Clustering, Gaussian mixture model and EM

MATH-412 - Statistical Machine Learning

K-means

Key assumption: Data composed of K "roundish" clusters of similar sizes with centroids $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$.

Problem can be formulated as: $$\min_{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_K} \frac{1}{n}\sum_{i=1}^{n} \min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2.$$

Difficult (NP-hard) nonconvex problem.

K-means algorithm

1. Draw centroids at random (or use the "k-means++" initialization)
2. Assign each point to the closest centroid
$$C_k \leftarrow \left\{ i \mid \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 = \min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \right\}$$
3. Recompute centroid as center of mass of the cluster
$$\boldsymbol{\mu}_k \leftarrow \frac{1}{\mid C_k \mid} \sum_{i \in C_k} \mathbf{x}_i$$
4. Go to 2

K-means properties

Three remarks:

- K-means is a greedy algorithm
- It can be shown that K-means converges in a finite number of steps.
- The algorithm however typically gets stuck in local minima and in practice it is necessary to try several restarts of the algorithms with different initialization to have chances to obtain a better solution.
- Will typically fail if the clusters are not round or of too different sizes.

The EM algorithm for the Gaussian mixture model

Jensen's Inequality

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$

1. if f is **convex** and if X is a random variable (with $\mathbb{E}[X] \in \mathbb{R}$), then

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

2. When f is **strictly convex**, we have equality in the previous inequality if and only if X is constant almost surely.

The Kullback-Leibler divergence

Definition Let $\mathcal{X}$ a finite (or countable) state space and p and q two distributions on $\mathcal{X}$

$$KL(p \,\|\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{X\sim p}\Big[\log \frac{p(X)}{q(X)}\Big]$$

Entropy: $H(p) = -\sum_x p(x) \log p(x) \geq 0$

So $$KL(p \,\|\, q) = \mathbb{E}_{X\sim p}\big[-\log q(X)\big] - H(p).$$

Property: $\forall p, q,\ KL(p \,\|\, q) \geq 0$ (could be infinite)

Proof:

$$KL(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{X\sim p}\left[\log \frac{p(X)}{q(X)}\right] = \mathbb{E}_{X\sim p}\left[-\log \frac{q(X)}{p(X)}\right] =$$

The function $f(y) = -\log y$ is strictly convex so

$$\geq -\log \mathbb{E}_{X\sim p}\left[\frac{q(X)}{p(X)}\right] = -\log \sum_x p(x) \frac{q(X)}{p(X)} = -\log \sum_x q(x) = 0$$

with equality if and only if $p = q$ almost surely

Differential KL and entropies

Let P and Q two probability distributions with densities p and q with respect to a measure μ. Then, we can define

$$KL(p \,\|\, q) = \int_x \left(\log \frac{p(x)}{q(x)} \right) p(x) d\mu(x) = \mathbb{E}_{X \sim P} \left[\log \frac{p(X)}{q(X)} \right]$$

Differential entropy

$$\mathcal{H}(p) = -\int_x p(x) \log p(x) d\mu(x)$$

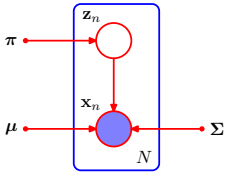
Caveats: the differential entropy is dangerous

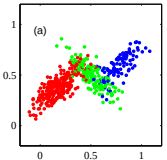
- $\mathcal{H}(p) \not\geq 0$
- $\mathcal{H}$ depends on the choice of μ...!

Gaussian mixture model

- K components
- $\boldsymbol{z} = (z_1, \ldots, z_K)^\top \in \{0,1\}^K$ indicator variable (one hot encoding)

- $\boldsymbol{z} \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$
- $p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$, so $p(e_k) = \pi_k$
- $p(\mathbf{x}|\boldsymbol{z}; (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_k) = \sum_{k=1}^{K} z_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $p(\mathbf{x}) = \sum_{z} p(\mathbf{x}|\boldsymbol{z}) p(\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Estimation: $\underset{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}{\operatorname{argmax}} \log \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]$





