CS 7641 CSE/ISYE 6740 Homework 1

Deadline: Sep. 26 Monday, 11:55pm

- Submit your answers as an electronic copy on T-square.
- No unapproved extension of deadline is allowed. Zero credit will be assigned for late submissions. Email request for late submission may not be replied.
- For typed answers with LaTeX (recommended) or word processors, extra credits will be given. If you handwrite, try to be clear as much as possible. No credit may be given to unreadable handwriting.
- Explicitly mention your collaborators if any.
- Recommended reading: PRML[1] Section 9.1, 12.1

1 Probability [15 pts]

(a) Stores A, B, and C have 50, 75, and 100 employees and, respectively, 50, 60, and 70 percent of these are women. Resignations are equally likely among all employees, regardless of stores and sex. Suppose an employee resigned, and this was a woman. What is the probability that she has worked in store C? [5 pts]

(a)Answer:

Number of female employee in store A, B and C:

$$N_A = 50 \times 50\% = 25, N_B = 75 \times 60\% = 45, N_C = 100 \times 70\% = 70$$

Given an female employee resigned, then we have:

$$P(Woman) = \frac{25+45+70}{50+75+100} = \frac{28}{45}$$

$$P(Store_C, Woman) = \frac{70}{50+75+100} = \frac{14}{45}$$

$$P(Store_C|Woman) = \frac{P(Store_C, Woman)}{P(Woman)} = \frac{\frac{14}{45}}{\frac{28}{45}} = \frac{1}{2}$$

(b) A laboratory blood test is 95 percent effective in detecting a certain disease when it is, in fact, present. The test also yields a false positive result for 1 percent of the healthy persons tested. That is, if a healthy person is tested then with probability 0.01 the test result will imply he has the disease. If 0.5 percent of the population actually has the disease, what is the probability a person has the disease given that his test result is positive? [5 pts]

(b)Answer:

$$P(Positive|True) = 95\%, \quad P(Positive|False) = 1\%, \quad P(True) = 0.5\%$$

[1]Christopher M. Bishop, Pattern Recognition and Machine Learning, 2006, Springer.

Given a test result is positive, then we have:

$$P(Positive) = P(Positive|True)P(True) + P(Positive|False)(1 - P(True)) = \frac{147}{10000}$$

$$P(True|Positive) = \frac{P(True, Positive)}{P(Positive)} = \frac{P(True)P(Positive|True)}{P(Positive)} = \frac{0.5\% \cdot 95\%}{\frac{147}{10000}} = \frac{95}{294} \approx 32.31\%$$

[c-d] On the morning of September 31, 1982, the won-lost records of the three leading baseball teams in the western division of the National League of the United States were as follows:

Team	Won	Lost
Atlanta Braves	87	72
San Francisco Giants	86	73
Los Angeles Dodgers	86	73

Each team had 3 games remaining to be played. All 3 of the Giants games were with the Dodgers, and the 3 remaining games of the Braves were against the San Diego Padres. Suppose that the outcomes of all remaining games are independent and each game is equally likely to be won by either participant. If two teams tie for first place, they have a playoff game, which each team has an equal chance of winning.

(c) What is the probability that Atlanta Braves wins the division? [2 pts]

(c)Answer:

A stands for Atlanta Braves, S stands for San Francisco Giants and L stands for Los Angeles Dodgers.
(1)Atlanta Braves win 3 games:

$$P(A = 90) = (\frac{1}{2})^3 = \frac{1}{8}$$

(2)Atlanta Braves win 2 games and San Francisco Giants or Los Angeles Dodgers win 2 games:

$$P(A = 89, L = 88 or 87) = (3 \cdot (\frac{1}{2})^3 \cdot 3 \cdot (\frac{1}{2})^3) \cdot 2 = \frac{9}{32}$$

(3)Atlanta Braves win 2 games and San Francisco Giants or Los Angeles Dodgers win 3 games, additional playoff game is needed:

$$P(A = 89, L = 89 or S = 89) = (3 \cdot (\frac{1}{2})^3 \cdot (\frac{1}{2})^3 \cdot \frac{1}{2}) \cdot 2 = \frac{3}{64}$$

