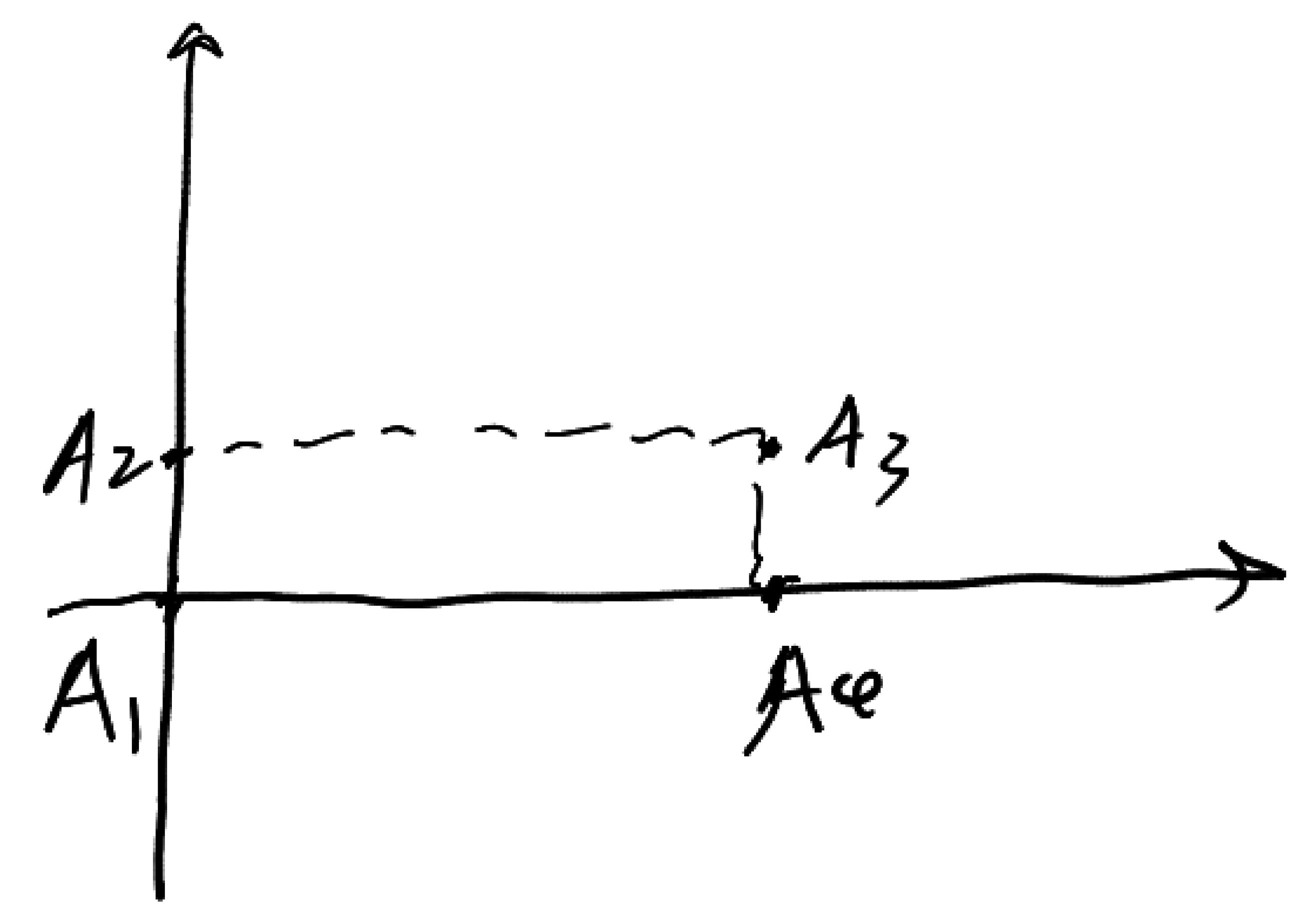
1 k-means may converge to bad local minima

We explicitly write the objective function:

Let $A_1 = (0,0)$ $A_2 = (0,1)$ $A_3 = (4,1)$ $A_4 = (4,0)$



(a) if $A_1, A_2, A_3, A_4 \in C_1$ $\alpha_1 = (2, \frac{1}{2})$

$I(C_1, C_2) = \sum_{i=1}^{4} \|A_i - \alpha_1\|^2 = 4 \cdot (4 + \frac{1}{4}) = 17$

(b) If one of $A_1 \sim A_4$ in C_2, others in C_1

WLOG $A_1, A_2, A_3 \in C_1$, $A_4 \in C_2$

$\alpha_1 = (\frac{4}{3}, \frac{2}{3})$

$I(C_1, C_2) = \sum_{i=1}^{3} \|A_i - \alpha_1\|^2 + 0 = 2 \times \frac{20}{9} + \frac{65}{9} = \frac{35}{3}$

