Topics for Today

- EM Example: Mixtures
- Bayesian alternatives to MLE

An example of apply the general EM algorithm

Mixture distribution:

\[ p(x) = \sum_{j=1}^{m} a_j p_j(x) \quad \sum_{j=1}^{m} a_j = 1 \]

Treat mixture component index as a hidden variable \( x = \text{observed}, \ (x, z) = \text{complete} \)

\[ p(x) = \sum_{j=1}^{m} p(z_j) p(x | z_j) \quad p(z_j) = \alpha_j \]

\[ p(x | z_j) = \beta_j(x) \]

Let’s consider the case where:

\[ p(x | z_j) \sim N(\mu_j, \sigma^2) \quad \text{(1 \sigma^2 shared by all } j \text{)} \]

Then \( \Theta = \frac{1}{2} \sum \alpha_j \beta_j \mu_j^2, \gamma, \nu \) \( \text{let } \nu = \frac{1}{\sigma^2} \)

E-Step

Find \( Q(\Theta | \Theta^{(n)}) = E\left[ \log p(x, z | \Theta) | x, \Theta^{(n)} \right] \)

\( E\left[ \log p(x, z | \Theta) | x, \Theta^{(n)} \right] \) \quad \text{Assuming} \ (x_i, z_i) \text{are i.i.d.}

\[ = \sum_{j=1}^{m} E\left[ \log p(x, z | \Theta) | x_i, \Theta^{(n)} \right] \text{ hidden} \]

\[ = \sum_{i=1}^{n} E\left[ \log p(x, z | \Theta) | x_i, \Theta^{(n)} \right] \text{ observed} \]
\[
= \sum_{i=1}^{N} \left[ \mathbb{E}_{\theta_i^{(p)}} \left[ \log p(x_i \mid z_i, \theta) \mid x_i, \theta^{(p)} \right] + \mathbb{E}_{\theta_i^{(p)}} \left[ \log p(z_i \mid \theta) \mid x_i, \theta^{(p)} \right] \right]
\]

Recall \( \mathbb{E}_q(z) = \sum_z q(z) p(z) \)

\[
= \sum_{i=1}^{N} \left[ \log p(x_i \mid z_i, \theta) \mid x_i, \theta^{(p)} \right] \]

\[
\approx \frac{1}{2} \log r - \frac{1}{2} (x_i - \mu_j)^2 = \gamma_{ij}
\]

\[
E_{\theta_i^{(p)}} \left[ \log p(z_i \mid \theta) \mid x_i, \theta^{(p)} \right] \approx \gamma_{ij}
\]

Putting it all together: At iteration \( (p) \)

\[
\phi_{ij} \text{ compute } \gamma_{ij} = p(z_i = j \mid x_i, \theta^{(p)}) \]

\[
= \frac{p(x_i \mid z_i = j, \theta^{(p)}) \alpha_{ij}^{(p)}}{\sum_k p(x_i \mid z_i = k, \theta^{(p)}) \alpha_{ik}^{(p)}}
\]

\[
Q(\theta | \theta^{(p)}) = \sum_{i=1}^{N} \sum_{j=1}^{m} \left[ \frac{1}{2} \log r - \frac{1}{2} (x_i - \mu_j)^2 + \log \gamma_{ij} \right]
\]

\[
- \lambda \left( 1 - \sum_{j=1}^{m} \gamma_{ij} \right)
\]

Lagrangian to meet constraint that \( \sum_{j=1}^{m} \gamma_{ij} = 1 \)
**M-step**

Find $\alpha_j$, $\mu_j$, $r^{(p+1)} = \frac{1}{r^{(p)}}$ to max $Q(\theta|\theta^{(p)})$

\[
\frac{d}{d\mu_j} Q = \frac{n}{2} \sum_{i=1}^{n} \delta_{ij} r(x_i - \mu_j) = 0
\]

\[
\sum_{i=1}^{n} \delta_{ij} x_i = \mu_j \sum_{i=1}^{n} \delta_{ij} \Rightarrow \mu_j = \frac{\sum_{i=1}^{n} \delta_{ij} x_i}{\sum_{i=1}^{n} \delta_{ij}}
\]

\[
\frac{d}{dr} Q = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij} \left[ \frac{1}{2r} - \frac{1}{2} (x_i - \mu_j)^2 \right] = 0
\]

\[
+ \frac{1}{r} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij} x_i = \frac{1}{r} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij} (x_i - \mu_j) = \frac{1}{r} \sum_{i=1}^{n} \sum_{j=1}^{m} \delta_{ij} (x_i - \mu_j)
\]

\[
\frac{d^2}{d\alpha_j} Q = \frac{n}{2r} \delta_{ij} \frac{1}{\alpha_j} - 1 = 0
\]

\[
\frac{1}{\alpha_j} \sum_{i=1}^{n} \delta_{ij} = 1 \Rightarrow \alpha_j = \frac{\sum_{i=1}^{n} \delta_{ij}}{r^{(p)}}
\]

Choose $r$ s.t. $\sum_j \alpha_j = 1 \Rightarrow r^{(p+1)} = \frac{\sum_{i=1}^{n} \delta_{ij} x_i}{\sum_{i=1}^{n} \delta_{ij}}$

(Note: $\alpha_j$ solution does not depend on form of $p(x_i|\theta_j)$)

**Interpretation:**

**E-step:** Compute weights (belief) that $x_i$ came from mixture $\frac{\delta_{ij}}{r^{(p)}} = p(z_i = j|x_i, \theta^{(p)})$

**M-step:** Use weighted data & weighted counts in ML estimate.
Alternatives to MLE

\[ \hat{\theta}_{\text{MLE}} = \arg\max_{\theta} p(x|\theta) = \arg\max_{\theta} \log p(x|\theta) \]

point estimate, assume true model form
a rely entirely on data

What if...

