College of Engineering & Applied Sciences

CSPB 2820

Linear Algebra With Computer Science Applications

Study Guide 4 - Clustering

Taylor Larrechea

2023

Sections

Study Guide 4

Clustering

Study Guide Instructions

- Submit your work in Gradescope as a PDF - you will identify where your “questions are.”
- Identify the question number as you submit. Since we grade "blind" if the questions are NOT identified, the work WILL NOT BE GRADED and a 0 will be recorded. Always leave enough time to identify the questions when submitting.
- One section per page (if a page or less) - We prefer to grade the main solution in a single page, extra work can be included on the following page.
- Long instructions may be removed to fit on a single page.
- **Do not start a new question in the middle of a page.**
- Solutions to book questions are provided for reference.
- You may NOT submit given solutions - this includes minor modifications - as your own.
- Solutions that do not show individual engagement with the solutions will be marked as no credit and can be considered a violation of honor code.
- If you use the given solutions you must reference or explain how you used them, in particular...

Method Selection

For full credit, EACH book exercise in the Study Guides must use one or more of the following methods and FOR EACH QUESTION. Identify the number the method by number to ensure full credit.

- **Method 1** - Provide original examples which demonstrate the ideas of the exercise in addition to your solution.
- **Method 2** - Include and discuss the specific topics needed from the chapter and how they relate to the question.
- **Method 3** - Include original Python code, of reasonable length (as screenshot or text) to show how the topic or concept was explored.
- **Method 4** - Expand the given solution in a significant way, with additional steps and comments. All steps are justified. This is a good method for a proof for which you are only given a basic outline.
- **Method 5** - Attempt the exercise without looking at the solution and then the solution is used to check work. Words are used to describe the results.
- **Method 6** - Provide an analysis of the strategies used to understand the exercise, describing in detail what was challenging, who helped you or what resources were used. The process of understanding is described.

Problem 1

Problem Statement

Describe the K-Means algorithm in your own words to someone who has never heard of it. Explain each step.

Solution

Overview

The k-means algorithm is an algorithm that helps group or 'cluster' vectors together based off of their distances from one another. The algorithm works with a fixed number of vectors usually represented with N. These vectors are of an arbitrary size, lets call the size of the vectors n. The goal of this algorithm is to cluster the vectors in question into k groups where they are being clustered together based off of their similarities between one another. Lets dive into this algorithm with a little more detail.

Initialization

Initially, when the k-means algorithm is going to be run, the programmer must choose the number of clusters (this number is represented with k) that they wish their vectors are to be grouped in. The algorithm then chooses an arbitrary starting point called the 'centroid' for each cluster. These centroids are usually picked at random.

Calculation

Once the number of centroids has been chosen for the given data set, we then calculate the distance for a given data point with each centroid. For each individual data point in the data set, a distance calculation is performed to determine what is the smallest distance between the given data point and the centroids. The centroid that has the smallest distance between the given data set is the centroid that the data point will be assigned to.

Centroid Update

After each iteration of assigning a new data point to a centroid, we calculate the 'mean' of these data points and update the centroid location for that specific cluster. Essentially, we are pinpointing the new center of the centroid by finding the optimal location of it by finding the average of the co-ordinates of each data point for a given cluster.

Completion

After the calculation and centroid update step has been completed for every data point in the data set, we call this convergence meaning that the data points have been correctly assigned to their correct centroid and the centroid is in the optimal location for all data sets assigned to it. At this point, the algorithm has finished running and we have successfully clustered all of our data points in the given data set.

Summary

Overall, the k-means algorithm is a very efficient way to group data points in a data set together based off of similar characteristics. Initially, the centroid location is randomly positioned within the data set and it is constantly updated based off of the mean of the data that it is assigned to. After this process repeats multiple times, we end up with a clustering of data where the data in the data set usually share some common characteristic with one another.

