Hands on session 2 : Clustering

Exercise 2 : On the consistency of K-Means

$$crit(G) = \sum_{k=1}^{K} \sum_{a \in G_k} \|X_a - \bar{X}_{G_k}\|^2 \quad \text{with} \quad \bar{X}_{G_k} = \frac{1}{|G_k|} \sum_{b \in G_k} X_b$$

1. $$crit(G) = \sum_{k=1}^{K} \sum_{a \in G_k} \left\langle X_a - \frac{1}{|G_k|} \sum_{b \in G_k} X_b \,,\, X_a - \bar{X}_{G_k} \right\rangle$$

$$= \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \langle X_a - X_b \,,\, X_a - \bar{X}_{G_k} \rangle$$

$$= \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \left(\langle X_a - X_b \,,\, X_a \rangle - \langle X_a - X_b \,,\, \bar{X}_{G_k} \rangle \right)$$

and $$\sum_{a,b \in G_k} \langle X_a - X_b \,,\, X_{G_k} \rangle = \sum_{a \in G_k} \langle |G_k| X_a - |G_k| \bar{X}_{G_k} \,,\, X_{G_k} \rangle$$

$$= 0$$

So, $$\boxed{crit(G) = \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \langle X_a \,,\, X_a - X_b \rangle}$$

$$\forall a, b \in G_k, \quad \|X_b - X_a\|^2 = \langle X_a \,,\, X_a - X_b \rangle + \langle X_b \,,\, X_b - X_a \rangle$$

So, $$2\, crit(G) = \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_b - X_a\|^2$$

So $$\boxed{crit(G) = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{|G_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2}$$

2. $X_a = \mu_a + \varepsilon_a$, $\varepsilon_1, \dots, \varepsilon_n$ centered and independent

$$\forall a, b, \quad \|X_a - X_b\|^2 = \|\mu_a - \mu_b\|^2 + 2\langle \varepsilon_a \,,\, \mu_a - \mu_b \rangle + 2\langle \varepsilon_b \,,\, \mu_a - \mu_b \rangle + \|\varepsilon_a\|^2 + \|\varepsilon_b\|^2 - 2\langle \varepsilon_a \,,\, \varepsilon_b \rangle$$

So, $\forall a,b$, $E(\|X_b - X_a\|^2) = \|\mu_a - \mu_b\|^2 + 2\langle E(\varepsilon_a), \mu_a - \mu_b\rangle + 2\langle E(\varepsilon_b), \mu_a - \mu_b\rangle$
$+ E(\|\varepsilon_a\|^2) + E(\|\varepsilon_b\|^2) - 2\langle E(\varepsilon_a), E(\varepsilon_b)\rangle$
since ε_a and ε_b are indept if $a \neq b$ (0 otherwise)

So, $\forall a,b$, $E(\|X_b - X_a\|^2) = (\|\mu_a - \mu_b\|^2 + v_a + v_b)\mathbb{1}_{a \neq b}$
where $v_a = \mathrm{tr}(\mathrm{cov}(X_a)) = E(\|\varepsilon_a\|^2)$
since $E(\varepsilon_a) = E(\varepsilon_b) = 0$

So, $$E(\mathrm{crit}(G)) = \frac{1}{2}\sum_{\ell=1}^{K} \frac{1}{|G_\ell|} \sum_{a,b \in G_\ell} (\|\mu_a - \mu_b\|^2 + v_a + v_b)\mathbb{1}_{a \neq b}$$

3 MM: $\exists\, m_1, \dots, m_K \in \mathbb{R}^p$, $\exists\, r_1, \dots, r_K > 0$

a) St $\forall \ell \in [\![1, K]\!]$, $\forall a \in G_\ell^*$, $\mu_a = m_\ell$ and $v_a = r_\ell$

then $E(\mathrm{crit}(G^*)) = \frac{1}{2}\sum_{\ell=1}^{K} \frac{1}{|G_\ell^*|} \sum_{a,b \in G_\ell^*} (\|m_\ell - m_\ell\|^2 + r_\ell + r_\ell)\mathbb{1}_{a \neq b}$

$$= \frac{1}{2}\sum_{\ell=1}^{K} \frac{1}{|G_\ell^*|} 2 r_\ell (|G_\ell^*| - 1)\,|G_\ell^*|$$

so $$E(\mathrm{crit}(G^*)) = \sum_{\ell=1}^{K} r_\ell (|G_\ell^*| - 1)$$

b) If $r_1 = \dots = r_K$, $G = G^*$ minimizes $E(\mathrm{crit}(G))$

c) $\|m_2 - m_3\|^2 = \|(0,1,0) - (0, 1-\tau, \sqrt{1-(1-\tau)^2})\|^2$

$= (1-(1-\tau))^2 + 1 - (1-\tau)^2 = 1 + \tau^2 - (1-\tau)^2 = 2\tau$

So, $\|m_2 - m_3\|^2 = 2\tau$

d) $E(\mathrm{crit}(G^*)) = (r_+ + 2r_-)(S-1)$

e) $\sum_{a,b \in G_1} (\|\mu_a - \mu_b\|^2 + v_a + v_b) \mathbb{1}_{a \neq b} = \sum_{a,b \in G_1} (v_a + v_b) \mathbb{1}_{a \neq b}$

$= \left(\frac{s}{2}\right)\left(\frac{s}{2} - 1\right)(2\gamma_+)$

$\sum_{a,b \in G_2} (\|\mu_a - \mu_b\|^2 + v_a + v_b) \mathbb{1}_{a \neq b} = \left(\frac{s}{2}\right)\left(\frac{s}{2} - 1\right)(2\gamma_+)$

$\sum_{a,b \in G_3} (\|\mu_a - \mu_b\|^2 + v_a + v_b) \mathbb{1}_{a \neq b} =$

$s(s-1)2\gamma_- + s(s-1)2\gamma_- + s^2(2\tau + 2\gamma_-) + s^2(2\tau + 2\gamma_-)$

$S_0, E(\text{crit}(G')) = \frac{1}{2}[2(s-2)\gamma_+ + (s-1)\gamma_- + (s-1)\gamma_- + 2s(\tau + \gamma_-)]$

$$E(\text{crit}(G')) = s(\gamma_+ + 2\gamma_- + \tau) - (2\gamma_+ + \gamma_-)$$

f) $E(\text{crit}(G^*)) < E(\text{crit}(G'))$ iff $(\gamma_+ + 2\gamma_-)(s-1) < s(\gamma_+ + 2\gamma_- + \tau) - (2\gamma_+ + \gamma_-)$

iff $\tau > \frac{1}{s}[(\gamma_+ + 2\gamma_-)(s-1) + 2\gamma_+ + \gamma_-] - \gamma_+ - 2\gamma_-$

iff $\boxed{\tau > (\gamma_+ - \gamma_-)/s}$

g) The optimal solution of k-means for the MM is the generative model only if the model is well-separable.