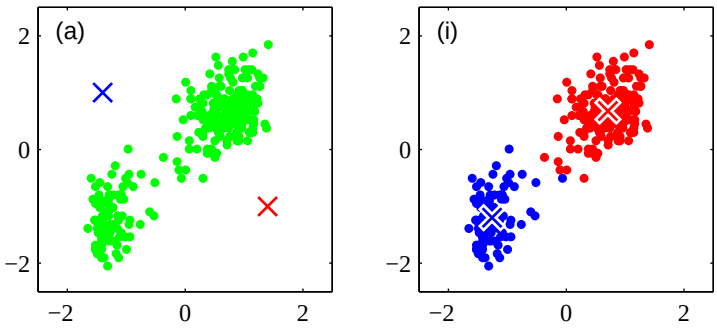


Question 1. How is clustering different from other machine learning techniques we have covered?

Why Cluster?

Question 2. What are some applications for clustering that interest you?

The k-Means Clustering

Basic k-Means Algorithm (Lloyd's):

Goal: Minimize distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|_2^2$$

- **Step 0** Initialize $\{\boldsymbol{\mu}_k\}$ to some values

- **Step 1** Optimize $\{r_{nk}\}$ values.

- **Step 2** Optimize $\{\boldsymbol{\mu}_k\}$ values.

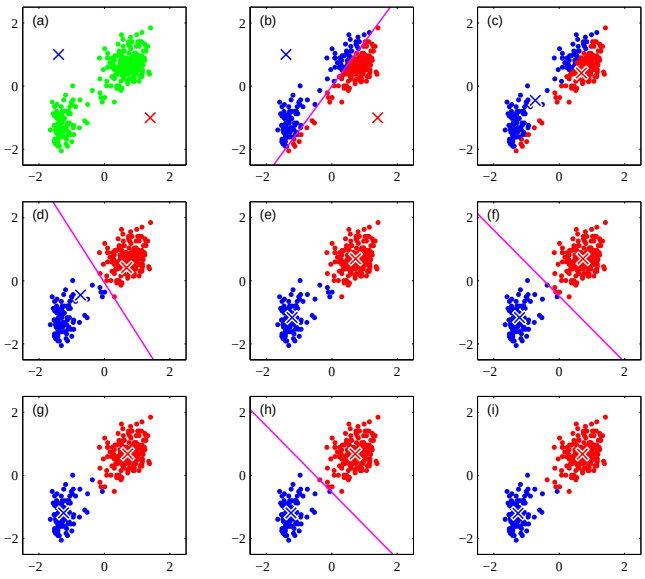
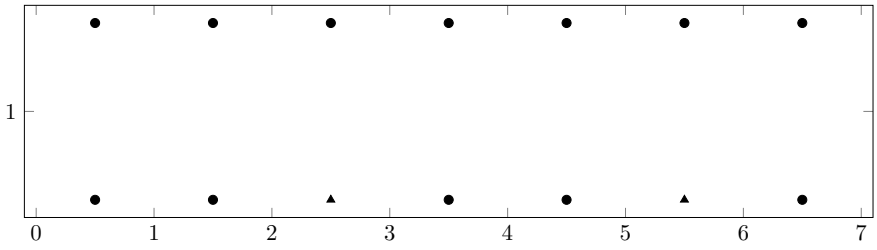


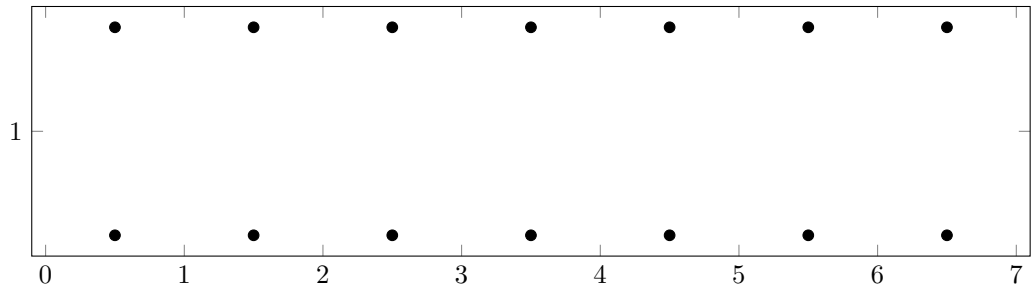
Figure 1: An example nine steps (four iterations) for Lloyd's algorithm

Question 3. Consider the following dataset. All points are unlabeled and part of the same set. Use the points indicated by triangles as the initial values of $\boldsymbol{\mu}$ for $k = 2$. When the algorithm converges, there will be a clear dividing line between the two clusters. Draw the line clearly in the diagram.

