Exercises
Introduction to Machine Learning
FS 2020

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s20`
For questions, please refer to Piazza.

Series 5, May 1st, 2020
(K-means convergence)

Problem 1 (K-means convergence):

In the K-means clustering algorithm, you are given a set of n points $x_i \in \mathbb{R}^d$, $i \in 1, \ldots, n$ and you want to find the centers of k clusters $\mu = (\mu_1, \ldots, \mu_k)$ by minimizing the average distance from the points to the closest cluster center.

Formally, you want to minimize the following loss function

$L(\mu) = \sum_{i=1}^{n} \min_{j \in 1,\ldots,k} \|x_i - \mu_j\|_2^2$

To approximate the solution, we introduce new assignment variables $z_i \in \arg\min_{j \in \{1,\ldots,k\}} \|x_i - \mu_j\|_2^2$ for each data point x_i.

The K-means algorithm iterates between updating the variables z_i (assignment step) and updating the centers $\mu_j = \frac{1}{|\{i:z_i=j\}|} \sum_{i:z_i=j} x_i$ (refitting step). The algorithm stops when no change occurs during the assignment step.

Through the following questions, we show that K-means is guaranteed to converge (to a local optimum). We need to prove that the loss function is guaranteed to decrease monotonically in each iteration until convergence. We prove this separately for the assignment step and the refitting step.

First, we show the decrease in loss function for the assignment step.

Assignment step:

1. Let us consider a data point x_i, and let z_i be the assignment and μ_j be the centers from the previous iteration. Choose the expression for the new assignment z_i^*.

   (a) $z_i^* \in \arg\min_{j \in \{1,\ldots,k\}} \|x_i - \mu_j\|_2^2$
   (b) $z_i^* \in \arg\max_{j \in \{1,\ldots,k\}} \|x_i - \mu_j\|_2^2$
   (c) $z_i^* \in \arg\min_{j \in \{1,\ldots,k\}} \|x_i - \mu_j^*\|_2^2$
   (d) $z_i^* \in \arg\max_{j \in \{1,\ldots,k\}} \|x_i - \mu_j^*\|_2^2$

   Solution:
   The correct answer is (a).
   $z_i^* \in \arg\min_{j \in \{1,\ldots,k\}} \|x_i - \mu_j\|_2^2$

2. Let z^* denote the change in assignment for all datapoints. What is the change in the loss function after the update in assignment? Select all the options that apply.

   (a) $L(\mu, z^*) - L(\mu, z) = \sum_{i=1}^{n} \|x_i - \mu_{z_i^*}\|_2^2 - \|x_i - \mu_{z_i}\|_2^2 \leq 0$
   (b) $L(\mu, z) - L(\mu, z^*) = \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_2^2 - \|x_i - \mu_{z_i^*}\|_2^2 \geq 0$
   (c) $L(\mu, z^*) - L(\mu, z) = \sum_{i=1}^{n} \|x_i - \mu_{z_i^*}\|_2^2 - \|x_i - \mu_{z_i}\|_2^2 \geq 0$
   (d) $L(\mu, z) - L(\mu, z^*) = \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|_2^2 - \|x_i - \mu_{z_i^*}\|_2^2 \leq 0$

(e) $L(\mu, z^*) + L(\mu, z) = \sum_{i=1}^{n} \|x_i - \mu_{z_i^*}\|_2^2 + \|x_i - \mu_{z_i}\|_2^2 \geq 0$

(f) $L(\mu, z^*) + L(\mu, z) = \sum_{i=1}^{n} \|x_i - \mu_{z_i^*}\|_2^2 + \|x_i - \mu_{z_i}\|_2^2 \geq 0$

Solution:
The correct answers are (a) and (b).
$L(\mu, z^*) - L(\mu, z) = \sum_{i=1}^{n} \|x_i - \mu_{z_i^*}\|_2^2 - \|x_i - \mu_{z_i}\|_2^2 \leq 0$

Refitting step:

3. Select the rewritten loss function needed to proceed with the proof.

(a) $L(\mu, z) = \sum_{j=1}^{k} \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2$

(b) $L(\mu, z) = \sum_{j=1}^{n} \sum_{i=1}^{k} \|x_i - \mu_j\|_2^2$

(c) $L(\mu, z) = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - \mu_j\|_2^2$

Solution:
The correct answer is (a).
$L(\mu, z) = \sum_{j=1}^{k} \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2$

4. Now, we show the decrease in loss function for the refitting step. Let us consider the j^{th} cluster, and let μ_j be the cluster center from the previous iteration. Select the expression for the new cluster center, μ_j^*.

