Ira A. Fulton Schools of Engineering
Arizona State University

CSE 575: Statistical Machine Learning

K-means-Strategy Project

Purpose

In this project, you are required to implement the K-means algorithm and apply your implementation on the given dataset (AllSamples.npy), which contains a set of 2-D points. You are required to implement two different strategies for choosing the initial cluster centers.

Objectives

Learners will be able to:

- Implement the K-Means algorithm
- Evaluate its performance with two different strategies for choosing initial cluster centers.
- Compute final coordinated of the centroids and loss values
- Test the various cluster counts

Technology Requirements

Algorithms:

- k-Means Clustering

Resources:

- A 2-D dataset to be provided

Workspace:

- Any Python programming environment
- Ed Lessons

Software:

- Python environment

Language:

- Python

Project Description

In this project, you will be implementing the K-Means algorithm and applying it to a given dataset containing a set of 2-D points. You are required to implement two different strategies for choosing the initial cluster centers and evaluate the performance of each strategy.

In part 1 of this project, your task is to implement the K-Means algorithm with the first strategy involving randomly picking the initial centers from the given samples and test your implementation on the provided dataset, varying the number of clusters from 2 to 10. For each cluster count, compute the final coordinates of the centroids and calculate the loss based on the objective function.

In part 2 of this project, your task is to implement the K-means algorithm with a second strategy for choosing the initial cluster centers involving randomly picking the first center and selecting the subsequent centers based on the sample that maximizes the average distance to all previous centers. Again, test your implementation on the provided dataset, varying the number of clusters from 2 to 10. For each cluster count, compute the final coordinates of the centroids and calculate the loss based on the objective function.

Accessing Ed Lessons

You will complete and submit your work through Ed Lessons. Follow the directions to correctly access the provided workspace:

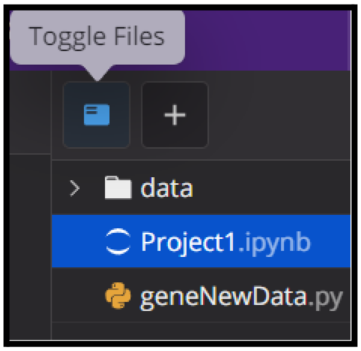
1. Go to the Canvas Assignment, "**Submission: K-means-Strategy Project**".
2. Click the “**Load Submission…in new window**” button.
3. Once in Ed Lesson, select the assignment titled "**K-means-Strategy Project**".
4. Select a code challenge to work on:
    a. To start part 1, click on the “**Part 1: K-means-Strategy Project**”
    b. To start part 2, click on the “**Part 2: K-means-Strategy Project**”

5. In a code challenge, first review the directions and resources provided in the description.
6. When ready, start working in the notebook titled “**project2_part#.ipynb**”.

Part 1 Directions

Download Mat File: “**CSE 575_K-means-Strategy Project_AllSamples**” (attached in the Project Overview page in the course). This file contains datasets to complete your project work.

The Ed notebook files and data can be downloaded by selecting the “**Toggle Files**” icon in the workspace (first option in the right corner).



Lab Directions

You are required to implement the following strategy for choosing the initial cluster centers.

- Part 1 is to randomly pick the initial centers from the given samples.
- You need to test your implementation on the given data, with the number *k* of clusters ranging from 2-10, output the final coordinate of the centroids and compute the loss based on the objective function.
  - (Referring to the course notes: When clustering the samples into k clusters/sets D_i, with respective center/mean vectors μ_1, , $\mu_2,\ldots\mu_k$, the objective function is defined as $\Sigma_{i=1}^{k}\Sigma_{x\in D_i}\left|\left|x-\mu_i\right|\right|^2$

Required Tasks

1. Write code to implement the k-means algorithm with Strategy 1.

2. Use your code to do clustering on the given data; compute the objective function as a function of k (k = 2, 3, …, 10).
3. Repeat the above step with another initialization.
4. Plot the following graphs:
    a. Loss (objective) function vs number of cluster
    b. Plot to show the final clustering and centroids for all the clusters.
5. Submit a report summarizing the results under different settings described above. You can also include the plots in your report to support your analysis

Note: You should implement your own K-means algorithm.

