CSC 4020 Fundamentals of Machine Learning:

Expectation Maximization

Baoyuan Wu

Apirl 26/28

[Slide credit: Mengye Ren and Matthew MacKay]

A Generative View of Clustering

- Last time: introduced EM algorithm as a way of fitting a Gaussian Mixture Model
  - E-step: Compute probability each datapoint came from certain cluster, given model parameters
  - M-step: Adjust parameters of each cluster to maximize probability it would generate data it is currently responsible for
- This lecture: derive EM from principled approach and see how EM can be applied to general latent variable models

Latent Variable Models

- Recall: variables which are always unobserved are called **latent variables** or sometimes hidden variables
- In a mixture model, the identity of the component that generated a given datapoint is a latent variable
- Why use latent variables if introducing them complicates learning?
  - We can build a complex model out of simple parts - this can simplify the description of the model
  - We can sometimes use the latent variables as a representation of the original data (e.g. cluster assignments in a GMM model)

Preliminaries: Jensen's Inequality

- **Theorem:** Suppose f is a convex function and X is a random variable. Then:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

- If X takes on two values $\mathbf{x}_1$ and $\mathbf{x}_2$ with probabilities p_1 and p_2, just the defintion of a convex function:

$$f(p_1\mathbf{x}_1 + p_2\mathbf{x}_2) \leq p_1 f(\mathbf{x}_1) + p_2 f(\mathbf{x}_2)$$

  - This is a convenient way to remember which way the inequality goes

Preliminaries: Jensen's Inequality

Jensen's Inequality: For convex f:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

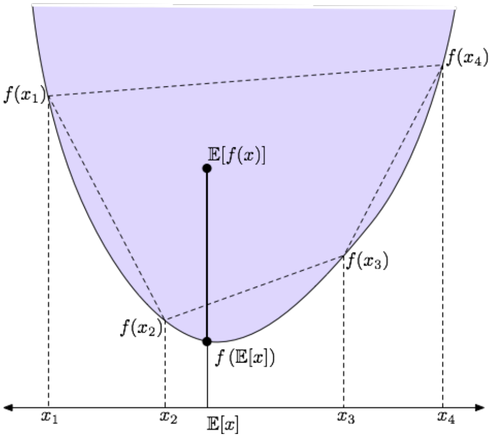


Image credit: Mark Reid

Preliminaries: Jensen's Inequality

Jensen's Inequality: For convex f:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

- Sufficient condition for equality: if X is a constant (i.e. the random variable takes on one value)
- If g is **concave**, the inequality changes directions:

$$g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)]$$

Preliminaries: Notation

- In this lecture, we'll be using $\mathbf{x}$ to denote **observed data** and z to denote **the latent variables**
- We'll let $p(z, \mathbf{x}; \boldsymbol{\theta})$ denote the probabilistic model we've defined
  - Anything following a semicolon denotes a parameter of the distribution
  - We're not treating the parameters as random variables
- We assume we have an observed dataset $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ and would like to fit $\boldsymbol{\theta}$ using maximum likelihood:

$$\log p(\mathcal{D}; \boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}^{(n)}; \boldsymbol{\theta})$$

- To compute $p(\mathbf{x}; \boldsymbol{\theta})$, we have to **marginalize** over z:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{z} p(z, \mathbf{x}; \boldsymbol{\theta})$$

EM

- Typically no closed form solution to the maximum likelihood problem

$$\log p(\mathcal{D};\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}^{(n)};\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \left(\sum_{z^{(n)}} p(z^{(n)}, \mathbf{x}^{(n)};\boldsymbol{\theta}) \right)$$

- Key difficulty: once z is marginalized out, $p(\mathbf{x};\boldsymbol{\theta})$ could be complex (e.g. a mixture distribution)
- We'd like to write an objective in terms of $\log p(z, \mathbf{x};\boldsymbol{\theta})$, which should be simpler to solve
- To accomplish this, we need to move the summation outside the log
- We introduce auxilliary distributions $q_n(z^{(n)})$ over each of the latent variables

EM

$$
\begin{aligned}
\sum_{n=1}^{N} \log\left(\sum_{z^{(n)}} p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})\right) &= \sum_{n=1}^{N} \log\left(\sum_{z^{(n)}} q_n(z^{(n)}) \frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right) \\
&= \sum_{n=1}^{N} \log\left(\mathbb{E}_{q_n(z^{(n)})}\left[\frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right]\right) \\
&\geq \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\log \frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right]
\end{aligned}
$$

- In the last step, we use Jensen's Inequality. Since log is concave:

$$
\log\left(\mathbb{E}_{q_n(z^{(n)})}\left[\frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right]\right) \geq \mathbb{E}_{q_n(z^{(n)})}\left[\log \frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right]
$$

