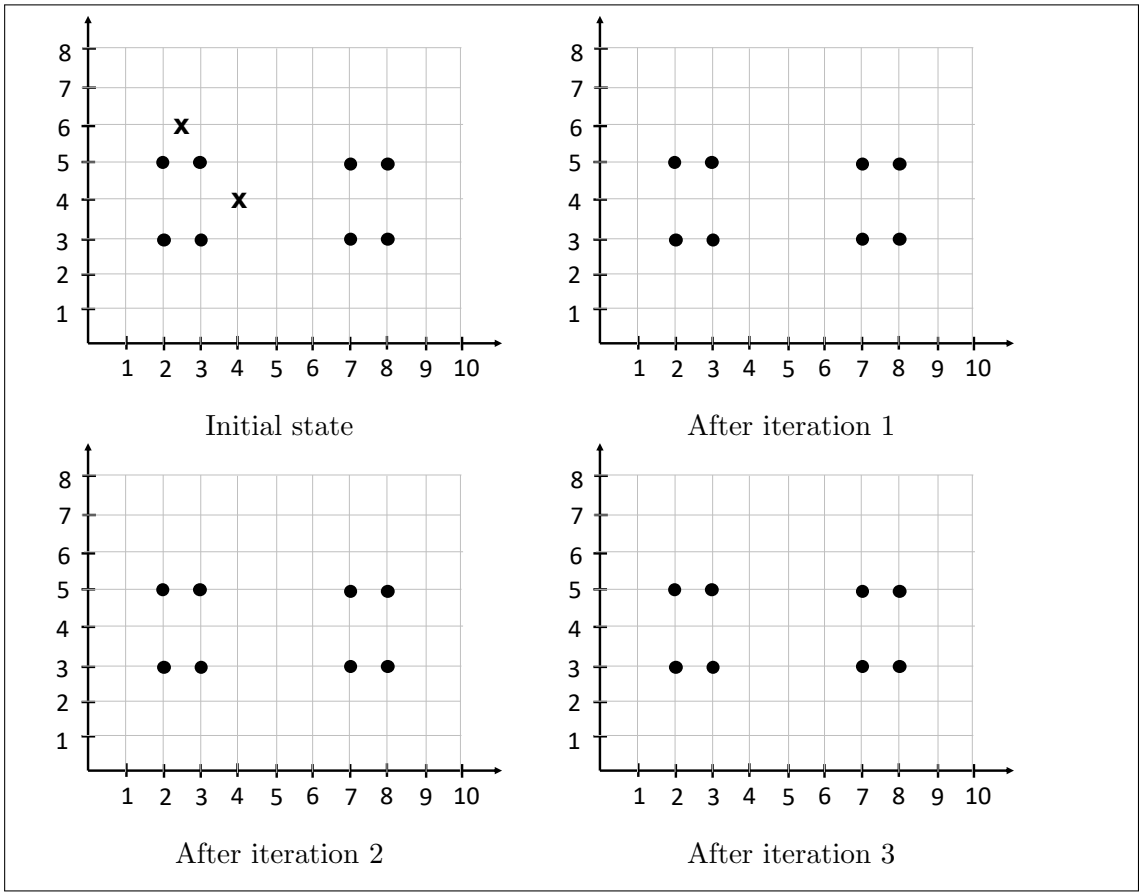
**Name:** _____

## Clustering

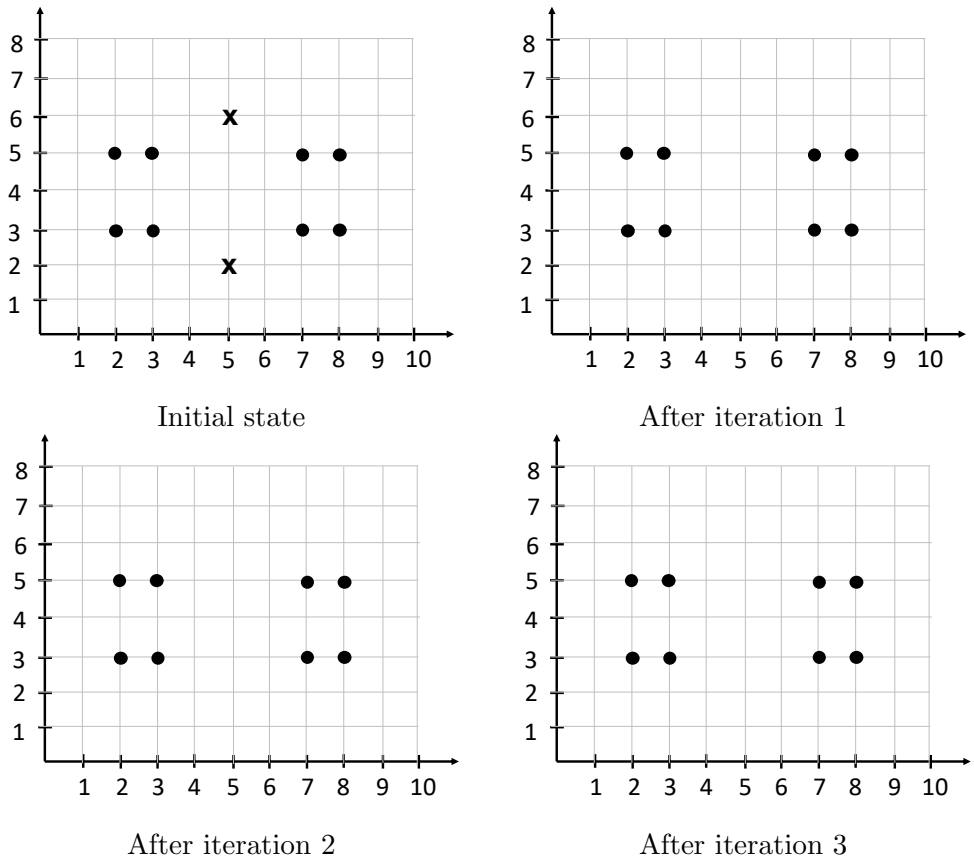
2. (12 points) Assume that the number of clusters $k = 2$ for all of the following questions.

   (a) Walk through each step of the $k$-means algorithm, beginning with the initialization shown in the plot in the top left of the box below. Dots show the observed data. In each plot (go left to right, top to down), mark with two '**x**' symbols where the cluster centers are in that iteration of $k$-means. These are already shown in the initial state. Once the $k$-means algorithm has converged, you can leave all subsequent plots unmarked.




