Applied Machine Learning

EM for Mixtures of Normal Distributions

EM for Mixtures of Normal Distributions

- Modeling clusters through Normal Distributions
- EM algorithm for mixture of normals: E-Step
- EM algorithm for mixture of normals: M-Step

EM Algorithm

1. Initialize probability distributions
2. While ($\theta^{(n)}$ has not reached convergence)
    1. E-step
        - $p(\delta \,|\, \theta^{(n)}, \mathbf{x})$
        - $$\begin{aligned} \mathcal{Q}(\theta; \theta^{(n)}) &= \sum_\delta \mathscr{L}(\theta; \mathbf{x}, \delta) p(\delta \,|\, \theta^{(n)}, \mathbf{x}) \\ &= \mathbb{E}_{p(\delta|\theta^{(n)}, \mathbf{x})}[\mathscr{L}(\theta; \mathbf{x}, \delta)] \end{aligned}$$
        - $w_{i,j}$ to associate each item $\mathbf{x}_i$ to cluster center j
    2. M-step
        - $\theta^{(n+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta; \theta^{(n)})$

Blobs with Normal Distributions

- 1 blob
  - mean
  - covariance matrix
- several blobs (t), one normal distribution for each blob j
  - mean (and center) for j: μ_j
  - Covariance matrix for j: Σ can be factorized as $\Sigma = AA^\top$
    - A: full rank, thus Σ is positive definite, A^{-1} exists
    - $\Sigma = I$

Probability Model for Mixture of Normals

- data item: $\mathbf{x}_i$
- Mixture of t normal distributions
    - $p(\mathbf{x}_i \,|\, \mu_1, \ldots, \mu_t, \pi_1, \ldots, \pi_t) = \sum_j \pi_j \left[\frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{1}{2}(\mathbf{x}_i - \mu_j)^\top (\mathbf{x}_i - \mu_j)} \right]$
- Parameters $\theta = (\mu_1, \ldots, \mu_t, \pi_1, \ldots, \pi_t)$

- $\delta_{i,j} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ comes from blob } j \\ 0 & \text{otherwise} \end{cases}$
- $p(\delta_{i,j} = 1 \,|\, \theta) = \pi_j$
- $p(\delta_i \,|\, \theta) = \prod_j [\pi_j]^{\delta_{i,j}}$
- $p(\mathbf{x}_i, \delta_i \,|\, \theta) = \prod_j \left[\pi_j \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{1}{2}(\mathbf{x}_i - \mu_j)^\top (\mathbf{x}_i - \mu_j)} \right]^{\delta_{i,j}}$
- $\mathscr{L}(\theta; \mathbf{x}, \delta) = \sum_{i,j} \left[log \pi_j - \frac{1}{2}(\mathbf{x}_i - \mu_j)^\top (\mathbf{x}_i - \mu_j) \right] \delta_{i,j} + K$

EM for Mixture of Normals

1. Initialize probability distributions
2. While ($\theta^{(n)}$ has not reached convergence)
   1. E-step
      - $p(\delta \mid \theta^{(n)}, \mathbf{x})$
      - $$\begin{aligned} \mathcal{Q}(\theta; \theta^{(n)}) &= \textstyle\sum_{\delta} \mathcal{L}(\theta; \mathbf{x}, \delta) p(\delta \mid \theta^{(n)}, \mathbf{x}) \\ &= \mathbb{E}_{p(\delta \mid \theta^{(n)}, \mathbf{x})}[\mathcal{L}(\theta; \mathbf{x}, \delta)] \end{aligned}$$
      - $w_{i,j}$ to associate each item $\mathbf{x}_i$ to cluster center j
   2. M-step
      - $\theta^{(n+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta; \theta^{(n)})$

- E-Step: find weights $\mathcal{Q}(\theta; \theta^{(n)})$ from data items and $\theta^{(n)}$

$$\mathcal{Q}(\theta; \theta^{(n)}) = \textstyle\sum_{\delta} \mathcal{L}(\theta; \mathbf{x}, \delta) p(\delta \mid \theta^{(n)}, \mathbf{x})$$