(a) $\mu_j^* = \frac{1}{|\{i:z_i=j\}|} \sum_{i:z_i=j} x_i$

(b) $\mu_j^* = \frac{1}{n} \sum_{i:z_i=j} x_i$

(c) $\mu_j^* = \frac{1}{k} \sum_{i:z_i=j} x_i$

Solution:
The correct answer is (a).
$\mu_j^* = \frac{1}{|\{i:z_i=j\}|} \sum_{i:z_i=j} x_i$

5. Let μ^* denote the updated cluster centers after the refitting step. What is the change in loss function after the update? Select all options that apply.

(a) $L(\mu^*, z) - L(\mu, z) = \sum_{j=1}^{k} (\sum_{i:z_i=j} \|x_i - \mu_j^*\|_2^2 - \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2) \leq 0$

(b) $L(\mu, z) - L(\mu^*, z) = \sum_{j=1}^{k} (\sum_{i:z_i=j} \|x_i - \mu_j\|_2^2 - \sum_{i:z_i=j} \|x_i - \mu_j^*\|_2^2) \geq 0$

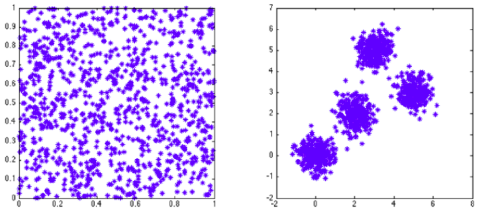
(c) $L(\mu^*, z) - L(\mu, z) = \sum_{j=1}^{k} (\sum_{i:z_i=j} \|x_i - \mu_j^*\|_2^2 - \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2) \geq 0$

(d) $L(\mu, z) - L(\mu^*, z) = \sum_{j=1}^{k} (\sum_{i:z_i=j} \|x_i - \mu_j\|_2^2 - \sum_{i:z_i=j} \|x_i - \mu_j^*\|_2^2) \leq 0$

Solution:
The correct answers are (a) and (b).
$L(\mu^*, z) - L(\mu, z) = \sum_{j=1}^{k} (\sum_{i:z_i=j} \|x_i - \mu_j^*\|_2^2 - \sum_{i:z_i=j} \|x_i - \mu_j\|_2^2) \leq 0$
The inequality holds because the update rule of μ_j^* essentially minimizes this quantity.
Thus, we see that the loss function is guaranteed to decrease monotonically in each iteration, for both the assignment step and the refitting step, until convergence. Thus, K-means is guaranteed to converge (to a local optimum).

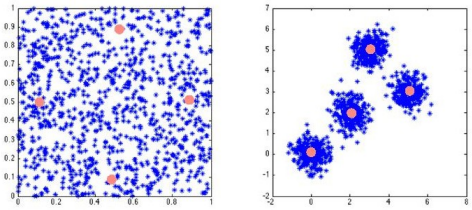
Problem 2 (K-means initialization):

6. You are given two example datasets consisting of 1000 two-dimensional points each. We want to find 4 clusters in each of them.

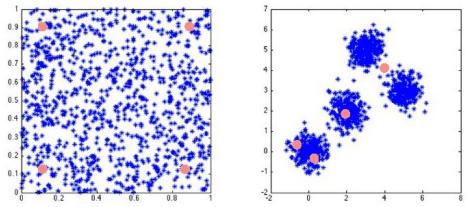


We know that K-means is not robust to initialization. Below, you are given two different initializations for each of the datasets. Which initialization schemes would result in qualitatively different clusters?

