CS195-5 : Introduction to Machine Learning
Lecture 20

Greg Shakhnarovich

November 1, 2006
Announcements
• Locally-weighted regression:
  Fit a parametric model to neighbors, weighted by a kernel.

• Parametric mixture models:

\[ p(x; \mathbf{p}) = \sum_{c=1}^{k} p(y = c) p(x | y = c). \]
Generative model for a Gaussian mixture

\[ p(x; \theta, p) = \sum_{c=1}^{k} p_c \cdot \mathcal{N}(x; \mu_c, \Sigma_c), \]

- The graphical model

- The plate notation is a shorthand for
Plan for today

- The Expectation Maximization (EM) algorithm.
Likelihood of a mixture model

- The log-likelihood of $p, \theta$:

$$
\log p(X_N; p, \theta) = \sum_{i=1}^{N} \log \sum_{c=1}^{k} p_c \mathcal{N}(x_i; \mu_c, \Sigma_c).
$$

- No closed-form solution because of the sum inside log.
  - Since we need to take into account all possible components that could have generated $x_i$. 
Mixture density estimation

- Suppose that we do observe $y_i \in \{1, 2\}$ for each $i = 1, \ldots, N$.

- Let us introduce a set of binary indicator variables $z_i = [z_{i1}, \ldots, z_{ik}]$ where

$$z_{ic} = 1 = \begin{cases} 1 & \text{if } y_i = c, \\ 0 & \text{otherwise}. \end{cases}$$

- The count of examples from $c$-th component:

$$N_c = \sum_{i=1}^{N} z_{ic}.$$
Mixture density estimation: known labels

- If we know \( z_i \), the ML estimates of the Gaussian components, just like in class-conditional model, are

\[
\hat{p}_c = \frac{N_c}{N},
\]

\[
\hat{\mu}_c = \frac{1}{N_c} \sum_{i=1}^{N} z_{ic} x_i,
\]

\[
\hat{\Sigma}_c = \frac{1}{N_c} \sum_{i=1}^{N} z_{ic} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T.
\]
Credit assignment

• When we don’t know \( z \), we face a credit assignment problem: which component is responsible for \( x_i \)?

• Suppose we do know component parameters \( \theta = [\mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k] \) and mixing probabilities \( p = [p_1, \ldots, p_k] \).

• The posterior of indicators using Bayes’ theorem:

\[
\gamma_{ic} \triangleq p(z_{ic} | x; \theta, p) = \frac{p_c \cdot p(x; \mu_c, \Sigma_c)}{\sum_{l=1}^{k} p_l \cdot p(x; \mu_l, \Sigma_l)}
\]

• We will call \( \gamma_{ic} \) the responsibility of the \( c \)-th component for \( x \).
  – Note: \( \sum_{c=1}^{k} \gamma_{ic} = 1 \).
Expected likelihood

- The “complete data” likelihood (when \( z \) are known):

\[
p(X_N, Z_N; \ p, \theta) \propto \prod_{i=1}^N \prod_{c=1}^k (p_c \mathcal{N}(x_i; \mu_c, \Sigma_c))^{z_{ic}}.
\]

and the log:

\[
\log p(X_N, Z_N; \ p, \theta) = \text{const} + \sum_{i=1}^N \sum_{c=1}^k z_{ic} (\log p_c + \log \mathcal{N}(x_i; \mu_c, \Sigma_c)).
\]

- We can’t compute it, but can take the expectation w.r.t. the posterior of \( z \):

\[
E_{z_{ic} \sim p(\cdot|X_N, \theta, p)} [\log p(X_N, Z_N; \ p, \theta)].
\]
Expected likelihood

- Expectation of $z_{ic}$:

  $$E_{z_{ic} \sim p(\cdot | X_N, \theta, p)}[z_{ic}] = \sum_{z \in \{0, 1\}} z p(z_{ic} = z \mid x_i; \theta, p) = \gamma_{ic}.$$ 