Problem 1 Summary

Procedure

- Answer the prompts with a concise summary of the k-means algorithm

Key Concepts

- k-means is algorithm that is used in clustering to group data points together based off of similar characteristics
- k-means first finds a spot for the centroids in the algorithm to be placed
- k-means then calculates the distance between the centroid and each of its data points to determine the smallest distance between remaining data points
- k-means then updates the location of the centroid based off the mean distance between data points that have been assigned to the centroid
- k-means then is complete when the centroid is placed in the lowest average distance between data points for a given cluster

Variations

- We could be asked to summarize a different process in linear algebra
  - We then would have to talk about aspects that pertain to that specific process

Problem 2

Problem Statement

Select one page or section of Chapter One of VMLS to annotate. Include a screenshot of your annotation here. (not example 4.4.1 since that is the next question)

For annotations this week, I have chosen to annotate section 4.2 of VMLS. The annotations for this problem can be seen on the following page.

4.2 A clustering objective

In this section we formalize the idea of clustering, and introduce a natural measure of the quality of a given clustering.

Specifying the cluster assignments. We specify a clustering of the vectors by saying which cluster or group each vector belongs to. We label the groups $1, \ldots, k$, and specify a clustering or assignment of the N given vectors to groups using an N-vector c, where c_i is the group (number) that the vector x_i is assigned to. As a simple example with $N = 5$ vectors and $k = 3$ groups, $c = (3, 1, 1, 1, 2)$ means that x_1 is assigned to group 3, x_2, x_3, and x_4 are assigned to group 1, and x_5 is assigned to group 2. We will also describe the clustering by the sets of indices for each group. We let G_j be the set of indices corresponding to group j. For our simple example above, we have

$$G_1 = \{2, 3, 4\}, \quad G_2 = \{5\}, \quad G_3 = \{1\}.$$

(Here we are using the notation of sets; see appendix A.) Formally, we can express these index sets in terms of the group assignment vector c as

$$G_j = \{i \mid c_i = j\},$$

which means that G_j is the set of all indices i for which $c_i = j$.

Group representatives. With each of the groups we associate a *group representative* n-vector, which we denote $z_1, \ldots, z_k$. These representatives can be any n-vectors; they do not need to be one of the given vectors. We want each representative to be close to the vectors in its associated group, *i.e.*, we want the quantities

$$\|x_i - z_{c_i}\|$$

to be small. (Note that x_i is in group $j = c_i$, so z_{c_i} is the representative vector associated with data vector x_i.)

A clustering objective. We can now give a single number that we use to judge a choice of clustering, along with a choice of the group representatives. We define

$$J^{\text{clust}} = \left(\|x_1 - z_{c_1}\|^2 + \cdots + \|x_N - z_{c_N}\|^2\right)/N, \tag{4.1}$$

which is the mean square distance from the vectors to their associated representatives. Note that J^{clust} depends on the cluster assignments (*i.e.*, c), as well as the choice of the group representatives $z_1, \ldots, z_k$. The smaller J^{clust} is, the better the clustering. An extreme case is $J^{\text{clust}} = 0$, which means that the distance between every original vector and its assigned representative is zero. This happens only when the original collection of vectors only takes k different values, and each vector is assigned to the representative it is equal to. (This extreme case would probably not occur in practice.)

Our choice of clustering objective J^{clust} makes sense, since it encourages all points to be near their associated representative, but there are other reasonable

choices. For example, it is possible to use an objective that encourages more balanced groupings. But we will stick with this basic (and very common) choice of clustering objective.

Optimal and suboptimal clustering. We seek a clustering, *i.e.*, a choice of group assignments $c_1, \ldots, c_N$ and a choice of representatives $z_1, \ldots, z_k$, that minimize the objective J^{clust}. We call such a clustering *optimal*. Unfortunately, for all but the very smallest problems, it is practically impossible to find an optimal clustering. (It can be done in principle, but the amount of computation needed grows extremely rapidly with N.) The good news is that the k-means algorithm described in the next section requires far less computation (and indeed, can be run for problems with N measured in billions), and often finds a very good, if not the absolute best, clustering. (Here, 'very good' means a clustering and choice of representatives that achieves a value of J^{clust} near its smallest possible value.) We say that the clustering choices found by the k-means algorithm are *suboptimal*, which means that they might not give the lowest possible value of J^{clust}.

