Mathematics 327
Lab 1: k-means clustering
Due: Friday, January 18

Instructions: The exercises here should be completed in groups of 2 or 3 students with one write-up submitted from each group.

This lab explores a technique called *k-means clustering*, which is a useful tool in data science. In many cases, we have a large collection of data and believe that it breaks naturally into groups. For instance, we could be looking at medical data from a group of patients who have either type 1 or type 2 diabetes. We would expect that data from a type 1 patient would be similar to other type 1 patients and somewhat distinct from type 2 patients.

Now that we can measure the length of vectors, we can develop a measure of when two data points are similar. More specifically, we imagine representing the individual pieces of information corresponding to a single patient as a vector. If patients are similar, their vectors should lie close to one another. In the case of diabetes patients, we might expect that the vectors from the two types would cluster together. The problem is that we may have a lot of data about the patients—age, weight, etc—so the vectors may be high dimensional and hence not easily visualized.

In this lab, we will learn about an algorithm that automates the detection of clusters. We will look at a simple example first to understand the idea and then move on to a more realistic example.

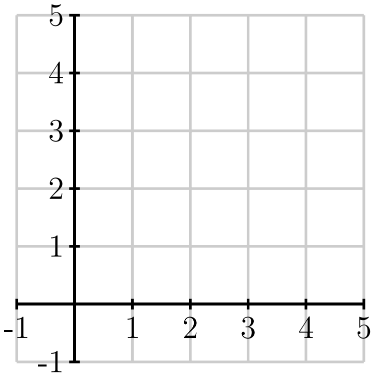
1. To begin, we need to introduce the *centroid* of a collection of vectors, which is just the average of the vectors. For instance, the centroid of $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is

$$\frac{1}{n}\left(\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_n\right).$$

   Find the centroid of the vectors

$$\mathbf{v}_1 = \left[\begin{array}{c} 1 \\ 1 \end{array}\right], \qquad \mathbf{v}_2 = \left[\begin{array}{c} 4 \\ 1 \end{array}\right], \qquad \mathbf{v}_3 = \left[\begin{array}{c} 4 \\ 4 \end{array}\right].$$

   Now sketch the vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ and their centroid below. (Rather than sketching the vectors, you may wish to just draw the points at the tips of the vectors. In this way, we can think of vectors as points.)

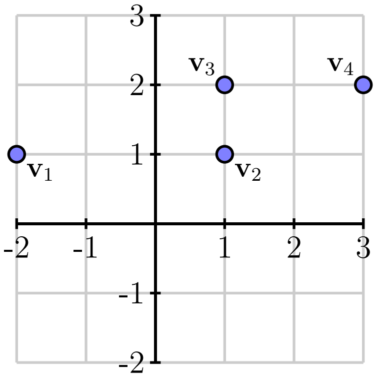


Notice how the centroid is roughly in the center of the points.

2. We will now look at a simple example to learn how to find a clustering. Suppose we have the following 4 vectors, which we think of as a small data set,

$$\mathbf{v}_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \qquad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \mathbf{v}_3 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \qquad \mathbf{v}_4 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

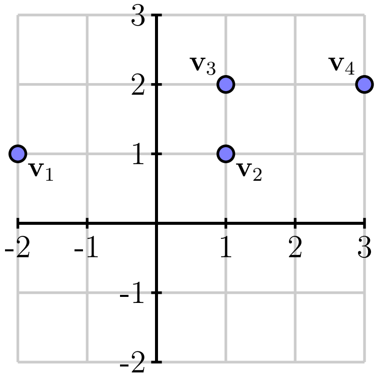
as shown below.



Suppose that we would like to group these points into two clusters (for the time being, don't worry about why we are looking for two clusters). Remember that we cannot usually see our data so we need an algorithm to find the clusters. We therefore begin by choosing two points c_1 and c_2 at random and declaring them to be the "centers" of the two clusters. Suppose we randomly choose $c_1 = \mathbf{v}_2$ and $c_2 = \mathbf{v}_3$ as the center of two clusters. The cluster centered on $c_1 = \mathbf{v}_2$ will be the set of points that are closer to $c_1 = \mathbf{v}_2$ than to $c_2 = \mathbf{v}_3$. State which of the four data points are in this cluster, which we denote by C_1, and circle them in the figure above.

