Marcus Rose
Math 123 HW 4
Due 2/21/23

---

1. Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$. Let $F : \mathbb{R}^D \to [0, \infty)$ be

$$F(y) = \sum_{i=1}^{n} \|x_i - y\|_2^2.$$

Prove that F is minimized for $y = \frac{1}{n}\sum_i^n x_i$, i.e. at the mean of the sequence.

Proof A minimum is found via the root of the first order partial derivative. That is,

$$\frac{d}{dy}F(y) = \frac{d}{dy}\sum_{i=1}^{n} \|x_i - y\|_2^2 = \frac{d}{dy}\sum_{i}^{n} \|x_i^2 - 2xy - y^2\|_2 = \sum_{i}^{n} \| - 2x_i + 2y\| = 2(\sum_{n}^{i} -x_i + y)$$

Setting the derivative equal to zero finds the minimization.

$$0 = -2(\sum_{i}^{n} x_i) + 2ny \implies 2ny = \sum_{i}^{n} x_i \implies y = \frac{1}{n}\sum_{i}^{n} x_i,$$

which is equal to the mean of the sequence.

□

2. Suppose $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ are data, and outlier x^o is added with the property that for $\delta > 0$ fixed, $\|x_i - x^o\|_2 > \delta$ for all $i = 1, ..., n$. Suppose we run K-means on this data with $K = 2$.

   (a) Argue that as $\delta \to +\infty$, one of the clusters learned by K-means will consist only of x^o.

   K-means learns the data by creating k similar partitions of data grouped around k centroids. Setting δ arbitrarily large (even approaching ∞) will cause the norm defined between each x_i and the outlier x^o to be greater than that upper bound. So with only $k = 2$ clusters, there will be one cluster around the centroid nearest to all x_i, but because this upper bound is so large, the outlier will consist entirely of a cluster in its own right.

   (b) This lack of robustness to outliers is sometimes considered a defect of K-means. Suggest changes to the algorithm to improve its robustness to outliers.

   A simple (perhaps trivial) first approach would be to remove the outlier x^o in the first place, and then apply k-means. However, if for some reason you were unable to remove the outlier, you could perhaps modify k-means by calculating the distance between points on the basis of medians instead of means. Because the cluster's distance is squared, this has the amplifying effect for each cluster involved. If instead the 1-norm was used instead of the classic Euclidean distance measure, this would calculate k-medians. However, this has problems in its own right, though robustness to outliers would have some degree of success.

(c) Instead of thinking of the lack of robustness to outliers as a defect, are there any virtues?

The outliers are easily removable/identifiable if they are in a cluster of their own. That is, if clusters are a mechanism to identify, then clusters containing outliers are identifiable in their own right.

3. K-means is often combined with a feature extraction step in which the data to be clustered is first transformed to a more convenient form. As the course progresses, we will consider some data-dependent feature extraction methods, but for now, let us consider a very particular feature extraction method: converting Cartesian to polar coordinates in $\mathbb{R}^2$.

(a) Load the data in and run K-means with K = 2, displaying your labels as colors on the plotted data. In terms of the K-means functional, why does this method produce the "incorrect" clusters it does?

```
% Concatenate data and plot
X=vertcat(X1,X2);

close all;
scatter(X(:,1),X(:,2));
title('Circular data to cluster');


% Run K-means with K=2
data_1 = X;
data_2 = X;

[theta, rho] = cart2pol(data(:,1), data(:,2));
pdata = [rho, theta];

% [idx, centroids] = kmeans(data_1, 2);
[idx, centroids] = kmeans(pdata, 2);

% Plot the data with different colors for each cluster
figure;
hold on;
scatter(data(idx==1,1), data(idx==1,2), 'r');
scatter(data(idx==2,1), data(idx==2,2), 'b');
scatter(centroids(:,1), centroids(:,2), 'k', 'filled');
legend('Cluster 1', 'Cluster 2', 'Centroids');
xlabel('x');
ylabel('y');
title('K-means Clustering with K=2 (polar)');
```

Listing 1: Code for k-means on concentric ellipses

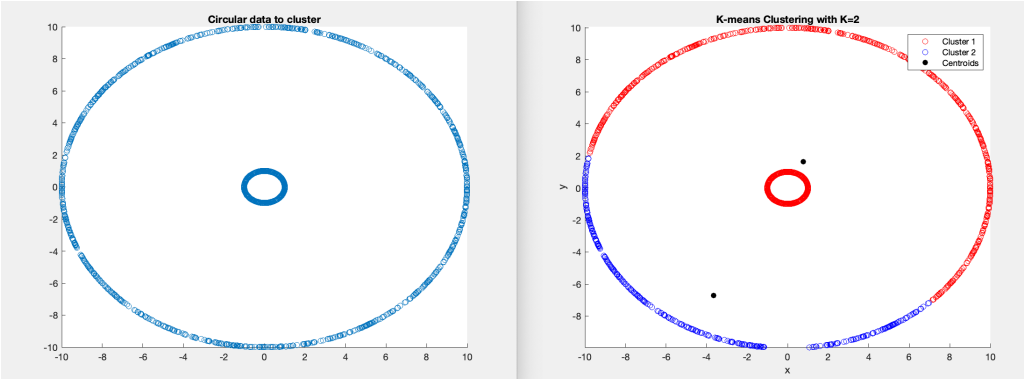

Figure 1: K-means on two ellipses

(b) Convert the data to polar coordinates and run K-means again to show the data can be correctly labeled in this case.