(4)Atlanta Braves win 1 games and San Francisco Giants or Los Angeles Dodgers win 2 games, additional playoff game is needed:

$$P(A = 88, L = 88 or S = 88) = (3 \cdot (\frac{1}{2})^3 \cdot 3 \cdot (\frac{1}{2})^3 \cdot \frac{1}{2}) \cdot 2 = \frac{9}{64}$$

Therefore:

$$P(A) = \frac{1}{8} + \frac{9}{32} + \frac{3}{64} + \frac{9}{64} = \frac{19}{32}$$

(d) What is the probability to have an additional playoff game? [3 pts]

(d)Answer:
(1)Atlanta Braves win 2 games and San Francisco Giants or Los Angeles Dodgers win 3 games, additional playoff game is needed:

$$P(A = 89, L = 89 or S = 89) = ((\frac{1}{2})^3 \cdot 3 \cdot (\frac{1}{2})^3) \cdot 2 = \frac{3}{32}$$

(2)Atlanta Braves win 1 games and San Francisco Giants or Los Angeles Dodgers win 2 games, additional playoff game is needed:

$$P(A = 88, L = 88 or S = 88) = ((\frac{1}{2})^3 \cdot 3 \cdot (\frac{1}{2})^3 \cdot 3) \cdot 2 = \frac{9}{32}$$

Therefore:

$$P(A = 89, L = 89 or S = 89) + P(A = 88, L = 88 or S = 88) = \frac{3}{32} + \frac{9}{32} = \frac{3}{8}$$

2 Maximum Likelihood [15 pts]

Suppose we have n i.i.d (independent and identically distributed) data samples from the following probability distribution. This problem asks you to build a log-likelihood function, and find the maximum likelihood estimator of the parameter(s).

(a) Poisson distribution [5 pts]

The Poisson distribution is defined as

$$P(x_i = k) = \frac{\lambda^k e^{-\lambda}}{k!} (k = 0, 1, 2, ...).$$

What is the maximum likelihood estimator of λ?
(a)Answer:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} P(x_i; \theta) \\ \ln L(\theta) &= \sum_{i=1}^{n} \ln P(x_i; \theta) = \sum_{i=1}^{n} \ln(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}) = -n\lambda + \sum_{i=1}^{n} (x_i \ln \lambda - \ln(x_i!)) \\ Let: \frac{d \ln(L)}{d\lambda} &= -n + \sum_{i=1}^{n} \frac{x_i}{\lambda} = 0 \\ \Rightarrow \lambda = \hat{\theta} &= argmax((\ln(\theta))) = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x} \end{aligned}$$

(b) Multinomial distribution [5 pts]

The probability density function of Multinomial distribution is given by

$$f(x_1, x_2, \ldots, x_k; n, \theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{x_1! x_2! \cdots x_k!} \prod_{j=1}^{k} \theta_j^{x_j},$$

where $\sum_{j=1}^{k} \theta_j = 1, \sum_{j=1}^{k} x_j = n$. What is the maximum likelihood estimator of $\theta_j, j = 1, \ldots k$?
(b)Answer: The distribution function subjects to $\sum_{j=1}^{k} \theta_j = 1$, we need to satisfy both functions. Thus

we introduce a Lagrange multiplier λ.

$$
\begin{aligned}
l(\theta_j) &= f(x_1, x_2, \ldots, x_k; n, \theta_1, \theta_2, \ldots, \theta_k) \\
\ln(l(\theta_j)) &= \ln(\frac{n!}{x_1!x_2!\cdots x_k!}) + \ln(\prod_{j=1}^{k} \theta_j^{x_j}) = \ln(\frac{n!}{x_1!x_2!\cdots x_k!}) + \sum_{j=1}^{k} x_j \ln(\theta_j) \\
L(\theta_j) &= \ln(l(\theta_j)) - \lambda(\sum_{j=1}^{k} \theta_j - 1) = \ln(\frac{n!}{x_1!x_2!\cdots x_k!}) + \sum_{j=1}^{k} x_j \ln(\theta_j) - \lambda(\sum_{j=1}^{k} \theta_j - 1) \\
\Rightarrow \frac{\partial L(\theta_j)}{\partial \theta_j} &= \frac{x_j}{\theta_j} - \lambda = 0 \quad \Rightarrow \theta_j = \frac{x_j}{\lambda}
\end{aligned}
$$

