CMSC 25400 Homework 1

Deqing Fu

October 5, 2017

1 Problem 1

1.1 (a)

As both J_{IC} and J_{avg^2} sum over j from 1 to k, we only consider the j-th entry of J_{avg^2} and J_{IC}. That is we need to prove that $(J_{IC})_j = 2(J_{avg^2})$, which is

$$(J_{IC})_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \sum_{\mathbf{x}' \in C_j} d(\mathbf{x}, \mathbf{x}')^2 = 2 \sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mathbf{m_j})^2 = 2(J_{avg^2})_j$$

First, Let's talk about $(J_{IC})_j$.

$$\begin{aligned}
(J_{IC})_j &= \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \sum_{\mathbf{x}' \in C_j} d(\mathbf{x}.\mathbf{x}')^2 \\
&= \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \sum_{\mathbf{x}' \in C_j} (||\mathbf{x}||^2 + ||\mathbf{x}'||^2 - 2\mathbf{x}^T\mathbf{x}') \\
&= \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} (\sum_{\mathbf{x}' \in C_j} ||\mathbf{x}||^2 + \sum_{\mathbf{x}' \in C_j} ||\mathbf{x}'||^2 - 2 \sum_{\mathbf{x}' \in C_j} \mathbf{x}^T\mathbf{x}') \\
&= \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} (|C_j|||\mathbf{x}||^2 + \sum_{\mathbf{x}' \in C_j} ||\mathbf{x}'||^2 - 2\mathbf{x}^T \sum_{\mathbf{x}' \in C_j} \mathbf{x}') \\
&= \sum_{\mathbf{x} \in C_j} ||\mathbf{x}^2|| + \sum_{\mathbf{x}' \in C_j} ||\mathbf{x}'||^2 - \frac{2}{|C_j|} \sum_{\mathbf{x} \in C_j} \sum_{\mathbf{x}' \in C_j} \mathbf{x}^T\mathbf{x}' \\
&= 2 \sum_{\mathbf{x} \in C_j} ||\mathbf{x}^2|| - \frac{2}{|C_j|} \sum_{\mathbf{x} \in C_j} \sum_{\mathbf{x}' \in C_j} \mathbf{x}^T\mathbf{x}' \\
&= 2(\sum_{\mathbf{x} \in C_j} ||\mathbf{x}^2|| - \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \sum_{\mathbf{x}' \in C_j} \mathbf{x}^T\mathbf{x}')
\end{aligned}$$

On the other hand of $(J_{avg^2})_j$,

$$\begin{aligned}
(J_{avg^2})_j &= \sum_{\mathbf{x}\in C_j} d(\mathbf{x}, \mathbf{m_j})^2 \\
&= \sum_{\mathbf{x}\in C_j} (||\mathbf{x}||^2 + ||\mathbf{m_j}||^2 - 2\mathbf{x}^T\mathbf{m_j}) \\
&= \sum_{\mathbf{x}\in C_j} ||\mathbf{x}||^2 + \sum_{\mathbf{x}\in C_j} ||\frac{1}{|C_j|}\sum_{\mathbf{x}'\in C_j}\mathbf{x}'||^2 - \sum_{\mathbf{x}\in C_j} 2(\mathbf{x}^T(\frac{1}{|C_j|}\sum_{\mathbf{x}'\in C_j}\mathbf{x}')) \\
&= \sum_{\mathbf{x}\in C_j} ||\mathbf{x}||^2 + |C_j|\cdot(\frac{1}{|C_j|^2}(\sum_{\mathbf{x}'\in C_j}\mathbf{x}'\cdot\sum_{\mathbf{x}'\in C_j}\mathbf{x}')) - \frac{2}{|C_j|}\sum_{\mathbf{x}\in C_j}\sum_{\mathbf{x}'\in C_j}\mathbf{x}^T\mathbf{x}' \\
&= \sum_{\mathbf{x}\in C_j} ||\mathbf{x}||^2 + \frac{1}{|C_j|}\sum_{\mathbf{x}\in C_j}\mathbf{x}\cdot\sum_{\mathbf{x}'\in C_j}\mathbf{x}' - \frac{2}{|C_j|}\sum_{\mathbf{x}\in C_j}\sum_{\mathbf{x}'\in C_j}\mathbf{x}^T\mathbf{x}' \\
&= \sum_{\mathbf{x}\in C_j} ||\mathbf{x}||^2 + \frac{1}{|C_j|}\sum_{\mathbf{x}\in C_j}\sum_{\mathbf{x}'\in C_j}\mathbf{x}^T\mathbf{x}' - \frac{2}{|C_j|}\sum_{\mathbf{x}\in C_j}\sum_{\mathbf{x}'\in C_j}\mathbf{x}^T\mathbf{x}' \\
&= \sum_{\mathbf{x}\in C_j} ||\mathbf{x}||^2 - \frac{1}{|C_j|}\sum_{\mathbf{x}\in C_j}\sum_{\mathbf{x}'\in C_j}\mathbf{x}^T\mathbf{x}'
\end{aligned}$$