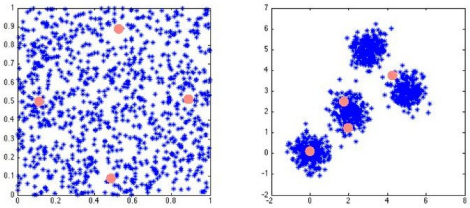
(a) **Initialization Scheme 1:**



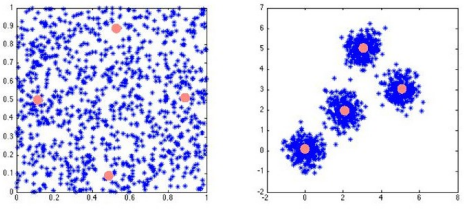
Initialization Scheme 2:



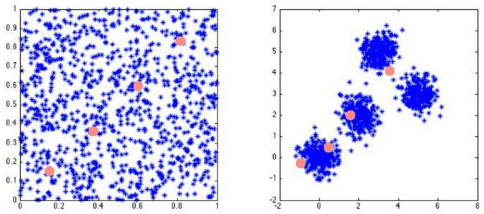
(b) **Initialization Scheme 1:**



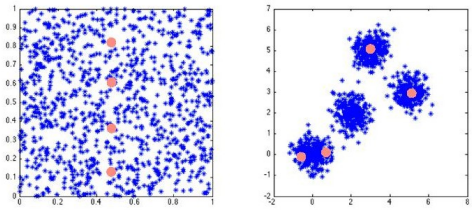
Initialization Scheme 2:



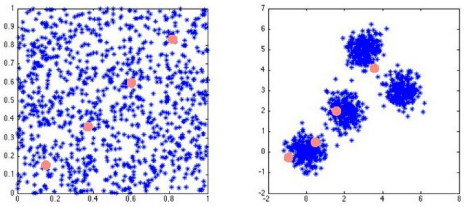
(c) **Initialization Scheme 1:**



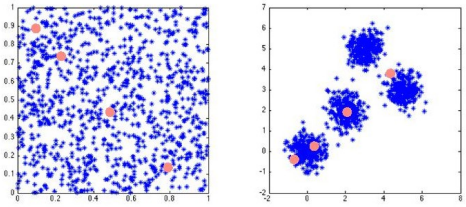
Initialization Scheme 2:



(d) **Initialization Scheme 1:**

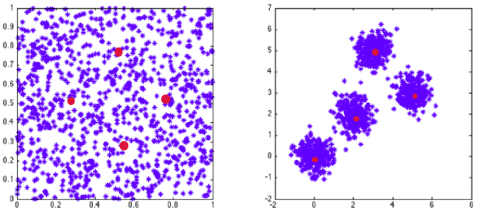


Initialization Scheme 2:

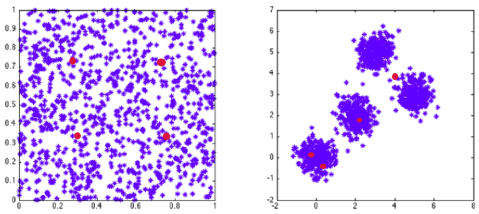


Solution:
The correct answers are (a), (c) and (d).
The following are the estimated of final centroid positions:

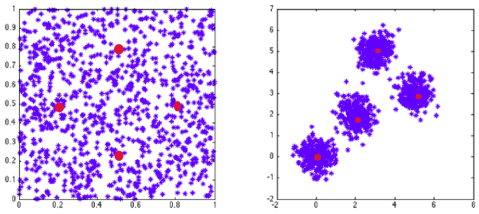
(a) **Initialization Scheme 1:**



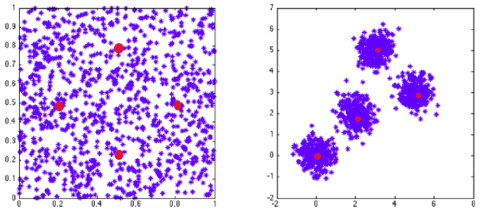
Initialization Scheme 2:



