EXPECTATION MAXIMIZATION

Slides based on *Pattern Recognition and Machine learning* by Christopher Bishop (chapter 9) and slides by Shai Fine
Tutorial outline

• Demonstrations
  • K-means
  • GMM

• Expectation-Maximization
  • Formalization

• GMM
  • Overview

• Binomial Mixture Model

• K-means as “hard GMM”
DEMONSTRATIONS
K-means demonstration

- Goal: Hard clustering the data to K clusters
Gaussian Mixture Models

• Goal: Soft clustering our data under the assumption that it is generated by a mixture of Gaussians
K-means animation
GMM animation
EXPECTATION-MAXIMIZATION
Expectation Maximization (EM)

- Probabilistic method for soft clustering.
  - “Soft version of $k$-means”
- Assumes a probabilistic model of clusters that allows computing $\Pr(c_j|x)$ for each cluster, $c_j$, for a given example, $x$
  - If we would know for each data instance from what distribution it came, could use parametric estimation
- Introduce unobservable (latent) variables which indicate source distribution
- Run iterative process
  - Estimate latent variables from data and current estimation of distribution parameters
  - Use current values of latent variables to refine parameter estimation
More Formally

• Log likelihood for a mixture model
\[
\mathcal{L}(X|\Theta) \overset{i.i.d}{=} \log \prod_i \Pr(x_i|\Theta) = \sum_i \log \sum_{j=1}^k \Pr(x_i|C_j; \Theta) \Pr(C_j; \Theta)
\]

• Assume latent variables \(z\), which when known make the optimization simpler
  • Complete likelihood, \(\mathcal{L}_c(X, Z|\Theta)\), in terms of \(x\) and \(z\)
  • Incomplete likelihood, \(\mathcal{L}(X|\Theta)\), in terms of \(x\)

• However, \(z\) is latent, so we can’t compute \(\mathcal{L}_c(X, Z|\Theta)\)
  • But we can compute its conditional expected value, given \(X\), and old \(\Theta^t\)
\[
Q(\Theta; \Theta^t) \triangleq \mathbb{E}_Z[\mathcal{L}_c(X, Z|\Theta)|X, \Theta^t] = \sum_Z \Pr(Z|X, \Theta^t) \log \Pr(X, Z; \Theta)
\]
More Formally – cont.

• From a computation viewpoint
  • The **E-step** computes the **posterior probability** \( \Pr(Z|X, \theta^t) \) using the current estimates (probability point \( i \) belongs to model \( j \))
  • The **M-step** updates the **parameter estimates** to get \( \theta^{t+1} \) by maximizing \( Q(\theta; \theta^t) \)

• The EM Algorithm requires an initial guess \( \theta^0 \) for the parameters
• Each iteration of E-step and M-step is guaranteed to increase the log-likelihood of the observed data, \( \log \Pr(X|\theta) \) until a local maximum is reached
E-Step and M-Step

- Iterate the two steps
  - **E-step**: Estimate $Z$ given $X$ and current $\Theta$
    - $Q(\Theta; \Theta^t) = E[\mathcal{L}_c(X, Z|\Theta)|X, \Theta^t]$
  - **M-step**: Find new $\Theta$ given $Z$, $X$, and old $\Theta$
    - $\Theta^{t+1} = \arg\max_\Theta Q(\Theta; \Theta^t)$

- An increase in $Q$ increases the incomplete likelihood
  $$\mathcal{L}(X|\Theta^t) \geq Q(\Theta|\Theta^t)$$
EM Recipe – E step

\[
\Pr(z_i = j | x_i, \Theta) = \frac{\Pr(x_i, z_i = j | \Theta)}{\sum_{j'} \Pr(x_i, z_i = j' | \Theta)}
\]

Substitute the probabilities with the desired distribution
EM Recipe – Derive Q, M-step

\[ Q(\Theta; \Theta^t) = E_Z [\mathcal{L}_c(X, Z|\Theta)|X, \Theta^t] = \sum_Z \Pr(Z|X, \Theta^t) \log \Pr(X, Z; \Theta) \]

\[ Q(\Theta; \Theta^t) = \sum_i \sum_{j=1}^k \Pr(z_i = j|x_i, \Theta^t) \log \Pr(x_i, z_i = j|\Theta) \]

\[ = \sum_i \sum_{j=1}^k r_{ij} [\log \Pr(z_i = j|\Theta) + \log \Pr(x_i|z_i = j, \Theta)] \]

