ISYE 6740 Fall 2020
Homework 1 solution

1 Clustering [25 points]

Given m data points x^i, $i = 1, \ldots, m$, K-means clustering algorithm groups them into k clusters by minimizing the distortion function over $\{r^{ij}, \mu^j\}$

$$J = \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} \|\mathrm{x}^i - \mu^j\|^2,$$

where $r^{ij} = 1$ if x^i belongs to the j-th cluster and $r^{ij} = 0$ otherwise.

1. (5 points) Prove (using mathematical arguments) that using the squared Euclidean distance $\|\mathrm{x}^i - \mu^j\|^2$ as the dissimilarity function and minimizing the distortion function, we will have

$$\mu^j = \frac{\sum_i r^{ij} \mathrm{x}^i}{\sum_i r^{ij}}.$$

That is, μ^j is the center of j-th cluster.
Hint: consider taken derivative of J with respect to μ^j.

The minimizer μ^j, $\forall j$, satisfies the first-order condition, i.e.,

$$\begin{aligned} \frac{\partial J}{\partial \mu^j} &= -\sum_{i=1}^{m} 2 r^{ij} (\mathrm{x}^i - \mu^j) \\ &= -2 \sum_i r^{ij} \mathrm{x}^i + 2 \sum_i r^{ij} \mu^j = 0 \\ \Rightarrow \quad \sum_i r^{ij} \mathrm{x}^i &= \sum_i r^{ij} \mu^j \\ \mu^j &= \frac{\sum_i r^{ij} \mathrm{x}^i}{\sum_i r^{ij}} \end{aligned}$$

we hence conclude the statement.

rubric: any reasonable attempt 2pts, correct proof 5pts

2. (5 points) Now suppose we replace the similarity function (the squared ℓ_2 distance here: $\|\mathrm{x}^i - \mu^j\|^2$) by another distance measure, the quadratic distance (also known as the Mahalanobis distance) $d(x, y) = (\mathrm{x} - y)^T \Sigma (\mathrm{x} - y)$, where the given weight matrix Σ is symmetrical and positive definite (meaning that the corresponding $d(x, y) > 0$ when $x \neq y$). (i) Show (prove) that the centroid in this case will be the same

$$\mu^j = \frac{\sum_i r^{ij} \mathrm{x}^i}{\sum_i r^{ij}}.$$

(ii) However, the assignment function will be different – comment how the assignment function r^{ij} should be in this case. Thus, the point is here that, depending on the choice of the similarity function (for generalized k-means, the corresponding centroid will be different as well.)
Hint: consider taken derivative of J with respect to μ^j, and use the multivariate calculus rule that the derivative of $z^T \Sigma z$ with respect to z is given by $2\Sigma z$.

(i) Now the distortion function over $\{r^{ij}, \mu^j\}$ becomes

$$J = \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} (\mathrm{x}^i - \mu^j)^T \Sigma (\mathrm{x}^i - \mu^j),$$

where $r^{ij} = 1$ if x^i belongs to the j-th cluster and $r^{ij} = 0$ otherwise. The minimizer μ^j, $\forall j$, satisfies the first-order condition, i.e.,

$$\begin{aligned} \frac{\partial J}{\partial \mu^j} &= -\sum_{i=1}^{m} 2 r^{ij} \Sigma (\mathrm{x}^i - \mu^j) \\ &= -2 \sum_i r^{ij} \Sigma \mathrm{x}^i + 2 \sum_i r^{ij} \Sigma \mu^j = 0 \\ \Rightarrow \quad \Sigma \left(\sum_i r^{ij} \mathrm{x}^i \right) &= \Sigma \left(\sum_i r^{ij} \mu^j \right) \\ \mu^j &= \frac{\sum_i r^{ij} \mathrm{x}^i}{\sum_i r^{ij}} \end{aligned}$$

where the last equality follows the positive definiteness of the matrix Σ, and we hence conclude the statement.

(ii) The "cluster assignment" step will assign each data point x^i to the nearest cluster center:

$$\pi(i) = \arg\min_{j=1,\ldots,k} (\mathrm{x}^i - \mu^j)^T \Sigma (\mathrm{x}^i - \mu^j)$$

rubric: any reasonable attempt 2pts, correct proof and comment 5pts

3. (5 points) Prove (using mathematical arguments) that K-means algorithm converges to a local optimum in finite steps.

Proof sketch:
1.There are limit number of total possible combination of the cluster assignments to the certain number of data points.
2. During each iteration, the cost function decreases monotonically.

rubric: any reasonable attempt 1pts, address each of the above points 2pts

4. (10 points) Calculate k-means by hands. Given 5 data points configuration in Figure 1. Assume $k = 2$ and use Manhattan distance (a.k.a. the ℓ_1 distance: given two 2-dimensional points (x_1, y_1) and (x_2, y_2), their distance is $|x_1 - x_2| + |y_1 - y_2|$). Assuming the initialization of centroid as shown, after one iteration of k-means algorithm, answer the following questions.

