Lecture 19: More EM

Machine Learning
April 15, 2010
Last Time

• Expectation Maximization
• Gaussian Mixture Models
Today

• EM Proof
  – Jensen’s Inequality

• Clustering sequential data
  – EM over HMMs
  – EM in any Graphical Model
    • Gibbs Sampling
Gaussian Mixture Models
How can we be sure GMM/EM works?

• We’ve already seen that there are multiple clustering solutions for the same data.
  – Non-convex optimization problem

• Can we prove that we’re approaching some maximum, even if many exist.
Bound maximization

• Since we can’t optimize the GMM parameters directly, maybe we can find the maximum of a lower bound.

• Technically: optimize a convex lower bound of the initial non-convex function.
EM as a bound maximization problem

- Need to define a function $Q(x, \Theta)$ such that
  - $Q(x, \Theta) \leq l(x, \Theta)$ for all $x, \Theta$
  - $Q(x, \Theta) = l(x, \Theta)$ at a single point
  - $Q(x, \Theta)$ is concave

\[
l(\theta) \geq Q_t(\theta) \\
l(\theta_t) \geq Q_t(\theta_t) \\
Q_t(\theta_{t+1}) > Q_t(\theta_t) \\
l(\theta_{t+1}) \geq Q_t(\theta_{t+1}) \\
l(\theta_{t+1}) > l(\theta_t)
\]
EM as bound maximization

• Claim:
  – for GMM likelihood
    \[ l(\theta) = \sum_n \log \sum_z p(x_n, z|\theta) \]
  – The GMM MLE estimate is a convex lower bound
    \[ Q(\theta) = \arg\max_{\theta} \sum_n \sum_z \tau^z_n \log p(x_n, z|\theta) \]
EM Correctness Proof

• Prove that \( l(x, \Theta) \geq Q(x, \Theta) \)

\[
l(\theta) = \sum_n \log p(x_n | \theta) \quad \text{Likelihood function}
\]

\[
= \sum_n \log \sum_z p(x_n, z | \theta) \quad \text{Introduce hidden variable (mixtures in GMM)}
\]

\[
= \sum_n \log \sum_z p(x_n, z | \theta) \frac{p(z | x_n, \theta_t)}{p(z | x_n, \theta_t)}
\]

\[
= \sum_n \log \sum_z p(z | x_n, \theta_t) \frac{p(x_n, z | \theta)}{p(z | x_n, \theta)}
\]

\[
\geq \sum_n \sum_z p(z | x_n, \theta_t) \log \frac{p(x_n, z | \theta)}{p(z | x_n, \theta_t)} \quad \text{Jensen’s Inequality (coming soon...)}
\]

\[
\geq \sum_n \sum_z p(z | x_n, \theta_t) \log p(x_n, z | \theta) - \sum_z p(z | x_n, \theta_t) \log p(z | x_n, \theta_t)
\]
EM Correctness Proof

\[ Q(\theta) = \arg\max_{\theta} \sum_{n} \sum_{z} \tau_n^z \log p(x_n, z|\theta) \]

\[ l(\theta) \geq \sum_{n} \sum_{z} p(z|x_n, \theta_t) \log p(x_n, z|\theta) - \sum_{z} p(z|x_n, \theta_t) \log p(z|x_n, \theta_t) \]

\[ \geq Q(\theta|\theta_t) - \text{const} \]

\[ \theta_{t+1} = \arg\max_{\theta} Q(\theta|\theta_t) \]

\[ = \arg\max_{\theta} \sum_{n} \sum_{z} p(z|x_n, \theta_t) \log p(x_n, z|\theta) \]

\[ = \arg\max_{\theta} \sum_{n} \sum_{z} \tau_n^z \log p(x_n, z|\theta) \quad \text{GMM Maximum Likelihood Estimation} \]
The missing link: Jensen’s Inequality

• If $f$ is concave (or convex down): $f(\mathbb{E}\{x\}) \geq \mathbb{E}\{f(x)\}$

• Incredibly important tool for dealing with mixture models.

$$f \left( \sum_i \pi_i p(x|\mu_i, \Sigma_i) \right) \geq \sum_i \pi_i f (p(x|\mu_i, \Sigma_i))$$

if $f(x) = \log(x)$

$$\log \left( \sum_i \pi_i p(x|\mu_i, \Sigma_i) \right) \geq \sum_i \pi_i \log (p(x|\mu_i, \Sigma_i))$$

$$\log(\pi x_1 + (1 - \pi)x_2) \geq \pi \log(x_1) + (1 - \pi) \log(x_2)$$
Generalizing EM from GMM