Hence, we can see from above that,

$$(J_{IC})_j = 2(\sum_{\mathbf{x}\in C_j} ||\mathbf{x}^2|| - \frac{1}{|C_j|}\sum_{\mathbf{x}\in C_j}\sum_{\mathbf{x}'\in C_j}\mathbf{x}^T\mathbf{x}') = 2(J_{avg^2})_j, \forall j \in \mathbb{N} \text{ and } 1 \leq j \leq k$$

Hence,

$$J_{IC} = \sum_{j=1}^{k}(J_{IC})_j = \sum_{j=1}^{k} 2(J_{avg^2})_j = 2J_{avg^2}$$

1.2 (b)

1) On one hand, after step one of re-labelling each datapoint to a new cluster, $\forall \mathbf{x_i}, i \in \{1, 2, 3, \cdots, k\}$, note the newly assigned cluster label be γ_i'. Hence,

$$\forall j \in \{1, 2, \cdots, k\}, d(\mathbf{x_i}, \mathbf{m}_{\gamma_\mathbf{i}'}) \leq d(\mathbf{x_i}, \mathbf{m_j})$$

Hence, as $\gamma_i \in \{1, 2, 3, \cdots, k\}$,

$$\forall j \in \{1, 2, \cdots, k\}, d(\mathbf{x_i}, \mathbf{m}_{\gamma_\mathbf{i}'}) \leq d(\mathbf{x_i}, \mathbf{m}_{\gamma_\mathbf{i}})$$

Thus,

$$\begin{aligned} J_{avg^2}(\gamma'_1, \gamma'_2, \cdots, \gamma'_k, \mathbf{m_1}, \mathbf{m_2}, \cdots, \mathbf{m_k}) &= \sum_{i=1}^{k} d(\mathbf{x_i}, \mathbf{m}_{\gamma'_\mathbf{i}}) \\ &\leq \sum_{i=1}^{k} d(\mathbf{x_i}, \mathbf{m}_{\gamma_\mathbf{i}}) \\ &= J_{avg^2}(\gamma_1, \cdots, \gamma_k, \mathbf{m_1}, \cdots, \mathbf{m_k}) \end{aligned}$$

2) On the other hand, after step two of updating the centroids, suppose the mean of the points of cluster j is $\mathbf{m'_j} = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$. Hence

$$\begin{aligned} \sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mathbf{m_j}) &= \sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{m_j}||^2 \\ &= \sum_{\mathbf{x} \in C_j} ||(\mathbf{x} - \mathbf{m'_j}) + (\mathbf{m'_j} - \mathbf{m_j})||^2 \\ &= \sum_{\mathbf{x} \in C_j} (||\mathbf{x} - \mathbf{m'_j}||^2 + ||\mathbf{m_j} - \mathbf{m'_j}||^2 - 2(\mathbf{m_j} - \mathbf{m'_j})^T(\mathbf{x} - \mathbf{m'_j})) \\ &= \sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{m'_j}||^2 + \sum_{\mathbf{x} \in C_j} ||\mathbf{m_j} - \mathbf{m'_j}||^2 - 2(\mathbf{m_j} - \mathbf{m'_j})^T \sum_{\mathbf{x} \in C_j} (\mathbf{x} - \mathbf{m'_j}) \\ &= \sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{m'_j}||^2 + \sum_{\mathbf{x} \in C_j} ||\mathbf{m_j} - \mathbf{m'_j}||^2 - 2(\mathbf{m_j} - \mathbf{m'_j})^T (\sum_{\mathbf{x} \in C_j} \mathbf{x} - \sum_{\mathbf{x} \in C_j} \mathbf{m'_j}) \\ &= \sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{m'_j}||^2 + \sum_{\mathbf{x} \in C_j} ||\mathbf{m_j} - \mathbf{m'_j}||^2 - 2(\mathbf{m_j} - \mathbf{m'_j})^T (\sum_{\mathbf{x} \in C_j} \mathbf{x} - |C_j| \cdot (\frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x})) \\ &= \sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{m'_j}||^2 + \sum_{\mathbf{x} \in C_j} ||\mathbf{m_j} - \mathbf{m'_j}||^2 - 2(\mathbf{m_j} - \mathbf{m'_j})^T \cdot 0 \\ &= \sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{m'_j}||^2 + \sum_{\mathbf{x} \in C_j} ||\mathbf{m_j} - \mathbf{m'_j}||^2 \end{aligned}$$