(a) Show the cluster assignment;

(b) Show the location of the new center;

(c) Will it terminate in one step?

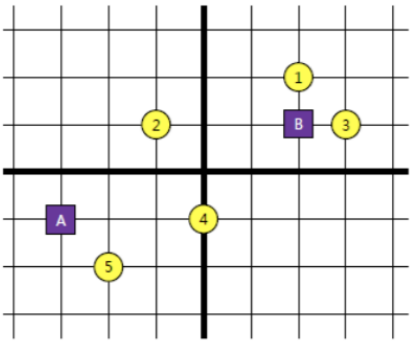


Figure 1: K-means.

(a) the new cluster assignment: $A = \{4, 5\}$, $B = \{1, 2, 3\}$

(b) the new centers: To solve the new centers, we follow the definition, and solve the optimization problem. For the first updated centroid:

$$C_A = \arg \min_{v_1 \in \mathbb{R}, v_2 \in \mathbb{R}} |v_1 - 0| + |v_2 + 1| + |v_1 + 2| + |v_2 + 2|$$

Note that the objective function is decoupled in v_1 and v_2, so we can solve two one-dimensional optimization problem with respect to each of them separately; the plot is shown in Figure 2. Note that the optimization is convex; in this case, they are simple and we can derive the solution. Note that any $v_1 \in [-2, 0]$ and $v_2 \in [-2, -1]$ will minimize the function. So you can pick one minimizer as the solution, for example: $C_A = (-1, -\frac{3}{2})$.

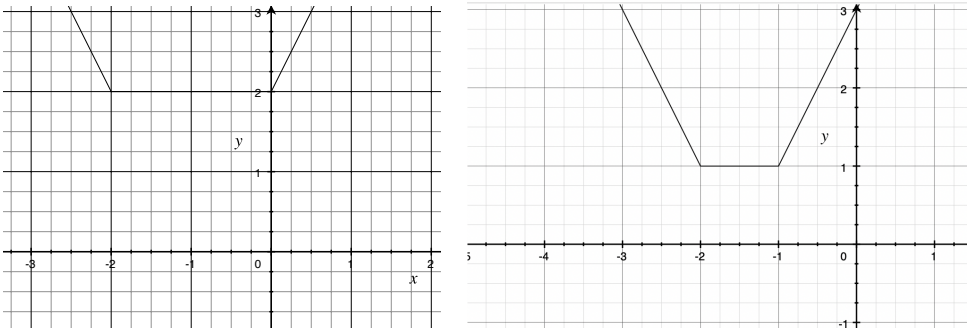


Figure 2: Left: Plot of $y = |x| + |x + 2|$; note that any $x \in [-2, 0]$ will minimize the function; Right: Plot of $y = |x + 1| + |x + 2|$; note that any $x \in [-2, -1]$ will minimize the function.

For the second update centriod:

$$C_B = \arg \min_{v_1 \in \mathbb{R}, v_2 \in \mathbb{R}} |v_1 - 2| + |v_2 - 2| + |v_1 + 1| + |v_2 - 1| + |v_1 - 3| + |v_2 - 1|$$

From the plot in Figure 3, we find that the minimizer is $v_1 = 2$ and $v_2 = 1$. Note that this anticipated, since the optimization problem here is a linear objective function, and the minimizer should happens at one of the vertices. As the result, the update centroid is: $C_B = (2, 1)$.

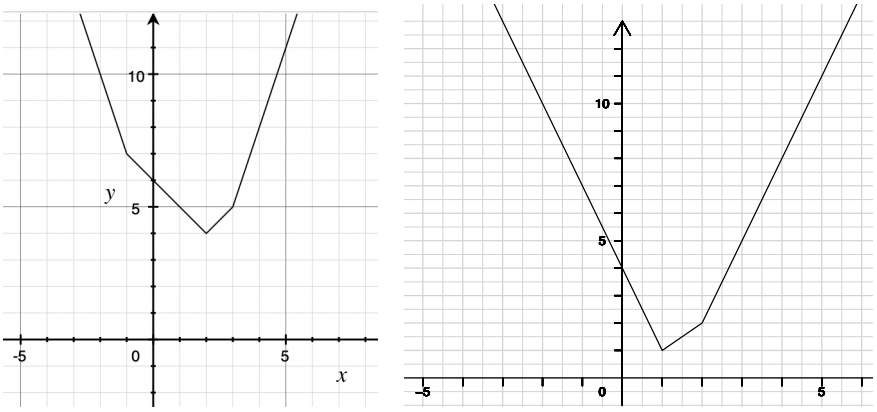


Figure 3: Left: Plot of $y = |x-2| + |x+1| + |x-3|$ and the minimizer is $x = 2$; Right: Plot of $y = |x-2| + |x-1| + |x-1|$ and the minimizer is $x = 1$.