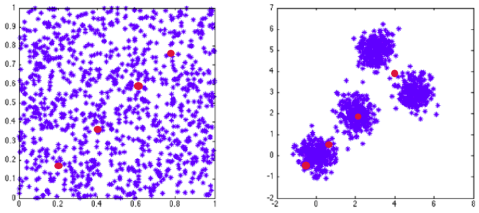
(b) **Initialization Scheme 1:**



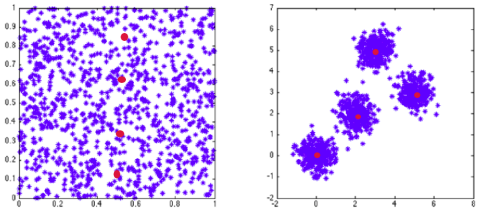
Initialization Scheme 2:



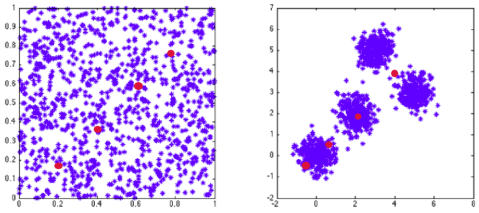
(c) **Initialization Scheme 1:**



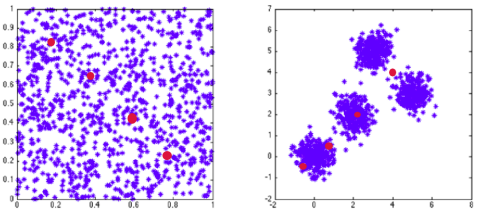
Initialization Scheme 2:



(d) **Initialization Scheme 1:**

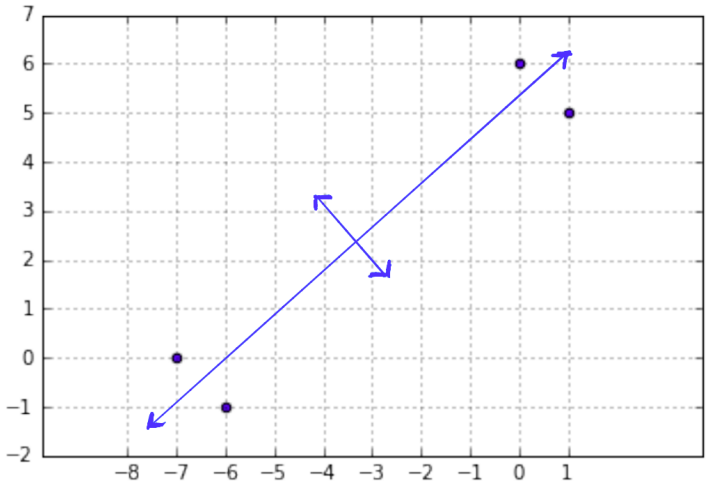


Initialization Scheme 2:



Problem 3 (PCA):

Suppose we have a dataset with 4 points: $D = \{(1,5),(0,6),(-7,0),(-6,-1)\}$. Following is a plot of the dataset with an estimate of two principal components.



Compute the empirical covariance matrix, its eigenvalues and eigenvectors. Do the eigenvectors correspond to your guess of principal components? Please do not forget the assumptions of PCA. (The dataset should be centered and we want unit eigenvectors.)