Note in Figure 2, the data is almost perfectly labeled.

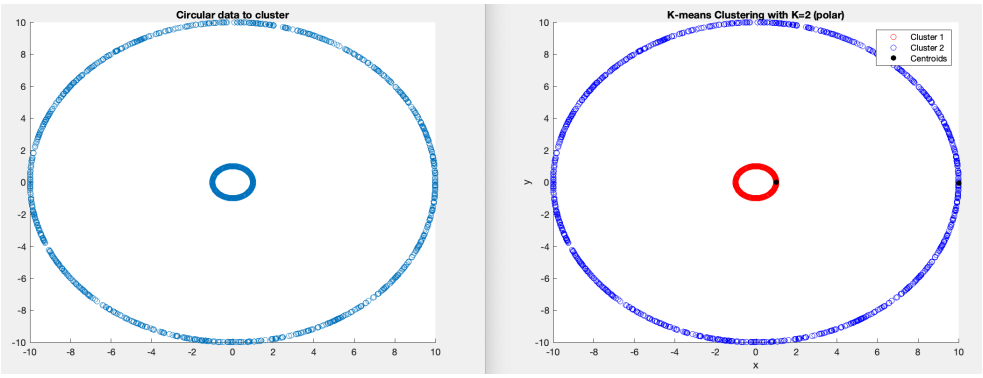

Figure 2: Same calculations, but with polar coordinates

(c) Explain what about the polar coordinate representation is convenient for this data.

Polar coordinates clusters the data in the correct way because polar coordinates represent the data in a grid of circles, while the Cartesian plane represents data as straight lines. This is due to the fact that radial distance better captures the spherical nature of the ellipse. Note that the shape of the ellipse respects the symmetry of the plane in which it lies (i.e. the polar plane). Hence, clusters around the data would also respect such symmetry by design: with a radial distance and angle formed from the center reaching to the edge of the ellipse.

4. (a) In MATLAB, create a data set in which single linkage and complete linkage hierarchical clustering differ substantially. Demonstrate this by computing the dendrograms using the built-in 'linkage' function in MATLAB, and arguing that they capture different structure in the data.

Below is a snapshot of the code used to produce a linkage function that effectively is data along a chain:

```
% create data in the shape of a chain
n = 20;
```

```
data = zeros(n,2);
for i = 1:n
    data(i,:) = [i, 0.1*i];
end

% data = rand(100, 2);
% Z_single = linkage(data, 'single');

% slc
figure;
link_single = linkage(data, 'single');
dendrogram(link_single);
title('Single Linkage Dendrogram');

% clc
figure;
link_complete = linkage(data, 'complete');
dendrogram(link_complete);
title('Complete Linkage Dendrogram');
```

Listing 2: Code for k-means on concentric ellipses

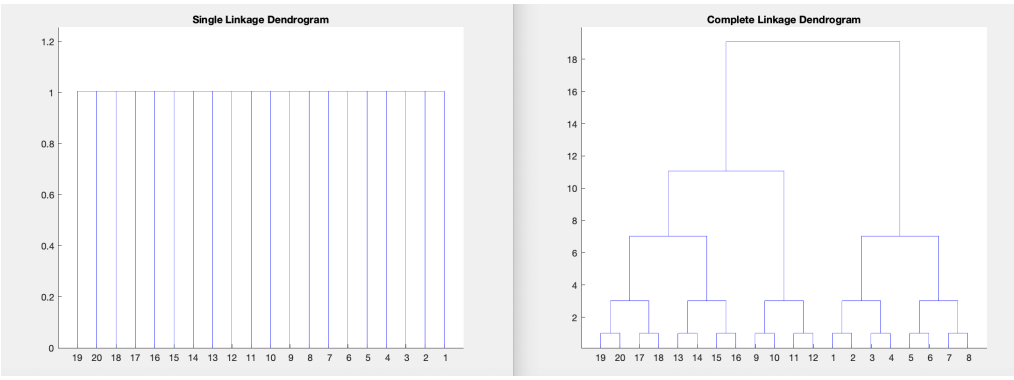

Figure 3: Same calculations, but with polar coordinates

(b) Argue why the two linkage methods differ on this data in terms of their mathematical formulation.

Note the stark differences in the dendograms in Figure 3. The data chosen for the linkage function models that of a chain. The single linkage computes the distance from from each "link" in the chain, while the complete linkage uses the maximum distance, that is, the entire length of the chain, these distances are extremely different from one another and hence their respective dendrograms are very different too.