(c) The possible region for the new centroid A is the polytope with vertices $(-2,-1)$, $(-2,-2)$, $(0,-2)$, $(0,-1)$.

- When the new centroid A is chosen to locate in the shade triangle in Figure 4, the distance between centroid A and point 2 is no greater than the distance between centroid B and point 2, hence the new class assignment: cluster A: $\{2,4,5\}$ and cluster B: $\{1,3\}$, hence the algorithm will not terminate;
- When the new centroid A is chosen to locate outside the shade triangle, the distance between centroid A and point 2 is greater than the distance between centroid B and point 2, hence the new class assignment: cluster A: $\{4,5\}$ and cluster B: $\{1,2,3\}$, hence the algorithm will terminate.

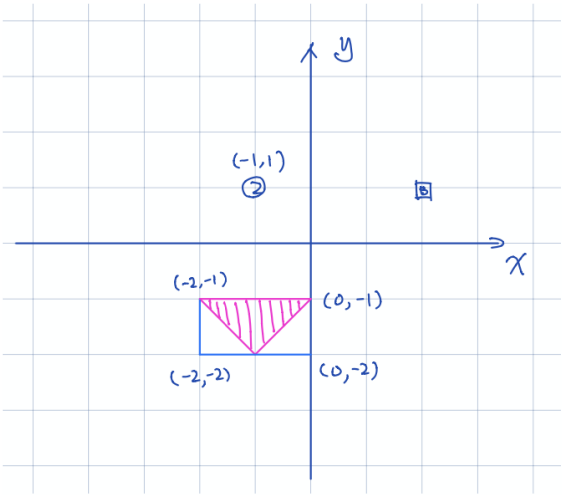


Figure 4:

rubric:

(a) any attempt 1pt, correct answer 4pts

(b) any attempt 1pt, correct answer 4pts

(c) correct terminating conclusion based on the new centroid A from part (b) 2pts

2 Image compression using clustering [40 points]

In this programming assignment, you are going to apply clustering algorithms for image compression. Your task is implementing the clustering parts with two algorithms: *K-means* and *K-medoids*. **It is required you implementing the algorithms yourself rather than calling from a package.**

K-medoids

In class, we learned that the basic K-means works in Euclidean space for computing distance between data points as well as for updating centroids by arithmetic mean. Sometimes, however, the dataset may work better with other distance measures. It is sometimes even impossible to compute arithmetic mean if a feature is categorical, e.g, gender or nationality of a person. With K-medoids, you choose a representative data point for each cluster instead of computing their average. Please note that K-medoid is different from generalized K-means: Generalized K-means still computes centre of a cluster is not necessarily one of the input data points (it is a point that minimizes the overall distance to all points in a cluster in a chosen distance metric).

Given m data points $\mathrm{x}^i (i = 1, \ldots, m)$, K-medoids clustering algorithm groups them into K clusters by minimizing the distortion function $J = \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} D(\mathrm{x}^i, \mu^j)$, where $D(\mathrm{x}, \mathrm{y})$ is a distance measure between two vectors x and y in same size (in case of K-means, $D(x, y) = \|\mathrm{x} - \mathrm{y}\|^2$), μ^j is the center of j-th cluster; and $r^{ij} = 1$ if x^n belongs to the k-th cluster and $r^{ij} = 0$ otherwise. In this exercise, we will use the following iterative procedure:

- Initialize the cluster center μ^j, $j = 1, ..., k$.
- Iterate until convergence:
  - Update the cluster assignments for every data point x^i: $r^{ij} = 1$ if $j = \arg\min_j D(\mathrm{x}^i, \mu^j)$, and $r^{ij} = 0$ otherwise.
  - Update the center for each cluster j: choosing another representative if necessary.

There can be many options to implement the procedure; for example, you can try many distance measures in addition to Euclidean distance, and also you can be creative for deciding a better representative of each cluster. We will not restrict these choices in this assignment. You are encouraged to try many distance measures as well as way of choosing representatives (e.g., ℓ_1 norm).

Formatting instruction

Input

- `pixels`: the input image representation. Each row contains one data point (pixel). For image dataset, it contains 3 columns, each column corresponding to Red, Green, and Blue component. Each component has an integer value between 0 and 255.
- `k`: the number of desired clusters. Too high value of K may result in empty cluster error. Then, you need to reduce it.

Output

- `class`: cluster assignment of each data point in pixels. The assignment should be 1, 2, 3, etc. For $k = 5$, for example, each cell of class should be either 1, 2, 3, 4, or 5. The output should be a column vector with `size(pixels, 1)` elements.
- `centroid`: location of k centroids (or representatives) in your result. With images, each centroid corresponds to the representative color of each cluster. The output should be a matrix with K rows and 3 columns. The range of values should be [0, 255], possibly floating point numbers.

