Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2018

Homework 2

YOUR NAME Guoqing Zhang November 26, 2018

- **Acknowledgments:** This template takes some materials from course CSE 547/Stat 548 of Washington University: `https://courses.cs.washington.edu/courses/cse547/17sp/index.html`.

  If you refer to other materials in your homework, please list here.

- **Collaborators:** I finish this homework by myself.

3.1 (a) i. First, it's obvious that we can just focus on the part: $\sum_{x\in C_j}\|x-\mu_j\|^2$ and $\frac{1}{2|C_j|}\sum_{x,x'\in C_j}\|x-x'\|^2$.

$$\begin{aligned}
\sum_{x\in C_j}\|x-\mu_j\|^2 &= \sum_{x\in C_j}\left(\|x\|^2 - 2x^T\mu_j + \|\mu_j\|^2\right)\\
&= \sum_{x\in C_j}\|x\|^2 - \frac{2}{|C_j|}\sum_{x\in C_j}x^T\sum_{x'\in C_j}x' + \frac{1}{|C_j|^2}\sum_{x\in C_j}\sum_{x' inC_j}x'^T\sum_{x''\in C_j}x''\\
&= \sum_{x\in C_j}\|x\|^2 - \frac{2}{|C_j|}\sum_{x,x'\in C_j}x^Tx' + \frac{|C_j|}{|C_j|^2}\sum_{x,x'\in C_j}x^Tx'\\
&= \sum_{x\in C_j}\|x\|^2 - \frac{1}{|C_j|}\sum_{x,x'\in C_j}x^Tx'
\end{aligned}$$

$$\begin{aligned}
\frac{1}{2|C_j|}\sum_{x,x'\in C_j}\|x-x'\|^2 &= \frac{1}{2|C_j|}\sum_{x,x'\in C_j}\left(\|x\|^2 - 2x^Tx' + \|x'\|^2\right)\\
&= \frac{1}{|C_j|}\sum_{x,x'\in C_j}\|x\|^2 - \frac{1}{|C_j|}\sum_{x,x'\in C_j}x^Tx'\\
&= \sum_{x\in C_j}\|x\|^2 - \frac{1}{|C_j|}\sum_{x,x'\in C_j}x^Tx'
\end{aligned}$$

So they are equivalent.
ii. Because of i, we could know:

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{x\in C_j}\|x-\mu_j\|^2 &= \sum_{j=1}^{k}\sum_{x\in C_j}\left(\|x\|^2 - \frac{1}{|C_j|}\sum_{x,x'\in C_j}x^Tx'\right)\\
&= \sum_{i=1}^{m}\|x\|^2 - \sum_{j=1}^{k}\frac{1}{|C_j|}\sum_{x,x'\in C_j}x^Tx'\\
&= A - \sum_{j=1}^{k}\frac{1}{|C_j|}\sum_{x,x'\in C_j}x^Tx'
\end{aligned}$$

Where $A = \sum_{i=1}^{m} \|x\|^2$ is a constant.
So
$argmin_C(A - \sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{x,x' \in C_j} x^T x') = argmax_c(\sum_{j=1}^{k} \frac{1}{|C_j|} \sum_{x,x' \in C_j} x^T x')$
First,let's just focus on the $\sum_{j=1}^{k} |C_i||C_j|\|\mu_i - \mu_j\|^2$.

$$\begin{aligned}
\sum_{j=1}^{k} |C_i||C_j|\|\mu_i - \mu_j\|^2 &= |C_i| \sum_{j=1}^{k} |C_j| \left(\|\mu_i\|^2 - 2\mu_i^T \mu_j + \|\mu_j\|^2\right) \\
&= \sum_{j=1}^{k} \left(\frac{|C_j|}{|C_i|} \sum_{x,x' \in C_i} x^T x' - 2 \sum_{x \in C_i, x' \in C_j} x^T x' + \frac{|C_i|}{|C_j|} \sum_{x,x' \in C_j} x^T x' \right)
\end{aligned}$$

Then, combine the whole fomulation:

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{k} |C_i||C_j|\|\mu_i - \mu_j\|^2 &= \sum_{i=1}^{k}\sum_{j=1}^{k} \left(\frac{|C_j|}{|C_i|} \sum_{x,x' \in C_i} x^T x' + \frac{|C_i|}{|C_j|} \sum_{x,x' \in C_j} x^T x' - 2 \sum_{x \in C_i, x' \in C_j} x^T x' \right) \\
&= 2\sum_{i=1}^{k}\sum_{j=1}^{k} \left(\frac{|C_j|}{|C_i|} \sum_{x,x' \in C_i} x^T x' - \sum_{x \in C_i, x' \in C_j} x^T x' \right) \\
&= 2\sum_{i=1}^{k} \frac{m}{|C_i|} \sum_{x,x' \in C_i} x^T x' - 2\sum_{i=1}^{k}\sum_{j=1}^{k} \sum_{x \in C_i, x' \in C_j} x^T x' \\
&= 2\sum_{i=1}^{k} \frac{m + |C_i|}{|C_i|} \sum_{x,x' \in C_i} x^T x' - 2\left(\sum_{i=1}^{k}\sum_{j=1}^{k} \sum_{x \in C_i, x' \in C_j} x^T x' + \sum_{i=1}^{k} \sum_{x,x' \in C_i} x^T x' \right) \\
&= 2\sum_{i=1}^{k} \frac{m + |C_i|}{|C_i|} \sum_{x,x' \in C_i} x^T x' - A \\
&= 2\sum_{i=1}^{k} (\frac{m}{|C_i|} + 1) \sum_{x,x' \in C_i} x^T x' - A
\end{aligned}$$

