Linear Algebra and Optimization for Machine Learning – Project 1

October 19, 2022

Attached, you can find the text file `EastWestAirlinesCluster.csv`, which includes data from 3 999 passengers who belong to an airline's frequent flier program. For each passenger, the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The column description is as follows.

Column	**Description**
ID	Unique ID
Balance	Number of miles eligible for award travel
Qual_mile	Number of miles counted as qualifying for Topflight status
cc1_miles	Number of miles earned with freq. flyer credit card in the past 12 months: 1 = under 5000, 2 = 5000 - 10000, 3 = 10001 - 25000, 4 = 25,001 - 50,000, 5 = over 50,000
cc2_miles	Number of miles earned with Rewards credit card in the past 12 months as above
cc3_miles	Number of miles earned with Small Business credit card in the past 12 months as above
Bonus_miles	Number of miles earned from non-flight bonus transactions in the past 12 months
Bonus_trans	Number of non-flight bonus transactions in the past 12 months
Flight_miles_12mo	Number of flight miles in the past 12 months
Flight_trans_12	Number of flight transactions in the past 12 months
Days_since_enrolled	Number of days since enrolled in flier program
Award	Whether that person had award flight (free flight) or not

Your task is to cluster samples of the data set into K clusters that, in your opinions, provide a meaningful and understandable division of the samples, in the sense that there is a reasonable interpretation for each cluster (for example: cluster 1: 'top 10% mile collectors', cluster 2: 'occasional flyers', etc.). You should decide for a value of K yourself and explain your choice.

Specifically, we ask you to apply the following clustering methods:

1. As a first part of the assignment, perform K-means clustering of this dataset using
   (a) standard squared Euclidean distance
   (b) kernelized distance using a kernel of your choice.
2. Now, perform unnormalized and normalized spectral clustering of this dataset:
   (a) Using a kernel of your choice, construct a suitable graph Laplacian matrix. Then, use a ready-made Python package to construct a matrix whose columns are the K eigenvectors corresponding to the K smallest eigenvalues of the Laplacian. Use the K-means algorithm you constructed in Part 1 to group the points according to the rows of this matrix.
   (b) Implement your own efficient solver for computing the (approximations of) the K eigenvectors corresponding to the K smallest eigenvalues of the Laplacian. Use the K-means algorithm you constructed in Part 1 to group the points according to the rows of the obtained matrix.
3. **For groups with 4 people:** Investigate whether all 12 features given in the columns of the csv file are necessary for a reasonable clustering of the data samples, or if the number can be reduced by omitting or combining features. You may think about using dimension reduction techniques introduced in the lectures (e.g., based on an SVD) or come up with your own ideas.

Explain and motivate all the steps/algorithmic choices made in your solution. Discuss and compare the results of the different methods regarding the accuracy of the obtained clustering and the efficiency of your methods (runtimes).

To select the proper number of clusters, you can use your own common sense and/or some more quantitative measure such as, for example, the error sum of squares of the final clustering obtained with a given method for a given K. This measure stands for the sum of squared distances between the individual observations and the centers of the respective clusters to which they belong. In other words:

$$\mathrm{ESS}(C_1, \ldots, C_K) = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - c_k\|_2^2,$$

where c_k is the center of the k-th cluster. Generally speaking, the ESS should decrease when increasing the number of clusters. However, the marginal effect of adding more clusters later on should be dropping (the more clusters the harder it is to get a meaningful understanding as to into what groups do they divide the data).

The ideal situation (grade 10) is that you implement all the requested methods (K-means algorithm and the spectral clustering using your own implementation of an eigensolver) on the entire dataset. If you are not successful, you may either use an (as large as possible) subset of the data set and/or implementations from Python packages instead. However, this will lead to some point deductions. As a point of reference, an assignment using only a subset of the entire dataset and which uses a ready package for all the algorithms will receive a grade 5.5 if all the results are discussed clearly.

Leave clear comments in your Python code – unclear assignments and their codes will result in a reduction of the grade. Also, explain if something does not work or works too slowly; if you can explain well the problem and how this could be addressed, we might subtract less points, if something does not work as expected.

The page limit for the entire assignment is 6 pages of text excluding images (font size 11). A submitted assignment consists of a PDF file with the assignment and a zipped folder including .py files + a short readme.txt file explaining how to use the code.

The deadline is Nov 21st, 2022, 23:59 CET. Please submit your projects through Brightspace.

As mentioned in the lectures, we expect you to work in groups of 3–4; for groups of 4, please be sure to also work on exercise 3.