• Notice, the EM optimization proof never introduced the exact form of the GMM
• Only the introduction of a hidden variable, \( z \).
• Thus, we can generalize the form of EM to broader types of latent variable models
General form of EM

• Given a joint distribution over observed and latent variables: \( p(X, Z|\theta) \)
• Want to maximize: \( p(X|\theta) \)

1. Initialize parameters \( \theta^{old} \)
2. E Step: Evaluate:
   \[
p(Z|X, \theta^{old})
   \]
3. M-Step: Re-estimate parameters (based on expectation of complete-data log likelihood)
   \[
   \theta^{new} = \arg\max_{\theta} \sum_{Z} p(Z|X, \theta^{old}) \ln p(X, Z|\theta)
   \]
4. Check for convergence of params or likelihood
Applying EM to Graphical Models

- Now we have a general form for learning parameters for latent variables.
  - Take a Guess
  - Expectation: Evaluate likelihood
  - Maximization: Reestimate parameters
  - Check for convergence
Clustering over sequential data

- Recall HMMs

- We only looked at training supervised HMMs.
- What if you believe the data is sequential, but you can’t observe the state.
EM on HMMs

- also known as Baum-Welch

\[\pi_i = q_0^i \quad \alpha_{ij} = \frac{\sum_{t=0}^{T-2} q_t^i q_t^{j+1}}{\sum_{k=0}^{M-1} \sum_{t=0}^{T-1} q_t^i q_t^{k+1}} \quad \eta_{ij} = \frac{\sum_{t=0}^{T-1} q_t^i x_t^j}{\sum_{k=0}^{N-1} \sum_{t=0}^{T-1} q_t^i x_t^k}\]

- Recall HMM parameters:

\[\pi_i = \mathbb{E}\left\{q_0^i\right\} \quad \alpha_{ij} = \frac{\sum_{t=0}^{T-2} \mathbb{E}\left\{q_t^i q_t^{j+1}\right\}}{\sum_{k=0}^{M-1} \sum_{t=0}^{T-1} \mathbb{E}\left\{q_t^i q_t^{k+1}\right\}} \quad \eta_{ij} = \frac{\sum_{t=0}^{T-1} \mathbb{E}\left\{q_t^i x_t^j\right\}}{\sum_{k=0}^{N-1} \sum_{t=0}^{T-1} \mathbb{E}\left\{q_t^i x_t^k\right\}}\]

- Now the training counts are estimated.
EM on HMMs

• Standard EM Algorithm
  – Initialize
  – E-Step: evaluate expected likelihood
  – M-Step: reestimate parameters from expected likelihood
  – Check for convergence
EM on HMMs

- **Guess:** Initialize parameters, $\theta = [\pi, \alpha, \eta]^T$

- **E-Step:** Compute $\mathbb{E}\{l(\theta)\} = \mathbb{E}\{\log p(x, q|\theta)\}$

  $\mathbb{E}\{\log p(x, q|\theta)\} = \mathbb{E}\left\{ \log \left( p(q_0) \prod_{t=1}^{T-1} p(q_t|q_{t-1}) \prod_{t=0}^{T-1} p(x_t|q_t) \right) \right\}$

  $= \mathbb{E}\left\{ \sum_{i=0}^{M-1} q_i^0 \log \pi_i + \sum_{t=1}^{T-1} \sum_{i,j=0}^{M-1} q_i^t q_{t-1}^j \log \alpha_{ij} + \sum_{t=0}^{T-1} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} q_i^t x_t^j \log \eta_{ij} \right\}$

  $= \sum_{i=0}^{M-1} \mathbb{E}\{q_0^i\} \log \pi_i + \sum_{t=1}^{T-1} \sum_{i,j=0}^{M-1} \mathbb{E}\{q_t^i q_{t-1}^j\} \log \alpha_{ij} + \sum_{t=0}^{T-1} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \mathbb{E}\{q_t^i x_t^j\} \log \eta_{ij}$
EM on HMMs

• But what are these $E\{\ldots\}$ quantities?

$$\begin{align*}
\mathbb{E}\{q_0^i\} \log \pi_i + \sum_{t=1}^{T-1} \sum_{i,j=0}^{M-1} \mathbb{E}\{q_t^i q_{t-1}^j\} \log \alpha_{ij} + \sum_{t=0}^{T-1} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \mathbb{E}\{q_t^i x_t^j\} \log \eta_{ij}
\end{align*}$$

$$\begin{align*}
\mathbb{E}\{x^i\} &= \sum_x p(x)x^i = \sum_x p(x)\delta(x = x^i) = p(x^i) \\
\text{so...} \quad \mathbb{E}\{q_0^i\} &= p(q_0^i | \bar{x}) \quad \mathbb{E}\{q_t^i q_{t-1}^j\} = p(q_t^i, q_{t-1}^j | \bar{x}) \quad \mathbb{E}\{q_t^i\} = p(q_t^i | \bar{x})
\end{align*}$$

