Honor code. This assignment is individual work. The goal of this assignment is for you to put in practice the concepts we learned in the video recordings, as well as explore complementary concepts. As mentioned in the synchronous lecture, academic integrity will be strictly enforced. If for any reason you are tempted to cheat (i.e., because you are facing personal hardship), contact the instructor immediately by email.

Instructions. To facilitate grading, please follow the following guidelines:

- You should submit a single PDF with all of your answers in the same order than this handout.
- You can use any tool you want to generate the PDF (e.g., word, LaTeX, scan your handwriting), but each page of the PDF should be easy to read and oriented properly.
- If you decide to handwrite your exam rather than typeset it, ensure your handwriting is readable otherwise TAs will have the discretion to not grade your answer.
- Clearly state any questions that you skip by writing down the question number along with "I skip this question"
- Graphs produced should be clearly interpretable. Include labels on axes and a legend.
- Attach a python script (.py) or an iPython notebook for the questions that require handing in code.
- Only your last Quercus submission will be graded. Make sure it contains a single PDF for your solutions and one .py/.ipynb script for your code,

Assignment structure. The assignment contains 8 questions worth a total of 20 points.

Problem 1 - Clustering with κ-means We will consider the UCI ML Breast Cancer Wisconsin dataset. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. You can download the dataset using the function:

`sklearn.datasets.load_breast_cancer`

Unless specified otherwise, questions in Problem 1 below refer to the UCI ML Breast Cancer Wisconsin dataset when the word "dataset" is used.

1. (5 points) Implement κ-means yourself. Your function should take in an array containing a dataset and a value of κ, and return the cluster centroids along with the cluster assignment for each data point. You may choose the initialization heuristic of your choice among the two we saw in class. Hand-in the code for full credit. **For this question, you should not rely on any library other than `numPy` in Python.**

2. (1 point) Run the κ-means algorithm for values of κ varying between 2 and 7, at increments of 1. Justify in your answer which data you passed as the input to the κ-means algorithm.

3. (2 points) Plot the distortion achieved by κ-means for values of κ varying between 2 and 7, at increments of 1. Hand-in the code and figure output for full credit. For this question, you may rely on plotting libraries such as `matplotlib`.

4. (1 point) If you had to pick one value of κ, which value would you pick? Justify your choice.

Problem 2 - Lack of optimality of κ-means

1. (3 points) Construct an analytical demonstration that κ-means might converge to a solution that is not globally optimal. *Hint:* consider the case where $\kappa = 2$ and the dataset is made up of 4 points in $\mathbb{R}$ as follows: $x^{(1)} = 1, x^{(2)} = 2, x^{(3)} = 3, x^{(4)} = 4$. Initialize κ-means with the centroids $\mu_1 = 2$ and $\mu_2 = 4$. *Note:* you may assume that if a point $x^{(i)}$ is equally distant to multiple centroids μ_k, the point will be assigned to the centroid whose index is smallest, i.e., k with the smallest value for $k \in \arg\min_k \|x^{(i)} - \mu_k\|^2$.

Problem 3 - SVD This problem will help you review background required to understand an upcoming lecture video on PCA, which involves singular value decomposition (SVD). Consider the following matrix:

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 1 \end{bmatrix}$$

All following questions should be answered "by hand", i.e., you should derive the results analytically—not using a computer program. You may only use your computer as a simple calculator: this means you can for instance use the computer to multiply two real numbers, but **not** to compute the matrices asked for in the question using MATLAB. In the following, we use the notation $U\Sigma V^\top$ to denote the SVD of matrix A

1. (1 point) Show that the matrix A is of rank 2.

2. (2 points) Show that the singular values of the matrix A are $\sigma_1 = \sqrt{10 + \sqrt{97}}$ and $\sigma_2 = \sqrt{10 - \sqrt{97}}$

3. (5 points) Derive the matrices U and V for the SVD of matrix A.

*
* *