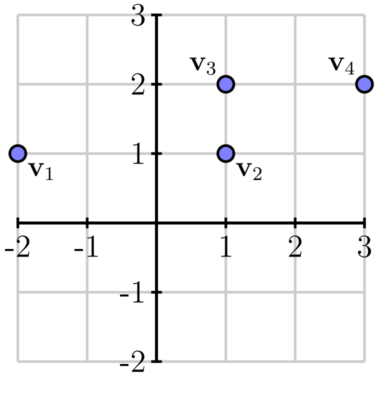
The second cluster will consist of the data points that are closer to $c_2 = \mathbf{v}_3$ than $c_1 = \mathbf{v}_2$. State which of the four data points are in this cluster, which we denote by C_2, and circle them in the figure above.

We have two clusters, but it may be that we can improve them so we will do the following. Now that we have clusters, we can find their centroids. So we redefine c_1 to be the centroid of cluster C_1 and c_2 to be the centroid of C_2. Find those centroids and indicate them on the diagram below.



Now we update the clusters C_1 and C_2 to be the set of data points closest to c_1 and c_2, respectively. State the data points in the clusters below and circle them in the figure above.

Let's perform the last step again. Find the centroids of the new clusters and then update the clusters. Indicate your centroids and clusters below.



You should have seen that we have the same clusters so repeating this step will continue to give us the same set of clusters so there is no point in repeating. We declare this to our final set of clusters.

So the algorithm is like this:

- Randomly choose k data points $c_1, c_2, \ldots, c_k$.
- Construct the cluster C_1 as the set of data points closest to c_1, C_2 as the set of data points closest to c_2, and so forth.
- Repeat the following until the clusters don't change:
  - Find the centroids $c_1, c_2, \ldots, c_k$ of the clusters.
  - Update the clusters.

3. It turns out that the clustering we find depends on the initial points $c_1, c_2, \ldots, c_k$ that we randomly choose. For instance, if we choose $c_1 = \mathbf{v}_3$ and $c_2 = \mathbf{v}_4$, we obtain the clusters $C_1 = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and $C_2 = \{\mathbf{v}_4\}$. This doesn't look like such a good clustering, but we would like to develop a quantity that assesses the quality of a set of clusters.

   We think of a set of clusters as being good if the points in a cluster are close to the cluster's centroid. For this reason, we measure the square of the distance from a data point to the centroid of its cluster. We say that the *objective* of a clustering is the average of the squares of these distances.

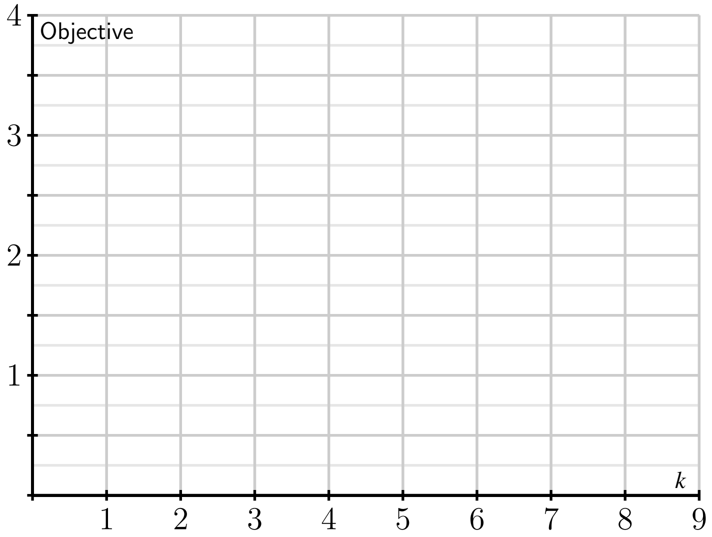
   Find the objective of the clustering that you found in the second part of this activity.

   The objective of the clustering with clusters $C_1 = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and $C_2 = \{\mathbf{v}_4\}$ is $5/3$ so we conclude that our original clustering is the better of the two. In fact, our original clustering has the smallest objective of any clustering, and you might think that it "feels" like the right one.

4. Given a value of k, we can now create clusterings with k clusters, and we can even assess how good the clusterings are with the objective measure. Our final question is to determine what the best value of k should be.

   Go to `cocalc` and open up the `k-means` Jupyter notebook. After you see the initial data set, state below how many clusters you think there should be.

5. Construct the plot below by finding the objective for the best clusterings when $k = 2, 3, 4, \ldots, 9$.



This plot is called an *elbow plot*. This name is meant to suggest something about its shape. Notice how the objective values decrease and then level out. We typically choose the value of k where the plot first levels out because this means that adding more clusters does not necessarily improve the clusterings. Given this elbow plot, what value of k would you choose for the final clustering? How does this compare to value of k you recorded above based on your visual observations?