(c) If $A_1, A_2 \in C_1$, $A_3, A_4 \in C_2$, $I(C_1, C_2) = 2 \times \frac{1}{4} \times 2 = 1$

(d) If $A_1, A_4 \in C_1$, $A_2, A_3 \in C_2$, $I(C_1, C_2) = 2 \times 2 \times 2^2 = 16$

(e) If $A_1, A_3 \in C_1$, $A_2, A_4 \in C_2$, $I(C_1, C_2) = 4 \times (4 + \frac{1}{4}) = 17$

We have exhausted all situations. Global minimum is case (c) with $I(C_1, C_2) = 1$

Case (d) is a bad local minima with $I(C_1, C_2) = 16$

Consider initialization $\alpha_1^{(0)} = (2, 0)$ $\alpha_2^{(0)} = (2, 1)$

Assignment step: $C_1^{(1)} = \{A_1, A_4\}$ $C_2^{(1)} = \{A_2, A_3\}$

Update step: $\alpha_1^{(1)} = \frac{A_1 + A_4}{2} = \alpha_1^{(0)}$ $\alpha_2^{(1)} = \frac{A_2 + A_3}{2} = \alpha_2^{(0)}$

Thus $(\alpha_1^{(t)}, \alpha_2^{(t)}) = (\alpha_1^{(t-1)}, \alpha_2^{(t-1)})$ for $t = 1, 2, \ldots$

and we converge to case (d), a bad local minima. #

2 Kernel k-means

$$d^2_{kernel}(x, C_k) := \|\Phi(x) - \alpha_k\|^2_{\mathcal{H}}$$

$$= \langle \Phi(x) - \alpha_k, \Phi(x) - \alpha_k \rangle_{\mathcal{H}}$$

$$= \langle \Phi(x) - \frac{1}{|C_k|}\sum_{i \in C_k} \Phi(x_i), \Phi(x) - \frac{1}{|C_k|}\sum_{j \in C_k} \Phi(x_j) \rangle_{\mathcal{H}}$$

$$= \frac{1}{|C_k|^2} \langle |C_k| \Phi(x) - \sum_{i \in C_k} \Phi(x_i), |C_k| \Phi(x) - \sum_{j \in C_k} \Phi(x_j) \rangle_{\mathcal{H}}$$

$$= \frac{1}{|C_k|^2} \langle \sum_{i \in C_k} (\Phi(x) - \Phi(x_i)), \sum_{j \in C_k} (\Phi(x) - \Phi(x_j)) \rangle$$

$$= \frac{1}{|C_k|^2} \sum_{i \in C_k} \sum_{j \in C_k} \langle \Phi(x) - \Phi(x_i), \Phi(x) - \Phi(x_j) \rangle$$

$$= \frac{1}{|C_k|^2} \sum_{i,j \in C_k} \Big(\langle \Phi(x), \Phi(x) \rangle - \langle \Phi(x_i), \Phi(x) \rangle - \langle \Phi(x), \Phi(x_j) \rangle + \langle \Phi(x_i), \Phi(x_j) \rangle \Big)$$

$$= \frac{1}{|C_k|^2} \sum_{i,j \in C_k} \big(k(x, x) - k(x_i, x) - k(x, x_j) + k(x_i, x_j) \big)$$

#

3 The EM algorithm for GMM of $d>1$

(a) We have

($\text{denote } \phi(x, \mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$.

n: training size

$X = (x^{(1)} \dots x^{(n)})$ data)

$$Q(\theta|\theta^t) := \mathbb{E}_{z|X,\theta^t}\left[\log p(X, z|\theta)\right]$$

$$= \sum_{i=1}^{n} \mathbb{E}_{z^{(i)}|x^{(i)},\theta^t}\left[\log p(x^{(i)}, z^{(i)}|\theta)\right]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} p(z^{(i)}=k|x^{(i)},\theta^t) \log\left[p(x^{(i)}, z^{(i)}|\theta)\right]$$

$$\gamma_k := p(z=k|x) = \frac{p(z=k)p(x|z=k)}{p(x)}$$

$$= \frac{p(z=k)p(x|z=k)}{\sum_{j=1}^{K} p(z=j)p(x|z=j)}$$

$$= \frac{\pi_k \phi(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \phi(x|\mu_j, \Sigma_j)}$$