- \( r_{ij} = \Pr(z_i = j|x_i, \Theta) \) (from E step)
- Substitute \( \Pr(x_i|z_i = j, \Theta) \) with desired probability
- Find MLE (differentiate and compare to 0)
GAUSSIAN MIXTURE MODELS
GMM reminder

By Dr. Seuss Gauss

One Gaussian
Two Gaussian
Red Gaussian
Blue Gaussian
GMM reminder – more formally

- One Gaussian:
  \[ \mathcal{N}(x|\mu, \Sigma) \equiv \Pr(x|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) \right\} \]

- Gaussian mixture:
  \[ \Pr(x) = \sum_{j=1}^{k} \alpha_j \mathcal{N}(x|\mu_j, \Sigma_j) \]

- The log-likelihood of a GMM:
  \[ \mathcal{L}(X|\Theta) = \log \prod_i \Pr(x_i|\Theta) = \sum_i \log \sum_{j=1}^{k} \alpha_j \mathcal{N}(x|\mu_j, \Sigma_j) \]

- No closed form solution and not convex
GMM as EM

• We introduce a latent random variable $z$
  • $z \in \{0,1\}^k$ – a one-hot random variable indicating to source Gaussian the sample belongs to
  • $\Pr(z_k) = \alpha_k$ - the probability of that source
  • Note that $\sum_{k=1}^{K} \alpha_k = 1$

• The marginal probability:
\[
\Pr(x) = \sum_{z} p(z)p(x|z) = \sum_{j=1}^{k} \alpha_j \mathcal{N}(x|\mu_j, \Sigma_j)
\]
GMM – E step

- The E-step computes the posterior probability of the missing data

\[
\Pr(z_i = j|x_i, \Theta) = \frac{\Pr(x_i, z_i = j|\Theta)}{\sum_{j'} \Pr(x_i, z_i = j'|\Theta)} = \frac{\alpha_j \Pr(x_i |\mu_j, \Sigma_j)}{\sum_{j'} \alpha_j' \Pr(x_i |\mu_j, \Sigma_j)}
\]

- Denote \( r_{ij} = \Pr(z_i = j|x_i, \Theta) \)
- What is \( \Pr(x_i |\mu_j, \Sigma_j) \)?
GMM – Calc. $Q(\Theta; \Theta^t)$

$$Q(\Theta; \Theta^t) = E_Z[\mathcal{L}_c(X, Z|\Theta)|X, \Theta^t] = \sum_{Z} \Pr(Z|X, \Theta^t) \log \Pr(X, Z; \Theta)$$

$$Q(\Theta; \Theta^t) = \sum_{i} \sum_{j=1}^{k} \Pr(z_i = j|x_i, \Theta) \log \Pr(x_i, z_i = j|\Theta)$$

$$= \sum_{i} \sum_{j=1}^{k} r_{ij} [\log \Pr(z_i = j|\Theta) + \log \Pr(x_i|z_i = j, \Theta)]$$

$$= \sum_{i} \sum_{j=1}^{k} r_{ij} [\log \alpha_j + \log \Pr(x_i|\mu_j, \Sigma_j)]$$

$$= \sum_{i} \sum_{j=1}^{k} r_{ij} \log \alpha_j + - \frac{1}{2} \sum_{j=1}^{k} \log |\Sigma_j| \sum_{i} r_{ij} +$$

$$- \frac{1}{2} \sum_{i} \sum_{j=1}^{k} r_{ij} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) + \text{Conts}$$
GMM – M-step

• To maximize $Q(\Theta; \Theta^t)$ with respect to $\mu_j$, set the gradient to zero

$$Q(\Theta; \Theta^t) = \sum_i \sum_{j=1}^{k} r_{ij} \log \alpha_j +$$
$$- \frac{1}{2} \sum_{j=1}^{k} \log |\Sigma_j| \sum_i r_{ij} + - \frac{1}{2} \sum_i \sum_{j=1}^{k} r_{ij} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) + \text{Conts}$$

$$\frac{\partial}{\partial \mu_j} Q(\Theta; \Theta^t) = \Sigma_{i=1}^{n} r_{ij} \Sigma_j^{-1} (x_i - \mu_j) = 0$$

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} r_{ij} x_i}{\sum_{i=1}^{n} r_{ij}}$$

• Similarly, we have

$$\hat{\Sigma}_j = \frac{\sum_{i=1}^{n} r_{ij} (x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n} r_{ij}}$$

Reminder:

$$\frac{\partial}{\partial s} (x - As)^T W (x - As) = -2A^T W (x - As)$$
GMM – M-step – cont.

