg.</td>
<td>80.4</td>
<td>92.0</td>
</tr>
</tbody>
</table>

**COIL20**

- Please check our papers for more details.

**TDT2**
How to use the local consistency idea?

- Matrix factorization
  - Non-negative matrix factorization

- Topic modeling
  - Probabilistic latent semantic analysis

- Clustering
  - Gaussian mixture model
What is Topic Modeling

- Powerful tool for text mining
  - Topic discovery,
  - Summarization,
  - Opinion mining,
  - Many more ...

Text Collection → Probabilistic Topic Modeling → Topic models (Multinomial distributions)

- term: 0.16
- relevance: 0.08
- weight: 0.07
- feedback: 0.04
- web: 0.21
- search: 0.10
- link: 0.08
- graph: 0.05
- ...
Language Model Paradigm in IR

- Probabilistic relevance model
  - Random variables
    \[ R_d \in \{0, 1\} : \text{relevance of document } d \]
    \[ q \subseteq \Sigma : \text{query, set of words} \]
  - Bayes’ rule
    \[
P(R_d = 1 | q) = \frac{P(q | R_d = 1) \cdot P(R_d = 1)}{P(q)}
    \]
    probability that document d is relevant for query q
    probability of generating a query q to ask for relevant d
    prior probability of relevance for document d (e.g. quality, popularity)

Language Model Paradigm

\[ P(R_d = 1|q) \propto P(q|R_d = 1) \times P(R_d = 1) \]

- **First contribution: prior probability of relevance**
  - simplest case: uniform (drops out for ranking)
  - **popularity**: document usage statistics (e.g. library circulation records, download or access statistics, hyperlink structure)

- **Second contribution: query likelihood**
  - query terms \( q \) are treated as a **sample** drawn from an (unknown) relevant document
Query Likelihood

- \( P(q|R_d = 1) \equiv P(q|d) \)

- \( q = (w_1, \ldots, w_q) \)

- Independent Assumption

\[
P(q|d) = \prod_{w \in q} P(w|d)
\]

\( P(w|d) \)?
**Naive Approach**

Maximum Likelihood Estimation

$$\hat{P}_{\text{ML}}(w|d) = \frac{n(d, w)}{\sum_{w'} n(d, w')}$$

Zero frequency problem: terms not occurring in a document get zero probability
Estimation Problem

- Crucial question: In which way can the document collection be utilized to improve probability estimates?
Probabilistic Latent Semantic Analysis

Documents

Terms

\[ P(z|d; \theta) \quad P(w|z; \pi) \]

TRADE

Latent Concepts

- economic
- imports
- trade

Model fitting

\[ \hat{P}_{LSA}(w|d) = \sum_z P(w|z; \theta) P(z|d; \pi) \]

pLSA via Likelihood Maximization

- **Log-Likelihood**

$$l(\theta, \pi; \mathbf{N}) = \sum_{d,w} n(d, w) \log \left( \sum_z P(w|z; \theta) P(z|d; \pi) \right)$$

- **Goal**: Find model parameters that maximize the log-likelihood, i.e. maximize the average predictive probability for observed word occurrences (non-convex optimization problem)
Expectation Maximization Algorithm

- **E step**: posterior probability of latent variables ("concepts")

\[
P(z|d, w) = \frac{P(z|d; \pi)P(w|z; \theta)}{\sum_{z'} P(z'|d; \pi)P(w|z'; \theta)}
\]

Probability that the occurrence of term \( w \) in document \( d \) can be "explained" by concept \( z \).

- **M step**: parameter estimation based on "completed" statistics

\[
P(w|z; \theta) \propto \sum_d n(d, w)P(z|d, w),
\]

\[
P(z|d; \pi) \propto \sum_w n(d, w)P(z|d, w)
\]

Local Consistency?

- Put edges between neighbors (nearby data points);
- Two nodes in the graph connected by an edge share similar properties.

Network data
- Co-author network, Facebook, webpage
Text Collections with Network Structure

- Literature + coauthor/citation network
- Email + sender/receiver network
- Web page + hyperlink structure
- Blog articles + friend network
- News + geographic network
Importance of Topic Modeling on Network

Computer Science Literature = Information Retrieval + Data Mining + Machine Learning, ...

or

Domain Review + Algorithm + Evaluation, ...