   (b) What is the numerical value of the $k$-means objective for the clustering found in (a), after the algorithm has finished running?

(c) Just as in (a), walk through each step of the $k$-means algorithm, beginning with the initialization shown in the plot in the top left. In each plot (go left to right, top to down), mark with two '**x**' symbols where the cluster centers are in that iteration of $k$-means. Once the $k$-means algorithm has converged, you can leave all subsequent figures unmarked.

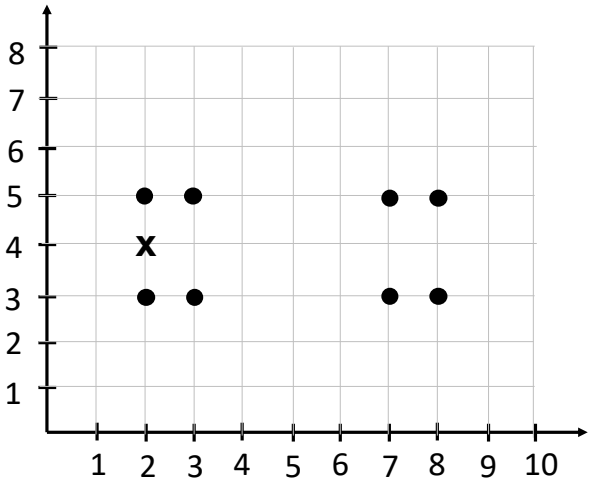



(d) What is the numerical value of the $k$-means objective for the clustering found in (c), after the algorithm has finished running?

