Recitation 10/8

Mixture Models, PCA

Slides borrowed from Prof. Seyoung Kim, Ryan Tibshirani. Thanks!
Gaussian Mixture Models (GMMs)

- Consider a mixture of $K$ Gaussian components:

$$p(x_n) = \sum_k p(x_n \mid z_n = k)p(z_n = k)$$

$$= \sum_k N(x_n \mid \mu_k, \Sigma_k)\pi_k$$

Law of Total Probability

- $z_n$ is the latent variable indicating the mixture component.
- $x_n$ is the observed variable.
- $\mu_k$ and $\Sigma_k$ are the mean and covariance of the $k$th Gaussian component.
- $\pi_k$ is the mixture proportion.
Since $z$ uses a 1-of-$K$ representation, we have

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}. \quad (9.10)$$

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}. \quad (9.11)$$

$$p(x) = \sum_z p(z)p(x|z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad (9.12)$$
MLE for GMM with fully observed data

- If we are doing MLE for completely observed data

- Data log-likelihood

\[
l(\theta; D) = \log \prod_n p(z_n, x_n) = \log \prod_n p(z_n | \pi) p(x_n | z_n, \mu, \sigma)
\]

\[
= \sum_n \log \prod_k \pi_k^{z_{nk}} + \sum_n \log \prod_k N(x_n; \mu_k, \sigma)^{z_{nk}}
\]

\[
= \sum_n \sum_k z_{nk} \log \pi_k - \sum_n \sum_k z_{nk} \frac{1}{2\sigma^2} (x_n - \mu_k)^2 + C
\]

- MLE

\[
\hat{\pi}_{k, MLE} = \arg \max_{\pi} l(\theta; D),
\]

\[
\hat{\mu}_{k, MLE} = \arg \max_{\mu} l(\theta; D)
\]

\[
\hat{\sigma}_{k, MLE} = \arg \max_{\sigma} l(\theta; D)
\]

- What if we do not know \(z_n\)?
What if we do not know $z_n$?

- Maximize the expected data log likelihood for $(x_i, z_i)$ based on $p(x_i, z_i)$
  - Expectation-Maximization (EM) algorithm
Complete vs. Expected Complete Log Likelihoods

- The complete log likelihood:

\[
I(\Theta; D) = \log \prod_n p(z_n, x_n) = \log \prod_n p(z_n | \pi)p(x_n | z_n, \mu, \sigma) \\
= \sum_n \log \prod_k \pi_k^{z_n^k} + \sum_n \log \prod_k N(x_n; \mu_k, \sigma)^{z_n^k} \\
= \sum_n \sum_k z_n^k \log \pi_k - \sum_n \sum_k z_n^k \frac{1}{2\sigma^2} (x_n - \mu_k)^2 + C
\]

- The expected complete log likelihood

\[
\langle l_c (\Theta; z, x) \rangle = \sum_n \langle \log p(z_n | \pi) \rangle_{p(z|x)} + \sum_n \langle \log p(x_n | z_n, \mu, \Sigma) \rangle_{p(z|x)} \\
= \sum_n \sum_k \langle z_n^k \rangle \log \pi_k - \frac{1}{2} \sum_n \sum_k \langle z_n^k \rangle (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) + \log |\Sigma_k| + C
\]

- EM optimizes the expected complete log likelihood
The Expectation-Maximization (EM) Algorithm

E step:

$$\tau_n^{k(i)} = \langle z_n^k \rangle_{q(i)} = p(z_n^k = 1 \mid x, \mu^{(i)}, \Sigma^{(i)})$$

M step:

$$\pi_k^{(i+1)} = \frac{\sum_n \tau_n^{k(i)}}{N} = \frac{\langle n_k \rangle}{N}$$

$$\mu_k^{(i+1)} = \frac{\sum_n \tau_n^{k(i)} x_n}{\sum_n \tau_n^{k(i)}}$$

$$\Sigma_k^{(i+1)} = \frac{\sum_n \tau_n^{k(i)} (x_n - \mu_k^{(i+1)})(x_n - \mu_k^{(i+1)})^T}{\sum_n \tau_n^{k(i)}}$$
Example 2-d data points coming from $K = 2$ Gaussian distributions

K=2 1-d Gaussian distributions:

$$G_1(\mu_1, \sigma_1^2), G_2(\mu_2, \sigma_2^2)$$

<x, y> pairs

$$x \in \mathcal{R}, y \in \{G_1, G_2\}$$

$$x = (2, 4, 7)$$
Example 2-d data points coming from $K = 2$ Gaussian distributions

K=2 1-d Gaussian distributions:

$G_1(\mu_1, \sigma_1^2), G_2(\mu_2, \sigma_2^2)$

$x, y$ pairs

$x \in \mathcal{R}, y \in \{G_1, G_2\}$

$x = (2, 4, 7)$

Initialize

$\mu^{(0)} = (3, 6)$

$\pi^{(0)} = \left(\frac{1}{2}, \frac{1}{2}\right)$

$\sigma^{2(0)} = \left(\frac{1}{2}, \frac{1}{2}\right)$
Example 2-d data points coming from $K = 2$ Gaussian distributions