- The expected likelihood of the data:

  $$E_{z_{ic} | x_i, \theta, p}[\log p(X_N, Z_N; p, \theta)] = \text{const}$$

  $$+ \sum_{i=1}^{N} \sum_{c=1}^{k} \gamma_{ic} \left( \log p_c + \log \mathcal{N}(x_i; \mu_c, \Sigma_c) \right).$$
Expectation maximization

\[ E_{z_{ic}|x_i, \theta, \mathbf{p}}[\log p(X_N, Z_N; \mathbf{p}, \theta)] = \sum_{i=1}^{N} \sum_{c=1}^{k} \gamma_{ic} (\log p_c + \log \mathcal{N}(x_i; \mu_c, \Sigma_c)). \]

- We can find \( p, \theta \) that maximize this expected likelihood – by setting derivatives to zero and, for \( p \), using Lagrange multipliers to enforce \( \sum_c p_c = 1 \).

\[
\hat{p}_c = \frac{\sum_{i=1}^{N} \gamma_{ic}}{N},
\]

\[
\hat{\mu}_c = \frac{1}{\sum_{i=1}^{N} \gamma_{ic}} \sum_{i=1}^{N} \gamma_{ic} x_i,
\]

\[
\hat{\Sigma}_c = \frac{1}{\sum_{i=1}^{N} \gamma_{ic}} \sum_{i=1}^{N} \gamma_{ic} (x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T.
\]
Summary so far

- If we know the parameters and indicators (assignments) we are done.
- If we know the indicators but not the parameters, we can do ML estimation of the parameters – and we are done.
- If we know the parameters but not the indicators, we can compute the posteriors of indicators;
  - With known posteriors, we can estimate parameters that maximize the expected likelihood – and then we are done.
- But in reality we know neither the parameters nor the indicators.
The EM algorithm

- Start with a guess of $\theta, p$.
  - Typically, random $\theta$ and $p_c = 1/k$.

- Iterate between:
  
  **E-step** Compute values of expected assignments, i.e. calculate $\gamma_{ic}$, using current estimates of $\theta, p$.
  
  **M-step** Maximize the expected likelihood, under current $\gamma_{ic}$.
EM for Gaussian mixture: an example

- Colors represent $\gamma_{ic}$ after the E-step.

1st iteration
EM for Gaussian mixture: an example

- Colors represent $\gamma_{ic}$ after the E-step.
EM for Gaussian mixture: an example

- Colors represent $\gamma_{ic}$ after the E-step.

1st iteration

2nd iteration

3rd iteration
EM for Gaussian mixture: an example

4th iteration

7th iteration

9th iteration
EM for Gaussian mixture: an example

4th iteration

7th iteration

9th iteration
EM for Gaussian mixture: an example

4th iteration

7th iteration

9th iteration
The EM for Gaussian mixtures- summary

- **Initialize:** random $\mu_{c}^{old}$, $\Sigma_{c}^{old}$, $p_{c}^{old} = 1/k$ for $c = 1, \ldots, k$.

- **Iterate until convergence:**
  
  **E-step** estimate responsibilities:
  \[
  \gamma_{ic} = \frac{p_{c}^{old} \mathcal{N}(x_i; \mu_{c}^{old}, \Sigma_{c}^{old})}{\sum_{l=1}^{k} p_{l}^{old} \mathcal{N}(x_i; \mu_{l}^{old}, \Sigma_{l}^{old})}
  \]

  **M-step** re-estimate mixture parameters:
  \[
  \hat{p}_{c}^{new} = \frac{\sum_{i=1}^{N} \gamma_{ic}}{N},
  \]
  \[
  \hat{\mu}_{c}^{new} = \frac{1}{\sum_{i=1}^{N} \gamma_{ic}} \sum_{i=1}^{N} \gamma_{ic} x_i,
  \]
  \[
  \hat{\Sigma}_{c}^{new} = \frac{1}{\sum_{i=1}^{N} \gamma_{ic}} \sum_{i=1}^{N} \gamma_{ic} (x_i - \hat{\mu}_{c}^{new})(x_i - \hat{\mu}_{c}^{new})^T.
  \]
Next time

More on the EM algorithm.
Model selection.