(e) According to the $k$-means objective of the learned clusters, which initialization was better?

◯ Initialization (a) ◯ Initialization (c)

(f) Consider the data in black dots shown in the plot below. We drew one cluster center with an **x** symbol at (2,4). Draw the second cluster center to satisfy the following property. When we initialize the clusters centers at the two **x**'s and run the k-means algorithm to convergence, the final state will be such that one cluster will have all the data points assigned to it, and the other cluster will have no data points assigned to it.



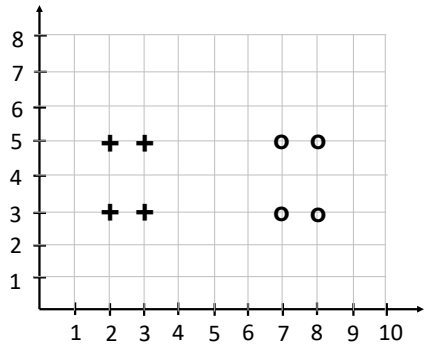

(g) Christy thinks she came up with a compelling new initialization method for the $k$-means algorithm. Looking at her code below, explain why it is unlikely to give good results.

```
def kmeans_init(X, n_clusters):
    centers = []
    for i in range(n_clusters):
        centers.append(X[:, X.shape[1]-1-i])
    return np.asarray(centers).T
```

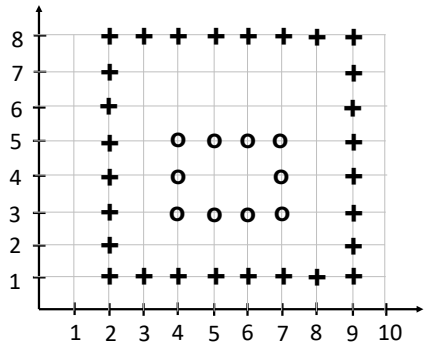
**Name:** _____

(h) Each of the following five data sets has two ground truth clusters, whose points are denoted as '+' and '**o**'. For which of these would the clustering with the smallest $k$-means objective value **not** recover the ground truth? Assume $k = 2$. (Select all that apply.)

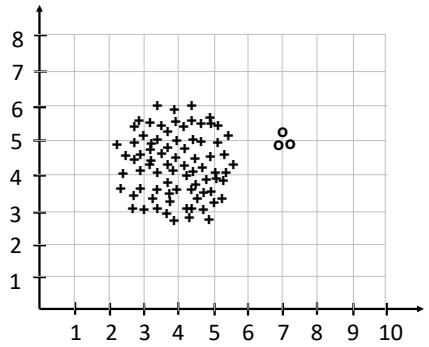
◯ (I)
◯ (II)
◯ (III)
◯ (IV)
◯ (V)



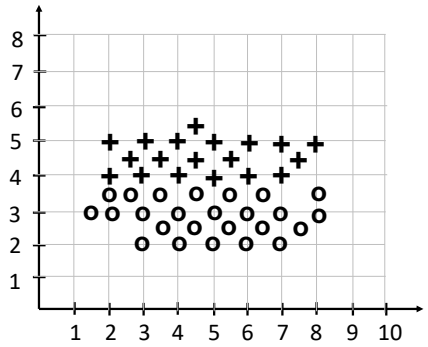
(I)



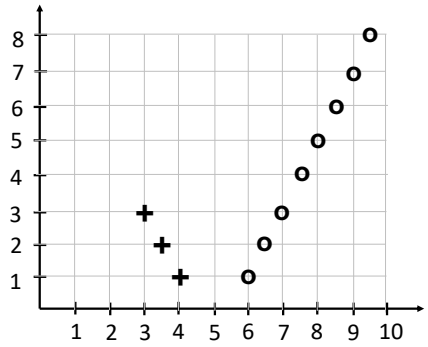
(II)



(III)



(IV)



(V)