Listen
(https://app.readspeaker.com/cgi-bin/rsent?customerid=11151&url=https%3A%2F%2Fe.centennialcollege.ca%2Fcontent%2Fenforced%2F694716-COMP257001_2021F%2FProject%25202%2520K-Means%2520Clustering.html&lang=en_us&voice=11151&voice=Kate&readid=d2l_read_element_1)

Assignment 2: K-Means Clustering

Discover & Learn

Important Information

School of Engineering Technology and Applied Science

Information and Communication Engineering Technology

Unsupervised and Reinforcement Learning (COMP257)

- **K-Means Clustering (10%)**
- **Due Date: Friday of Week 4 by 11:59 pm EST (late penalty at 10 points per day)**
- **Upload your assignment here: Assignment 2: K-Means Clustering (/d2l/common/dialogs/quickLink/quickLink.d2l?ou=694716&type=dropbox&rcode=CENCOL-3439542)**

Instructions

- You are free to choose any toolkits to solve the problems at hand (e.g., TensorFlow, Sci-Learn, etc.)
- All written reports and codes are to be maintained on a repository of your choice such as Github. The course instructor will discuss and exchange with you information to get access to your code.
- The video presentation will be required as part of the submission that documents the steps taken to obtain the results.
- **IMPORTANT NOTES:**
  - 1 point will be deducted for each incident that does not conform to the requirements (e.g., code not properly formatted, comments not relevant to support documentation of code, missing code documentation, etc.).
  - All points will be deducted for submission of *nonsensical code* (i.e., code that doesn't contribute to the relevancy of the task at hand). This is question-specific.

Questions

- **Read the question below carefully.**

Question 1

[100 points]

1. Retrieve and load the Olivetti faces dataset [5 points]
2. Split the training set, a validation set, and a test set using stratified sampling to ensure that there are the same number of images per person in each set. Provide your rationale for the split ratio [10 points]
3. Using k-fold cross validation, train a classifier to predict which person is represented in each picture, and evaluate it on the validation set. [30 points]
4. Use K-Means to reduce the dimensionality of the set. Provide your rationale for the similarity measure used to perform the clustering. Use the silhouette score approach to choose the number of clusters. [25 points]
5. Use the set from (4) to train a classifier as in (3) using k-fold cross validation. [30 points]