- You want to incorporate prior knowledge about \( \theta \)?

Example: \( \theta \) known to be small, near zero say \( \theta \sim N(0, \sigma_\theta^2) \)

![Normal distribution with mean 0 and variance \( \sigma_\theta^2 \)]

\( \Rightarrow \) use combined info from prior & data

Starting point = prior mean \( \hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|x) \)

- You want to incorporate uncertainty in your estimate?

![Comparison of narrower and wider distributions]

\( \Rightarrow \) \( \hat{\theta} \) estimate varies

\[ p(x|\theta) = \int p(x|\theta)p(\theta|x)\,d\theta \]

Bayesian Learning
Bayesian Learning

In Bayesian Learning, we are interested in estimating $p(\theta|\mathbf{x})$. By Bayes rule:

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta) p(\theta)$$

\[\text{posterior} \propto \text{likelihood} \cdot \text{prior}\]

- MAP: $\hat{\theta}_{\text{MAP}} = \arg \max \theta p(\theta|\mathbf{x}) = \arg \max \theta p(\mathbf{x}|\theta) p(\theta)$

Bayesian learning: estimate full $p(\mathbf{x}|\theta)$

- Prior can influence what $\theta$ is chosen (or it could be some MLE w/ uniform prior)

\underline{MAP example: estimate Gaussian mean, Prior given. (suppose $\sigma^2$ known)}

1. $p(\mathbf{x} | \theta) = \prod_{i=1}^{n} p(x_i | \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$

   (Assume prior is also Gaussian (this is a conjugate prior; makes posterior easy to compute)

2. $p(\theta) = p(\mu) = N(\mu_0, \sigma_0^2)$

   $= \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp \left( -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$

3. $p(\theta | \mathbf{x}) = p(\mu | \mathbf{x}) \propto p(\mathbf{x} | \theta) p(\mu)$

   $= \mathcal{K} \exp \left( -\frac{1}{2} \left( \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^2} + (\frac{\mu - \mu_0}{\sigma_0})^2 \right) \right)$

   $= \mathcal{K}' \exp \left( -\frac{1}{2} \left( \frac{\sum_{i=1}^{n} (x_i - \mu_0)^2 + 2\mu x_i - 2\mu x_0}{\sigma^2} \right) + \frac{\mu^2 + \mu_0^2 - 2\mu \mu_0}{\sigma_0^2} \right)$

   $= \mathcal{K}'' \exp \left( -\frac{1}{2} \left[ \mu \left( \frac{\mu}{\sigma^2} + \frac{1}{\sigma_0^2} \right) + M \left( \frac{2\mu x_i}{\sigma^2} - \frac{2\mu_0}{\sigma_0^2} \right) \right] \right)$
We identify that \( \theta \) is of Gaussian form and will try to match the coefficients \((\mu_n, \sigma_n^2)\) of 

\[
p(\theta | x) = \frac{1}{(2\pi \sigma_n^2)^{1/2}} \exp\left(-\frac{1}{2\sigma_n^2} (\theta - \mu_n)^2 \right)
\]

Completing the squares, we get to:

\[
\mu_n = \left( \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} \right) \bar{x} + \frac{\sigma^2}{\sigma_0^2 + \sigma^2} \mu_0
\]

\[
\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2}
\]

Observe:

(a) \( \mu_n \) is a weighted sum of sample mean and prior mean.

The weights are non-negative & sum to 1.

\[
\text{convex combination } \mu_0, \bar{x}
\]

(b) As \( n \) increases, we move toward \( \bar{x} \) (trust data more). If \( n \) small, we trust prior more.

(c) If \( \sigma_0^2 = 0 \), our prior of \( \mu_0 \) is so strong that \( \mu_n = \mu_0 \).

If \( \sigma^2 \ll \sigma_0^2 \), we're so uncertain of our prior that \( \bar{x} \) gets more weight.

(d) As \( n \) increases, our posterior \( p(\theta | x) \) peaks.

\[
\sigma_n^2 \to 0
\]
Remarks

- In MAP, we pick \( \hat{\Theta}_{\text{MAP}} = \arg \max \ p(\Theta|\mathbf{x}) \) point estimate.

- In Bayesian learning, we retain full \( p(\Theta|\mathbf{x}) \), use \( p(\mathbf{x}|\Theta) \).

- For simplicity, we often choose conjugate prior to get parametric form for posterior. This need not be the case. Full Bayesian learning can get general distributions (but requires complex computation).

- Application of MAP:
  - Prior knowledge from out-of-domain data, i.e. Adaptation problem)
  - Prior knowledge to enforce regularization (e.g. small parameter values)

- Bayesian learning has a sequential belief update property.

- There exists a Frequentist-Bayesian debate in statistics (But this may be less relevant for engineers —)
  just use whatever works.

- Another point estimate (another way of using \( p(\Theta|\mathbf{x}) \)) is to take the expected \( \Theta \) rather than most likely \( \Theta \)

\[ \hat{\Theta}_{\text{MMSE}} = \mathbb{E}[\Theta|\mathbf{x}] = \int \Theta p(\Theta|\mathbf{x}) d\Theta \]

Advantage: minimizes estimation error (better than MAP)
Disadvantage: can be harder to compute.