?
Intuitions

- People working on the same topic belong to the same “topical community”
- Good community: coherent topic + well connected
- A topic is semantically coherent if people working on this topic also collaborate a lot

More likely to be an IR person or a compiler person?
Social Network Context for Topic Modeling

- Context = author
- Coauthor = similar contexts
- Intuition: I work on similar topics to my neighbors

Smoothed Topic distributions \( \rightarrow \) \( P(\theta_j | \text{author}) \)

- e.g. coauthor network

D. Cai, X. Wang, and X. He, Probabilistic Dyadic Data Analysis with Local and Global Consistency, ICML’09.
Objective Function

\[
\ell(\theta, \pi; \mathbf{N}) = \sum_{d,w} n(d, w) \log\left( \sum_z P(w|z; \theta) P(z|d; \pi) \right) + \lambda R
\]

\[
\min \sum_{i,j} W_{ij} \left( f(x_i) - f(x_j) \right)^2 \quad f(x_i) = f(d_i) \equiv P(z|d_i)
\]

\[
D \left( P(z|d_i) \| P(z|d_j) \right) = \sum_z P(z|d_i) \log \frac{P(z|d_i)}{P(z|d_j)}
\]

\[
R = -\frac{1}{2} \sum_{i,j} W_{ij} \left( D \left( P(z|d_i) \| P(z|d_j) \right) + D \left( P(z|d_j) \| P(z|d_i) \right) \right)^2
\]

D. Cai, X. Wang, and X. He, Probabilistic Dyadic Data Analysis with Local and Global Consistency, ICML'09.
Parameter Estimation via EM

- **E step**: posterior probability of latent variables ("concepts")

\[ P(z_k | d_i, w_j) = \frac{P(w_j | z_k)P(z_k | d_i)}{\sum_{l=1}^{K} P(w_j | z_l)P(z_l | d_i)} \]

Same as PLSA

- **M step**: parameter estimation based on "completed" statistics

\[ P(w_j | z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k | d_i, w_j)}{\sum_{m=1}^{M} \sum_{i=1}^{N} n(d_i, w_m)P(z_k | d_i, w_m)} \]

Same as PLSA

\[ P(z_k | d_i) = ? \]

D. Cai, X. Wang, and X. He, Probabilistic Dyadic Data Analysis with Local and Global Consistency, ICML'09.
Parameter Estimation via EM

- **M step:** parameter estimation based on “completed” statistics

\[
\begin{bmatrix}
    P(z_k \mid d_1) \\
    P(z_k \mid d_2) \\
    \vdots \\
    P(z_k \mid d_N)
\end{bmatrix} = (\Omega + \lambda L)^{-1} \begin{bmatrix}
    \sum_{j=1}^{M} n(d_1, w_j) P(z_k \mid d_1, w_j) \\
    \sum_{j=1}^{M} n(d_2, w_j) P(z_k \mid d_2, w_j) \\
    \vdots \\
    \sum_{j=1}^{M} n(d_N, w_j) P(z_k \mid d_N, w_j)
\end{bmatrix}
\]

\[
\Omega = \begin{bmatrix}
    n(d_1) \\
    \vdots \\
    n(d_N)
\end{bmatrix} \quad L = D - W, \quad \text{Graph Laplacian}
\]

If \( \lambda = 0 \)

\[
P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) P(z_k \mid d_i, w_j)}{n(d_i)} \quad \text{Same as PLSA}
\]

D. Cai, X. Wang, and X. He, Probabilistic Dyadic Data Analysis with Local and Global Consistency, ICML'09.
Experiments

- Bibliography data and coauthor networks
  - DBLP: text = titles; network = coauthors
  - Four conferences (expect 4 topics): SIGIR, KDD, NIPS, WWW
Topical Communities with PLSA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>term</td>
<td>peer</td>
<td>visual</td>
<td>interface</td>
</tr>
<tr>
<td>question</td>
<td>patterns</td>
<td>analog</td>
<td>towards</td>
</tr>
<tr>
<td>protein</td>
<td>mining</td>
<td>neurons</td>
<td>browsing</td>
</tr>
<tr>
<td>training</td>
<td>clusters</td>
<td>vlsi</td>
<td>xml</td>
</tr>
<tr>
<td>weighting</td>
<td>stream</td>
<td>motion</td>
<td>generation</td>
</tr>
<tr>
<td>multiple</td>
<td>frequent</td>
<td>chip</td>
<td>design</td>
</tr>
<tr>
<td>recognition</td>
<td>e</td>
<td>natural</td>
<td>engine</td>
</tr>
<tr>
<td>relations</td>
<td>page</td>
<td>cortex</td>
<td>service</td>
</tr>
<tr>
<td>library</td>
<td>gene</td>
<td>spike</td>
<td>social</td>
</tr>
</tbody>
</table>
Topical Communities with NetPLSA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>retrieval</td>
<td>mining</td>
<td>neural</td>
<td>web</td>
</tr>
<tr>
<td>information</td>
<td>data</td>
<td>learning</td>
<td>services</td>
</tr>
<tr>
<td>document</td>
<td>discovery</td>
<td>networks</td>
<td>semantic</td>
</tr>
<tr>
<td>query</td>
<td>databases</td>
<td>recognition</td>
<td>services</td>
</tr>
<tr>
<td>text</td>
<td>rules</td>
<td>analog</td>
<td>peer</td>
</tr>
<tr>
<td>search</td>
<td>association</td>
<td>vlsi</td>
<td>ontologies</td>
</tr>
<tr>
<td>evaluation</td>
<td>patterns</td>
<td>neurons</td>
<td>rdf</td>
</tr>
<tr>
<td>user</td>
<td>frequent</td>
<td>gaussian</td>
<td>management</td>
</tr>
<tr>
<td>relevance</td>
<td>streams</td>
<td>network</td>
<td>ontology</td>
</tr>
</tbody>
</table>

