Assignment 1 – Clustering

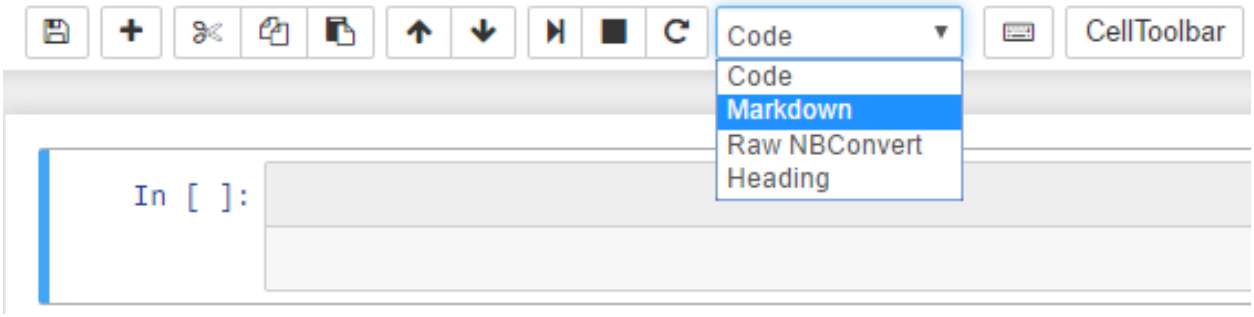
Objective

In this assignment, you will strengthen your programming skills in Python and numpy. You will implement and analyze the k-means algorithm. Additionally, you will discover applications to real engineering problems, and apply your algorithm to actual data to discover insights. Finally, you will practice your presentation and speaking skills, giving a 2-3 minute presentation to the class on your results.

Logistics

For this project, you should use an IPython notebook to write your code and analysis. Please ask the instructor if you have questions on installation. You should be able to launch an IPython notebook by typing "jupyter notebook" in a command prompt and pressing "Enter" on both Linux and Mac. Windows users should be able to launch directly from the Start menu.

To answer the text questions, change the cell from "Code" to "Markdown" as shown below:



If you run a Markdown cell (Shift-Enter), it will display regular text. For more details on formatting your text with Markdown, read through the following link:

http://jupyter-notebook.readthedocs.io/en/stable/examples/Notebook/Working With Markdown Cells.html

For all code, include descriptive comments.

Please submit BOTH your IPython notebook file (.ipynb) and a PDF of your IPython notebook to Moodle by **5 pm on 9/26**.

To save as PDF, go to File --> Download as --> PDF via LaTeX (.pdf)

Presentations will occur in class on 10/1. Slides should be uploaded to a shared Google Drive folder before class, see Moodle for the link.

Tasks

(1) Implement the k-means algorithm.

Implement the k-means algorithm as a function named **k_means**. Review Chapter 4 of *Introduction to Applied Linear Algebra* before you begin.

The inputs to your function should be *N n*-vectors, and *k*, the number of groups you would like to create. The output of your function should be the group representative vector for each of the groups, the cluster assignment vector *c*, and the optimization objective J^{clust}. Return all these quantities for every iteration. You may split this function up into several functions. Please comment your code.

You will have to choose how to initialize your algorithm. Use the information on page 76 of the text as a starting point for figuring out how to perform the initialization. Research outside references to determine at least 1 other method of initialization. Cite your source(s) and describe these alternative method(s) as well as the method you finally choose and why in a Markdown cell.

(2) Visualization for the k-means algorithm.

Create a visualization function for your k-means algorithm. The input of your visualization function should be *N n*-vectors, two integers corresponding to indices in the *n*-vectors, the number of groups *k*, and the cluster assignment vector *c*. The function should create a scatterplot of the *n*-vectors corresponding to the 2 input indices, and color each cluster with a different color. You may use the Python package matplotlib to create your visualizations.

(3) Test your algorithm and visualization.

Generate simulated datasets of 2-vectors with sklearn.datasets.make_blobs: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_blobs.html

Test your algorithm and visualization on 3 different simulated datasets with different values of *k*. Show the final visualizations and values for J^{clust} in your IPython notebook. Comment on your algorithm's results in a Markdown cell. Were you able to find the correct clustering and centroids? What happens if you assume the wrong *k*? What happens if you change the initialization of the group representative vectors?

(4) Create an animation of your visualizations.

Generate a simulated dataset as in (3) and run your algorithm. Create scatterplots as described in (2) at every iteration of your algorithm. Create an animation of these scatterplots. You can choose the file format of your choice for the animation. One suggestion is to create a gif animation using the Python package imageio.

(5) Use your algorithm to analyze real data.

In this task, you need to pick an application for k-means clustering. Use the examples in the text as a starting point. Pick an application topic that you are personally interested in that has publicly available data associated with it. Find literature references related to your topic. Create a Markdown cell describing your topic, why the topic is significant, why clustering is useful to solving a problem related to your topic, and cite literature references.

Publicly available datasets for clustering may be found in these following links (you are welcome and encouraged to search beyond these links):

http://archive.ics.uci.edu/ml/index.php
https://www.kaggle.com/datasets
https://www.ncbi.nlm.nih.gov/gds/
http://snap.stanford.edu/data/index.html
http://qwone.com/~jason/20Newsgroups/

Apply your algorithm to a publicly available dataset associated with your topic. You will have to figure out how to import your data into your notebook and format it appropriately. Internet search engines and Stack Overflow (https://stackoverflow.com/) will be your friend here.

Create static visualizations of the final results of your algorithm, as well as an animation as in (4). You will use these visualizations to create presentation slides in (7).

(6) Analyze the results of the clustering algorithm.

Answer the following questions:

(a) Describe any different initial conditions you tried to analyze your data. What worked and what didn't?

(b) How many iterations did it take until the algorithm converged?

(c) Can you extract any meaning from the clustering results?

(d) What value(s) of k are reasonable for your application and why?

(e) Explain any intuition behind the final clustering result. For example, if you had chosen the handwritten digits dataset shown in the text, you would analyze whether the clustering algorithm separated each digit into a different cluster or not. To figure this out, look at examples from your dataset, and how they were categorized.

(7) **Presentation.**

You will make a short 2-3 minute presentation on your work in this assignment. In your presentation, you should:

(a) Introduce your topic

(b) Explain why you choose your topic and why it's important

(c) Show an animation of your clustering algorithm.

(d) Show the final clustering visualization.

(e) What is the utility of clustering on your data? Did you find any meaning in the clustering results?

Please upload your slides to the shared Google Drive folder before class on 10/1. You may use Google Slides or Microsoft PowerPoint.