Case Study 2: Document Retrieval

Review:
Mixtures of Gaussians

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
April 21\textsuperscript{st}, 2015

Some Data

want to cluster
- unsup.
- generative approach
Gaussian Mixture Model

- Most commonly used mixture model
- Observations: \( x^1, \ldots, x^N \in \mathbb{R}^d \)
- Parameters: \( \pi = [\pi_1, \ldots, \pi_K] \) \# clusters
  \( \phi = \begin{bmatrix} \phi_1 \phi_2 \cdots \phi_K \end{bmatrix} \) \# per-cluster likelihood
  \( \phi_k = \begin{bmatrix} \mu_k \Sigma_k \end{bmatrix} \)
- Cluster indicator:
  \( z^i \in \{1, \ldots, K\} \)
- Per-cluster likelihood:
  \( \text{Pr}(z^i = k) = \pi_k \)

- Ex. \( z^i \) = country of origin, \( x^i \) = height of \( i \)th person
  - \( k \)th mixture component = distribution of heights in country \( k \)

Generative Model

- We can think of *sampling* observations from the model
- For each observation \( i \),
  - Sample a cluster assignment
    - Sample from \( \pi \)
  - Sample the observation from the selected Gaussian
    - \( x^i | z^i = k \sim N(\mu_k, \Sigma_k) \)

"conditioned upon" can "generate" obs.
Also Useful for Density Estimation

Contour Plot of Joint Density

Density as Mixture of Gaussians

- Approximate density with a mixture of Gaussians

Contour Plot of Joint Density
Density as Mixture of Gaussians

• Approximate density with a mixture of Gaussians

\[ p(x^i \mid \pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(x^i \mid \mu_k, \Sigma_k) \]

Summary of GMM Components

• Observations \( x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N \)
• Hidden cluster labels \( z_i \in \{1, 2, \ldots, K\}, \quad i = 1, 2, \ldots, N \)
• Hidden mixture means \( \mu_k \in \mathbb{R}^d, \quad k = 1, 2, \ldots, K \)
• Hidden mixture covariances \( \Sigma_k \in \mathbb{R}^{d \times d}, \quad k = 1, 2, \ldots, K \)
• Hidden mixture probabilities \( \pi_k, \quad \sum_{k=1}^{K} \pi_k = 1 \)

Gaussian mixture marginal and conditional likelihood:

\[
\begin{align*}
    p(x_i \mid \pi, \mu, \Sigma) &= \sum_{z_i=1}^{K} \pi_{z_i} N(x_i \mid \mu_{z_i}, \Sigma_{z_i}) \\
    p(x_i \mid z_i, \pi, \mu, \Sigma) &= N(x_i \mid \mu_{z_i}, \Sigma_{z_i})
\end{align*}
\]
Case Study 2: Document Retrieval

Application to Document Modeling

Machine Learning for Big Data
CSE547/STAT548, University of Washington
Emily Fox
April 21st, 2015

©Emily Fox 2015

Task 2: Cluster Documents

- Now:
  - Cluster documents based on topic
Document Representation

- Bag of words model

\[ X = \begin{bmatrix} x_1, \ldots, x_D \end{bmatrix} \]

previously \( X \) vector of word counts (e.g. tf-idf)

performed operations on this vector

now:

\[ X = \{ w_1, \ldots, w_N \} \]

unordered set of N words in doc.

\[ w \in V \ (\text{vocab}) \]

A Generative Model

- Documents:
- Associated topics:
- Parameters: \( \theta = \{ \pi, \beta \} \)

\[ \pi = [\pi_1, \ldots, \pi_K] \] topic prob.

\[ Pr(z^d = k) = \pi_k \]

\[ V = \frac{1}{D} \sum_{d=1}^{D} x_d \] avg. words in doc.

\[ \beta = \frac{1}{K} \sum_{k=1}^{K} \frac{V_k}{\beta_{kV}} \] avg. words in topic

\[ \beta_{kV} = \frac{1}{\beta_{V}} \] prob. of word \( V \) in topic \( k \)

\[ \beta = \frac{1}{K} \sum_{k=1}^{K} \beta_k V \] pmf
A Generative Model

- Documents: $x^1, \ldots, x^D$
- Associated topics: $z^1, \ldots, z^D$
- Parameters: $\theta = \{\pi, \beta\}$
- Generative model:
  
  Sample topic: $z^d \sim \pi$
  
  Sample words: $w^d_1 \sim \beta^{z^d}_d, i = 1, \ldots, N_d$
  
  Given topic $z^d = k$ for doc $d$, draw each word from $\beta_k$

Form of Likelihood

- Conditioned on topic...
  
