Luis Buonaventura

CMSC 25025
Assignment 2

Problem 1

a) We know that $\mathbb{E}(R(\hat{C}_n) - R(C^*)) \leq c\sqrt{\frac{k(d+1)\log n}{n}}$ when $\mathbb{P}(\|X\|^2 \leq B) = 1$ for some $B < \infty$ and for $(x_1, \dots, x_n)$ data for $x_i \in \mathbb{R}^d$.
So, we know that empirical risk $R(\hat{C})$ is close to optimal risk $R(C^*)$ ~~for~~ large k.
We see on slide 37 of our clustering slides that a regular implementation of k-means clustering can result in effectively poor clustering when there are unbalanced clusters.
So, even though the empirical risk may be close to optimal risk within some bound, the clustering may not be good.

b) Let $R^{(k)}$ be the minimal risk among all possible clusterings with k clusters.
We want show that $R^{(k)}$ is non-increasing in k.
So we want to prove that

$$R^{(k+1)} \leq R^{(k)}$$

$$\Rightarrow R^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} \min_{1 \leq j \leq k+1} \|x_i - c_j\|^2 \quad R^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \min_{1 \leq j \leq k} \|x_i - c_j\|^2$$

$$R^{(k+1)} = \sum_{i=1}^{n} \min_{1 \leq j \leq k+1} \|x_i - c_j\|^2, \quad R^{(k)} = \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|x_i - c_j\|^2$$

We can decompose $\{x_1, \dots, x_n\}$ into two sets,
$A := \{x_i \mid \min_{1 \leq j \leq k+1} \|x_i - c_j\|^2 = \min_{1 \leq j \leq k} \|x_i - c_j\|^2\}$ and
$B := \{x_i \mid \|x_i - c_{k+1}\|^2 < \min \|x_i - c_j\|^2\}$

$$R^{(k+1)} = \sum_{x \in A} \min_{1 \leq j \leq k} \|x_i - c_j\|^2 + \sum_{x \in B} \min \|x_i - c_{k+1}\|^2$$

Note: $\min_{1 \leq j \leq k+1} \|x_i - c_j\|^2 \neq \min_{1 \leq j \leq k} \|x_i - c_j\|^2$

So, $R^{(k+1)} - R^{(k)}$

$$= \left(\sum_{x\in A} \min_{1\le j\le k} \|x_i - c_j\|^2 + \sum_{x\in B} \|x_i - c_{k+1}\|^2 \right) - \left(\sum_{x\in A} \min_{1\le j\le k} \|x_i - c_j\|^2 + \sum_{x\in B} \min_{1\le j\le k} \|x_i - c_j\|^2 \right)$$

$$= \sum_{x\in B} \left(\|x_i - c_{k+1}\|^2 - \min_{1\le j\le k} \|x_i - c_j\|^2 \right) \le 0 \quad \text{by construct}$$

So, $R^{(k+1)} \le R^{(k)}$.

c) We wish to consider $\min_{\mu, \{d_i\}, V_k} \sum_{i=1}^{n} \|x_i - \mu - V_k d_i\|^2$

for $x_i \in \mathbb{R}^n$, $\mu \in \mathbb{R}^n$, $V_k \in \mathbb{R}^{d\times k}$, $d_i \in \mathbb{R}^k$.

First we wish find the optimums given by $\hat{\mu}$ and $\hat{d}_i$.

$$\min_{\mu, \{d_i\}} \sum_{i=1}^{n} \|x_i - \mu - V_k d_i\|^2 = \min_{\mu, \{d_i\}} \sum_{i=1}^{n} (x_i - \mu - V_k d_i)^T (x_i - \mu - V_k d_i)$$

$$\rightarrow \min_{\mu, \{d_i\}} \sum_{i=1}^{n} (x_i^T x_i - x_i^T \mu - x_i^T V_k d_i - \mu^T x_i + \mu^T \mu + \mu^T V_k d_i - d_i^T V_k^T x_i + d_i^T V_k^T \mu + d_i^T d_i)$$

$[\mu]$ $\sum_{i=1}^{n} -x_i - x_i + 2\mu + 2 V_k d_i = 0$

$\Rightarrow n\mu = \sum_{i=1}^{n} x_i - V_k d_i$

$\Rightarrow n\mu = n\bar{x} - \sum_{i=1}^{n} V_k d_i$

$[d_i]$ $-V_k^T x_i + V_k^T \mu - V_k^T x_i + V_k^T \mu + 2 d_i = 0$

$\Rightarrow d_i = V_k^T \mu - V_k^T x_i$

So, $n\mu = n\bar{x}_n - \sum_{i=1}^{n} V_k (V_k^T \mu - V_k^T x_i)$

$\rightarrow n\mu = n\bar{x}_n - n\mu^k + n\bar{x}_k$

$\Rightarrow 2n\mu = n(\bar{x}_n + \bar{x}_k)$ Supposing $\bar{x}_n = \bar{x}_k$

$\mu = \bar{x}_n$

and $d_i = V_k^T(\mu - x_i)$

Now we wish to discuss the uniquess of $\hat{\mu}$.

Since $V_k^T : \mathbb{R}^k \to \mathbb{R}^d$ and $V_k : \mathbb{R}^d \to \mathbb{R}^k$,

We know $V_k V_k^T : \mathbb{R}^d \to \mathbb{R}^d$ but $\dim(V_k V_k^T x) \le k \le d$

$$x \longmapsto V_k V_k^T x$$

Since this is not full rank when $k < d$,

$\hat{\mu}$ is not unique.

$\hat{\mu} = \frac{1}{2}(\bar{x}_n + \bar{x}_k)$.