Optimal clustering usually only happens in theory, this gives best value of J_c

Suboptimal gives not the lowest value of J_c

Even though it is a hard problem to choose the best clustering and the best representatives, it turns out that we *can* find the best clustering, if the representatives are fixed, and we can find the best representatives, if the clustering is fixed. We address these two topics now.

Partitioning the vectors with the representatives fixed. Suppose that the group representatives $z_1, \ldots, z_k$ are fixed, and we seek the group assignments $c_1, \ldots, c_N$ that achieve the smallest possible value of J^{clust}. It turns out that this problem can be solved exactly.

The objective J^{clust} is a sum of N terms. The choice of c_i (*i.e.*, the group to which we assign the vector x_i) only affects the ith term in J^{clust}, which is $(1/N)\|x_i - z_{c_i}\|^2$. We can choose c_i to minimize just this term, since c_i does not affect the other $N-1$ terms in J^{clust}. How do we choose c_i to minimize this term? This is easy: We simply choose c_i to be the value of j that minimizes $\|x_i - z_j\|$ over j. In other words, we should assign each data vector x_i to its nearest neighbor among the representatives. This choice of assignment is very natural, and easily carried out.

c_i must be smallest dist. between x_i and z_j

So when the group representatives are fixed, we can readily find the best group assignment (*i.e.*, the one that minimizes J^{clust}), by assigning each vector to its nearest representative. With this choice of group assignment, we have (by the way the assignment is made)

Smallest distance → $$\|x_i - z_{c_i}\| = \min_{j=1,\ldots,k} \|x_i - z_j\|,$$

so the value of J^{clust} is given by

$$\left(\min_{j=1,\ldots,k} \|x_1 - z_j\|^2 + \cdots + \min_{j=1,\ldots,k} \|x_N - z_j\|^2\right)/N.$$

This has a simple interpretation: It is the mean of the squared distance from the data vectors to their closest representative.

$$J^c = \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{1}{N} \min(\|x_i - z_j\|^2)$$

J^c is the mean square distance of the data points with their closest representative to that data point

Optimizing the group representatives with the assignment fixed. Now we turn to the problem of choosing the group representatives, with the clustering (group assignments) fixed, in order to minimize our objective J^{clust}. It turns out that this problem also has a simple and natural solution.

We start by re-arranging the sum of N terms into k sums, each associated with one group. We write

$$J^{\text{clust}} = J_1 + \cdots + J_k,$$

where

$$J_j = (1/N) \sum_{i \in G_j} \|x_i - z_j\|^2$$

is the contribution to the objective J^{clust} from the vectors in group j. (The sum here means that we should add up all terms of the form $\|x_i - z_j\|^2$, for any $i \in G_j$, *i.e.*, for any vector x_i in group j; see appendix A.)

Summing $\|x_i - z_j\|^2$ for vectors x_i in group j

The choice of group representative z_j only affects the term J_j; it has no effect on the other terms in J^{clust}. So we can choose each z_j to minimize J_j. Thus we should choose the vector z_j so as to minimize the mean square distance to the vectors in group j. This problem has a very simple solution: We should choose z_j to be the average (or mean or centroid) of the vectors x_i in its group:

$$z_j = (1/|G_j|) \sum_{i \in G_j} x_i,$$

$|G_j|$ → cardinality of set G_j

where $|G_j|$ is standard mathematical notation for the number of elements in the set G_j, *i.e.*, the size of group j. (See exercise 4.1.)

So if we fix the group assignments, we minimize J^{clust} by choosing each group representative to be the average or centroid of the vectors assigned to its group. (This is sometimes called the *group centroid* or *cluster centroid*.)