Where A is a constant, it's equal to all the pair's product in X. Now I have some trouble to eliminate '1'. It's a little embarrassing. I am not sure which step is wrong, or I can't proof it in this way.
I also see something about the law of total variance could help to prove it. But I am not very clear about it, so I won't write down.
(b) i. If the algorithm has converged, than the μ would be never change, so it's obvious that the distortion will increase.
The distortion could be compute at two states. First, from x we got μ, compute J, then, we reassign the clusters, compute J.
It's obvious that in the second state, the J will not increase, because:

$$J(\{c^{(i)}\}_{i=1}^{m}, \{\mu_j\}_{j=1}^{k}) = \sum_{i=1}^{m} \|x^{(i) - \mu_{c^i}}\|^2$$

And the x is reassigned to the closest μ_j, which means:$\|x^{(i) - \mu'_{c^{(i)}}}\|^2 \leq \|x^{(i) - \mu_{c^{(i)}}}\|^2$. Each term will not increase, so the distortion will also not increase.

Then what we want to prove is:

$$argmin_p \sum_{x \in C_j} \|x - p\|^2 = \mu_j$$

Assume $l(p) = \sum_{x \in C_j} \|x - p\|^2$.

$$\begin{aligned}
\frac{\partial l}{\partial p} &= \frac{\partial \sum_{x \in C_j} \left(\|x\|^2 - 2x^T p + \|p\|^2\right)}{\partial p} \\
&= \sum_{x \in C_j} 2(p - x) \\
&= 2|C_j|p - 2\sum_{x \in C_j} x
\end{aligned}$$

Then let $\frac{\partial l}{\partial p} = 0$, we could get:

$$p = \frac{1}{|C_j|} \sum_{x \in C_j} x = \mu_j$$

This means J will not increase in the first state.
The Lloyd's algorithm just iterates the two steps, so the distortion will not increase.
ii. It will always converge. First, the distortion J have the lower bound:$J \geq 0$. If we consider the valuse of J computed from each iteration as a sequence, we know that Monotonous bounded sequence has a convergence. Frome i, we know that this sequence is monotonous, which means this algorhithm will converge.

3.2 (a) i.

$$\begin{aligned}
\mu_T \operatorname{Cov}(x)\mu &= \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_d \end{bmatrix} \begin{bmatrix} \operatorname{Cov}(x_1, x_1) & \operatorname{Cov}(x_1, x_2) & \cdots & \operatorname{Cov}(x_1, x_d) \\ \operatorname{Cov}(x_2, x_1) & \operatorname{Cov}(x_2, x_2) & \cdots & \operatorname{Cov}(x_2, x_d) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(x_d, x_1) & \operatorname{Cov}(x_d, x_2) & \cdots & \operatorname{Cov}(x_d, x_d) \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} \\
&= \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_d \end{bmatrix} \begin{bmatrix} D(x_1) & \operatorname{Cov}(x_1, x_2) & \cdots & \operatorname{Cov}(x_1, x_d) \\ \operatorname{Cov}(x_2, x_1) & D(x_2) & \cdots & \operatorname{Cov}(x_2, x_d) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(x_d, x_1) & \operatorname{Cov}(x_d, x_2) & \cdots & D(x_d) \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix} \\
&= \sum_{i=1}^{d} \mu_i^2 D(x_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} 2\mu_i \mu_j \operatorname{Cov}(x_i, x_j) \\
&= \sum_{i=1}^{d} D(\mu_i x_i) + \sum_{i=1}^{d-1} \sum_{j=i+1}^{d} 2 \operatorname{Cov}(\mu_i x_i, \mu_j x_j) \\
&= D(\mu_1 x_1 + \mu_2 x_2 + \cdots \mu_d x_d) \\
&\geq 0
\end{aligned}$$

We could do this because:

$$D(x) + D(y) + 2\operatorname{Cov}(x, y) = D(x + y), \operatorname{Cov}(x, z) + \operatorname{Cov}(y, z) = \operatorname{Cov}(x + y, z)$$

So:

$$\begin{aligned}
&D(x) + D(y) + D(z) + 2\operatorname{Cov}(x, y) + 2\operatorname{Cov}(x, z) + 2\operatorname{Cov}(y, z) \\
&= D(x + y) + D(z) + 2\operatorname{Cov}(x, z) + 2\operatorname{Cov}(y, z) \\
&= D(x + y) + D(z) + 2\operatorname{Cov}(x + y, z) \\
&= D(x + y + z)
\end{aligned}$$

In the same way, we could increase 3 variables to n variables, so until now I finish the proof.
ii. From i, we could get:

$$tr(\operatorname{Cov}(x)) = \sum_{i=1}^{d} D(x_i)$$

$$\begin{aligned}
\mathbb{E}[\|x - \mathbb{E}[x]\|^2] &= \mathbb{E}[(x_1 - \mathbb{E}[x_1])^2 + ... + (x_d - \mathbb{E}[x_d])^2] \\
&= \sum_{i=1}^{d} \mathbb{E}[(x_i - \mathbb{E}[x_i])^2] \\
&= \sum_{i=1}^{d} D(x_i)
\end{aligned}$$

So, we get that $tr(\operatorname{Cov}(x)) = \mathbb{E}[\|x - \mathbb{E}[x]\|^2]$.
(b) If we want to the $\hat{C}$ is non-singular, where $\hat{C}$ is a $d \times d$ matrix. The intuition is that $m \geq d$.
And the intuition is true. Because the rank of the matrix $\hat{C}$ is limited by the rank of X, where

$$X_{m \times d} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_m^T \end{bmatrix}.$$

So we could see: $r(X) \leq \min(m, d)$. So we could get that $m \geq d$.