Since $\sum_{j=1}^{k} \theta_j = 1, \quad \sum_{j=1}^{k} x_j = n,$

$$
\begin{aligned}
\Rightarrow \lambda &= \sum_{j=1}^{k} \frac{x_j}{\theta_j} = \frac{n}{1} = n \\
\theta_j &= \frac{x_j}{n} \quad (x_i = 0, 1, 2, \ldots)
\end{aligned}
$$

(c) Gaussian normal distribution [5 pts]

Suppose we have n i.i.d (Independent and Identically Distributed) data samples from a univariate Gaussian normal distribution $\mathcal{N}(\mu, \sigma^2)$, which is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

What is the maximum likelihood estimator of μ and σ^2?
(c)Answer:
Since we have ni.i.d data samples, then the likelihood function is:

$$
\begin{aligned}
L(x_1, x_2 \cdots x_n; \mu, \sigma^2) &= \prod_{i=1}^{n} N(x; \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{(-n/2)} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)
\end{aligned}
$$

$$\ln(L(x_1, x_2 \cdots x_n; \mu, \sigma^2)) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\Rightarrow \frac{\partial \ln(L)}{\partial \mu} = -\frac{1}{2\sigma^2} \cdot 2\sum_{i=1}^{n}(x_i - \mu)(-1) = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma^2} = 0$$

$$\Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Let $\sigma^2 = y$,

$$\ln(L(x_1, x_2 \cdots x_n; \mu, \sigma^2)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(y) - \frac{1}{2y}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\Rightarrow \frac{\partial \ln(L)}{\partial y} = -\frac{n}{2y} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2} \cdot y^{-2} = 0$$

$$\Rightarrow y = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

$$\Rightarrow \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

3 Principal Component Analysis [20 pts]

In class, we learned that Principal Component Analysis (PCA) preserves variance as much as possible. We are going to explore another way of deriving it: minimizing reconstruction error.

Consider data points $\mathrm{x}^n (n = 1, ..., N)$ in D-dimensional space. We are going to represent them in $\{\mathrm{u}_1, ..., \mathrm{u}_D\}$ orthonormal basis. That is,

$$\mathrm{x}^n = \sum_{i=1}^{D} \alpha_i^n \mathrm{u}_i = \sum_{i=1}^{D} (\mathrm{x}^{n^T} \mathrm{u}_i)\mathrm{u}_i.$$

Here, α_i^n is the length when x^n is projected onto u_i.

Suppose we want to reduce the dimension from D to $M < D$. Then the data point x^n is approximated by

$$\tilde{\mathrm{x}}^n = \sum_{i=1}^{M} z_i^n \mathrm{u}_i + \sum_{i=M+1}^{D} b_i \mathrm{u}_i.$$

In this representation, the first M directions of u_i are allowed to have different coefficient z_i^n for each data point, while the rest has a constant coefficient b_i. As long as it is the same value for all data points, it does not need to be 0.