These can be efficiently calculated from JTA potentials and separators.
EM on HMMs

\[ p(q_i, q_{i-1} | \bar{x}_0 \ldots \bar{x}_n) \]
\[ p(q_0 | \bar{x}_0 \ldots \bar{x}_n) \quad p(q_i | \bar{x}_0 \ldots \bar{x}_n) \]
\[ \psi(q_0, \bar{x}_0) \quad \phi(q_0) \quad \psi(q_{i-1}, q_i) \quad \phi(q_i) \quad \psi(q_i, q_{i+1}) \]
\[ \zeta(q_i) \quad \zeta(q_{i+1}) \quad p(q_{i+1} | \bar{x}_0 \ldots \bar{x}_n) \]
\[ \psi(q_i, \bar{x}_i) \quad \psi(q_{i+1}, x_{i+1}) \]
EM on HMMs

• Standard EM Algorithm
  – Initialize
  – E-Step: evaluate expected likelihood
    • JTA algorithm.
  – M-Step: reestimate parameters from expected likelihood
    • Using expected values from JTA potentials and separators

\[
\pi_i = \mathbb{E}\{q_0^i\} \quad \alpha_{ij} = \frac{\sum_{t=0}^{T-2} \mathbb{E}\{q_t^i q_{t+1}^j\}}{\sum_{k=0}^{M-1} \sum_{t=0}^{T-1} \mathbb{E}\{q_t^i q_t^k\}} \quad \eta_{ij} = \frac{\sum_{t=0}^{T-1} \mathbb{E}\{q_t^i x_t^j\}}{\sum_{k=0}^{N-1} \sum_{t=0}^{T-1} \mathbb{E}\{q_t^i x_t^k\}}
\]

– Check for convergence
Training latent variables in Graphical Models

• Now consider a general Graphical Model with latent variables.
EM on Latent Variable Models

• Guess
  – Easy, just assign random values to parameters

• E-Step: Evaluate likelihood.
  – We can use JTA to evaluate the likelihood.
  – And marginalize expected parameter values

• M-Step: Re-estimate parameters.
  – This can get trickier.
Maximization Step in Latent Variable Models

• Why is this easy in HMMs, but difficult in general Latent Variable Models?

• Many parents graphical model
Junction Trees

• In general, we have no guarantee that we can isolate a single variable.

• We need to estimate marginal separately.

• “Dense Graphs”
M-Step in Latent Variable Models

- M-Step: Reestimate Parameters.
  - Keep $k-1$ parameters fixed (to the current estimate)
  - Identify a better guess for the free parameter.
M-Step in Latent Variable Models

• M-Step: Reestimate Parameters.
  – Keep k-1 parameters fixed (to the current estimate)
  – Identify a better guess for the free parameter.
M-Step in Latent Variable Models

• M-Step: Reestimate Parameters.
  – Keep k-1 parameters fixed (to the current estimate)
  – Identify a better guess for the free parameter.
M-Step in Latent Variable Models

• M-Step: Reestimate Parameters.
  – Keep k-1 parameters fixed (to the current estimate)
  – Identify a better guess for the free parameter.
M-Step in Latent Variable Models

• M-Step: Reestimate Parameters.
  – Keep k-1 parameters fixed (to the current estimate)
  – Identify a better guess for the free parameter.
M-Step in Latent Variable Models

• M-Step: Reestimate Parameters.
  – Gibbs Sampling.
  – This is helpful if it’s easier to sample from a conditional than it is to integrate to get the marginal.

\[
\begin{align*}
x^{(t+1)} & \sim p(x|y^{(t)}) \\
y^{(t+1)} & \sim p(y|x^{(t+1)})
\end{align*}
\]
EM on Latent Variable Models

• Guess
  – Easy, just assign random values to parameters

• E-Step: Evaluate likelihood.
  – We can use JTA to evaluate the likelihood.
  – And marginalize expected parameter values

• M-Step: Re-estimate parameters.
  – Either JTA potentials and marginals OR
  – Sampling
Today

• EM as bound maximization
• EM as a general approach to learning parameters for latent variables
• Sampling
Next Time

• Model Adaptation
  – Using labeled and unlabeled data to improve performance.

• Model Adaptation Application
  – Speaker Recognition
    • UBM-MAP