Applying maximum likelihood to the Gaussian mixture

Let $\mathcal{Z} = \{z \in \{0,1\}^K \mid \sum_{k=1}^K z_k = 1\}$

$$p(\mathbf{x}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} p(\mathbf{x}, \boldsymbol{z}) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]^{z_k} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Issue

- The marginal log-likelihood $\tilde{\ell}(\theta) = \sum_i \log(p(\mathbf{x}^{(i)}))$ with $\theta = \left(\boldsymbol{\pi}, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{1 \leq k \leq K}\right)$ is now complicated
- $\log p(\mathbf{x}, \boldsymbol{z}) = \sum_k \log p(\mathbf{x}, e_k) 1(\boldsymbol{z} = e_k) = \sum_k z_k \log(\pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$
- By contrast the complete log-likelihood has a rather simple form:

$$\tilde{\ell}(\theta) = \sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \boldsymbol{z}^{(i)}) = \sum_{i,\ k} z_k^{(i)} \log \mathcal{N}(x^{(i)}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k),$$

Principle of the Expectation-Maximization Algorithm

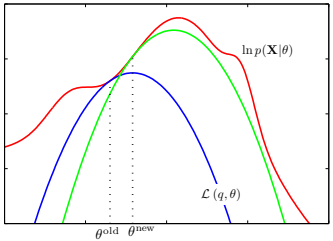
$$\begin{aligned}\log p(\mathbf{x};\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{z}} p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta}) = \log \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \frac{p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})} \\ &\geq \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \frac{p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})}{q(\boldsymbol{z})} = \mathbb{E}_q[\log p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})] + H(q) =: \mathcal{L}(q,\boldsymbol{\theta})\end{aligned}$$

- This shows that $\mathcal{L}(q,\boldsymbol{\theta}) \leq \log p(\mathbf{x};\boldsymbol{\theta})$
- $\boldsymbol{\theta} \mapsto \mathcal{L}(q,\boldsymbol{\theta})$ is *often*[a] concave or easy to maximize
- It is possible to show that

$$\mathcal{L}(q,\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta}) - KL(q \,\|\, p(\cdot|\mathbf{x};\boldsymbol{\theta}))$$

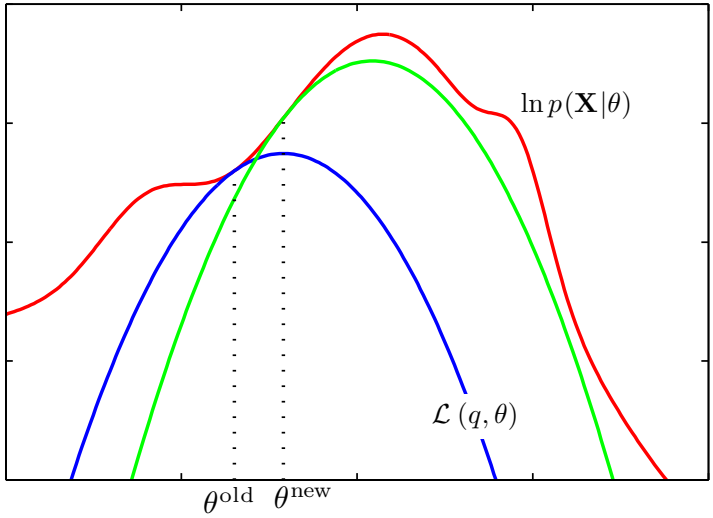
So that if we set $q(\boldsymbol{z}) = p(\boldsymbol{z} \mid \mathbf{x};\boldsymbol{\theta}^{(t)})$ then

$$\mathcal{L}(q,\boldsymbol{\theta}^{(t)}) = \log p(\mathbf{x};\boldsymbol{\theta}^{(t)}).$$



[a] It is concave if $p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta})\}_{\theta\in\Theta}$ is a canonical exponential family, with θ its natural parameter, i.e., $p(\mathbf{x},\boldsymbol{z};\boldsymbol{\theta}) = h(\mathbf{x},\boldsymbol{z}) \exp(\phi(\mathbf{x},\boldsymbol{z})^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}))$.