EM

$$\sum_{n=1}^{N} \log p(\mathbf{x}^{(n)}; \boldsymbol{\theta}) \geq \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[\log \frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})} \right]$$
$$\equiv \mathcal{L}(q, \boldsymbol{\theta}) \text{ where } q = \{q_1, \dots, q_N\}$$

- We expect $\mathcal{L}(q, \boldsymbol{\theta})$ might be easier to optimize w.r.t. $\boldsymbol{\theta}$, since it only appears in $\log p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})$, so we'll use this as our new objective
- For **any** auxilliary distributions q_n, we obtain a lower bound on the log likelihood
- Which q_n should we choose? Want to make the bound as tight as possible

EM

- We know this bound is tight (i.e. the inequality becomes an equality) if there are constants c_n such that:

$$\frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})} = \text{constant} \implies q_n(z^{(n)}) = c_n p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})$$

- Using $\sum_{z^{(n)}} q_n(z^{(n)}) = 1$, we have:

$$1 = \sum_{z^{(n)}} q_n(z^{(n)}) = c_n \sum_{z^{(n)}} p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta}) = c_n p(\mathbf{x}^{(n)}; \boldsymbol{\theta})$$

$$\implies c_n = \frac{1}{p(\mathbf{x}^{(n)}; \boldsymbol{\theta})}$$

- Hence:

$$q_n(z^{(n)}) = \frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{p(\mathbf{x}^{(n)}; \boldsymbol{\theta})} = p(z^{(n)} | \mathbf{x}^{(n)}; \boldsymbol{\theta})$$

EM

- For fixed $\boldsymbol{\theta}_0$, if we set $q_n(z^{(n)}) = p(z^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta}_0)$ the bound is tight:

$$\sum_{n=1}^{N} \log p(\mathbf{x}^{(n)};\boldsymbol{\theta}_0) = \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\log \frac{p(z^{(n)},\mathbf{x}^{(n)};\boldsymbol{\theta}_0)}{q_n(z^{(n)})}\right]$$

- Written another way:

$$\log p(\mathcal{D};\boldsymbol{\theta}_0) = \mathcal{L}(q;\boldsymbol{\theta}_0) \text{ if } \forall\ n\,,\, q_n(z^{(n)}) = p(z^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta}_0)$$

EM

- The EM algorithm alternates between making the bound tight at the current parameter values and then optimizing the lower bound
- If the current parameter value is $\boldsymbol{\theta}^{\text{old}}$:
  - **E-step**: For all n, set $q_n(z^{(n)}) = p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}})$ and form the lower bound $\mathcal{L}(q; \boldsymbol{\theta})$
    - Remember: $\log p(\mathcal{D}; \boldsymbol{\theta}^{\text{old}}) = \mathcal{L}(q; \boldsymbol{\theta}^{\text{old}})$ after this step
  - **M-step**: Optimize the lower bound:

$$\begin{aligned}
\boldsymbol{\theta}^{\text{new}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[\log \frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})} \right]
\end{aligned}$$

M-Step

- **M-step**: Optimize the lower bound:

$$\sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\log \frac{p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})}{q_n(z^{(n)})}\right] =$$

$$\sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})}\left[\log p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})\right] - \underbrace{\mathbb{E}_{q_n(z^{(n)})}\left[\log q_n(z^{(n)})\right]}_{\text{constant w.r.t.}\boldsymbol{\theta}}$$

- Substitute in $q_n(z^{(n)}) = p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}})$:

$$\boldsymbol{\theta}^{\text{new}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}})}\left[\log p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta})\right]$$

- This is the expected complete data log-likelihood.

EM Alternative Description

- **E-step**: For all n, set $q_n(z^{(n)}) = p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}})$ and form the lower bound $\mathcal{L}(q; \boldsymbol{\theta})$
- **M-step**: Optimize the lower bound:

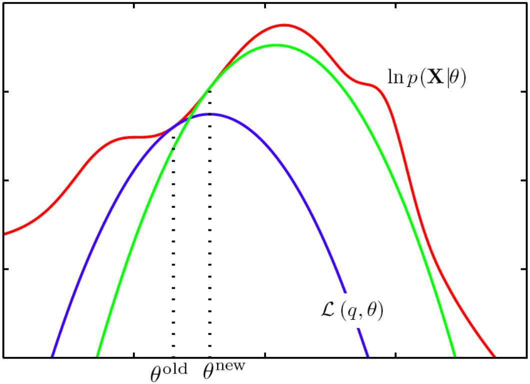
$$\begin{aligned}\boldsymbol{\theta}^{\text{new}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}})} \left[\log p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta}) \right]\end{aligned}$$