Let

$$f(\mathbf{m_j}) = \sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mathbf{m_j}) = \sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{m'_j}||^2 + \sum_{\mathbf{x} \in C_j} ||\mathbf{m_j} - \mathbf{m'_j}||^2$$

It's clear that $f(\mathbf{m_j})$ is minimized when $\mathbf{m_j} = \mathbf{m'_j}$ as $\sum_{\mathbf{x} \in C_j} ||\mathbf{x} - \mathbf{m'_j}||^2$ is a constant in terms of $\mathbf{m_j}$. Hence, $\sum_{\mathbf{x} \in C_j} d(\mathbf{x}, \mathbf{m_j})$ is minimized when

$$\mathbf{m_j} = \mathbf{m'_j} = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

1.3 (c)

This is an immediate corollary of b), as for every iteration, the distortion function decreases. Thus, the distortion function decreases monotonically.

1.4 (d)

There are only finite numbers of ways of arranging data points into groups. And the number of ways of grouping can be expressed as the following formula, if there are k clusters and n data points: the number of ways is k^n. Thus, k^n is the upper bound.

2 Problem 2

2.1 k-means

Figure 1: Result of k-means Algorithm

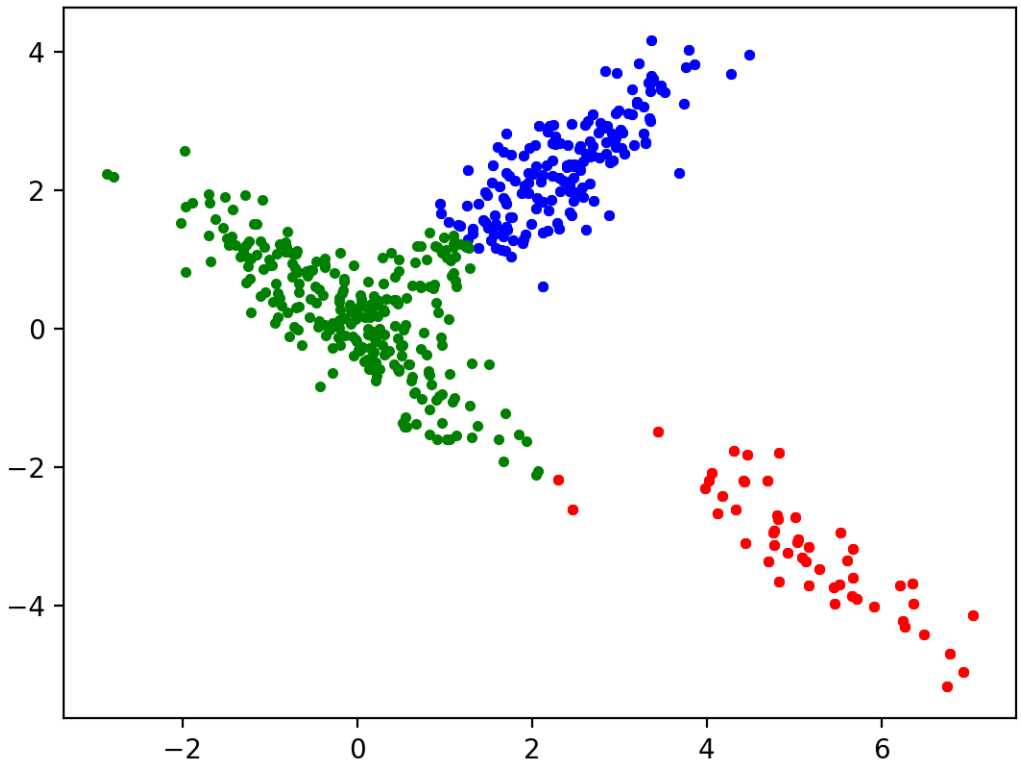
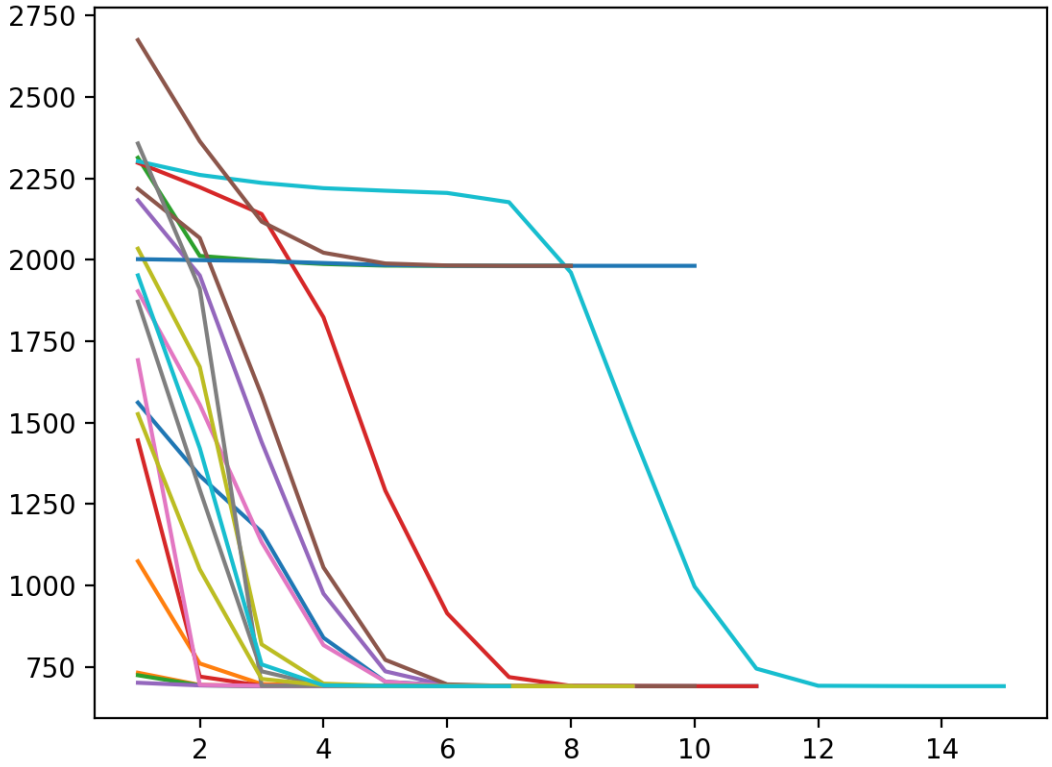


Figure 2: Distortion Function of k-means Algorithm



For the k-means algorithm, the distortion function converges to two values, the lower one is the global minimum, which should be the correct clustering and the other one is the local minimum, which generates the incorrect clustering pattern.

