Project 1

Due Date: Thur, April 29, 2021 at 11:59 PM.

Instructions:

1. For details, refer to Piazza instructions.
2. The 3-page report on the project must be in `.pdf` format.
3. **DO NOT** submit the dataset along with your source code.
4. Source codes will be checked for plagiarism using automated tools.
5. Clarity and efficiency of source codes will be taken into consideration during grading.

Overview

We are asking you to perform some data-analysis tasks on a sample review dataset. Overall, the goal of this exercise is not to get the best possible results. Thus, do not focus on investing time designing complicated methods and do not try to bring into your solutions heavy machinery.

We would like to see solutions that build on simple intuitions about the geometry of your data space. We also want to see how you build on simple building blocks, how you combine them, how you troubleshoot a solution that does not work and how you try to understand why it does not work. Thus in addition to your code, we want to see a writeup with plots, discussion of your methods and your results. We want to check your intuitive understanding of the problem, the transformation of this intuitive understanding into algorithms and your way of evaluating those algorithms, even if your conclusion is that they did not work well – as long as you know why they did work or not.

Clearly, in order to perform Tasks 1 and 2 below (which correspond to clustering and nearest-neighbor classification respectively) you will first need to bring the data in the right vector form in order to analyze it accordingly. Therefore, in your analysis we want to see why you chose a particular vector representation of your data, or any other data manipulation you did in order to bring the data in the form that was appropriate for your analysis.

The tasks

The description of the dataset and the two tasks are provided below:

Dataset: Yelp academic dataset (`https://www.yelp.com/dataset`) In this dataset you can find reviews as well as other information associated with individual restaurants. Focus on one culinary district, preferably one with a lot of reviews. We recommend that you work with the Las Vegas restaurants for which you have the most reviews.

Some documentation for the dataset (and how to get started) can be found here: `https://www.yelp.com/dataset/documentation/main`

Task 1: (50 points) Using only the review text again cluster the Las Vegas restaurants and visualize your results. Compare the results you get with the clustering results that you would have got had you only relied on the longitude and latitude features or the "categories" of a restaurant.

Task 2: (50 points) Now use nearest-neighbor classification to predict the rating of each restaurant from the reviews text of individual reviews. Evaluate your methodology and describe why it succeeds or fails.

Clarifications: Whenever we ask you to visualize or compare the results of different methods, we urge you to think of clever, clear, easy to understand visualizations that will help you convey the strengths and the weaknesses of your approaches as well as the key differences and similarities between approaches you are trying to compare.