Preparing the Deliverables

Part 1 Results Submission & Output

Submitting your work through Ed Lessons will create your results submission. As an output from your Notebook(project2_part1.ipynb), you should have your final centroids and loss values.

1. **Final centroids output format:**

    [[x1, y1], [x2, y2]]

    [[x1, y1], [x2, y2], [x3, y3]]

    .

    .

    .

    [[x1,y1],[x2, y2],...[x9,y9]]

2. **Loss function output format:**

    L1

    L2

    .

    .

.

L9

Part 1 Report Submission

Draft a report to go with your Part 1 Results Submission. The report must contain:

- Your full name and student ID number on the first page in the upper left corner
- A detailed description of your observations and analysis of each strategy
  - You may include the plots for the objective function in your report to support your analysis

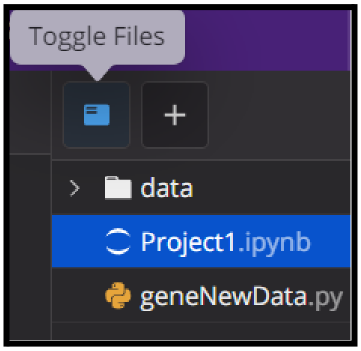
The report must also follow the required format:

- A maximum font size of 12pt
- A maximum length of two (2) pages (8x11 or A4 paper).
- Saved as a PDF (.pdf) file type

Directions: Part 2

Download Mat File: "**CSE 575_K-means-Strategy Project_AllSamples**" (attached in the Project Overview page in the course). This file contains datasets to complete your project work.

The Ed notebook files and data can be downloaded by selecting the "**Toggle Files**" icon in the workspace (first option in the right corner).



Lab Directions

You are required to implement the following strategy for choosing the initial cluster centers:

- Part 2 is to pick the first center randomly; for the i-th center (i >1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i -1) centers is maximal.
- You need to test your implementation on the given data, with the number k of clusters ranging from 2-10, output the final coordinate of the centroids and compute the loss based on the objective function.
    - (Referring to the course notes: When clustering the samples into k clusters/sets D_i, with respective center/mean vectors μ_1, , μ_2,…μ_k, the objective function is defined as $\Sigma_{i=1}^{k}\Sigma_{x\in D_i}\left|\left|x - \mu_i\right|\right|^2$

Required Tasks

1. Write code to implement the k-means algorithm with Strategy 2.
2. Use your code to do clustering on the given data; compute the objective function as a function of k (k = 2, 3, …, 10).
3. Repeat the above step with another initialization.
4. Plot the following graphs:
    a. Loss (objective) function vs number of cluster
    b. Plot to show the final clustering and centroids for all the clusters.
5. Submit a report summarizing the results under different settings described above. You can also include the plots in your report to support your analysis

Note: You should implement your own K-means algorithm.

Preparing the Deliverables

Part 2 Results Submission & Output

Submitting your work through Ed Lessons will create your results submission. As an output from your Notebook(project2_part2.ipynb), you should have your final centroids and loss values.

1. **Final centroids output format: Final centroids output format:**

    [[x1, y1], [x2, y2]]

[[x1, y1], [x2, y2], [x3, y3]]

.

.

.

[[x1,y1],[x2, y2],...[x9,y9]]

2. **Loss function output format:**

   L1

   L2

   .

   .

   .

   L9

Part 2 Report Submission

Draft a report to go with your Part 2 Results Submission. The report must contain:

- Your full name and student ID number on the first page in the upper left corner
- A detailed description of your observations and analysis of each strategy
  - You may include the plots for the objective function in your report to support your analysis

The report must also follow the required format:

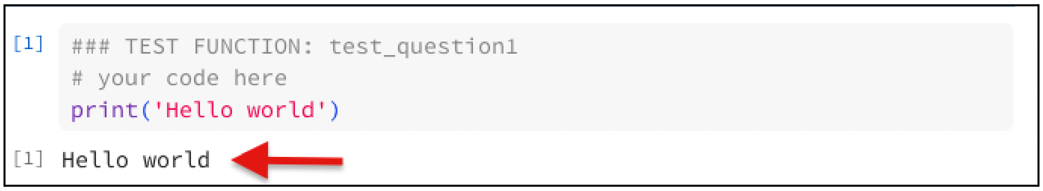
- A maximum font size of 12pt
- A maximum length of two (2) pages (8x11 or A4 paper).
- Saved as a PDF (.pdf) file type