Our goal is setting u_i, z_i^n, and b_i for $n = 1, ..., N$ and $i = 1, ..., D$ so as to minimize reconstruction error. That is, we want to minimize the difference between x^n and $\tilde{\mathrm{x}}^n$ over $\{\mathrm{u}_i, z_i^n, b_i\}$:

$$J = \frac{1}{N}\sum_{n=1}^{N} \|\mathrm{x}^n - \tilde{\mathrm{x}}^n\|^2.$$

(a) What is the assignment of z_j^n for $j = 1, ..., M$ minimizing J? [5 pts]

(a)Answer:

$$
\begin{aligned}
J &= \frac{1}{N}\sum_{n=1}^{N} \|\mathrm{x}^n - \tilde{\mathrm{x}}^n\|^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}(\sum_{i=1}^{M}(\alpha_i^n - z_i^n)^2 + \sum_{i=M+1}^{D}(\alpha_i^n - b_i)^2) \\
\text{Let} \quad \frac{\partial J}{\partial z_i^n} &= \frac{1}{N}\cdot(-2)\cdot(\alpha_i^n - z_i^n) = 0 \\
\Rightarrow z_i^n &= \alpha_i^n = (x^{nT}u_i)
\end{aligned}
$$

(b) What is the assignment of b_j for $j = M + 1, ..., D$ minimizing J? [5 pts]

(b)Answer:

$$
\begin{aligned}
\text{Let} \quad \frac{\partial J}{\partial b_i} &= \frac{-2}{N}\sum_{n=1}^{N}(\alpha_i^n - b_i) = 0 \\
\Rightarrow b_i &= \frac{1}{N}\sum_{n=1}^{N}\alpha_i^n = \bar{a_i} = \frac{1}{N}\sum_{n=1}^{N}(x^{nT}u_i)
\end{aligned}
$$

(c) Express optimal $\tilde{\mathrm{x}}^n$ and $\mathrm{x}^n - \tilde{\mathrm{x}}^n$ using your answer for (a) and (b). [2 pts]

(c)Answer:

$$
\begin{aligned}
\tilde{\mathrm{x}}^n &= \sum_{i=1}^{M}\alpha_i^n u_i + \sum_{i=M+1}^{D}\bar{a_i}u_i \\
x^n - \tilde{\mathrm{x}}^n &= \sum_{i=1}^{M}(\alpha_i^n - z_i^n)u_i + \sum_{i=M+1}^{D}(\alpha_i^n - \bar{a_i})u_i = \sum_{i=M+1}^{D}(\alpha_i^n - \bar{a_i})u_i
\end{aligned}
$$

(d) What should be the $\mathbf{u}_i$ for $i = 1, ..., D$ to minimize J? [8 pts]

Hint: Use $S = \frac{1}{N}\sum_{n=1}^{N}(\mathrm{x}^n - \bar{\mathrm{x}})(\mathrm{x}^n - \bar{\mathrm{x}})^T$ for sample covariance matrix.

4 Clustering [20 pts]

[a-b] Given N data points $\mathrm{x}^n (n = 1, \ldots, N)$, K-means clustering algorithm groups them into K clusters by minimizing the distortion function over $\{r^{nk}, \mu^k\}$

$$
J = \sum_{n=1}^{N}\sum_{k=1}^{K} r^{nk}\|\mathrm{x}^n - \mu^k\|^2,
$$

where $r^{nk} = 1$ if x^n belongs to the k-th cluster and $r^{nk} = 0$ otherwise.

(a) Prove that using the squared Euclidean distance $\|\mathbf{x}^n - \mu^k\|^2$ as the dissimilarity function and minimizing the distortion function, we will have

$$\mu^k = \frac{\sum_n r^{nk}\mathbf{x}_n}{\sum_n r^{nk}}.$$

That is, μ^k is the center of k-th cluster. [5 pts]

(a)Proof:
According to EM algorithm, in the M steps phase, we minimize J with respect to the μ^k, keeping r^{nk} fixed. Therefore, the objective function J is a quadratic function of μ^k, so let the derivative with respect to μ^k to zero will give us the minimum of J.

$$2\sum_{n=1}^{N} r^{nk}(x^n - \mu^k) = 0$$
$$\Rightarrow \mu^k = \frac{\sum_n r^{nk}x^n}{\sum_n r^{nk}}$$

(b) Prove that K-means algorithm converges to a local optimum in finite steps. [5 pts]

(b)Answer:

(1)Since we have to partition N data point x^n into k clusters, then we can have k^N ways. Therefore, there will be finite iterations to converge.