A graphical idea of the EM algorithm



Expectation Maximization algorithm

Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

WHILE (Not converged)

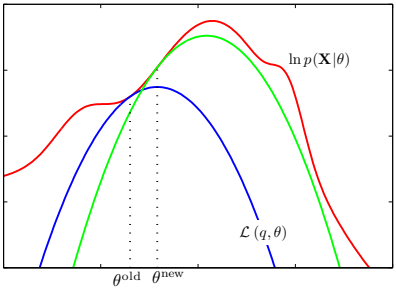
Expectation step

1. $q(\boldsymbol{z}) = p(\boldsymbol{z} \mid \mathbf{x}; \boldsymbol{\theta}^{(t-1)})$
2. $\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_q\big[\log p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta})\big] + H(q)$

Maximization step

1. $\boldsymbol{\theta}^{(t)} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \mathbb{E}_q\big[\log p(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta})\big]$

ENDWHILE



$$\boldsymbol{\theta}^{\text{old}} = \boldsymbol{\theta}^{(t-1)}$$

$$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{(t)}$$

Expected complete log-likelihood

With the notation: $q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$, we have

$$
\begin{aligned}
\mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right] &= \mathbb{E}_{q^{(t)}}\left[\log p(\boldsymbol{X}, \boldsymbol{Z}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i=1}^{M} \log p(\mathbf{x}^{(i)}, \boldsymbol{z}^{(i)}; \theta)\right] \\
&= \mathbb{E}_{q^{(t)}}\left[\sum_{i,k} z_k^{(i)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} z_k^{(i)} \log(\pi_k)\right] \\
&= \sum_{i,\,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right] \log(\pi_k) \\
&= \sum_{i,\,k} q_{ik}^{(t)} \log \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i,k} q_{ik}^{(t)} \log(\pi_k)
\end{aligned}
$$

Expectation step for the Gaussian mixture

We computed previously $q_i^{(t)}(\boldsymbol{z}^{(i)})$, which is a multinomial distribution defined by

$$q_i^{(t)}(\boldsymbol{z}^{(i)}) = p(\boldsymbol{z}^{(i)}|\mathbf{x}^{(i)}; \theta^{(t-1)})$$

Abusing notation we will denote $(q_{i1}^{(t)}, \dots, q_{iK}^{(t)})$ the corresponding vector of probabilities defined by

$$q_{ik}^{(t)} = \mathbb{P}_{q_i^{(t)}}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$$

$$q_{ik}^{(t)} = p(z_k^{(i)} = 1 \mid \mathbf{x}^{(i)}; \theta^{(t-1)}) = \frac{\pi_k^{(t-1)} \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{j=1}^{K} \pi_j^{(t-1)} \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_j^{(t-1)}, \boldsymbol{\Sigma}_j^{(t-1)})}$$

Maximization step for the Gaussian mixture

$$\left(\boldsymbol{\pi}^t, (\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})_{1\leq k\leq K}\right) = \operatorname*{argmax}_{\theta} \mathbb{E}_{q^{(t)}}\left[\tilde{\ell}(\theta)\right]$$

This yields the updates:

$$\boldsymbol{\mu}_k^{(t)} = \frac{\sum_i \mathbf{x}^{(i)}\, q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}, \qquad \boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_i \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}$$

and

$$\pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{\sum_{i,k'} q_{ik'}^{(t)}}$$

Final EM algorithm for the Gaussian mixture model

Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

WHILE (Not converged)

Expectation step

$$q_{ik}^{(t)} \leftarrow \frac{\pi_k^{(t-1)} \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{j=1}^{K} \pi_j^{(t-1)} \mathcal{N}(\mathbf{x}^{(i)}, \boldsymbol{\mu}_j^{(t-1)}, \boldsymbol{\Sigma}_j^{(t-1)})}$$

Maximization step

$$\boldsymbol{\mu}_k^{(t)} = \frac{\sum_i \mathbf{x}^{(i)} q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}, \quad \boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_i \left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)\left(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t)}\right)^\top q_{ik}^{(t)}}{\sum_i q_{ik}^{(t)}}$$

and $$\pi_k^{(t)} = \frac{\sum_i q_{ik}^{(t)}}{\sum_{i,k'} q_{ik'}^{(t)}}$$

ENDWHILE