Boston University
Department of Electrical and Computer Engineering

ENG EC 500 B1 (Ishwar) Introduction to Learning from Data

Matlab Exercise 4

Issued: Wed 18 Nov 2015 **Due:** 4pm Tue 1 Dec 2015 PHO440 box + Blackboard
Required reading: Your notes from lectures and additional notes on website on spectral clustering.

- This homework assignment requires some programming background in MATLAB. Please refer to the following link for an introduction (or review) of MATLAB:
`http://www.math.ucsd.edu/~bdriver/21d-s99/matlab-primer.html`
- You will be making two submissions: (1) A paper submission in the box outside PHO440. (2) An electronic submission of all your matlab code to blackboard (in a single zipped file appropriately named as described below).
- **Paper submission:** This must include all plots, figures, tables, numerical values, derivations, explanations (analysis of results and comments), and also printouts of all the matlab .m files that you either created anew or modified. Submit color printouts of figures and plots whenever appropriate. Color printers are available in PHO307 and PHO305. Be sure to annotate figures, plots, and tables appropriately: give them suitable *titles* to describe the content, label the *axes*, indicate *units* for each axis, and use a *legend* to indicate multiple curves in the plots. Please also explain each figure properly in your solution.
- **Blackboard submission:** All the matlab .m files (and only .m files) that you either create anew or modify must be appropriately named and placed into a **single** directory which should be zipped and uploaded into the course website. Your directory must be named as follows: `<yourBUemailID>_matlab4`. For example, if your BU email address is `mary567@bu.edu` you would submit a single directory named: `mary567_matlab4.zip` which contains all the matlab code (and only the code).
- **File naming convention:** Instructions for file names to use are provided for each problem. As a general rule, each file name must begin with your BU email ID, e.g., `mary567_<filename>.m`. The file name will typically contain the problem number and subpart, e.g., for problem 4.1b, the file name would be `mary567_matlab4_1b.m`. Note that the dot . in 4.1 is replaced with an underscore (this is important).

Problem 4.1 *k*-means vs. Spectral Clustering
In this problem we will use a circle-shaped dataset and a spiral-shaped dataset. Figure 1 shows examples for circle and spiral shaped datasets with 2 clusters. These synthetic examples can be generated using the provided functions `sample_circle.m` and `sample_spiral.m`.

(a) Use `sample_circle.m` to sample a circle-shaped dataset with $k = 3$ clusters and 500 points for each cluster (denoted as $\mathcal{D}_1$). Similarly, use `sample_spiral.m` to sample a spiral-shaped dataset with $k = 3$ clusters and 500 points for each cluster (denoted as $\mathcal{D}_2$). Use MATLAB's built-in function `kmeans` to cluster $\mathcal{D}_1$ and $\mathcal{D}_2$. **Important:** in any part of this assignment which requires running `kmeans`, before running it, set: `rng(2)`. For both datasets, set `'Replicates'` to 20 and `'Distance'` to `'sqeuclidean'`. We will explore k-means clustering with $k = 2, 3, 4$ to understand how results change when the specified value of k differs from the true value of k. For each choice of k:

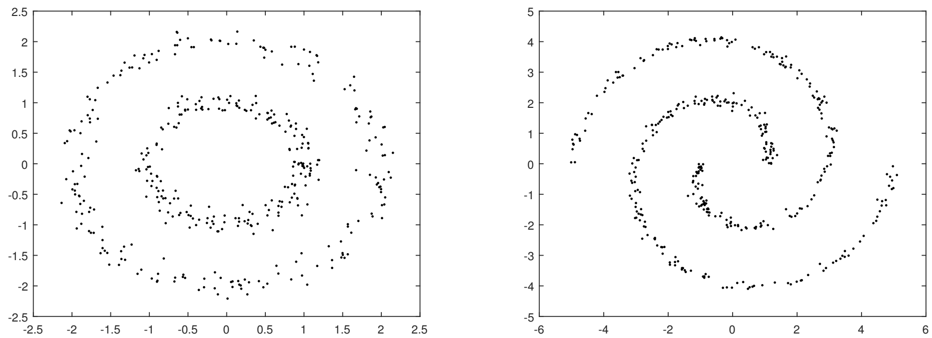

Figure 1: *Example circle-shaped (Left) and spiral-shaped (Right) dataset, each with $k = 2$ clusters and 200 points per cluster.*