(2)The goal of k means clustering is to minimize the cost. Therefore, for each iteration, the new clustering is produced based on the merging of old clustering and closest data point. If there is nothing to merge into the old cluster, then it keeps the cost value. If the old cluster is replaced by the new one, then the cost should be lower than the old one, such that there will be a local optimal after finite iteration.

[c-d] In class, we discussed bottom-up hierarchical clustering. For each iteration, we need to find two clusters $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_m\}$ and $\{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_p\}$ with the minimum distance to merge. Some of the most commonly used distance metrics between two clusters are:

- Single linkage: the minimum distance between any pairs of points from the two clusters, i.e.

$$\min_{\substack{i=1,\ldots,m \\ j=1,\ldots,p}} \|\mathrm{x}_i - \mathrm{y}_j\|$$

- Complete linkage: the maximum distance between any parts of points from the two clusters, i.e.

$$\max_{\substack{i=1,\ldots,m \\ j=1,\ldots,p}} \|\mathrm{x}_i - \mathrm{y}_j\|$$

- Average linkage: the average distance between all pair of points from the two clusters, i.e.

$$\frac{1}{mp}\sum_{i=1}^{m}\sum_{j=1}^{p} \|\mathrm{x}_i - \mathrm{y}_j\|$$

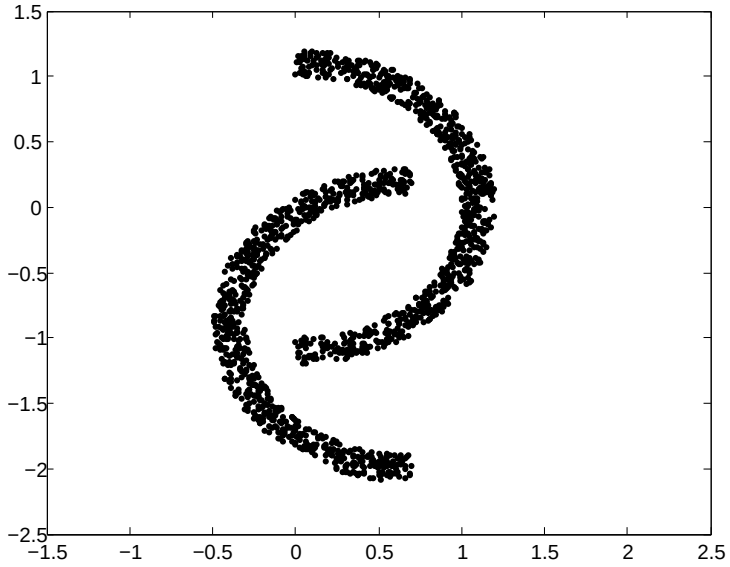
(c) When we use the bottom up hierarchical clustering to realize the partition of data, which of the three cluster distance metrics described above would most likely result in clusters most similar to those given by K-means? (Suppose K is a power of 2 in this case). [5 pts]

(c)Answer:
The average linkage will most likely result in clusters most similar to those given by K-means. For example,

if there are some extreme outliers in the dataset, then it will highly affect the other two clustering method except average linkage clustering. Because average linkage can decrease the effect of outlier data points by averaging.

(d) For the following data (two moons), which of these three distance metrics (if any) would successfully separate the two moons? [5 pts]

(d)Answer:
The single linkage will successfully separate the two moons. According to the data point layout of the moons, we can easily notice that the point is very compact in respective moon. However, the distance between pair of points in different moon is larger than the distance between pair of points in the same moon. As to single linkage clustering, the initial clustering will start clustering between points pair in the same moon first.(Since it is closer than in different moons.) Therefore, when we choose $k = 2$, single linkage clustering can separate the two moons.



