Ira A. Fulton Schools of Engineering
Arizona State University

CSE 575: Statistical Machine Learning

Project 2: K-means-Strategy

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

Software:

- Python environment

Language(s):

- 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 part1 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 part2 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.

Directions

Download Mat File: "**CSE 575_Project 2_AllSamples**" (attached in the Course Project Overview page in the Welcome and Start Here section of the course)

Lab: Project 2: K-mean-Strategy, Part 1

K-means_algortihm_Strategy Part 1 (K-mean) Overview:

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 $\mu_1, , \mu_2, \ldots \mu_k$, the objective function is defined as $\Sigma_{i=1}^{k} \Sigma_{x \in D_i} \left|\left| x - \mu_i \right|\right|^2$

You are highly suggested to use the built-in Jupyter Notebook to implement your algorithm. Another python environment is okay but you must be responsible for any numerical error caused by a different programming environment.

Lab: Project 2: K-mean-Strategy, Part 2

K-means_algortihm_Strategy2 (K-mean ++) Overview:

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 $\mu_1, , \mu_2, \ldots \mu_k$, the objective function is defined as $\Sigma_{i=1}^{k} \Sigma_{x \in D_i} \left|\left| x - \mu_i \right|\right|^2$

You are highly suggested to use the built-in Jupyter Notebook to implement your algorithm. Another python environment is okay but you must be responsible for any numerical error caused by a different programming environment.

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. Write code to implement the k-means algorithm with Strategy 2.
5. Use your code to do clustering on the given data; compute the objective function as a function of k (k = 2, 3, …, 10).
6. Repeat the above step with another initialization.
7. Submit a short report summarizing the results of both parts, including the plots for the objective function values under different settings described above.

Submission Directions for Project Deliverables

What to Submit:

1. **Result Submissions**: Code file with comments explaining what you do for each part as directed in their respective result submission quiz.
2. **Report Submission**: A report on both parts that summarizes the results and includes the plots for each of the objective function values.

Result Submission

Quiz: K-means-Strategy, Part 1 Result Submission

Once you launch the lab, you will be able to find two sets of initial k and points, which are associated with your ID. Please test your algorithm with this initialization.

The 1) final coordinate of the centroids and 2) the loss computed by the objective function should be submitted to the quiz titled **"Quiz: K-means-Strategy, Part 1 Result Submission"** located under “Week 5 Graded Coursework”.

Note:

1. You should implement your own K-means algorithm.
2. You will not get any points for the project by simply programming here. Please remember you need to submit the results in the result submission quiz.

Quiz: K-means-Strategy, Part 2 Result Submission

Once you launch the lab, you will be able to find two sets of initial k and points, which are associated with your ID. Please test your algorithm with this initialization.

The 1) final coordinate of the centroids and 2) the loss computed by the objective function should be submitted to the quiz titled "**Quiz: K-means-Strategy, Part 2 Result Submission**" located under "Week 5 Graded Coursework".

Note:

1. You should implement your own K-means algorithm.
2. You will not get any points for the project by simply programming in the Lab. Please remember you need to submit the results in the result submission quiz.

Report Submission

Graded Assignment: K-means-Strategy, Report Submission

Please submit your code files and report for both parts regarding Project 2 to the item titled "**Graded Assignment: K-means-Strategy, Report Submission**" located under "Week 5 Graded Coursework".

- Acceptable file types for report: .pdf or .doc/docx.
- Length of the report: no more than 2 A4 pages.
- Content of the report: (The following must be included)
    - Please include the *k* and initial points assigned to you as well as the final clustering centroid and loss in the report.
    - Your observation and analysis about the two strategies.
- Code files: submit as a ZIP together

Evaluation

Part 1 and Part 2 Result Submissions and Reports

Each part's result submission and the report will be evaluated together for the following.

- Final clustering centroid and the loss
- Your analysis