4.3 The k-means algorithm

It might seem that we can now solve the problem of choosing the group assignments and the group representatives to minimize J^{clust}, since we know how to do this when one or the other choice is fixed. But the two choices are circular, *i.e.*, each depends on the other. Instead we rely on a very old idea in computation: We simply *iterate* between the two choices. This means that we repeatedly alternate between updating the group assignments, and then updating the representatives, using the methods developed above. In each step the objective J^{clust} gets better (*i.e.*, goes down) unless the step does not change the choice. Iterating between choosing the group representatives and choosing the group assignments is the celebrated *k-means algorithm* for clustering a collection of vectors.

The k-means algorithm was first proposed in 1957 by Stuart Lloyd, and independently by Hugo Steinhaus. It is sometimes called the Lloyd algorithm. The name 'k-means' has been used since the 1960s.

Problem 2 Summary

Procedure

- Annotate a section from the textbook by adding comments and insights for the given section

Key Concepts

- This section of the textbook covers the topic of clustering an objective
- Clustering involves grouping data points together based upon their similarities
- The lower the value of J_C the better the clustering of data points
- J_c is calculated with

$$J_C = \sum_{i=1}^{N} \sum_{j=1}^{k} \frac{1}{N} \mathbf{min}(\|x_i - z_j\|^2)$$

- The term z_j is calculated with

$$z_j = \frac{1}{|G_j|} \sum_{i \in G_j}^{n} x_i$$

  where G_j is group j for a given data set and $|G_j|$ is the number of elements in that group

Variations

- We could be asked to annotate a different section from the textbook

Problem 3

Problem Statement

Explain example 4.4.1 in detail. Why is this an interesting example? What is k in this example?

Solution

Overview

This example demonstrates a use of the k-means algorithm to cluster data that represents hand written digits. There are a total of 60,000 images in the MNIST database of handwritten images. Each image is an image with 784 pixels, where each pixel is represented with a value that is indicative of the brightness of the pixel in question. One can interpret each pixel to be a vector of length 1 (or a scalar value for that matter). The images in question are 28×28 with a total of 784 pixels. Each image is a vector of length 784.

In the context of the k-means algorithm, the clustering that is done on these images is with a k value of 20. This means that when the algorithm is ran, the pixels are clustered into 20 separate clusters that are assembled to resemble a digit that was initially hand drawn.

Process

Each individual image is ran through the k-means algorithm with $k = 20$. The images in the example are then processed three times and there is a plot that represents each iteration of the algorithm for three different initial partitions of the data. Figure 4.7 has three colored lines, the line that is red represents the initial partition with the worst clustering of the data. The line that is blue represents the partition that had the best clustering of the data. The brown line in the graph is the partition that is the median valued partition of the data.

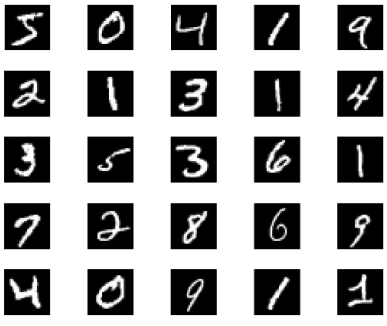


Figure 4.6 25 images of handwritten digits from the MNIST data set. Each image has size 28×28, and can be represented by a 784-vector.

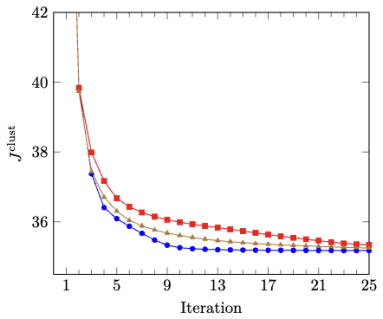


Figure 4.7 Clustering objective J^{clust} after each iteration of the k-means algorithm, for three initial partitions, on digits of the MNIST set.

Figure 4.6 represents some examples of the images found in MNIST database and Figure 4.7 represents the comparison of clustering that was done on the images with three different random partitions of the centroids in the k-means algorithm.

Comparison

Figure 4.8 below is the partition that had the worst clustering. Consequently, the digits in these are very blurry and that is indicative visually with a clustering of data that is poor. Figure 4.9 on the other hand is the partition that had the best clustering. Although some of the images are blurry, most of the images in this figure are easily identifiable and this is indicative of better clustering.

