

# Machine Learning for Data Science (CS 4786)

## Lecture 15: EM Algorithm for Gaussian Mixture Models and Mixture of Multinomials

### 1 Gaussian Mixture Models

Each  $\theta \in \Theta$  consist of mixture distribution  $\pi$  which is a distribution over the choices of the  $K$  clusters,  $\mu_1, \dots, \mu_K \in \mathbb{R}^d$  the choices of the  $K$  means for the corresponding gaussians and  $\Sigma_1, \dots, \Sigma_K$  the choices of the  $K$  covariance matrices. The latent variables are  $c_1, \dots, c_n$  the cluster assignments for the  $n$  points and  $x_1, \dots, x_n$  are the  $n$  observations.

#### 1.1 E-step

On iteration  $i$ , for each data point  $t \in [n]$ , set

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

Note that

$$\begin{aligned} Q_t^{(i)}(c_t) &= P(c_t|x_t, \theta^{(i-1)}) \\ &\propto p(x_t|c_t, \theta^{(i-1)}) \times P(c_t|\theta^{(i-1)}) \\ &\propto \frac{1}{\sqrt{(2\pi)^d |\Sigma_{c_t}|}} \exp\left(-\frac{(x_t - \mu_{c_t})^\top \Sigma_{c_t}^{-1} (x_t - \mu_{c_t})}{2}\right) \pi_{c_t} \end{aligned}$$

#### 1.2 M-step for GMM

For the M-step (for MLE) we would like to find

$$\theta = \operatorname{argmax}_{\theta \in \Theta} \sum_{t=1}^n \sum_{c_t=1}^K Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta)$$

To this end note that

$$\begin{aligned} \sum_{t=1}^n \sum_{c_t=1}^K Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta) &= \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) (\log \phi(x_t|\mu_k, \Sigma_k) + \log \pi_k) \\ &= \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \left( \frac{1}{2} \log \left( \frac{1}{(2 * 3.14)^d |\Sigma_k|} \right) - \frac{1}{2} (x_t - \mu_k)^\top \Sigma_k^{-1} (x_t - \mu_k) + \log \pi_k \right) \\ &= \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \left( -\frac{1}{2} \log \det(\Sigma_k) - \frac{1}{2} (x_t - \mu_k)^\top \Sigma_k^{-1} (x_t - \mu_k) + \log \pi_k \right) + \text{constant terms} \end{aligned}$$

For notational convenience define:

$$L(\mu_{1:K}, \Sigma_{1:K}, \pi) = \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \left( -\frac{1}{2} \log \det(\Sigma_k) - \frac{1}{2} (x_t - \mu_k)^\top \Sigma_k^{-1} (x_t - \mu_k) + \log \pi_k \right)$$

Our goal is to find parameters that maximize  $L(\mu_{1:K}, \Sigma_{1:K}, \pi)$ .

**M-step for mean:** To optimize with respect to mean we take derivative and equate to 0,

$$\begin{aligned} \frac{\partial}{\partial \mu_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) &= -\frac{1}{2} \frac{\partial}{\partial \mu_k} \left( \sum_{t=1}^n Q_t^{(i)}(k) (x_t - \mu_k)^\top \Sigma_k^{-1} (x_t - \mu_k) \right) \\ &= -\sum_{t=1}^n Q_t^{(i)}(k) \Sigma_k^{-1} (x_t - \mu_k) = -\Sigma_k^{-1} \left( \sum_{t=1}^n Q_t^{(i)}(k) (x_t - \mu_k) \right) \end{aligned}$$

To maximize over  $\mu_k$  we set derivative equal to 0. Hence

$$\sum_{t=1}^n Q_t^{(i)}(k) (x_t - \mu_k) = \sum_{t=1}^n Q_t^{(i)}(k) x_t - \mu_k \left( \sum_{t=1}^n Q_t^{(i)}(k) \right) = 0$$

Or equivalently:

$$\mu_k = \frac{\sum_{t=1}^n Q_t^{(i)}(k) x_t}{\sum_{t=1}^n Q_t^{(i)}(k)}$$

Qt: n\*k  
 $\mu_k$  = sum of weight\*location  
divided by  
sum of weight

**M-step for mixture distribution:** Since we want to optimize over  $\pi$  subject to the constraint  $\sum_{k=1}^K \pi_k = 1$  (ie. its a distribution), we do so by introducing Lagrange variables. That is we want to optimize the following term w.r.t.  $\pi_k$

$$L(\mu_{1:K}, \Sigma_{1:K}, \pi) + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right)$$

Hence taking derivative of above w.r.t.  $\pi$  we get,

$$\frac{\partial}{\partial \pi_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right) = \frac{\partial}{\partial \pi_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) + \lambda$$

But,

$$\frac{\partial}{\partial \pi_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) = \frac{\partial}{\partial \pi_k} \sum_{t=1}^n Q_t^{(i)}(k) \log(\pi_k) = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{\pi_k}$$

Hence,

$$\frac{\partial}{\partial \pi_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right) + \sum_{i=1}^K \lambda_i \pi_i = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{\pi_k} + \lambda$$

?

