ECE-GY 6143 / Spring 2020

Homework 6

Please scan and upload your assignments on or before April 30, 2020.

- You are encouraged to discuss ideas with each other; but
- you **must acknowledge** your collaborator, and
- you **must compose your own** writeup and/or code independently.
- We **strongly** encourage answers to theory questions in Latex, and answers to coding questions in Python (Jupyter notebooks).
- Please upload your solutions in the form of a single .pdf or .zip file on NYUClasses.
- Maximum score: 50 points.

---

1. **(10 points)** Assume that you have 4 samples each with dimension 3, described in the data matrix X,

$$X = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 5 \\ 1 & 2 & 3 \\ 0 & 2 & 5 \end{bmatrix}$$

   For the problems below, you may do the calculations in python (or R or Matlab). Explain your calculations in each step.

   a. Find the sample mean.

   b. Zero-center the samples, and find the eigenvalues and eigenvectors of the data covariance matrix Q.

   c. Find the PCA coefficients corresponding to each of the samples in X.

   d. Reconstruct the original samples from the top two principal components, and report the reconstruction error for each of the samples.

2. **(10 points)** In class, we analyzed the per-iteration complexity of k-means. Here, we will prove that the k-means algorithm will terminate in a finite number of iterations. Consider a data set $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^d$.

   a. Show that the k-means loss function can be re-written in the form:

$$F(\eta, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{k} \eta_{ij} \|x_i - \mu_j\|^2$$

   where $\eta = (\eta_{ij})$ is a suitable binary matrix (with 0/1 entries). Provide a precise interpretation of η.

   b. Show that each iteration of Lloyd's algorithm can only decrease the value of F.

   c. Conclude that the algorithm will terminate in no more than T iterations, where T is some finite number. Give an upper bound on T in terms of the number of points n.

3. **(10 points)** Using the Senate Votes dataset demo'ed in Lecture 11, perform k-means clustering with $k = 2$ and show that you can learn (most of) the Senators' parties *in a completely unsupervised manner*. Which Senators did your algorithm make a mistake on, and why?

4. **(20 points)** The *Places Rated Almanac*, written by Boyer and Savageau, rates the livability of several US cities according to nine factors: climate, housing, healthcare, crime, transportation, education, arts, recreation, and economic welfare. The ratings are available in tabular form, available as a supplemental text file. Except for housing and crime, higher ratings indicate better quality of life. Let us use PCA to interpret this data better.

   a. Load the data and construct a table with 9 columns containing the numerical ratings. (Ignore the last 5 columns – they consist auxiliary information such as longitude/latitude, state, etc.)

   b. Replace each value in the matrix by its base-10 logarithm. (This pre-processing is done for convenience since the numerical range of the ratings is large.) You should now have a data matrix X whose rows are 9-dimensional vectors representing the different cities.

   c. Perform PCA on the data. Remember to center the data points first by computing the mean data vector μ and subtracting it from every point. With the centered data matrix, do an SVD and compute the principal components.

   d. Write down the first two principal components v_1 and v_2. Provide a qualitative interpretation of the components. Which among the nine factors do they appear to correlate the most with?

   e. Project the data points onto the first two principal components. (That is, compute the highest 2 scores of each of the data points.) Plot the scores as a 2D scatter plot. Which cities correspond to outliers in this scatter plot?

   f. Repeat Steps 2-5, but with a slightly different data matrix – instead of computing the base-10 logarithm, use the z-scores. (The z-score is calculated by computing the mean μ and standard deviation σ for each feature, and normalizing each entry x by $\frac{x-\mu}{\sigma}$). How do your answers change?