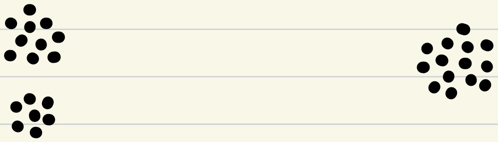
Exercises for K-means

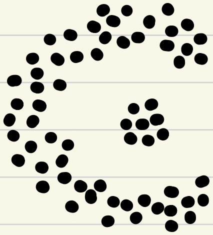
1. Consider the following dataset in the plane $\mathbb{R}^2$.



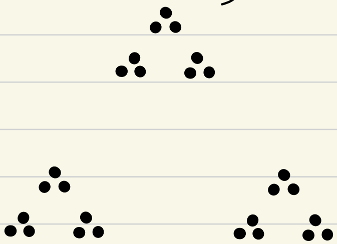
(i) How would you assign K=2 clusters to minimize error?
(ii) How would you assign K=3 clusters to minimize error?
(iii) Find an initial assignment of K=3 initial seed centers that shows that the iterative K-means algorithm might converge to a local min that is not global, that is, to clusters that don't minimize energy.

2. (i) How many clusters do you see?

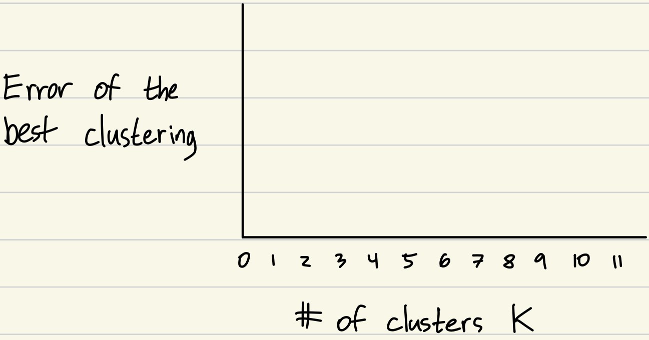
(ii) What results might K-means give you on this dataset?



3. (i) How many clusters would you say that the following dataset has?



(ii) Explain why as K increases, the error of the best clustering (with K clusters) never decreases.



(iii) Given a new dataset, what strategies might you propose for choosing the # of clusters K?

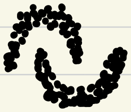
4. What are various pros and various cons that you can think of for the K-means clustering method?

<u>Pros:</u> <u>Cons:</u>

Coding exercises for K-means

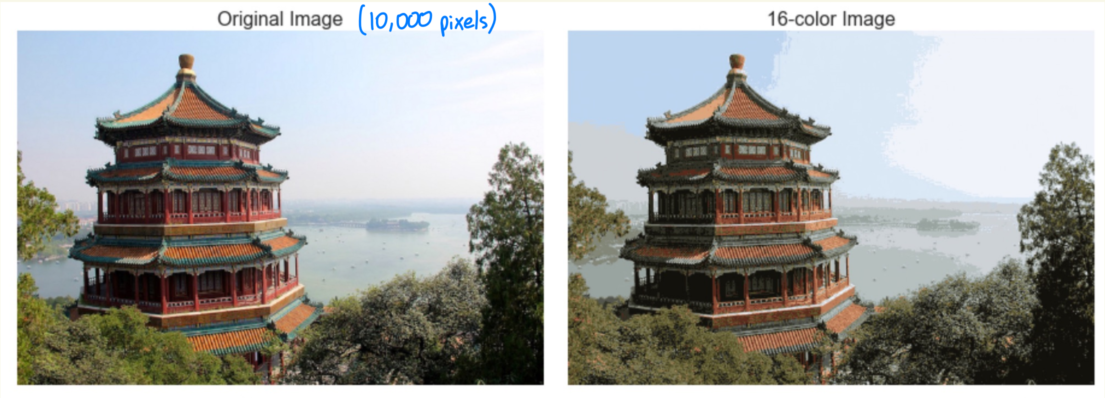
(i) From the course GitHub page, download and run the jupyter notebook 05.11 - K-Means.ipynb.

(ii) In the "two moons" example, how many clusters are needed until no cluster contains points from both moons?



(iii) In the digits example (#'s 0 - 9), which cluster center looks the least like a number? Why do you think this is?

(iv) In the color reduction application, change the # of colors in the simplified image. What does the K-means error represent in this application?



Original Image (10,000 pixels)

16-color Image