7. Enter your computation of the covariance matrix rounded off to the second decimal point.

(a) $\begin{pmatrix} 12.5 & 10.25 \\ 10.25 & 9.25 \end{pmatrix}$

(b) $\begin{pmatrix} 10.15 & 9.25 \\ 12.5 & 10.25 \end{pmatrix}$

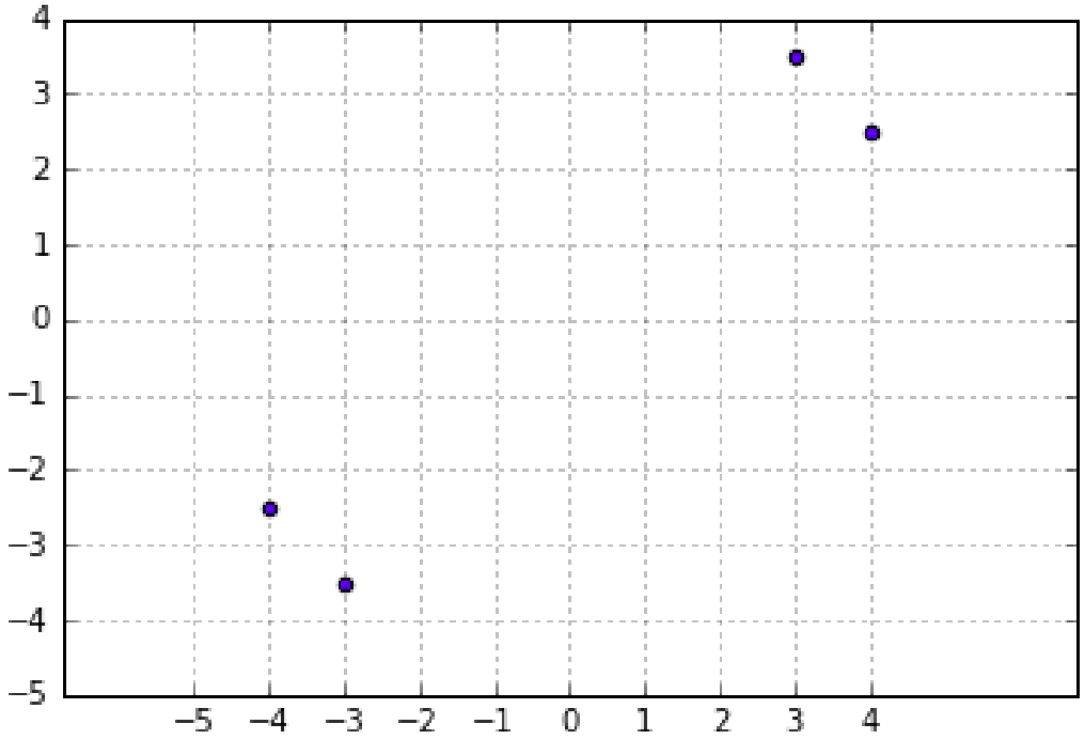
(c) $\begin{pmatrix} 12.5 & 10.15 \\ 10.25 & 9.25 \end{pmatrix}$

(d) $\begin{pmatrix} 10.25 & 12.5 \\ 9.25 & 10.15 \end{pmatrix}$

Solution:
The correct answer is (a).
We first need to center the data by subtracting from it its mean $(-3,\ 2.5)^T$, obtaining

$$\mathbf{x_1} = (4,\ 2.5)^T$$
$$\mathbf{x_2} = (3,\ 3.5)^T$$
$$\mathbf{x_3} = (-4,\ -2.5)^T$$
$$\mathbf{x_4} = (-3,\ -3.5)^T.$$

Plot of the centered dataset:



For the empirical covariance matrix, we obtain

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x_i}\mathbf{x_i}^T = \frac{1}{4} \cdot \begin{pmatrix} 50 & 41 \\ 41 & 37 \end{pmatrix} = \begin{pmatrix} 12.5 & 10.25 \\ 10.25 & 9.25 \end{pmatrix}.$$

8. Enter your computation of the eigenvectors rounded off to the second decimal point.

(a) $v_1 = \begin{pmatrix} 0.76 & 0.65 \end{pmatrix}^T$
$v_2 = \begin{pmatrix} -0.65 & 0.76 \end{pmatrix}^T$

(b) $v_1 = \begin{pmatrix} 0.76 & -0.65 \end{pmatrix}^T$
$v_2 = \begin{pmatrix} -0.65 & 0.76 \end{pmatrix}^T$

(c) $v_1 = \begin{pmatrix} 0.76 & 0.65 \end{pmatrix}^T$
$v_2 = \begin{pmatrix} 0.65 & 0.76 \end{pmatrix}^T$

(d) $v_1 = \begin{pmatrix} -0.76 & 0.65 \end{pmatrix}^T$
$v_2 = \begin{pmatrix} 0.65 & 0.76 \end{pmatrix}^T$

Solution:
The correct answer is (a).

