Machine Learning and Computational Physics
Fall 2020

Assignment 6

Due: Nov 17th 2020, 11:59:59 PM PDT

Spectral clustering algorithms

In this assignment, the objective is to find the number of clusters in a (pre-generated) dataset using the spectral structure of the graph-Laplacian. The dataset consists of N nodes sampled from a distribution in $\mathbb{R}^3$.

Perform the following steps:

1. Download the `data.npy` file from blackboard and load the data to your Google colab. (Check this stackoverflow question and links within it to learn how to load data to Google Colab).

2. Write a function that takes the dataset and the parameter value `sigma` as input. The function should:

   (a) Find the weight matrix $\boldsymbol{W}$,

   $$W_{ij} = \exp\left(-\frac{\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|^2}{\sigma^2}\right).$$

   The function `scipy.spatial.distance_matrix` might be useful here. Also, use $\sigma = 0.1$ in your experiments.

   (b) Construct the graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$.

   (c) Evaluate the eigenvalues and eigenvectors of $\boldsymbol{L}$. The function `np.linalg.eig` can help you with this.

   (d) The function should return the eigenvalues and eigenvectors.

3. Sort the eigenvalues in ascending order and generate the (semilogy) plot the first few to determine the spectral gap. This should give you an estimate of the number of clusters in the dataset. Note that you might need to consider the absolute real part of the eigenvalues.

4. Apply a K-means clustering algorithm on coordinates of the first k eigenvectors (where k=no. of clusters as determined in the previous step) to determine the labels for each datapoints. You should use `sklearn.cluster.Kmeans` for this.

5. Use these labels to annotate each data point with color corresponding to that label and create three scatter plots (showing X-Y, X-Z, and Y-Z plane view). If you want (not mandatory) you may also create an additional 3d plot for better visualization (see https://matplotlib.org/gallery/mplot3d/rotate_axes3d.html and https://matplotlib.org/3.1.1/gallery/mplot3d/scatter3d.html for reference).

Instructions:

- At the very beginning of your notebook insert a text cell and write your name and **USC email address**.
- You need to submit your work as a single notebook saved as `A1_FirstName_LastName.ipnyb` (for example `A1_Tommy_Trojan.ipnyb`). You can create this notebook locally (on your computer using Jupyter notebook) or on cloud using Google Colab (which we recommend). If you are using Google Colab, then please make sure that you are signed in to your USC Google account before starting. This will make sharing your saved work little easier.
- Make sure that your entire notebook runs successfully on Google Colab before submitting it. It is your responsibility to ensure this.
- Once you finish the assignment save it and share it with `dhruvvpa@usc.edu`. (If you are using Google Colab, then the notebook will automatically be saved to your Google Drive. Once you locate it in your Google Drive, right click on it and share it with `dhruvvpa@usc.edu`). While sharing make sure that you enable "editor" option, so that we can run your notebook on our end while grading it.