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).
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 \( x_1 \) and \( x_2 \) with probabilities \( p_1 \) and \( p_2 \), just the definition of a convex function:

\[
f(p_1 x_1 + p_2 x_2) \leq p_1 f(x_1) + p_2 f(x_2)
\]

- This is a convenient way to remember which way the inequality goes.
Jensen’s Inequality: For convex $f$:

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

Image credit: Mark Reid
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:

$$g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)]$$
In this lecture, we’ll be using $x$ to denote observed data and $z$ to denote the latent variables.

We’ll let $p(z, x; \theta)$ 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 $D = \{x^{(n)}\}_{n=1}^N$ and would like to fit $\theta$ using maximum likelihood:

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

To compute $p(x; \theta)$, we have to marginalize over $z$:

$$p(x; \theta) = \sum_z p(z, x; \theta)$$
Typically no closed form solution to the maximum likelihood problem

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

Key difficulty: once \( z \) is marginalized out, \( p(x; \theta) \) could be complex (e.g. a mixture distribution)

We’d like to write an objective in terms of \( \log p(z, x; \theta) \), which should be simpler to solve

To accomplish this, we need to move the summation outside the log

We introduce auxiliary distributions \( q_n(z^{(n)}) \) over each of the latent variables
\[
\sum_{n=1}^{N} \log \left( \sum_{z^{(n)}} p(z^{(n)}, x^{(n)}; \theta) \right) = \sum_{n=1}^{N} \log \left( \sum_{z^{(n)}} q_n(z^{(n)}) \frac{p(z^{(n)}, x^{(n)}; \theta)}{q_n(z^{(n)})} \right) \\
= \sum_{n=1}^{N} \log \left( \mathbb{E}_{q_n(z^{(n)})} \left[ \frac{p(z^{(n)}, x^{(n)}; \theta)}{q_n(z^{(n)})} \right] \right) \\
\geq \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[ \log \frac{p(z^{(n)}, x^{(n)}; \theta)}{q_n(z^{(n)})} \right]
\]

- In the last step, we use Jensen’s Inequality. Since \(\log\) is concave:

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

- We expect \(\mathcal{L}(q, \theta)\) might be easier to optimize w.r.t. \(\theta\), since it only appears in \(\log p(z^{(n)}, x^{(n)}; \theta)\), 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.
We know this bound is tight (i.e. the inequality becomes an equality) if there are constants $c_n$ such that:

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

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

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

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

Hence:

$$q_n(z^{(n)}) = \frac{p(z^{(n)}, x^{(n)}; \theta)}{p(x^{(n)}; \theta)} = p(z^{(n)} | x^{(n)}; \theta)$$
For fixed $\theta_0$, if we set $q_n(z^{(n)}) = p(z^{(n)}|x^{(n)}; \theta_0)$ the bound is tight:

$$\sum_{n=1}^{N} \log p(x^{(n)}; \theta_0) = \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[ \log \frac{p(z^{(n)}, x^{(n)}; \theta_0)}{q_n(z^{(n)})} \right]$$

Written another way:

$$\log p(D; \theta_0) = \mathcal{L}(q; \theta_0) \text{ if } \forall \ n, q_n(z^{(n)}) = p(z^{(n)}|x^{(n)}; \theta_0)$$
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 \( \theta^{old} \):

- **E-step**: For all \( n \), set \( q_n(z^{(n)}) = p(z^{(n)}|x^{(n)}; \theta^{old}) \) and form the lower bound \( \mathcal{L}(q; \theta) \).
  - Remember: \( \log p(D; \theta^{old}) = \mathcal{L}(q; \theta^{old}) \) after this step.