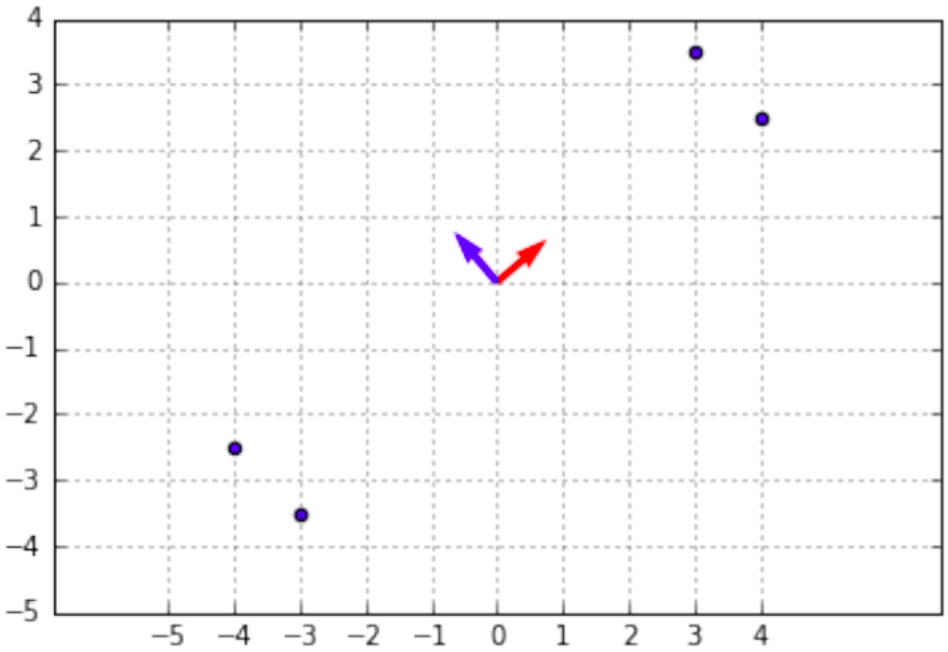
The unit-length eigenvectors of Σ are $v_1 = (0.76045416, 0.64939162)^T$ and $v_2 = (-0.64939162, 0.76045416)^T$.

9. Enter your computation of the corresponding eigenvalues rounded off to the second decimal point.

   (a) $w_1 = 21.25, w_2 = 0.50$
   (b) $w_1 = 0.50, w_2 = 21.25$
   (c) $w_1 = 0.50, w_2 = 10.25$
   (d) $w_1 = 10.25, w_2 = 0.50$

Solution:
The correct answer is (a).
The corresponding eigenvalues are $w_1 = 21.25301161$ and $w_2 = 0.49698839$, respectively.

Plot of the centered dataset with principal components 1 (red) and 2 (blue):



For a nice visualization of PCA, also see `http://setosa.io/ev/principal-component-analysis`.

Problem 4 (Another clustering approach):

In this exercise, you are asked to derive a new clustering algorithm that would use a different loss function given by

$L(\mu) = \sum_{i=1}^{n} min_{j \in 1,\ldots,k} \|x_i - \mu_j\|_1$. Through the next few questions, we find the update steps for both z_i and for μ_j in this case. In questions (10), (11), (12) and (13), fill in the blanks in the update step of z_i.

$z_i = \underline{(10)}_{\underline{(11)}} \|x_i - \mu_j\|^{\underline{(13)}}_{\underline{(12)}}$

10. Select one:

    (a) argmin
    (b) argmax
    (c) min
    (d) max
    (e) medoid
    (f) mean

    Solution:
    The correct answer is (a).

11. Select one:

    (a) $j \in \{1, \ldots, k\}$
    (b) $j \in \{1, \ldots, n\}$
    (c) $i \in \{1, \ldots, k\}$
    (d) $i \in \{1, \ldots, n\}$
    (e) $\mu_j \in \mathbb{R}^d$

    Solution:
    The correct answer is (a).

12. The answer is 1.

13. The answer is 1.

14. Fill in the blanks in update step of μ_j.
    $\mu_j = \underline{(14)}\,(\,\underline{(15)}\,)\forall j = 1, \ldots, k; \forall q = 1, \ldots, d$. Select one:

    (a) $\arg\min_j$
    (b) $\arg\max_j$
    (c) $\min_j$
    (d) $\max_j$
    (e) medoid
    (f) mean
    (g) median

    Solution:
    The correct answer is (g).