5 Programming: Image compression [30 pts]

Report

(1) K-medoids frameworks implementation:

1. Randomly select K data point from the pixels vector as my initial centroids.
2. Associate each data point to the closest centroid.
3. Within the same cluster, I choose the data points that are closest to the centroids as my representatives. Reassign this representatives as new centroids and repeat the previous steps until the centroids stop changing or the number of iterations is out of threshold (I set the maximum iteration as 100).
4. Also, I used Euclidean distance, Manhattan distance and Chebyshev distance to test the performance in my algorithms. And I chose Euclidean distance since it has a better performance.

(2) By applying my K-medoids and K-means algorithm with my own picture, with several different K, with small values 2 and 3, large value 16 and 32. Here are the results:

According to the above figures and table, we can easily conclude:

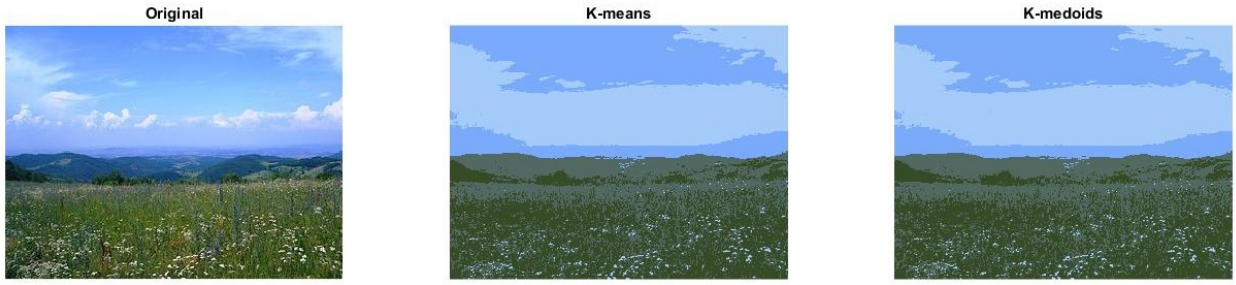


Figure 1: Result of Euclidean distance of K-medoids when k=4

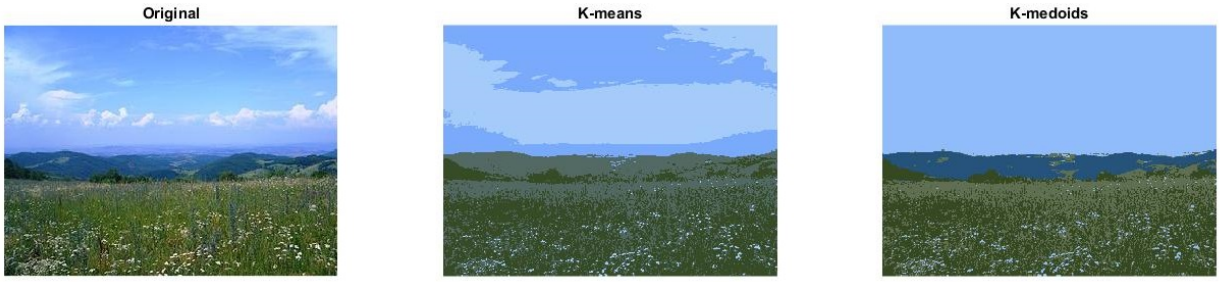


Figure 2: Result of Manhattan distance of K-medoids when k=4

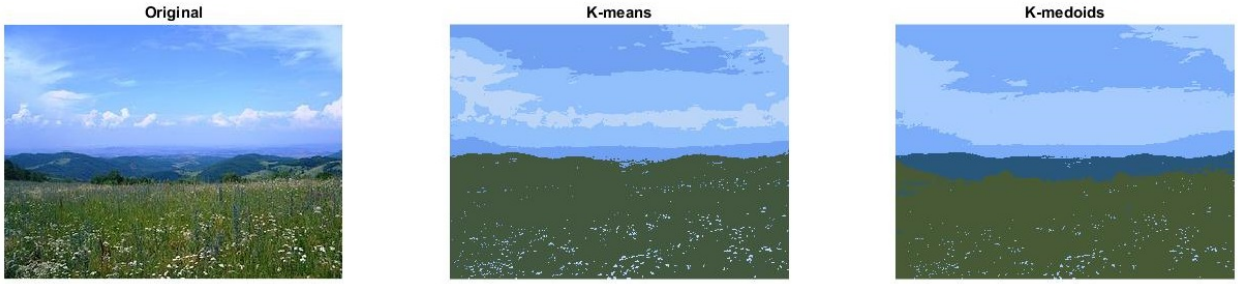


Figure 3: Result of Chebyshev distance of K-medoids when k=4

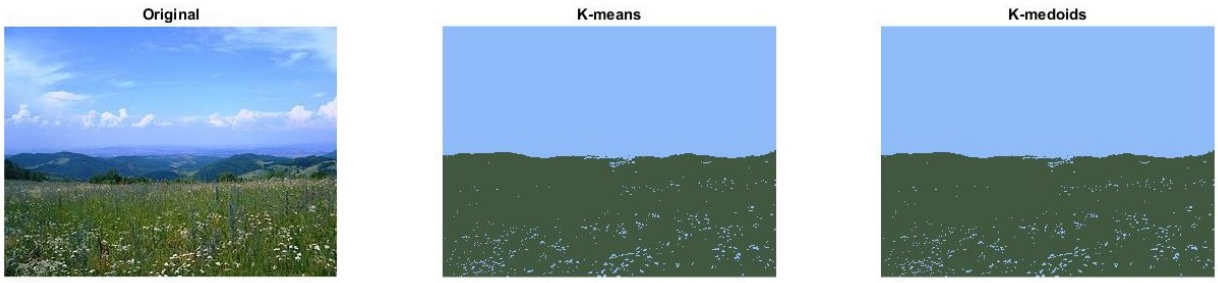


Figure 4: Result of my own pictures when k=2

1. As the K value increasing, the output images are gradually become more and more similar to the original one. Since the number of cluster, K represents the number of colors involved in the resemble