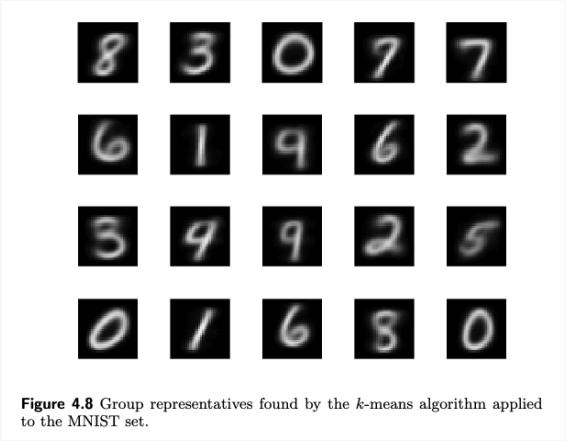


Figure 4.8 Group representatives found by the k-means algorithm applied to the MNIST set.

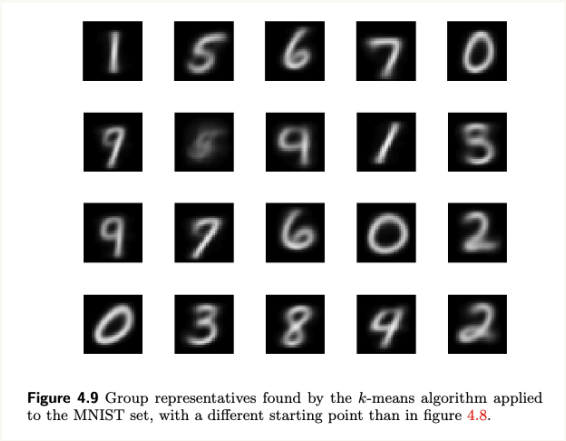


Figure 4.9 Group representatives found by the k-means algorithm applied to the MNIST set, with a different starting point than in figure 4.8.

The k-means algorithm appears to struggle with images that have the digits $4, 9, 3$ or 8 in them. This is because the representation of these images can be somewhat ambiguous when trying to cluster the data because of the similarity of the digits with others.

Insights

This is an interesting example, because of how blurry the digits may be in some cases, the k-means algorithm does a pretty good job of representing digits. Since the algorithm doesn't really know anything about how digits look to the human eye, it does an exceptionally good job at re-creating digits regardless of how blurry they might be to us who know what these digits look like if they are written clearly.

Problem 3 Summary

Procedure

- Analyze the process of how k-means algorithm works for this specific example of images

Key Concepts

- This problem showcases the k-means algorithm and how it works for a given set of images and how it can be used to re-create images with the use of clustering
- The k value in k-means is the number of clusters that the algorithm is attempting to cluster data into
- The lower the value of J_C the better the clustering of data

Variations

- We could be given a different set of images that we would need to analyze for how the k-means algorithm works
  - In this case we would use the same process of analyzing the images and see how the k-means algorithm performs in those scenarios

Problem 4

Problem Statement

Explain the solution to 4.1 here in your own words. (Since you are given a solution, you will be graded on your ability to explain).

Original Question

Minimizing mean square distance to a set of vectors. Let $x_1, \ldots, x_L$ be a collection of n-vectors. In this exercise you will fill in the missing parts of the argument to show that the vector z which minimizes the sum-square distance to the vectors,

$$J(z) = ||x_1 - z||^2 + \cdots + ||x_L - z||^2,$$

is the average or centroid of the vectors, $\bar{x} = (1/L)(x_1 + \cdots + x_L)$. (This result is used in one of the steps in the k-means algorithm. But here we have simplified the notation.)