- $$= \textstyle\sum_{i,j} \left(\left[log \pi_j - \frac{1}{2}(\mathbf{x}_i - \mu_j)^\top (\mathbf{x}_i - \mu_j) \right] w_{i,j} \right) + K$$

- where

- $$\begin{aligned} w_{i,j} &= p(\delta_{i,j} = 1 \mid \theta^{(n)}, \mathbf{x}) \\ &= \frac{p(\mathbf{x}, \delta_{i,j} = 1 \mid \theta^{(n)})}{\sum_l p(\mathbf{x}, \delta_{i,l} = 1 \mid \theta^{(n)})} \\ &= \frac{e^{-\frac{1}{2}(\mathbf{x}_i - \mu_j)^\top (\mathbf{x}_i - \mu_j)} \pi_j}{\sum_k e^{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^\top (\mathbf{x}_i - \mu_k)} \pi_k} \end{aligned}$$

- M-Step: parameters θ that maximize $\mathcal{Q}(\theta; \theta^{(n)})$

- $$\begin{aligned} \mu_j &= \frac{\sum_i \mathbf{x}_i w_{i,j}}{\sum_i w_{i,j}} \\ \pi_j &= \frac{\sum_i w_{i,j}}{N} \end{aligned}$$

EM for Mixture of Normals

1. Initialize probability distributions
2. While ($\theta^{(n)}$ has not reached convergence)
   1. E-step
      - weights to associate each item $\mathbf{x}_i$ to cluster centers j
      - $$w_{i,j}^{(n)} = \frac{e^{-\frac{1}{2}(\mathbf{x}_i-\mu_j)^\top(\mathbf{x}_i-\mu_j)}\pi_j^{(n)}}{\sum_k e^{-\frac{1}{2}(\mathbf{x}_i-\mu_k)^\top(\mathbf{x}_i-\mu_k)}\pi_k^{(n)}}$$
   2. M-step
      - parameters $\theta^{(n+1)}$
      - $$\mu_j^{(n+1)} = \frac{\sum_i \mathbf{x}_i w_{i,j}^{(n)}}{\sum_i w_{i,j}^{(n)}}$$ $$\pi_j^{(n+1)} = \frac{\sum_i w_{i,j}^{(n)}}{N}$$

- E-Step: find weights $\mathcal{Q}(\theta;\theta^{(n)})$ from data items and $\theta^{(n)}$
  $$\mathcal{Q}(\theta;\theta^{(n)}) = \sum_\delta \mathcal{L}(\theta;\mathbf{x},\delta)p(\delta\,|\,\theta^{(n)},\mathbf{x})$$
  - $$= \sum_{i,j}\left(\left[log\pi_j - \frac{1}{2}(\mathbf{x}_i-\mu_j)^\top(\mathbf{x}_i-\mu_j)\right]w_{i,j}\right) + K$$
  - where
  - $$w_{i,j} = p(\delta_{i,j}=1\,|\,\theta^{(n)},\mathbf{x}) = \frac{p(\mathbf{x},\delta_{i,j}=1\,|\,\theta^{(n)})}{\sum_l p(\mathbf{x},\delta_{i,l}=1\,|\,\theta^{(n)})} = \frac{e^{-\frac{1}{2}(\mathbf{x}_i-\mu_j)^\top(\mathbf{x}_i-\mu_j)}\pi_j}{\sum_k e^{-\frac{1}{2}(\mathbf{x}_i-\mu_k)^\top(\mathbf{x}_i-\mu_k)}\pi_k}$$
  - M-Step: parameters θ that maximize $\mathcal{Q}(\theta;\theta^{(n)})$
  - $$\mu_j = \frac{\sum_i \mathbf{x}_i w_{i,j}}{\sum_i w_{i,j}}$$ $$\pi_j = \frac{\sum_i w_{i,j}}{N}$$

EM for Mixtures of Normal Distributions

- Modeling clusters through Normal Distributions
- EM algorithm for mixture of normals: E-Step
- EM algorithm for mixture of normals: M-Step

Applied Machine Learning

EM for Mixtures of Normal Distributions