(i) **Plot** all the data points in $\mathcal{D}_1$ (resp. $\mathcal{D}_2$), and indicate the cluster assignment by coloring the data points in different colors: For $k = 2$, use red and blue. For $k = 3$, use red, blue and green. For $k = 4$, use red, blue, green, and black. On the same figure, plot the cluster centers. (You need to create 6 plots in total. You can use `subplot` to save space.)

(ii) **Report** the overall within-cluster sums of point-to-centroid (Euclidean) ℓ_2 squared distances.

(b) We will next **implement** three variants of spectral clustering. Given the $n \times n$ similarity matrix $\mathbf{S}$ and the adjacency matrix $\mathbf{W}$, we use the following three different spectral clustering algorithms:

Un-normalized spectral clustering (SC-1):
• Compute the unnormalized graph Laplacian $\mathbf{L}$.
• Compute the first k eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of $\mathbf{L}$.
• Let $\mathbf{V} \in \mathbb{R}^{n\times k}$ be the matrix containing the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ as columns.
• Cluster the n rows of $\mathbf{V}$ with the k-means algorithm into k clusters (set: `rng(2)` before running `kmeans`).

Normalized spectral clustering 1 (SC-2):
• Compute the normalized graph Laplacian $\mathbf{L}_{rw} = \mathbf{D}^{-1}\mathbf{L}$.
• Compute the first k eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of $\mathbf{L}_{rw}$.
• Let $\mathbf{V} \in \mathbb{R}^{n\times k}$ be the matrix containing the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ as columns.
• Cluster the n rows of $\mathbf{V}$ with the k-means algorithm into k clusters (set: `rng(2)` before running `kmeans`).

Normalized spectral clustering 2 (SC-3):
• Compute the normalized graph Laplacian $\mathbf{L}_{sym} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.
• Compute the first k eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ of $\mathbf{L}_{sym}$.
• Let $\mathbf{V} \in \mathbb{R}^{n\times k}$ be the matrix containing the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ as columns. Normalize the rows of $\mathbf{V}$ so that their ℓ_2 norms are 1.
• Cluster the n rows of $\mathbf{V}$ with the k-means algorithm into k clusters (set: `rng(2)` before running `kmeans`).

Implement the above three spectral clustering algorithms. Apply them to $\mathcal{D}_1$ and $\mathcal{D}_2$ created in part (a). Use the Gaussian similarity $S(i, j) = \exp\left\{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|_2^2}{2\sigma^2}\right\}$ to construct $\mathbf{S}$ with $\sigma = 0.2$. We will use the *fully-connected* graph and set $\mathbf{W} = \mathbf{S}$.

(i) **Plot** the eigenvalues of $\mathbf{L}$, $\mathbf{L}_{rw}$, and $\mathbf{L}_{sym}$ for $\mathcal{D}_1$ and $\mathcal{D}_2$ in ascending order. (You need to create

6 plots in total. You can use `subplot` to save space.)

(ii) Set $k = 2, 3, 4$ in your spectral clustering algorithms. For each choice of k, **plot** all the data points in $\mathcal{D}_1$ (resp. $\mathcal{D}_2$) and indicate the cluster assignment from SC-3 using different **colors** (For $k = 2$, use red and blue; For $k = 3$, use red, blue and green; For $k = 4$, use red, blue, green, and black. You need to create 6 plots in total. You can use `subplot` to save space.)

(iii) For $k = 3$, use `plot3` to plot the rows of $\mathbf{V}$ matrices in SC-1, SC-2, and SC-3 (after row-wise normalization) for $\mathcal{D}_1$ and $\mathcal{D}_2$ in 3-D figures. Indicate the corresponding cluster assignment using red, blue, and green. (You need to create 6 plots in total. You can use `subplot` to save space.)