Hand-in

Both of your code and report will be evaluated. Upload them together as a zip file. In your report, answer to the following questions:

1. (10 points) Within the k-medoids framework, you have several choices for detailed implementation. Explain how you designed and implemented details of your K-medoids algorithm, including (but not limited to) how you chose representatives of each cluster, what distance measures you tried and chose one, or when you stopped iteration.

   The general algorithm procedure:

   - Initialize the cluster center μ^j, $j = 1, \cdots, k$
   - iterate until convergence
     - Update the cluster assignments for every data point x^i: $r^{ij} = 1$ if $j = \arg\min_j D(x^i, \mu^j)$, and $r^{ij} = 0$ otherwise.
     - Update the center for each cluster j: choosing another representative if necessary

   Implementation details:

   (a) the representatives, i.e., the centroids, of a cluster, should be chosen as one of the data point that minimize the sum of the distance within the cluster. (This is the main difference from Kmeans algorithm, whose cluster centroid is the mean of the data point in the cluster, it would be different from any of the data point.)

   (b) the distance measure can be chosen from any reasonable distance metric, such as L_p norm (for continuous data), hamming distance (for categorical data), and etc

   (c) the convergence, i.e., iteration terminating criteria can be set as the cost no longer decrease

   $$R(X) = \sum_{i=1}^{m} D(x_i, C_{x_i})$$

   where $D(x_i, C_{x_i})$ is the distance metric function, C_{x_i} is the centroid associated with x_i

   rubric: any reasonable attempt 4pt. each answer to above three question 2pt

2. (10 points) Attach a picture of your own. We recommend size of 320×240 or smaller. Run your k-medoids implementation with the picture you chose, as well as two pictures provided (beach.bmp and football.bmp), with several different K. (e.g, small values like 2 or 3, large values like 16 or 32) What did you observe with different K? How long does it take to converge for each K? Please write in your report.

   reference result as below. the actual number of iteration/running time varies among different implementation efficiency, hardware, and testing image.

	Kmedoids		Kmeans	
K	iterations	times	iterations	times
2	4	0.62s	21	2.78s
4	3	0.47s	15	2.18s
8	5	0.89s	29	4.86s
16	11	2.70s	91	19.51s

rubric:
any program code that produce result without error, 2pts.
reasonable result for: at least three value of K, for at least one provided image and student's own image, 6pts.
proper analysis to the result including the running time 2pts

3. (10 points) Run your k-medoids implementation with different initial centroids/representatives. Does it affect final result? Do you see same or different result for each trial with different initial assignments? (We usually randomize initial location of centroids in general. To answer this question, an intentional poor assignment may be useful.) Please write in your report.

   In principle, k-medoids algorithm converges only at local optimum. Different initialization may end up with different result. Different initialization could lead to different running time. Some extreme initialization such as $[255, 255, 255]$ may cause the program to converge much slower. However, the output image would not have too much perceptible difference among the different initializations.

rubric:
reasonable effort 10pts

4. (10 points) Repeat question 2 and 3 with k-means. Do you see significant difference between K-medoids and k-means, in terms of output quality, robustness, or running time? Please write in your report.

Please refer to the next page for the output images

	Kmedoids		Kmeans	
K	iterations	times	iterations	times
2	4	0.62s	21	2.78s
4	3	0.47s	15	2.18s
8	5	0.89s	29	4.86s
16	11	2.70s	91	19.51s

rubric:
any program code that produce result without error 4pts.
reasonable result for: at least three value of K, for at least one provided image and student's own image, 4pts.
proper analysis 2pt

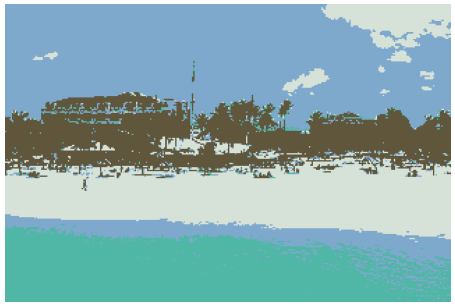
Note

- You may see some error message about empty clusters when you use too large k. Your implementation should treat this exception as well. That is, do not terminate even if you have an empty cluster, but use smaller number of clusters in that case.
- We will grade using test pictures which are not provided. We recommend you to test your code with several different pictures so that you can detect some problems that might happen occasionally.
- If we detect copy from any other student's code or from the web, you will not be eligible for any credit for the entire homework, not just for the programming part. Also, directly calling built-in functions or from other package functions is not allowed.

K-medoids K=2

K-means K=2

K-medoids K=4



K-means K=4

K-medoids K=8

K-means K=8

K-medoids K=16

K-means K=16