(a) Explain why, for any z, we have

$$J(z) = \sum_{i=1}^{L} ||x_i - \bar{x} - (z - \bar{x})||^2 = \sum_{i=1}^{L} (||x_i - \bar{x}||^2 - 2(x_i - \bar{x})^T(z - \bar{x}) + L||z - \bar{x}||^2).$$

(b) Explain why $\sum_{i=1}^{L}(x_i - \bar{x})^T(z - \bar{x}) = 0$. *Hint.* Write the left-hand side as

$$\left(\sum_{i=1}^{L}(x_i - \bar{x})\right)^T (z - \bar{x}),$$

and argue that the left-hand vector is 0.

(c) Combine the results of (a) and (b) to get $J(z) = \sum_{i=1}^{L} ||x_i - \bar{x}||^2 + L||z - \bar{x}||^2$. Explain why for any $z \neq \bar{x}$, we have $J(z) > J(\bar{x})$. This shows that the choice $z = \bar{x}$ minimizes $J(z)$.

Solution - Part (a)

For this problem, I will be using **Method 4**.

VMLS Solution:

(a) In the second expression, we are simply adding and subtracting the vector $\bar{x}$ from $x_i - z$, which has no effect. Expanding the norm squared gives the right-hand expression.

Explanation:

We first begin by cleverly adding 0 the LHS of the problem statement. This is done by adding and subtract the vector $\bar{x}$. Namely,

$$J(z) = \sum_{i=1}^{L} ||x_i - z + \bar{x} - \bar{x}||^2 = \sum_{i=1}^{L} ||x_i - \bar{x} - z + \bar{x}||^2 = \sum_{i=1}^{L} ||(x_i - \bar{x}) - (z - \bar{x})||^2 \quad (1)$$

where we have just regrouped the terms in equation (1) to get our final expression on the RHS. We know from VMLS that the norm of two vectors being summed is

$$||\alpha + \beta||^2 = ||\alpha||^2 + 2\alpha^T\beta + ||\beta||^2.$$

If we then apply the definition of a norm from VMLS with the RHS of equation (1) and make the substitution of $\alpha = (x_i - \bar{x})$ and $\beta = (z - \bar{x})$ we then have

$$J(z) = \sum_{i=1}^{L} ||(x_i - \bar{x}) - (z - \bar{x})||^2 = \sum_{i=1}^{L} (||x_i - \bar{x}||^2 - 2(x_i - \bar{x})^T (z - \bar{x}) + L||z - \bar{x}||^2). \tag{2}$$

The minus sign in equation (2) is coming from the fact that we are computing the distance and not the sum. The variable L comes from scaling the difference between z and $\bar{x}$.

Solution - Part (b)

For this problem, I will be using **Method 4**.

VMLS Solution:

(b) We follow the hint, which works since the inner product is a linear function, so

$$\sum_{i=1}^{L} (x_i - \bar{x})^T (z - \bar{x}) = \left(\sum_{i=1}^{L} (x_i - \bar{x}) \right)^T (z - \bar{x})$$

(Note that the first sum is over numbers, and the second is over vectors.) Note also that the right-hand vector in the inner product is the same for each i. The vector $\sum_{i=1}^{L} (x_i - \bar{x})$ is zero, since

$$\sum_{i=1}^{L} x_i = L\bar{x} = \sum_{i=1}^{L} \bar{x}.$$

Therefore the middle term in the right-hand expression for $J(z)$ in part (a) is zero.

Explanation:

We know from the problem statement that $\bar{x}$ can be written as

$$\bar{x} = \frac{1}{L}(x_1 + \cdots + x_L) = \frac{1}{L} \sum_{i=1}^{L} x_i. \tag{3}$$

This means if we isolate x_i from equation (3) we can then say

$$\sum_{i=1}^{L} x_i = L\bar{x}. \tag{4}$$

Going back to the problem statement for part (b), if we substitute equation (4) into it we have

$$\sum_{i=1}^{L} (x_i - \bar{x})^T (z - \bar{x}) = \sum_{i=1}^{L} (\bar{x} - \bar{x})^T (z - \bar{x}) = \sum_{i=1}^{L} (0)^T (z - \bar{x}) = 0 \tag{5}$$

and thus we have $\sum_{i=1}^{L} (x_i - \bar{x})^T (z - \bar{x}) = 0$.

