CS339: Artificial Intelligence
Spring 2021
Project 2: K-Means Algorithm

Overview:

The learning goal for this project is to implement the k-means algorithm and to apply it to a realistic situation. The project will involve multiple steps:

1. Basic implementation of the k-means clustering algorithm.
2. Apply k-means to an image for vector quantization.
3. Research an interesting question.
4. Write a research paper.

This is a larger project, with a larger scope, that will take more time to complete. Thus we will break the project down into the following deadlines:

Deadline1: basic k-means code

For this first deadline, you will complete a jupyter notebook implementing the k-means algorithm from scratch. You may use the algorithm in the book. I have given you a data set called `data.txt` which is a simple two-dimensional data set. You will run your hand crafted k-means algorithm on this dataset.

- handcraft the k-means algorithm. It should return the centers, the number of iterations to convergence, the final error (sum of squared distances from points to centers), and labels (the assignment of points to centers).
- run this algorithm on the `data.txt` data.
- create a graph of the original data and the centers (use a different color for the centers and use the optimal value for k).
- produce a graph of error vs k to find the elbow for the appropriate value for k
- use good commenting and notebook markup cells, but this does not have to be a polished research paper.

submission: a jupyter notebook containing the steps above

Deadline2: research paper draft

There are multiple steps in this part of the project.

- Find an image and load it into a jupyter notebook.
- You might want to use the sklearn version of k-means. It will almost certainly run faster than your native code. Go online and find some examples of how to use sklearn toolkit (and use the in-class example code).

- Perform k-means on your image. Try different values for k. Maybe produce another k vs error graph.
- Recreate your image using vector quantization and the centers produced from the previous step. Try creating different images with different values of k.
- Formulate a RESEARCH QUESTION. There are example questions below or you can choose your own.
- Complete the experimentation for your research question.
- Write a very good draft of your research paper. (more directions below).

submission: a jupyter notebook showing your research, and a pdf of your paper.

Deadline3: final research paper

This final step has you polish your previous draft and complete a final research paper for submission. We will do a peer-reviewed reading of your draft paper for you to receive feedback.

submission: the final pdf of your paper.

Research Questions

An important step in this project is to formulate a research question and then to conduct an experiment on your question. Here are some possible question ideas.

1. Create an image fidelity metric. You will have an original image. You will also have reconstructed images using vector quantization and various values of k. The reconstructed images will be of varying levels of quality. It would be nice to create a fidelity metric on a scale of 1 to 0 where 1 is an exact reconstruction of the original image and 0 is an extremely poor reconstruction. You might think about how to craft such a metric. Then apply the metric to the various reconstructed images. Compare your metric to "human subjective evaluation" of the image quality to see how well your metric captures human sentiment.

2. A graph of Error vs K for values of k-means allows you to see how well the error is reduced for each value of k. The elbow in this graph points to a good value of k. You might explore this graph on different kinds of images (choose several images) and see if there is a good elbow formation for some types of images but not others. You might compare the values of k which produce similar qualities of reproduction in each of the different kinds of images.

3. Cross Comparison. Grab two images. Perform k-means vector quantization on the first image. Then use the centers to reconstruct the second image. See how well these centers do at reconstructing images they were not trained on.

4. Apply k-means vector quantization to a black and white version of an image and also to a color version. Determine good values for k for each. Compare reconstructions of each image for the same value of k.

5. Formulate your own research question that involves k-means clustering and vector quantization of images. Visit an office hour or send an email to get approval of your research topic before you embark on it.

Research Paper

You will submit a full pdf (from LaTeX) paper for this project along with your code. The research paper should follow the conventions we discuss in class of good research papers. Your research paper should concentrate on your research question, and not on the whole k-means project or your learning experience. Keep the paper focused narrowly on your research question. Your question should form a thesis, and the structure of your paper should follow the experimentation that you performed to investigate that thesis.

Attributes for evaluation:

- Does your paper have a clear research-question driven thesis. Is that thesis immediately visible to the reader?
- Have you clearly thought about your intended audience? Does the paper address the needs of this audience?
- Does the paper use appropriate structures for a research paper of this size (I expect most research papers to be in the range of 3-7 pages, depending on your project and the number of graphics)?
- Does the paper present a clear data-story: question, experimental description, technique explanation, results, and analysis?
- Are there appropriate graphics to support the data story?
- Does the paper show good control over style, tone, and mechanics?