MATLAB functions: `eig, eigs, svd, kmeans`.

(c) Transform the Cartesian coordinates representation of each data point in $\mathcal{D}_1$ into polar coordinates using `cart2pol`. We denote this new dataset as $\mathcal{D}_3$. Now apply the k-means algorithm to $\mathcal{D}_3$ (set: `rng(2)` before running it). Also set `'Replicate'` to 20 and `'Distance'` to `'cityblock'`. For each choice of $k = 2, 3, 4$ in `kmeans`,

(i) **Plot** all the data points in $\mathcal{D}_3$, and indicate the cluster assignment by coloring the data points in different colors. (For $k = 2$, use red and blue; For $k = 3$, use red, blue and green; For $k = 4$, use red, blue, green, and black). On the same figure, plot the cluster centers. (You need to create 3 plots in total. You can use `subplot` to save space.)

(ii) **Report** the overall within-cluster sums of point-to-centroid (Euclidean) ℓ_2 squared distances.

Hint: when applying k-means to data whose attributes are of different units (say, radius and angle), a standard pre-processing step is to apply a linear transform to each attribute so that the minimum (resp. maximum) of each attribute is 0 (resp. 1). Apply this pre-processing step in part (c).

Problem 4.2 Spectral Clustering on Airbnb data

In this problem we will use a newly released real-world dataset obtained from

`http://insideairbnb.com/get-the-data.html`

It consists of $n = 2558$ houses listed on the Airbnb website in the Boston area in Oct. 2015. We have pre-processed the data into a MATLAB file "`BostonListing.mat`". For each listing, we will keep its **latitude**, **longitude**, and **neighborhood**. The "neighborhood" attribute indicates the region of each house listing such as "allston", "brighton", etc. We will use **latitude** and **longitude** as a 2-D feature vector for each listing and treat **neighborhood** as the ground-truth cluster label. Our goal is to cluster the listings into clusters that can reflect the neighborhood structure based on latitude and longitude. Figure 2 is a visualization of this dataset on Google Map.

(a) We will use the Gaussian similarity distance defined in problem 4.1 and construct a fully-connected graph. We set $\sigma = 0.01$ in this case and use the symmetrically normalized graph Laplacian (SC-3 defined in problem 4.1) for spectral clustering. For $k = 1, 2, \ldots, 25$, **calculate** the "purity" metric of the obtained cluster by treating the "neighborhood" label as the ground truth.[1] **Plot** the purity metric (y-axis) as a function of k (x-axis).

[1] The purity metric of a cluster assignment with k clusters is defined as follows. Let $n_{i,j}$ be the number of objects in cluster i that belong to class j, where $i = 1, \ldots, k$ and $j = 1, \ldots, m$. Here m is the number of ground-truth classes. Let $n_i = \max_{j=1,\ldots,m} n_{i,j}$. Purity is then defined as $\sum_{i=1}^{k} n_i / n$.

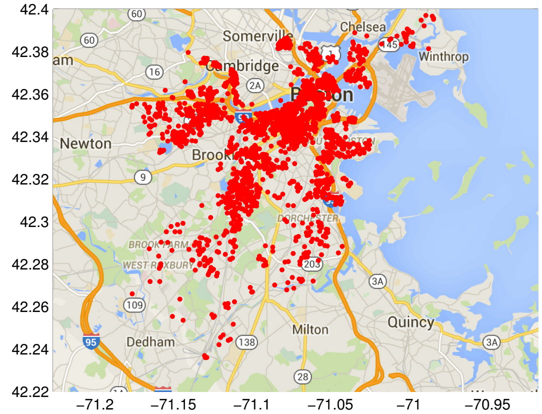

Figure 2: *House listings on Airbnb websites in the Boston area in Oct. 2015. Each point indicates a house listing. The dataset is obtained from Inside Airbnb.*

(b) Use the `plot_google_map.m` to **plot** all the data points on the map and indicate the cluster assignment with $k = 5$ using different colors.