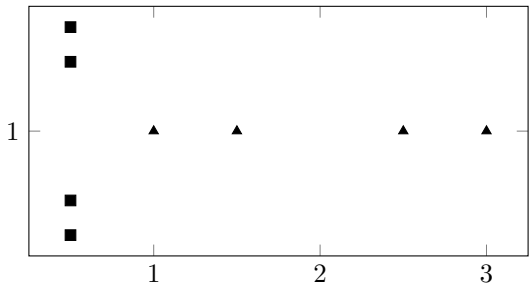


Question 4. On the diagram below, draw two triangles with the following properties:

- Do not draw the same triangles as the problem statement on the previous page.
- Choose two points such that, if we run Lloyd's k-means algorithm with these two as the initial points, we will get the same clustering you found for the previous problem. Draw a triangle at these points.
- You are not obligated to select a data point as a triangle, although you may do so.



Question 5. I might have run Lloyd's algorithm on the following data set. If I did so, I ran the algorithm until it fully converged (another iteration of the **while** loop would have the same partitioning / means). The points assigned to $\boldsymbol{\mu}_1$ are drawn as triangles and the points assigned to $\boldsymbol{\mu}_2$ are drawn as squares. I have drawn the result below:

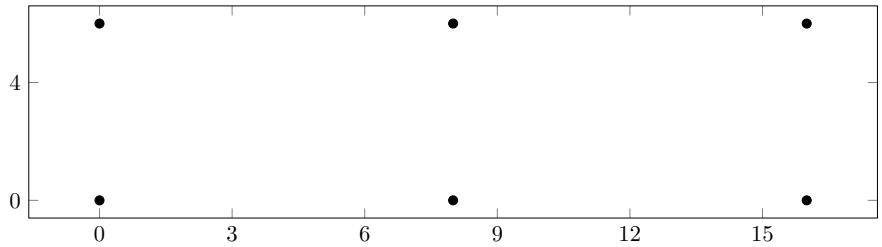


Could this actually be the result of a run of that algorithm? Indicate if you think the answer is "yes" or "no" and explain your answer briefly.

Question 6. There is a set S consisting of 6 points in the plane shown as below, $a = (0, 0)$, $b = (8, 0)$, $c = (16, 0)$, $d = (0, 6)$, $e = (8, 6)$, $f = (16, 6)$. Now we run Lloyd's Algorithm on those points with $k = 3$. The algorithm uses the Euclidean distance metric (i.e. the straight line distance between two points) to assign each point to its nearest centroid. Ties are broken in favor of the centroid to the left/down. Two definitions:

- A k-starting configuration is a subset of k starting points from S that form the initial centroids, e.g. $\{a, b, c\}$
- A k-partition is a partition of S into k non-empty subsets, e.g. $\{a, b, e\}$, $\{c, d\}$, $\{f\}$ is a 3-partition.

Clearly any k-partition induces a set of k centroids in the natural manner. A k-partition is called stable if another iteration of Lloyd's Algorithm with the induced centroids leaves it unchanged.



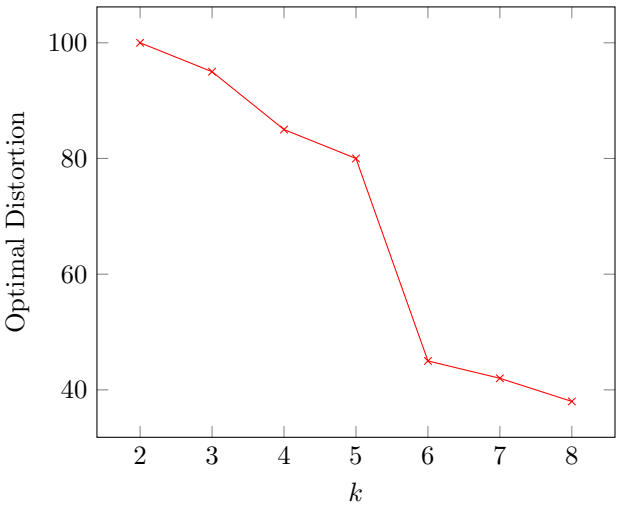
Note that there are 20 (6 choose 3) starting configurations for this data, and not all 3-partitions are listed below.

Fill in the table below.

3-partition	Stable?	An example 3-starting configuration that can arrive at the 3-partition after 0 or more iterations (or write "none" if no such 3-starting configuration exists)	How many 3-starting configurations arrive at this 3-partition?
$\{a, b, e\}$, $\{c, d\}$, $\{f\}$			
$\{a, b\}$, $\{d, e\}$, $\{c, f\}$			
$\{a, d\}$, $\{b, e\}$, $\{c, f\}$	Yes	a,b,c	8
$\{a\}$, $\{d\}$, $\{b, c, e, f\}$			
$\{a, b\}$, $\{d\}$, $\{c, e, f\}$		none	0
$\{a, b, d\}$, $\{c\}$, $\{e, f\}$			

How to choose k?

Suppose we don't have any idea what value of k to use for our data set; we think there might be some meaningful clustering, but we aren't sure (among other things) how many clusters to even look for. We run k-means several times for different values of k and find the following information:



Question 7. What do you suppose the "natural" cluster count for this data is?