Submission Directions for Project Deliverables

Results Submission

This assignment will be auto-graded. You must complete and submit your work through Ed Lesson's code challenges to receive credit for the course:

1. In order for your answers to be correctly registered in the system, you must place the code for your answers in the cell indicated for each question.
    a. You should submit the assignment with the output of the code in the cell's display area. The display area should contain only your answer to the question with no extraneous information, or else the answer may not be picked up correctly.
    b. Each cell that is going to be graded has a set of comment lines (ex: ### TEST FUNCTION: test_question1) at the beginning of the cell. **This line is extremely important and must not be modified or removed.**
2. After completing the notebook, run each code cell individually or click "**Run All**" at the top to print the outputs.



```
[1] ### TEST FUNCTION: test_question1
    # your code here
    print('Hello world')

[1] Hello world
```

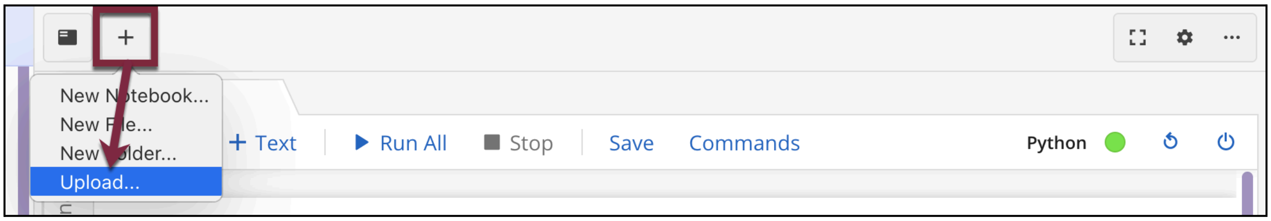
3. When you are ready to submit your completed work, click on "**Test**" at the bottom right of the screen.
4. You will know you have successfully completed the assignment when feedback appears for each test case with a score.
5. If needed: to resubmit the assignment in Ed Lesson
    a. Edit your work in the notebook
    b. Run the code cells again
    c. Click "**Test**" at the bottom of the screen

Your submission will be reviewed by the course team and then, after the due date has passed, your score will be populated from Ed Lesson into your Canvas grade.

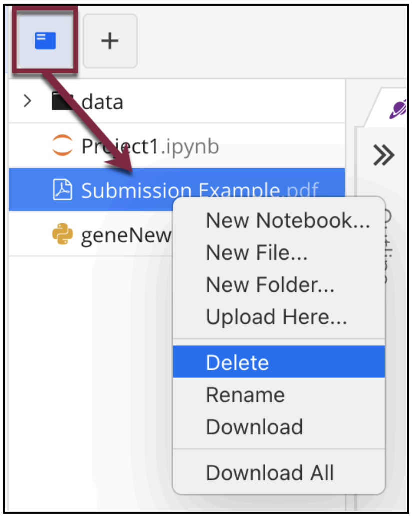
Report Submission

Each report for Part 1 and Part 2 will be manually graded by the course team. Each report must be submitted in their designated code challenge workspaces where you also submitted your results submission.

1. Go to "**Part #: K-means-Strategy Project**"
2. Click the **Plus (+)** icon in the upper left corner of the notebook workspace (second icon from the left)
3. Select "**Upload**"
4. Locate and select your report submission from your device (PDF file only)



5. Your file will appear in a left-pane menu that appears next to the notebook workspace
6. Click "**Submit**" in the upper right corner to submit your completed project.
7. If needed: to resubmit the report in Ed Lesson
    a. Click the "**Toggle Files**" icon in the upper left corner of the notebook (first icon from the left)
    b. Locate and right-click on your previous report submission file
    c. Click "**Delete**" to remove it from your attempt and then repeat the upload directions from Step 2



Your latest report submission will be reviewed by the course team and then, after the due date has passed, your score will be populated from Ed Lesson into your Canvas grade.

Evaluation

This project has two parts an autograded component and manually graded component. The assignments will be evaluated in Ed Lessons and grades will be automatically applied to the gradebook.