Solution - Part (c)

For this problem, I will be using **Method 4**.

VMLS Solution:

(c) For any $z \neq \bar{x}$, we have

$$J(z) = \sum_{i=1}^{L} ||x_i - \bar{x}||^2 + L||z - \bar{x}||^2 > \sum_{i=1}^{L} ||x_i - \bar{x}||^2 = J(\bar{x}),$$

where we use $||z - \bar{x}|| > 0$.

Explanation:

We know from part (b) that the term

$$\sum_{i=1}^{L}(x_i - \bar{x})^T = 0. \tag{6}$$

Substituting equation (6) into equation (2) we can then say

$$J(z) = \sum_{i=1}^{L}(||x_i - \bar{x}||^2 - 2(x_i - \bar{x})^T(z - \bar{x}) + L||z - \bar{x}||^2) \tag{7}$$

$$= \sum_{i=1}^{L}(||x_i - \bar{x}||^2 - 2(0)(z - \bar{x}) + L||z - \bar{x}||^2) \tag{8}$$

$$= \sum_{i=1}^{L}(||x_i - \bar{x}||^2 + L||z - \bar{x}||^2) \tag{9}$$

where equation (9) is the representation of $J(z)$ when $z \neq \bar{x}$. If we assume that $z = \bar{x}$ then equation (7) becomes

$$J(\bar{x}) = \sum_{i=1}^{L}(||x_i - \bar{x}||^2 - 2(x_i - \bar{x})^T(\bar{x} - \bar{x}) + L||\bar{x} - \bar{x}||^2) \tag{10}$$

$$= \sum_{i=1}^{L}(||x_i - \bar{x}||^2 - 2(0)^T(0) + L||0||^2) \tag{11}$$

$$= \sum_{i=1}^{L}||x_i - \bar{x}||^2. \tag{12}$$

Comparing equation (9) to that of equation (12), it is obvious that because of the extra term in equation (9), we can say

$$J(z) > J(\bar{x}). \tag{13}$$

Problem 4 Summary

Procedure

(a) Part (a)

- Add the term $(\bar{x} - \bar{x})$ to the original expression and carry out the algebra to get to the final expression

(b) Part (b)

- Carry out the algebra for what the inner product will evaluate to and show that the inner product in the term is zero

(c) Part (c)

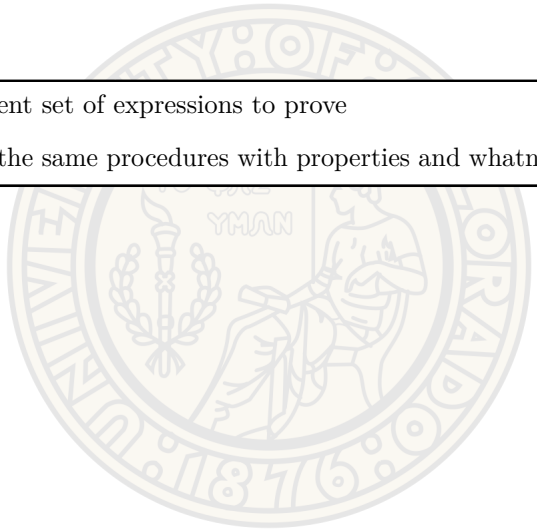
- Carry out the algebra again with this specific scenario to arrive to the final expression

Key Concepts

- Based off of specific scenarios, we can show that J_C is equivalent to certain values

Variations

- We can be given a different set of expressions to prove
  - We would then use the same procedures with properties and whatnot to prove the new expression



Problem 5

Problem Statement

Explain the solution to 4.2 here in your own words. (Since you are given a solution, you will be graded on your ability to explain).

Original Question

k-means with nonnegative, proportions, or Boolean vectors. Suppose that the vectors $x_1, \ldots, x_N$ are clustered using k-means, with group representatives $z_1, \ldots, z_k$.

(a) Suppose the original vectors x_i are nonnegative, i.e., their entries are nonnegative. Explain why the representatives z_j are also nonnegative.