Coherent Topical Communities

Semantics of community: “machine learning (NIPS)”

Semantics of community: “Data Mining (KDD)”

For More Details

- Please check our papers

- [http://www.zjucadcg.cn/dengcai/LapPLSA/index.html](http://www.zjucadcg.cn/dengcai/LapPLSA/index.html)

D. Cai, X. Wang, and X. He, Probabilistic Dyadic Data Analysis with Local and Global Consistency, ICML'09.
Q. Mei, D. Cai, D. Zhang, and C. Zhai, Topic Modeling with Network Regularization, WWW'08.
How to use the local consistency idea?

- Matrix factorization
  - Non-negative matrix factorization

- Topic modeling
  - Probabilistic latent semantic analysis

- Clustering
  - Gaussian mixture model
Gaussian Mixture Model (GMM) is one of the most popular clustering methods which can be viewed as a linear combination of different Gaussian components.
Gaussian Mixture Model

- **Multivariate Gaussian**
  - \( \mu \): mean of the distribution
  - \( \Sigma \): covariance of the distribution

\[
p(x) = \mathcal{N}(x; \mu, \Sigma) = \frac{1}{d} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}
\]

- **Maximum likelihood estimation**

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]
\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T
\]
Gaussian Mixture Model

- Linear combination of Gaussians
  - Assumption: $K$ Gaussians, each has a contribution of $\pi_k$ to the data points

\[
p(x; \Theta) = \sum_{k=1}^{K} \pi_k p_k(x; \theta_k)
\]

\[
\Theta = \{\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K\}, \quad \sum_{k=1}^{K} \pi_k = 1, \pi_k \in [0, 1]
\]

\[
p_k(x; \theta_k) = \mathcal{N}(x; \mu_k, \Sigma_k)
\]

- Parameters to be estimated: $\pi_k, \mu_k, \Sigma_k$
Gaussian Mixture Model

- The process of generating a data point
  - first pick one of the components with probability $\pi_k$
  - then draw a sample $x_i$ from that component distribution
- Each data point is generated by one of $k$ components
Gaussian Mixture Model

- The log-likelihood function:

\[
\log \prod_{i=1}^{N} p(x^{(i)}; \Theta) = \sum_{i=1}^{N} \log \left( \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}; \mu_k, \Sigma_k) \right)
\]

- Using EM algorithm:

\[
l(\theta) = \sum_{i=1}^{M} \sum_{z^{(i)}} Q^i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q^i(z^{(i)})}
\]

\[
eq \sum_{i=1}^{M} \sum_{k=1}^{K} Q^i(z_k^{(i)}) \log \pi_k \mathcal{N}(x^{(i)}; \mu_k, \Sigma_k)
\]
E-step:

\[ Q^i \left( z_k^{(i)} \right) = p \left( z_k^{(i)} | x^{(i)} ; \Theta \right) \]

\[ = \frac{\pi_k \mathcal{N}(x^{(i)}; \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}; \mu_k, \Sigma_k)} \]

M-step:

- Take the derivative of the complete log likelihood to obtain estimates for \( \pi_k, \mu_k, \Sigma_k \) directly

\[ \pi_k = \frac{\sum_{i=1}^{M} Q^i \left( z_k^{(i)} \right)}{M} \]

\[ \mu_k = \frac{\sum_{i=1}^{M} x^{(i)} Q^i \left( z_k^{(i)} \right)}{\sum_{i=1}^{M} Q^i \left( z_k^{(i)} \right)} \]

\[ \Sigma_k = \frac{\sum_{i=1}^{M} (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^T Q^i \left( z_k^{(i)} \right)}{\sum_{i=1}^{M} Q^i \left( z_k^{(i)} \right)} \]