  $p(x^d | z^d, \beta) = \prod_{i=1}^{N_d} p(w^d_i | z^d, \beta) = \prod_{i=1}^{N_d} \beta^{z^d}_d, w^d_i$

- Marginalizing latent topic assignment:
  
  $p(x^d | \beta, \pi) = \sum_{k=1}^K \pi_k p(x^d | z^d = k, \beta_k)$
Case Study 2: Document Retrieval

Review:
EM Algorithm

Learning Model Parameters

- Want to learn model parameters

Mixture of 3 Gaussians

Our actual observations

C. Bishop, Pattern Recognition & Machine Learning
ML Estimate of Mixture Model Params

- Log likelihood
  \[ L_x(\theta) \triangleq \log p\{\{x^i\}\mid \theta) = \sum_i \log \sum_{z^i} p(x^i, z^i \mid \theta) \]

- Want ML estimate
  \[ \hat{\theta}^{ML} = \arg \max_{\theta} L_x(\theta) \]

- Assume exponential family
  \[ p(x, z \mid \theta) = \frac{1}{Z(\theta)} e^{\theta^T \phi(x, z)} \]
  \[ L_x(\theta) = \sum_x \log \left( \sum_{z^i} e^{\theta^T \phi(x^i, z^i)} \right) - N \log Z(\theta) \]

- Neither convex nor concave and local optima

Complete Data

- Imagine we have an assignment of each \( x^i \) to a cluster

[Diagrams showing complete data labeled by true cluster assignments and our actual observations]

©Emily Fox 2015
Cluster Responsibilities

- We must infer the cluster assignments from the observations.

**Poserior probabilities of assignments to each cluster**

\[ r_{ik} = p(z^i = k \mid x^i, \pi, \phi) = \frac{\pi_k p(x^i \mid \phi_k)}{\sum_{j=1}^{K} \pi_j p(x^i \mid \phi_j)} \]

\[ = \frac{\text{e.g. } N(x^i \mid m_j, \Sigma_j)}{\sum_{k=1}^{K} \pi_k p(x^i \mid \phi_k)} \]

\[ \Rightarrow \pi_k = \frac{\sum_{i=1}^{N} r_{ik}}{N} \]

\[ \phi_k = \frac{1}{N} \sum_{i=1}^{N} r_{ik} x^i \]

Iterative Algorithm

- Motivates a coordinate ascent-like algorithm:
  1. Infer missing values \( z^i \) given estimate of parameters \( \hat{\Theta} \)
  2. Optimize parameters to produce new \( \hat{\Theta} \) given “filled in” data \( z^i \)
  3. Repeat

**Example: MoG (derivation soon... + HW)**

1. Infer "responsibilities"

\[ r_{ik}^{(t)} = p(z^i = k \mid x^i, \hat{\Theta}^{(t-1)}) = \frac{\pi_k^{(t-1)} p(x^i \mid \phi_k^{(t-1)})}{\sum_{j=1}^{K} \pi_j^{(t-1)} p(x^i \mid \phi_j^{(t-1)})} \]

2. Optimize parameters

\[ \text{max w.r.t. } \pi_k : \quad \pi_k^{(t)} = \frac{1}{N} \sum_{i=1}^{N} r_{ik}^{(t)} = \frac{c_k^{(t)}}{N} \leq \text{soft counts!} \]

\[ \text{max w.r.t. } \phi_k : \quad \mu_k^{(t)} = \frac{1}{N} \sum_{i=1}^{N} r_{ik}^{(t)} x^i \]

\[ \text{max w.r.t. } \Sigma_k^{(t)} : \quad \Sigma_k^{(t)} = \frac{1}{N} \sum_{i=1}^{N} r_{ik}^{(t)} (x^i - \mu_k^{(t)}) (x^i - \mu_k^{(t)})^T \]
Gaussian Mixture Example: Start

- Initialize \( \pi^{(0)} \) and \( \phi^{(0)} \)
- Compute \( r_{ik}^{(1)} \)

After first iteration

- Max like. given soft counts
- \( \pi^{(1)} \), \( \phi^{(1)} \)
- New \( r_{ik}^{(2)} \)
After 2nd iteration

After 3rd iteration
After 4th iteration

After 5th iteration
After 6th iteration

After 20th iteration
Expectation Maximization (EM) – Setup

- More broadly applicable than just to mixture models considered so far

- Model: \( x \) observable – “incomplete” data
  \( y \) not (fully) observable – “complete” data
  \( \theta \) parameters

- Interested in maximizing (wrt \( \theta \)):
  \[
p(x \mid \theta) = \sum_y p(x, y \mid \theta)
  \]