15. Select one:

    (a) $\sum_{i:z_i=j} |x_{i,q} - \mu_{i,q}|$
    (b) $\sum_{i=1}^{n} |x_{i,q} - \mu_{i,q}|$
    (c) $x_{i,q}, i : z_i = j$
    (d) $\mu_{j,q}, j \in \{1, \ldots, k\}$

Solution:
The correct answer is (c).
As in the K-means algorithm, let's again introduce hidden variables $z_i = \arg\min_{j\in 1,\dots,k} ||x_i - \mu_j||_1$ for each data point x_i. Then the initial problem

$$\mu = \arg\min_{\mu} \sum_{i=1}^{n} \min_{j\in 1,\dots,k} ||x_i - \mu_j||_1$$

can be rewritten in a different form (because we know where exactly the minimum is achieved):

$$\mu = \arg\min_{\mu} \sum_{i=1}^{n} ||x_i - \mu_{z_i}||_1$$

In order to find the solution with respect to μ_j with fixed z_i, let's leave only the data points that correspond to the j^{th} component:

$$\mu_j = \arg\min_{\mu_j} \sum_{i:z_i=j} ||x_i - \mu_j||_1$$

$$\mu_j = \arg\min_{\mu_j} \sum_{i:z_i=j} \sum_{q=1}^{d} |x_{i,q} - \mu_{j,q}|$$

This can again be separated component-wise:

$$\mu_{j,q} = \arg\min_{\mu_{j,q}} \sum_{i:z_i=j} |x_{i,q} - \mu_{j,q}|$$

Again, as in the K-means algorithm, we proceed by finding the derivative of the functional and setting it to zero. In order to get rid of the L_1 norm, we also separate the functional into the sum over those $x_{i,q}$ that are smaller than $\mu_{j,q}$ and those that are larger:

$$\sum_{i:z_i=j, x_{i,q}\le\mu_{j,q}} |x_{i,q}-\mu_{j,q}| + \sum_{i:z_i=j, x_{i,q}>\mu_{j,q}} |x_{i,q}-\mu_{j,q}| = \sum_{i:z_i=j, x_{i,q}\le\mu_{j,q}} (\mu_{j,q}-x_{i,q}) + \sum_{i:z_i=j, x_{i,q}>\mu_{j,q}} (x_{i,q}-\mu_{j,q})$$

The derivative of every bracket in the sum is either $+1$ or -1, and the number of $+1$'s is exactly $|\{i : z_i = j, x_{i,q} \le \mu_{j,q}\}|$. Therefore, we need to set

$$|\{i : z_i = j, x_{i,q} \le \mu_{j,q}\}| - |\{i : z_i = j, x_{i,q} > \mu_{j,q}\}| = 0$$

This means that $\mu_{j,q}$ is nothing but the *median* of all the numbers $x_{i,q}$, $i : z_i = j$.

The resulting algorithm then iterates between two steps:

- $z_i = \arg\min_{j\in 1,\dots,k} ||x_i - \mu_j||_1$
- $\mu_{j,q} = \text{median}(x_{i,q}, i : z_i = j), \forall j = 1,\dots,k; \forall q = 1,\dots,d.$

16. What can you say about the convergence of the algorithm? Select the true statements.

    (a) The same convergence properties as for k-means can be proved.
    (b) The algorithm does not converge.
    (c) The same convergence properties as for k-means CANNOT be proved.
    (d) The algorithm is guaranteed to converge to local minimum.
    (e) The algorithm is guaranteed to converge to global minimum.

Solution:
(a) and (d) are True.

17. In which situation would you prefer to use this clustering method instead of K-means clustering? Select the true statements.

    (a) If the data contains many outliers.

    (b) If the data does not contain clearly separable clusters.

    (c) If we have more data points than features.

    (d) If the Euclidean metric is not the proper distance measure.

    (e) If we are dealing with large datasets.

    (f) If we are dealing with high-dimensional data.

    Solution:
    Only (a) is True.