We have $\gamma_k^{(i)} := p(z^{(i)}=k|x^{(i)},\theta^t) = \frac{\pi_k^t \phi(x^{(i)}, \mu_k^t, \Sigma_k^t)}{\sum_{j=1}^{K} \pi_j^t \phi(x^{(i)}, \mu_k^t, \Sigma_k^t)}$

$$Q(\theta|\theta^t) = \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_k^{(i)}\left[\log(p(z^{(i)}=k|\theta)) + \log p(x^{(i)}|z^{(i)}=k,\theta)\right]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_k^{(i)}\left(\log \pi_k + \log \phi(x^{(i)}, \mu_k, \Sigma_k)\right)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_k^{(i)} \log \pi_k + \sum_{k=1}^{K}\sum_{i=1}^{n} \gamma_k^{(i)} \log \phi(x^{(i)}, \mu_k, \Sigma_k)$$

#

(b) Optimize with μ_k and Σ_k are unconstrained optimization:

$$Q(\theta|\theta^t) = \sum_{k=1}^{K}\sum_{i=1}^{n} \gamma_k^{(i)}\left[-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x^{(i)}-\mu_k)^T \Sigma_k^{-1} (x^{(i)}-\mu_k)\right]$$

$$0 = \frac{\partial Q(\theta|\theta^t)}{\partial \mu_k} \Rightarrow \sum_{i=1}^{n} r_k^{(i)} \Sigma_k^{-1} (x^{(i)} - \hat{\mu}_k) = 0$$

$$\Rightarrow \sum_{i=1}^{n} r_k^{(i)} (x^{(i)} - \hat{\mu}_k) = 0 \quad (\Sigma_k^{-1} \text{ is PD})$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} r_k^{(i)} x^{(i)}}{\sum_{i=1}^{n} r_k^{(i)}}$$

$$0 = \frac{\partial Q(\theta|\theta^t)}{\partial \Sigma_k^{-1}} \Rightarrow \sum_{i=1}^{n} r_k^{(i)} \left(\frac{1}{2} \Sigma_k^{-1} - \frac{1}{2} (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^T \right) = 0$$

$$\sum_{i=1}^{n} r_k^{(i)} \Sigma_k = \sum_{i=1}^{n} r_k^{(i)} (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^T$$

$$\hat{\Sigma}_k = \frac{\sum_{i=1}^{n} r_k^{(i)} (x^{(i)} - \hat{\mu}_k)(x^{(i)} - \hat{\mu}_k)^T}{\sum_{i=1}^{n} r_k^{(i)}}$$

(used derivation w.r.t. matrices)

Optimization w.r.t. π_k is constrained optimization

$$\max \sum_{i=1}^{n} \sum_{k=1}^{K} r_k^{(i)} \log \pi_k$$

$$\text{s.t.} \sum_k \pi_k = 1$$

Lagranrian $\mathcal{L}(\pi, \mu) = - \sum_{i=1}^{n} \sum_{k=1}^{K} r_k^{(i)} \log \pi_k + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right), \lambda \geq 0$

KKT condition reads

$$\begin{cases} \nabla_{\pi_k} \mathcal{L}(\pi, \mu) = 0 \Rightarrow - \sum_{i=1}^{n} r_k^{(i)} \frac{1}{\pi_k} + \lambda = 0 \\ \sum_k \pi_k = 1 \\ \lambda \geq 0 \end{cases}$$

This implies $\lambda = \sum_{k=1}^{K} \sum_{i=1}^{n} r_k^{(i)}$

$$\text{and } \pi_k = \frac{\sum_{i=1}^{n} r_k^{(i)}}{\sum_{k=1}^{K} \sum_{i=1}^{n} r_k^{(i)}} = \frac{\sum_{i=1}^{n} r_k^{(i)}}{\sum_{i=1}^{n} \sum_{k=1}^{K} r_k^{(i)}} = \frac{\sum_{i=1}^{n} r_k^{(i)}}{\sum_{i=1}^{n} 1} = \frac{\sum_{i=1}^{n} r_k^{(i)}}{n}$$

Thus, algorithm is:

E-step : $$\gamma_k^{(i)t} = \frac{\pi_k^t \phi(x^{(i)}, \mu_k^t, \Sigma_k^t)}{\sum_{j=1}^{K} \pi_j^t \phi(x^{(i)}, \mu_k^t, \Sigma_k^t)}$$

M-step: $$N_k^t = \sum_{i=1}^{n} \gamma_k^{(i)t}$$

$$\mu_k^{t+1} = \frac{1}{N_k^t} \sum_{i=1}^{n} \gamma_k^{(i)t} x^{(i)}$$

$$\Sigma_k^{t+1} = \frac{1}{N_k^t} \sum_{i=1}^{n} \gamma_k^{(i)t} (x^{(i)} - \mu_k^{t+1})(x^{(i)} - \mu_k^{t+1})^T$$

$$\pi_k^{t+1} = \frac{N_k^t}{n}$$

We are done. #