CSE 416: Introduction to Machine Learning
Assignment #6
Instructor: Pemi Nguyen
due: Friday, May 13, 2022, 11:59 pm.

Instructions:

- **Answers:** Please provide detailed yet concise explanations for your work.
- **Turn-in:** Do not write your name on your pages (your Gradescope account will identify you to us) and do not include a copy of the exercise's question in what you turn in. You must use Gradescope to upload your written homework solutions. You will submit a single PDF file containing your solutions to all the exercises in the homework. You must follow the Gradescope prompts that have you link exercise numbers to your pages. **We do not accept handwritten solutions**, so please typeset them, such as Microsoft Word or LaTeX editor.

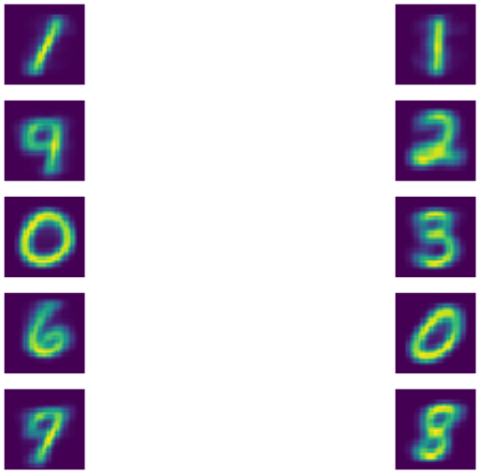
1. *(2 points)* Briefly describe two main differences between clustering and other classification methods.

2. *(3 points)* Clustering is more preferred than classifications when you don't know how many possible categories a dataset can have. Please describe how you can apply clustering to solve a real-life problem in 2-3 sentences. Some possible applications are market segmentation, social network analysis, search result grouping, medical imaging, image segmentation, anomaly detection, etc. Please do not use the examples mentioned in class.

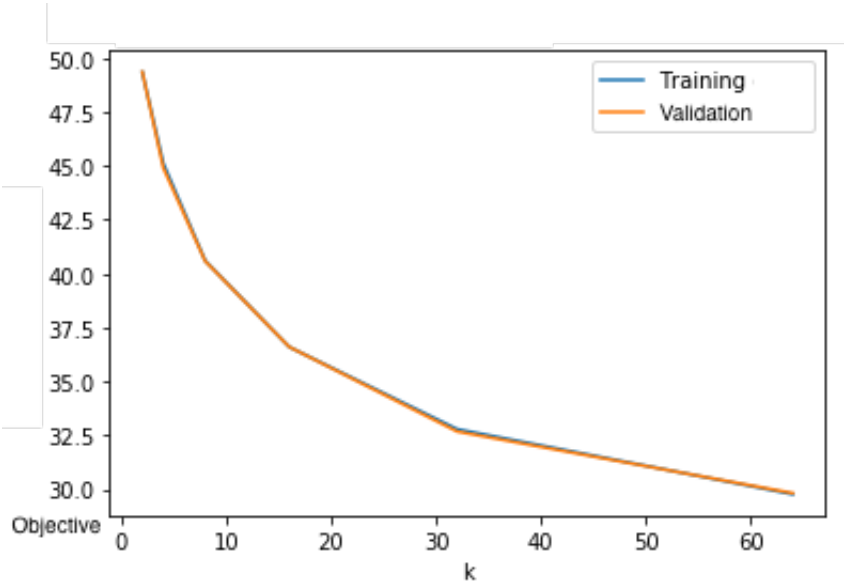
3. *(3 points)* MNIST is a dataset consisting of handwritten digits from 0 to 9 and commonly used for training various image processing systems. Below are some images from this dataset.

(a) *(1 point)* Even though this dataset has labels for the images, assume you want to run k-means clustering algorithm on this dataset without accessing the labels. What is a good choice of k, the number of clusters?

(b) *(1 point)* Below are the images visualized at the cluster centers when using $k = 10$. Does the k-means algorithm do a good job at clustering the dataset?



(c) *(2 points)* Suppose we split the dataset into a training and validation set and plot the objective values on both sets with increasing values of k.



Remember that we can use validation set approach to fine-tune hyperparameters for supervised learning methods. From the above plot, does it seem to be a good idea to use the validation set to fine-tune k for k-means clustering? Briefly explain why.

4. **Extra credit**

(a) *(2 points)* Prove that the objective function for the k-means algorithm $L(\mu, y) = \sum_{i=1}^{n} \left\| x^{(i)} - \mu_{y^{(i)}} \right\|_2^2$ converges. You can prove that the difference between the objective values between a current iteration and the next iteration is always greater than 0.

(b) *(1 point)* In practice, people sometimes use k-medians clustering instead of k-means clustering. What is the benefit of using the median metric instead of mean?

Hint: Assume a cluster has the following datapoints in 1 dimension [1, 4, 4, 9, 10]. If we change 10 to 100, what is the new centroid when using mean and when using median?