$x = (2, 4, 7)$

Initialize

$\mu^{(0)} = (3, 6)$

$\pi^{(0)} = \left( \frac{1}{2}, \frac{1}{2} \right)$

$\sigma^{(0)}_2 = \left( \frac{1}{2}, \frac{1}{2} \right)$

iteration $t = 1$

$\tau_1 = p(z_1 = 1 | x_1) = \frac{p(x_1 | \mu_1)p(\mu_1)}{p(x_1 | \mu_1)p(\mu_1) + p(x_1 | \mu_2)p(\mu_2)} = \frac{\frac{1}{2}N(2, 3, \frac{1}{\sqrt{2}})}{\frac{1}{2}N(2, 3, \frac{1}{\sqrt{2}}) + \frac{1}{2}N(2, 6, \frac{1}{\sqrt{2}})} = 1 - 10^{-7}$
Example 2-d data points coming from $K = 2$ Gaussian distributions

$$x = (2, 4, 7)$$

Initialize

$$\mu^{(0)} = (3, 6)$$

$$\pi^{(0)} = \left(\frac{1}{2}, \frac{1}{2}\right)$$

$$\sigma^2(0) = \left(\frac{1}{2}, \frac{1}{2}\right)$$

iteration $t = 1$

$$\tau^1_i = p(x_i = 1|x_1) = \frac{p(x_1|\mu_1)p(\mu_1)}{p(x_1|\mu_1)p(\mu_1) + p(x_1|\mu_2)p(\mu_2)} = \frac{\frac{1}{2}N(2, 3, \frac{1}{\sqrt{2}})}{\frac{1}{2}N(2, 3, \frac{1}{\sqrt{2}}) + \frac{1}{2}N(2, 6, \frac{1}{\sqrt{2}})} = 1 - 10^{-7}$$

$$\tau^2_i = 10^{-7}$$

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>2</th>
<th>4</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau^1_i$</td>
<td>$1 - 10^{-7}$</td>
<td>0.953</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>$\tau^2_i$</td>
<td>$10^{-7}$</td>
<td>0.047</td>
<td>$1 - 10^{-7}$</td>
</tr>
</tbody>
</table>
Example 2-d data points coming from $K = 2$ Gaussian distributions

$x = (2, 4, 7)$

Initialize

$\mu^{(0)} = (3, 6)$

$\pi^{(0)} = \left(\frac{1}{2}, \frac{1}{2}\right)$

$\sigma^{2(0)} = \left(\frac{1}{2}, \frac{1}{2}\right)$

iteration $t = 1$

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>2</th>
<th>4</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_i^1$</td>
<td>$1 - 10^{-7}$</td>
<td>0.953</td>
<td>$10^{-7}$</td>
</tr>
<tr>
<td>$\pi_i^2$</td>
<td>$10^{-7}$</td>
<td>0.047</td>
<td>$1 - 10^{-7}$</td>
</tr>
</tbody>
</table>

$\pi_1 = \frac{1.953}{3} = 0.651$, $\pi_2 = 0.349$

$\mu_1 \approx \frac{2 + 0.953 \times 4 + 0}{1.953} = 2.978$

$\mu_2 \approx 6.88$

$\sigma^2 \approx ...$
PCA

Principal components are a sequence of projections of the data, mutually uncorrelated and ordered in variance.
Principal Component Analysis (PCA)

Principal Components (PC) are orthogonal directions that capture most of the variance in the data

1\textsuperscript{st} PC – direction of greatest variability in data

2\textsuperscript{nd} PC – Next orthogonal (uncorrelated) direction of greatest variability

(removing all variability in first direction, then find next direction of greatest variability)

And so on …
Assume $X$ is a normalized $N \times p$ data matrix for $N$ samples and $p$ features.

Assume data is normalized. $\iff$ each column of $X$ is normalized.

Variance of projected data

$$\frac{1}{N} \sum_{n=1}^{N} (v^T x_n - v^T \bar{x}_n)^2 = v^T S v$$

$\iff$ Want to maximize this over $v$

where $S = \frac{1}{n} \sum_{i} (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T = \frac{1}{n} \sum_{i} x_i x_i^T$
Computing the Components

• Projection of vector $\mathbf{x}$ onto an axis (dimension) $\mathbf{u}$ is $\mathbf{u}^T \mathbf{x}$
• Assume $\mathbf{X}$ is a normalized $n \times p$ data matrix for $n$ samples and $p$ features. Direction of greatest variability is that in which the average square of the projection is greatest:

$$\text{Maximize} \quad \frac{1}{n} \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u}$$
$$\text{s.t.} \quad \mathbf{u}^T \mathbf{u} = 1$$

Construct Langrangian $$(1/n) \mathbf{u}^T \mathbf{X}^T \mathbf{X} \mathbf{u} - \lambda \mathbf{u}^T \mathbf{u}$$

Vector of partial derivatives set to zero

$$\frac{1}{n} \mathbf{X}^T \mathbf{X} \mathbf{u} - \lambda \mathbf{u} = 0$$

or equivalently $\mathbf{S} \mathbf{u} - \lambda \mathbf{u} = 0$ ($\mathbf{S} = 1/n \mathbf{X}^T \mathbf{X}$: covariance matrix)

As $\mathbf{u} \neq \mathbf{0}$ then $\mathbf{u}$ must be an eigenvector of $\mathbf{S}$ with eigenvalue $\lambda$

– $\lambda$ is the principal eigenvalue of the covariance matrix $\mathbf{S}$
– The eigenvalue denotes the amount of variability captured along that dimension
The proportion of variance explained is a nice way to quantify how much structure is being captured.