- Do the iterations until convergence, then \( Q^i \left( z_k^{(i)} \right) \) can be used for clustering
Objective Function

\[
\min \sum_{i,j} W_{ij} \left(f(x_i) - f(x_j)\right)^2 \quad f(x_i) \equiv P(z|x_i)
\]

\[
D \left( P(z|x_i) \| P(z|x_j) \right) = \sum_z P(z|x_i) \log \frac{P(z|x_i)}{P(z|x_j)}
\]

\[
R = -\frac{1}{2} \sum_{i,j} W_{ij} \left( D \left( P(z|x_i) \| P(z|x_j) \right) + D \left( P(z|x_j) \| P(z|x_i) \right) \right)^2
\]

\[
\sum_{i=1}^N \log \left( \sum_{k=1}^K \pi_k N \left( x^{(i)} ; \mu_k, \Sigma_k \right) \right) + \lambda R
\]

J. Liu, D. Cai, and X. He, Gaussian Mixture Model with Local Consistency, AAAI’10.
EM Equations

- **E-step:**
  \[
P(c_k | x_i) = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j)}
  \]

- **M-step:**
  \[
  S_{i,k} = (x_i - \mu_k)(x_i - \mu_k)^T
  \]
  \[
  N_k = \sum_{i=1}^{N} P(c_k | x_i)
  \]
  \[
  \mu_k = \frac{\sum_{i=1}^{N} x_i P(c_k | x_i)}{N_k}
  \]
  \[
  \sigma_k^2 = \frac{\sum_{i=1}^{N} P(c_k | x_i) S_{i,k}}{N_k}
  \]

  \[
  \pi_k \propto \frac{\sum_{i=1}^{N} P(c_k | x_i)}{N}
  \]

  \[
  \lambda \sum_{i,j=1}^{N} \left( P(c_k | x_i) - P(c_k | x_j) \right) (x_i - x_j) W_{ij}
  \]
  \[
  \lambda \sum_{i,j=1}^{N} \left( P(c_k | x_i) - P(c_k | x_j) \right) (S_{i,k} - S_{j,k}) W_{ij}
  \]

  \[
  \frac{2N_k}{2N_k}
  \]

original GMM part
Experiment

7 Real Data sets:

- The Yale face image database.
- The Waveform model described in “The Elements of Statistical Learning”.
- The Vowels data set which has steady state vowels of British English.
- The Libras movement data set containing hand movement pictures.
- The Control Charts data set consisting control charts.
- The Cloud data set is a simple 2 classes problem.
- The Breast Cancer Wisconsin data set computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
## Clustering Results

<table>
<thead>
<tr>
<th>Data set</th>
<th>LCGMM</th>
<th>GMM</th>
<th>K-means</th>
<th>Ncut</th>
<th>size</th>
<th># of features</th>
<th># of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yale</td>
<td>54.3</td>
<td>29.1</td>
<td>51.5</td>
<td>54.6</td>
<td>165</td>
<td>4096</td>
<td>15</td>
</tr>
<tr>
<td>Libras</td>
<td>50.8</td>
<td>35.8</td>
<td>44.1</td>
<td>48.6</td>
<td>800</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>Chart</td>
<td>70.0</td>
<td>56.8</td>
<td>61.5</td>
<td>58.8</td>
<td>990</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Cloud</td>
<td>100.0</td>
<td>96.2</td>
<td>74.4</td>
<td>61.5</td>
<td>360</td>
<td>90</td>
<td>15</td>
</tr>
<tr>
<td>Breast</td>
<td>95.5</td>
<td>94.7</td>
<td>85.4</td>
<td>88.9</td>
<td>600</td>
<td>60</td>
<td>6</td>
</tr>
<tr>
<td>Vowel</td>
<td>36.6</td>
<td>31.9</td>
<td>29.0</td>
<td>29.1</td>
<td>2048</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Waveform</td>
<td>75.3</td>
<td>76.3</td>
<td>51.9</td>
<td>52.3</td>
<td>569</td>
<td>30</td>
<td>2</td>
</tr>
</tbody>
</table>
The Take-home Messages

- Local consistency is a very useful idea.
- It is very simple.
  - Nearby points (neighbors) share similar properties.
    \[
    \min \sum_{i,j} W_{ij} \left( f(x_i) - f(x_j) \right)^2
    \]
- It can be put everywhere (with a lot of unlabeled data)
  - The key: how to optimize the regularized objective function.
Thanks!