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

\[
\theta^{new} = \arg\max_{\theta} \mathcal{L}(q, \theta)
\]

\[
= \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[ \log \frac{p(z^{(n)}, x^{(n)}; \theta)}{q_n(z^{(n)})} \right]
\]
M-Step

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

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

constant w.r.t. \(\theta\)

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

\[
\theta^{\text{new}} = \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{p(z^{(n)}|x^{(n)};\theta^{\text{old}})} \left[ \log p(z^{(n)}, x^{(n)}; \theta) \right]
\]

- This is the expected complete data log-likelihood.
**E-step:** For all $n$, set $q_n(z^{(n)}) = p(z^{(n)}|x^{(n)}; \theta^{\text{old}})$ and form the lower bound $\mathcal{L}(q; \theta)$

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

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q, \theta)$$

$$= \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{p(z^{(n)}|x^{(n)}; \theta^{\text{old}})} \left[ \log p(z^{(n)}, x^{(n)}; \theta) \right]$$
We can deduce that an iteration of EM will improve the log-likelihood by using the fact that the bound is tight at $\theta^{\text{old}}$ after the E-step.

Let $q$ denote the $q_n$’s after the E-step i.e. $q_n(z^{(n)}) = p(z^{(n)}|x^{(n)}; \theta^{\text{old}})$

$$\log p(D; \theta^{\text{new}}) \geq \mathcal{L}(q, \theta^{\text{new}})$$
$$\geq \mathcal{L}(q, \theta^{\text{old}})$$
$$= \log p(D; \theta^{\text{old}})$$

since $\log p(D; \theta) \geq \mathcal{L}(q, \theta)$ always

since $\theta^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q, \theta)$

since $\log p(D; \theta^{\text{old}}) = \mathcal{L}(q; \theta^{\text{old}})$
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.
Let’s revisit the mixture of Gaussians example from last lecture and derive the updates using our general EM algorithm.

Recall our model was:

\[
p(z = k; \theta) = \pi_k
\]

\[
p(x|z = k; \theta) = \mathcal{N}(x; \mu_k, \Sigma_k)
\]

In this scenario, we have \( \theta = \{\mu_k, \pi_k, \Sigma_k\}_{k=1}^K \)
E-Step for Mixture of Gaussians

- Let the current parameters be $\theta^{\text{old}} = \{\mu_k^{\text{old}}, \pi_k^{\text{old}}, \Sigma_k^{\text{old}}\}_{k=1}^K$

- **E-step**: For all $n$, set $q_n(z^{(n)}) = p(z^{(n)}|x^{(n)}; \theta^{\text{old}})$

$$r_k^{(n)} := q_n(z^{(n)} = k) = p(z^{(n)} = k|x^{(n)}; \theta^{\text{old}}) = \frac{\pi_k^{\text{old}} \mathcal{N}(x^{(n)}|\mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{j=1}^K \pi_j^{\text{old}} \mathcal{N}(x^{(n)}|\mu_j^{\text{old}}, \Sigma_j^{\text{old}})}$$
M-Step for Mixture of Gaussians

M-step:

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

Substitute in:

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

$$\theta^{\text{new}} = \arg\max_{\theta} \sum_{n=1}^{N} \mathbb{E}_{q_n(z^{(n)})} \left[ \sum_{k=1}^{K} \mathbb{I}[z^{(n)} = k] \left( \log \pi_k + \log \mathcal{N}(x^{(n)}; \mu_k, \Sigma_k) \right) \right]$$

$$= \arg\max_{\theta} \sum_{n=1}^{N} \sum_{k=1}^{K} r_k^{(n)} \left( \log \pi_k + \log \mathcal{N}(x^{(n)}; \mu_k, \Sigma_k) \right)$$
M-Step for Mixture of Gaussians

\[
\theta^{\text{new}} = \arg\max_{\theta} \sum_{n=1}^{N} \sum_{k=1}^{K} r_k^{(n)} \left( \log \pi_k + \mathcal{N}(x^{(n)}; \mu_k, \Sigma_k) \right)
\]

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

\[
\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} x^{(n)}
\]

\[
\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} r_k^{(n)} (x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T
\]

\[
\pi_k = \frac{N_k}{N} \quad \text{with} \quad N_k = \sum_{n=1}^{N} r_k^{(n)}
\]
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 $p(z|x; \theta)$, not possible for more complicated models.