Setting derivative to 0 we discover that

$$\pi_k \propto \sum_{t=1}^n Q_t^{(i)}(k) \quad \begin{array}{l} \pi(k) = \text{分子} / \text{lamda} \\ \text{lamda} = \sum_{k=1 \sim K} [\text{分子}] \end{array}$$

**M-step for Covariance:** This one needs being able to take derivative w.r.t. matrices and so I will only sketch the proof here. Let us consider optimizing w.r.t. some  $\Sigma_k$ . It makes the problem easier if we instead think of the problem as optimizing over  $\Sigma_k^{-1}$  and then invert the solution.

Here are two facts that come in handy:

$$\frac{\partial}{\partial \mathbf{X}} \log \det(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}$$

and for any vector  $v$ ,

$$\frac{\partial}{\partial \mathbf{X}} v^\top \mathbf{X} v = v v^\top$$

Now note that

$$\begin{aligned} \frac{\partial}{\partial \Sigma_k} L(\mu_{1:K}, \Sigma_{1:K}, \pi) &= \frac{\partial}{\partial \Sigma_k} \left( \sum_{t=1}^n Q_t^{(i)}(k) \left( -\frac{1}{2} \log \det(\Sigma_k) - \frac{1}{2} (x_t - \mu_k)^\top \Sigma_k^{-1} (x_t - \mu_k) \right) \right) \\ &= \left( \sum_{t=1}^n Q_t^{(i)}(k) \left( \frac{1}{2} (\Sigma_k^{-1})^{-1} - \frac{1}{2} (x_t - \mu_k)(x_t - \mu_k)^\top \right) \right) \end{aligned}$$



Hence equating to 0 we get that

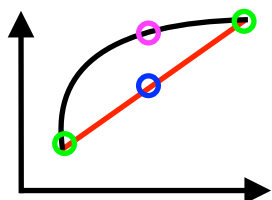
$$\Sigma_k = \frac{\sum_{t=1}^n Q_t^{(i)}(k) (x_t - \mu_k)(x_t - \mu_k)^\top}{\sum_{t=1}^n Q_t^{(i)}(k)}$$

that is the weighted sample variance.

## 2 EM Algorithm: Why it works?

Log likelihood only decreases after one iteration of EM algorithm. Why?

Jen sen's lemma:  
  $f(\mathbf{E} \cdot \mathbf{x}) \geq \mathbf{E} \cdot f(\mathbf{x})$    
 $f$  is concave function



$$\begin{aligned} \log P(x_1, \dots, x_n | \theta^{(i+1)}) &= \sum_{t=1}^n \log P(x_t | \theta^{(i+1)}) \\ &\geq \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left( \frac{P(x_t, c_t | \theta^{(i+1)})}{Q^{(i)}(c_t)} \right) \\ &\geq \sum_{t=1}^n \sum_{c_t=1}^K Q^{(i)}(c_t) \log \left( \frac{P(x_t, c_t | \theta^{(i)})}{Q^{(i)}(c_t)} \right) \\ &= \sum_{t=1}^n \log P(x_t | \theta^{(i)}) \\ &= \log P(x_1, \dots, x_n | \theta^{(i)}) \end{aligned}$$

### 3 Mixture of Multinomials

Each  $\theta \in \Theta$  consist of mixture distribution  $\pi$  which is a distribution over the choices of the  $K$  clusters or types,  $p_1, \dots, p_K$  the choices of the  $K$  multinomial distributions over the  $d$  items. The latent variables are  $c_1, \dots, c_n$  the cluster assignments for the  $n$  points and  $x_1, \dots, x_n$  are the  $n$  observations given as index of one of the  $d$  choices for t'th data point (product bought).

#### 3.1 E-step

On iteration  $i$ , for each data point  $t \in [n]$ , set

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

Note that

$$\begin{aligned} Q_t^{(i)}(c_t) &= P(c_t|x_t, \theta^{(i-1)}) \\ &\propto p(x_t|c_t, \theta^{(i-1)}) \times P(c_t|\theta^{(i-1)}) \\ &\propto p_{c_t}[x_t] \pi_{c_t} \end{aligned}$$

#### 3.2 M-step

To be filled in soon ...