- Special case:
  \[x = g(y)\]

  e.g. \( y = [x] \triangleq \text{class labels} \) in standard mix model

EM Algorithm

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

- **E-Step**
  Compute 
  \[
  U(\theta, \hat{\theta}^{(t)}) = E[\log p(y \mid \theta) \mid x, \hat{\theta}^{(t)}] 
  \]

- **M-Step**
  Compute
  \[
  \hat{\theta}^{(t+1)} = \arg \max_{\theta} U(\theta, \hat{\theta}^{(t)}) 
  \]
  \[
  \Rightarrow L_x(\hat{\theta}^{(t+1)}) \geq L_x(\hat{\theta}^{(t)}) \]
  \[\text{mild assump} \Rightarrow \hat{\theta} \text{ converges to a local mode}\]
Example – Mixture Models

- **E-Step** Compute
  \[ U(\theta, \hat{\theta}(t)) = E[\log p(y \mid \theta) \mid x, \hat{\theta}(t)] \]

- **M-Step** Compute
  \[ \hat{\theta}(t+1) = \arg \max_{\theta} U(\theta, \hat{\theta}(t)) \]

Consider \( y^i = \{z^i, x^i\} \) i.i.d.

\[
p(x^i, z^i \mid \theta) = \pi_z p(x^i \mid \phi_z) = \sum_k \pi_k p(x^i \mid \phi_k)
\]

\[
E_{q_t}[\log p(y \mid \theta)] = \sum_i E_{q_t}[\log p(x^i, z^i \mid \theta)] = \sum_k \sum_i r_{ik} \log \pi_k + \sum_k \sum_i r_{ik} \log p(x^i \mid \phi_k)
\]

\[ \text{M-step: maximize w.r.t. } \pi_k, \phi_k \]

\[ \text{E-step: compute the } r_{ik} \text{ based on } \hat{\theta}(t) \]

Initialization

- In mixture model case where \( y^i = \{z^i, x^i\} \), there are many ways to initialize the EM algorithm

- Examples:
  - Choose \( K \) observations at random to define each cluster. Assign other observations to the nearest “centroid” to form initial parameter estimates
  - Pick the centers sequentially to provide good coverage of data
  - Grow mixture model by splitting (and sometimes removing) clusters until \( K \) clusters are formed

- Can be quite important to convergence rates in practice

©Emily Fox 2015
What you need to know

- Mixture model formulation
  - Generative model
  - Likelihood
- Expectation Maximization (EM) Algorithm
  - Derivation
  - Concept of non-decreasing log likelihood
  - Application to standard mixture models

Case Study 2: Document Retrieval

Review:
Connection to k-means
K-means

1. Ask user how many clusters they’d like. (e.g. \( k=5 \))
2. Randomly guess \( k \) cluster Center locations
3. Each datapoint finds out which Center it’s closest to.
4. Each Center finds the centroid of the points it owns

K-means

- Randomly initialize \( k \) centers
  \[ \mu^{(0)} = \mu_1^{(0)}, \ldots, \mu_k^{(0)} \]
- **Classify**: Assign each point \( j \in \{1,\ldots,N\} \) to nearest center:
  \[ z^j \leftarrow \arg \min_i ||\mu_i - x^j||^2_2 \]  
  **hard assign.**
- **Recenter**: \( \mu_i \) becomes centroid of its point:
  \[ \mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j:z^j=i} ||\mu - x^j||^2_2 \]  
  – Equivalent to \( \mu_i \leftarrow \text{average of its points!} \)
Special Case: Spherical Gaussians + hard assignments

\[ P(z' = k, x') = \frac{1}{(2\pi)^{d/2} \|\Sigma_k\|^{1/2}} \exp \left[ -\frac{1}{2} \left( x' - \mu_k \right)^T \Sigma_k^{-1} \left( x' - \mu_k \right) \right] P(z' = k) \]

- If \( P(x|z=k) \) is spherical, with same \( \sigma \) for all classes:
  \[ P(x' | z' = k) \propto \exp \left[ -\frac{1}{2\sigma^2} \| x' - \mu_k \|^2 \right] \]
- Then, compare EM objective with k-means:

\[ \text{EM: max } \prod_{i} \sum_{z} p(x_i | z^i | \theta) \quad \text{max } \prod_{i} \sum_{z} p(x_i | z^i | \theta) \]

\[ \text{max } \prod_{i} \sum_{z} p(x_i | z^i | \theta) \quad \text{max } \prod_{i} \sum_{z} p(x_i | z^i | \theta) \]

© Emily Fox 2015