2.2 k-means ++

Figure 3: Result of k-means ++ Algorithm

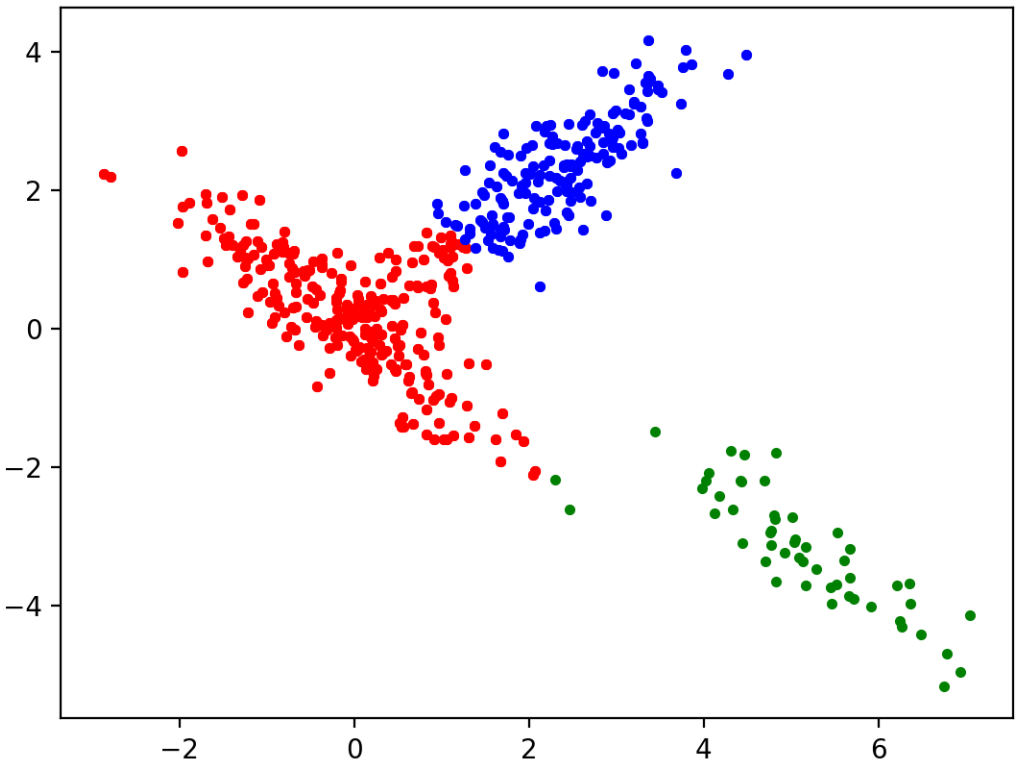
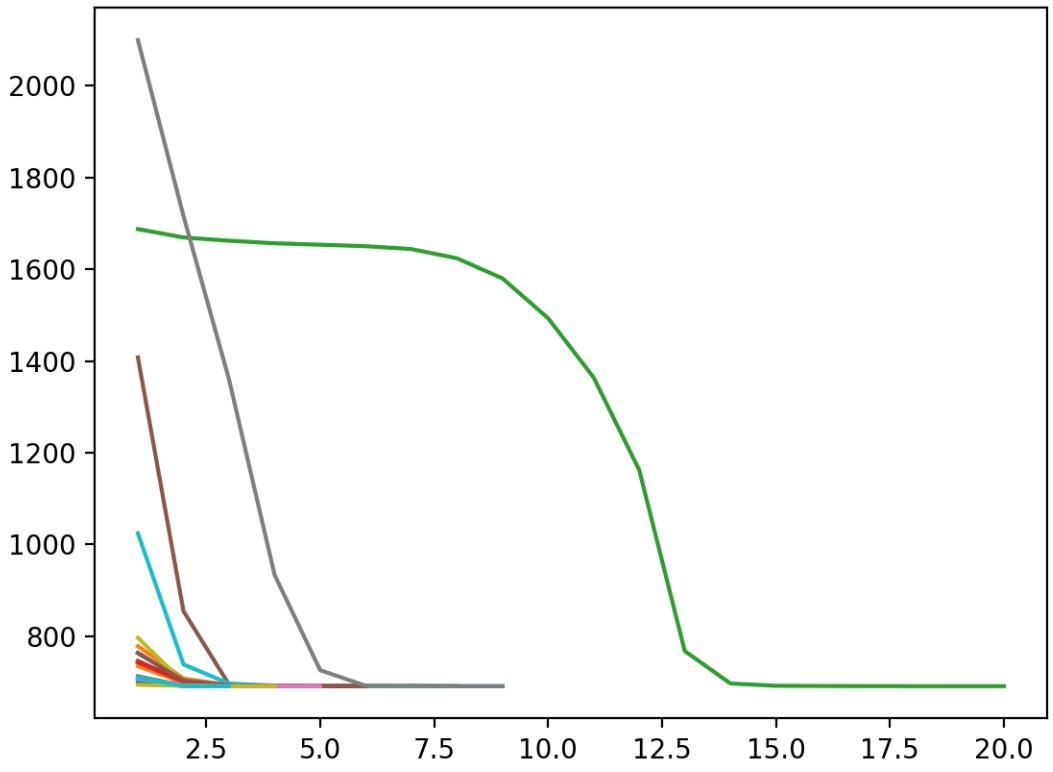


Figure 4: Distortion Function of k-means++ Algorithm



For the k-means ++ algorithm, the distortion function is more likely to converge to the global minimum because of the optimized initialization of centroids.