(b) Suppose the original vectors x_i represent proportions, i.e., their entries are nonnegative and sum to one. (This is the case when x_i are word count histograms, for example.) Explain why the representatives z_j also represent proportions, i.e., their entries are nonnegative and sum to one.

(c) Suppose the original vectors x_i are Boolean, i.e., their entries are either 0 or 1. Give an interpretation of $(z_j)_i$, the i^{th} entry of the j group representative.

Hint. Each representative is the average of some of the original vectors. The group representation z_j is the average of the vectors x_k in the group:

$$z_j = \frac{1}{|G_j|} \sum_{k \in G_j} x_k.$$

Solution - Part (a)

For this problem, I will be using **Method 4**.

VMLS Solution:

(a) If the vectors x_k are nonnegative, the average z_j is a nonnegative vector.

Explanation:

With the equation for z_j, we have

$$z_j = \frac{1}{|G_j|} \sum_{k \in G_j} x_k. \tag{1}$$

In equation (1), since $|G_j|$ is nonnegative, it is impossible for z_j to be nonnegative as long as x_k is nonnegative. Therefore z_j is nonnegative.

Solution - Part (b)

For this problem, I will be using **Method 4**.

VMLS Solution:

(b) If each vector sums to one, $\mathbf{1}^T x_k = 1$ for all k, then the same is true for the average:

$$\mathbf{1}^T z_j = \frac{1}{|G_j|} \sum_{k \in G_j} \mathbf{1}^T x_k = \frac{|G_j|}{|G_j|} = 1.$$

Explanation:

For each vector to sum to one, a vector α_i would look like

$$\mathbf{1}^T \alpha_i = 1. \tag{2}$$

Applying equation (2) to z_j, z_j would then look like

$$\mathbf{1}^T z_j = \frac{1}{|G_j|} \sum_{k \in G_j} \mathbf{1}^T x_k. \tag{3}$$

According to VMLS, $|G_j|$ is the number of elements that belong to group j. This then means the term $\sum_{k \in G_j} \mathbf{1}^T x_k$ is going to have the same number of elements as group j. In turn, this means we can simplify equation (3) to now be

$$\mathbf{1}^T z_j = \frac{1}{|G_j|} \sum_{k \in G_j} \mathbf{1}^T x_k = \frac{|G_j|}{|G_j|} = 1. \tag{4}$$

z_j represents the proportions because the entries are nonnegative and sum to one.

Solution - Part (c)

For this problem, I will be using **Method 4**.

VMLS Solution:

(c) The i^{th} entry of group representative z_j is the fraction of the vectors in group j that have i^{th} entry one. If it is equal to one, all vectors in the group have i^{th} entry one. If it is close to one, most vectors in the group have i^{th} entry one. If it is zero, no vectors in the group have i^{th} entry one.

Explanation:

If the entries are boolean, this means that the entries are either 0 or 1. If an index $i = 1$, this means for $(z_j)_i$ we have

$$(z_j)_i = 1. \tag{5}$$

If the entry is 0, namely $i = 0$, we then have

$$(z_j)_i = 0. \tag{6}$$

This means for the vectors x_i, the entries in them are either going to be 0 or 1. When you sum the vectors to get a final value, the final value is going to represent a percentage of what vectors are 0 or 1. Essentially, z_j then represents what fraction of the individual vectors x_i are 1. If each individual vector x_i is 0, then z_j will be 0. If each individual vector x_i is 1, then $z_j = 1$.

Problem 5 Summary

Procedure

(a) Part (a)

- Reason that the value cannot be nonnegative due to the properties of the sum

(b) Part (b)

- Reason that the value is essentially the number of elements in a group divided by the number of elements in a group

(c) Part (c)

- Show what it means for a boolean vector and how it applies to this scenario

Key Concepts

- For a given set of vectors that are to be used in a k-means algorithm, there are specific conclusions that can be drawn from that specific example

Variations

- We could be given a different set of expressions to prove
  - This would require us to use properties from the given specific