• Maximize $Q(\Theta; \Theta^t)$ with respect to $\alpha_j$. Use Lagrange multiplier:
  
  \[ \mathcal{L} = Q(\Theta; \Theta^t) + \lambda (1 - \sum_j \alpha_j) \]

  \[ \frac{\partial \mathcal{L}}{\partial \alpha_j} = \sum_i \frac{r_{ij}}{\alpha_j} - \lambda = 0 \quad (\ast) \quad \Leftrightarrow \quad s.t. \sum_j \alpha_j = 1 \]

• Find an expression for $\lambda$ by summing all partial derivatives of $\alpha_j$

  \[ \sum_i r_{ij}^{(t)} = \lambda \alpha_j \Rightarrow \sum_j \sum_i r_{ij}^{(t)} = \lambda \sum_j \alpha_j \Rightarrow \lambda = n \]

• Substituting $\lambda$ back in the expression $(\ast)$, we get the update expression for $\alpha_j$

  \[ \hat{\alpha}_j = \frac{\sum_{i=1}^n r_{ij}}{n} \]
BERNOULLI MIXTURE MODELS
Bernoulli Mixture Model

Problem Setting

- We have $k$ coins
  - The probability of observing a head with the $j$-th coin is $p_j$
  - We do not observe which coin was used
  - We only observe $x_i \in \{0, 1\}$, which records whether we see a heads or tails

- Let $z_i \in \{1, \ldots, k\}$ be the missing information of which coin was used on each flip
  - The probability of using the $j$-th coin is $\Pr(z_i = j) = \alpha_j$

- The complete data is given by $(X, Z)$
  - Using the law of total probability, the (marginal) probability of the observed data $X$
    \[
    \Pr(X) = \sum_j \Pr(X | Z = j) \Pr(Z = j)
    \]
  - Thus, the likelihood of the full data set (incomplete likelihood) is
    \[
    \mathcal{L}(X | \Theta) = \prod_i \sum_j \Pr(x_i | z_i = j) \Pr(z_i = j) = \prod_i \sum_j \alpha_j p_j^{x_i} (1 - p_j)^{1-x_i}
    \]
  - Where $\Theta = (\alpha, p)$
YOUR TURN
EM Recipe – E step

\[ \Pr(z_i = j|x_i, \Theta) = \frac{\Pr(x_i, z_i = j|\Theta)}{\sum_{j'} \Pr(x_i, z_i = j'|\Theta)} \]

Substitute the probabilities with the desired distribution
EM Recipe – Derive Q, M-step

\[ Q(\Theta; \Theta^t) = E_Z [\mathcal{L}_c (X, Z|\Theta)|X, \Theta^t] = \sum_Z \Pr(Z|X, \Theta^t) \log \Pr(X, Z; \Theta) \]

\[ Q(\Theta; \Theta^t) = \sum_i \sum_{j=1}^k \Pr(z_i = j|x_i, \Theta^t) \log \Pr(x_i, z_i = j|\Theta) \]

\[ = \sum_i \sum_{j=1}^k r_{ij} [\log \Pr(z_i = j|\Theta) + \log \Pr(x_i|z_i = j, \Theta)] \]

- \( r_{ij} = \Pr(z_i = j|x_i, \Theta) \) (from E step)
- Substitute \( \Pr(x_i|z_i = j, \Theta) \) with desired probability
- Find MLE (differentiate and compare to 0)
BMM SOLUTION
BMM – E step

• The E-step computes the posterior probability of the missing data

\[
\Pr(z_i = j|x_i, \Theta) = \frac{\Pr(x_i, z_i = j|\Theta)}{\sum_{j'} \Pr(x_i, z_i = j'|\Theta)}
\]

\[
= \frac{\alpha_j \Pr(x_i | p_j)}{\sum_{j'} \alpha_{j'} \Pr(x_i | p_j)}
\]

\[
= \frac{\alpha_j p_j^{x_i} (1 - p_j)^{1-x_i}}{\sum_{j'} \alpha_{j'} p_{j'}^{x_i} (1 - p_{j'})^{1-x_i}}
\]