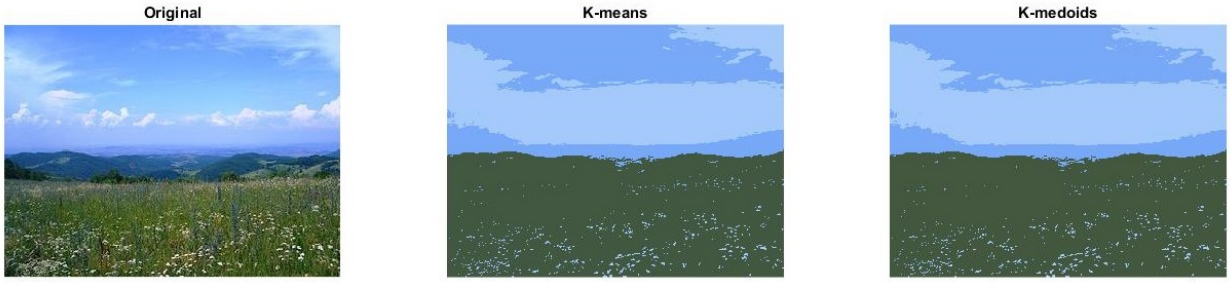


Figure 5: Result of my own pictures when k=3

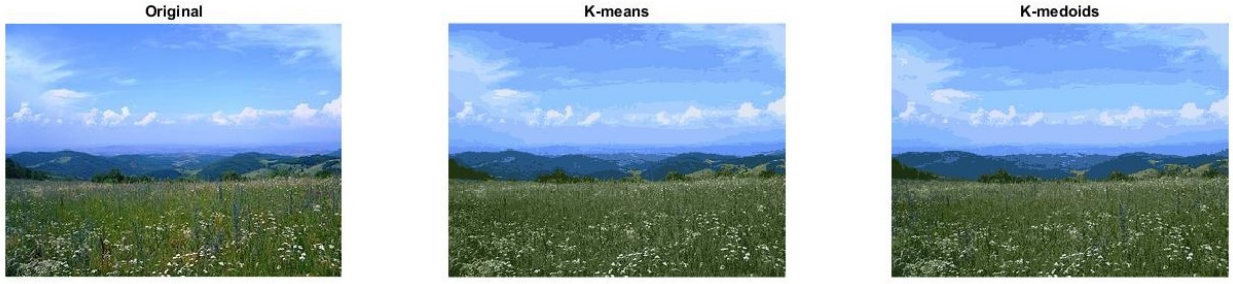


Figure 6: Result of my own pictures when k=16

Figure 7: Result of my own pictures when k=32

of the output image. Therefore, the larger the K, the more similar to the original image.

2. Besides, as the K value increasing, the total time to converge of both K-means and K-medoids are increasing. And the time to converge of K-means is longer than K-medoids.

3. Run my K-medoids implementation with different initial centroids doesn't have a huge impact to final result. I tested my K-medoids with some poor initialization. However the output image doesn't appear too much difference to the previous outcome. Besides, we can have a better outcome by tuning the maximum number of iteration even if the initialization is poor.

4. By running my K-means algorithm, I can notice the significant difference between K-means and K-medoids is the consuming time. K-means algorithm runs a longer time than K-medoids. Moreover, K-means algorithm might get a relatively poor performance since it might be affected by the outlier data points. And the K-mediods algorithm is more robust to outliers.

K	K-medoids	K-means
2	2.0907	2.1179
3	4.327	4.3491
16	10.9428	24.3967
32	17.5913	41.8031

Table 1: Running time in different K with different algorithm (sec.)

Metrics	K-medoids	K-means
Euclidean Distance	3.7696	6.2902
Chebyshev Distance	4.9426	9.2706
Manhattan Distance	5.3758	8.2425

Table 2: Running time in $K = 4$ with different distance metrics (sec.)

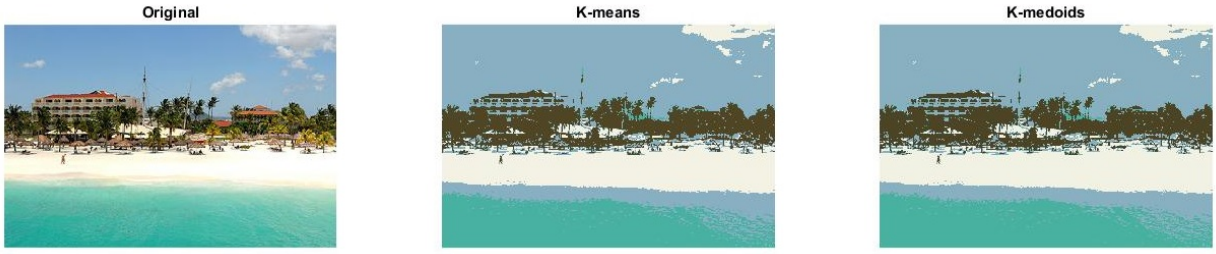


Figure 8: Result of beach.bmp when k=4

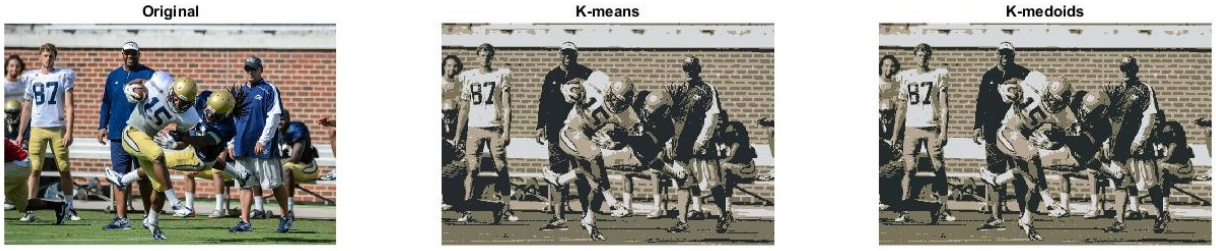


Figure 9: Result of football.bmp when k=4