EM Convergence

- We can deduce that an iteration of EM will improve the log-likelihood by using the fact that the bound is tight at $\boldsymbol{\theta}^{\text{old}}$ after the E-step
- Let q denote the q_n's after the E-step i.e. $q_n(z^{(n)}) = p(z^{(n)}|\mathbf{x}^{(n)};\boldsymbol{\theta}^{\text{old}})$

$$
\begin{aligned}
\log p(\mathcal{D};\boldsymbol{\theta}^{\text{new}}) &\geq \mathcal{L}(q,\boldsymbol{\theta}^{\text{new}}) && \text{since } \log p(\mathcal{D};\boldsymbol{\theta}) \geq \mathcal{L}(q,\boldsymbol{\theta}) \text{ always} \\
&\geq \mathcal{L}(q,\boldsymbol{\theta}^{\text{old}}) && \text{since } \boldsymbol{\theta}^{\text{new}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q,\boldsymbol{\theta}) \\
&= \log p(\mathcal{D};\boldsymbol{\theta}^{\text{old}}) && \text{since } \log p(\mathcal{D};\boldsymbol{\theta}^{\text{old}}) = \mathcal{L}(q;\boldsymbol{\theta}^{\text{old}})
\end{aligned}
$$

EM Visualization



- The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values

Revisiting Mixture of Gaussians

- Let's revisit the mixture of Gaussians example from last lecture and derive the updates using our general EM algorithm

- Recall our model was:

$$p(z = k; \boldsymbol{\theta}) = \pi_k$$

$$p(\mathbf{x}|z = k; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)$$

- In this scenario, we have $\boldsymbol{\theta} = \{\mu_k, \pi_k, \Sigma_k\}_{k=1}^{K}$

E-Step for Mixture of Gaussians

- Let the current parameters be $\boldsymbol{\theta}^{\text{old}} = \{\mu_k^{\text{old}}, \pi_k^{\text{old}}, \Sigma_k^{\text{old}}\}_{k=1}^K$
- **E-step**: For all n, set $q_n(z^{(n)}) = p(z^{(n)}|\mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}})$

$$r_k^{(n)} := q_n(z^{(n)} = k) = p(z^{(n)} = k|\mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}}) = \frac{\pi_k^{\text{old}} \mathcal{N}(\mathbf{x}^{(n)}|\mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{j=1}^K \pi_j^{\text{old}} \mathcal{N}(\mathbf{x}^{(n)}|\mu_j^{\text{old}}, \Sigma_j^{\text{old}})}$$

M-Step for Mixture of Gaussians

M-step:

$$\boldsymbol{\theta}^{\text{new}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[\log p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta}) \right]$$

- Substitute in:
  - $\log p(z^{(n)}, \mathbf{x}^{(n)}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{I}[z^{(n)} = k] \left(\log \pi_k + \log \mathcal{N}(\mathbf{x}^{(n)}; \mu_k, \Sigma_k) \right)$
  - $q_n(z^{(n)}) = p(z^{(n)} | \mathbf{x}^{(n)}; \boldsymbol{\theta}^{\text{old}})$:

$$\begin{aligned}
\boldsymbol{\theta}^{\text{new}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[\sum_{k=1}^{K} \mathbb{I}[z^{(n)} = k] \left(\log \pi_k + \log \mathcal{N}(\mathbf{x}^{(n)}; \mu_k, \Sigma_k) \right) \right] \\
&= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_k^{(n)} \left(\log \pi_k + \log \mathcal{N}(\mathbf{x}^{(n)}; \mu_k, \Sigma_k) \right)
\end{aligned}$$

M-Step for Mixture of Gaussians

$$\boldsymbol{\theta}^{\text{new}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_k^{(n)} \left(\log \pi_k + \mathcal{N}(\mathbf{x}^{(n)}; \mu_k, \Sigma_k) \right)$$

- Taking derivatives and setting to zero, we get the updates from last lecture:

$$\begin{aligned}
\mu_k &= \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} \mathbf{x}^{(n)} \\
\Sigma_k &= \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} (\mathbf{x}^{(n)} - \mu_k)(\mathbf{x}^{(n)} - \mu_k)^T \\
\pi_k &= \frac{N_k}{N} \quad \text{with} \quad N_k = \sum_{n=1}^{N} r_k^{(n)}
\end{aligned}$$

EM Recap

- A general algorithm for optimizing many latent variable models.
- Iteratively computes a lower bound then optimizes it.
- Converges but maybe to a local minima.
- Can use multiple restarts.
- Can initialize from k-means for mixture models
- Limitation - need to be able to compute $p(z|\mathbf{x};\boldsymbol{\theta})$, not possible for more complicated models.