• Denote \( r_{ij} = \Pr(z_i = j|x_i, \Theta) \)
BMM – Calc. $Q(\Theta; \Theta^t)$

$$Q(\Theta; \Theta^t) = \sum_i \sum_{j=1}^k \Pr(z_i = j|x_i, \Theta) \log \Pr(x_i, z_i = j|\Theta)$$

$$= \sum_i \sum_{j=1}^k r_{ij} \left[ \log \Pr(z_i = j|\Theta) + \log \Pr(x_i|z_i = j, \Theta) \right]$$

$$= \sum_i \sum_{j=1}^k r_{ij} \left[ \log \alpha_j + \log \Pr(x_i|p_j) \right]$$

$$= \sum_i \sum_{j=1}^k r_{ij} \log \alpha_j + \sum_i \sum_{j=1}^k \log \left( p_j^{x_i} (1 - p_j)^{1-x_i} \right)$$

$$= \sum_i \sum_{j=1}^k r_{ij} \log \alpha_j + \sum_i \sum_{j=1}^k r_{ij} x_i \log(p_j) + r_{ij} (1 - x_i) \log(1 - p_j)$$
BMM – M-step

- To maximize $Q(\Theta; \Theta^t)$ with respect to $p_j$, set the gradient to zero

$$Q(\Theta; \Theta^t) = \sum_i \sum_{j=1}^k r_{ij} \log \alpha_j + \sum_i \sum_{j=1}^k r_{ij} x_i \log(p_j) + r_{ij} (1 - x_i) \log(1 - p_j)$$

$$\frac{\partial}{\partial p_j} Q(\Theta; \Theta^t) = \sum_{i=1}^n r_{ij} \left( \frac{x_i}{p_j} - \frac{1-x_i}{1-p_j} \right) = 0$$

$$\hat{p}_j = \frac{\sum_{i=1}^n r_{ij} x_i}{\sum_{i=1}^n r_{ij}}$$
BMM – M-step – cont.

- To maximize $Q(\Theta; \Theta^t)$ with respect to $\alpha_j$, use Lagrange multiplier:
  - $\mathcal{L} = Q(\Theta; \Theta^t) + \lambda(1 - \sum_j \alpha_j)$
  - $\frac{\partial \mathcal{L}}{\partial \alpha_j} = \sum_i \sum_{j=1}^k r_{ij} \alpha_j - \lambda = 0$

- We already solved this for GMM:
  - $\hat{\alpha}_j = \frac{\sum_{i=1}^n r_{ij}}{n}$

- If all $r_{ij} = \{0,1\}$, i.e. deterministic
  - The component labels, $z_i$, are known
  - These update formulae reduce to the standard formulae for binomial distribution
K-MEANS AS “HARD GMM”
Relation to K-means

• Let’s re-examine the E-step:

\[
\Pr(z_i = j|x_i, \Theta) = \frac{\Pr(x_i, z_i = j|\Theta)}{\sum_{j'} \Pr(x_i, z_i = j'|\Theta)}
\]

\[
= \frac{c_j \exp \left\{ -\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right\}}{\sum_j' c_j' \exp \left\{ -\frac{1}{2} (x_i - \mu_j')^T \Sigma_j'^{-1} (x_i - \mu_j') \right\}}
\]

• Let’s assume all Gaussian have the same \( \Sigma = \epsilon I \)

\[
\frac{\exp \left\{ -\frac{1}{2\epsilon} \|x_i - \mu_j\|^2 \right\}}{\sum_j' \exp \left\{ -\frac{1}{2\epsilon} \|x_i - \mu_j'\|^2 \right\}}
\]
Relation to K-means – cont.

• $\text{Pr}(z_i = j | x_i, \Theta) = \frac{\exp\left\{-\frac{1}{2\epsilon}\|x_i - \mu_j\|^2\right\}}{\sum_j' \exp\left\{-\frac{1}{2\epsilon}\|x_i - \mu_j\|^2\right\}}$

• At the limit $\epsilon \to 0$, $\text{Pr}(z_i = j | x_i, \Theta) = 1$ for $j = \text{argmin}\{x_i - \mu_j\}$ and $\text{Pr}(z_i = j | x_i, \Theta) = 0$ for all others.

• Thus $r_{ij} = \begin{cases} 1, & j = \text{argmin}_j \|x_i - \mu_j\|^2 \\ 0, & \text{else} \end{cases}$

• The GMM equations are now identical to K-means’ equations:

  - $\hat{\mu}_j = \frac{\sum_{i=1}^n r_{ij} x_i}{\sum_{i=1}^n r_{ij}}$
  - $\hat{\alpha}_j = \frac{\sum_{i=1}^n r_{ij}}{n}$ (though it isn’t required)