Math 123 **Problem Set 4** **Due: Sunday 3/27 at 11:59 p.m. EST**

Instruction: Read the homework policy. For problems 2 and 3, include copies of your code with your final homework submission. You should submit a PDF copy of the homework and any associated codes on Gradescope. Your PDF must be a single file, not multiple images.

1. Consider N data points $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ lying in $\mathcal{R}^d$. We apply the nonlinear transform $\phi : \mathcal{R}^d \to \mathcal{R}^M$ and obtain data points in feature space $\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_N)$. Define the kernel matrix $\boldsymbol{K} \in \mathcal{R}^{N\times N}$ as $\boldsymbol{K}_{i,j} = \kappa_{i,j} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$.

(a) Prove that $\boldsymbol{K}$ is a symmetric and positive semidefinite matrix.

(b) The projected dataset may not be centered. Let's center each projected data point as $\hat{\phi}(\mathbf{x}_i) = \phi(\mathbf{x}_i) - \frac{1}{N}\sum_{k=1}^{N}\phi(\mathbf{x}_k)$. Define a new kernel matrix $\hat{\boldsymbol{K}} \in \mathcal{R}^{N\times N}$ as $\hat{\boldsymbol{K}}_{i,j} = \hat{\kappa}_{i,j} = \hat{\phi}(\mathbf{x}_i)^T\hat{\phi}(\mathbf{x}_j)$. Prove that $\hat{\boldsymbol{K}} = \boldsymbol{K} - \mathbf{1_N}\boldsymbol{K} - \boldsymbol{K}\mathbf{1_N} + \mathbf{1_N}\boldsymbol{K}\mathbf{1_N}$ where $\mathbf{1_N} = \frac{1}{N}\begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$.

2. In this problem, we implement the K-means algorithm and test it on two dimensional datasets.

(a) Implement the K-means algorithm. Use the "farthest heuristic" to initialize your clusters and have the algorithm converge when the centroids of the clusters are not moving. Your code should take as input a dataset and the number of clusters and should output the labels for each data point (i.e., which cluster each data point belongs to). [Develop your own version of the kmeans algorithm using Lecture 13 as a basis].

(b) Load the data `solvents.csv` available in the HW4 folder. We will work with the Boiling-point and Solubility features.

i. Run K-means with K = 3 clusters. Plot your final optimal clustering result, i.e. show a scatter plot of the points in different colors corresponding to their K-means label.

ii. Replicate the result 10 times, saving the final clustering result each time. What do you observe? Do the data points all belong to the exact same cluster every time you run your algorithm?

iii. Create a process to measure the error (by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster). Repeat the clustering with K = 2, 3, 4, ..., 10. Plot the error measure for K = 2, 3, 4, ..., 10, what do you observe? What is a good number of clusters to use and why?

[**Remark**: For 2(b)(ii), replicates here is the number of times the core K-means algorithm is run each time from a different initialization of the centroids. For 2(b)(iii), the error of the K-means objective is defined as follows

$$E_n = \sum_{\mathbf{x}_i\in C_1}(\mathbf{x}_i - \mu_1)^2 + \sum_{\mathbf{x}_i\in C_2}(\mathbf{x}_i - \mu_2)^2 + ... + \sum_{\mathbf{x}_i\in C_n}(\mathbf{x}_i - \mu_n)^2,$$

where C_1, C_2, ..., C_n are the clusters we have for the specific case, and μ_1, μ_2, ..., μ_n are their respective centroids].

3. In this problem, we will investigate clustering to the latent features obtained from a PCA. The dataset we consider is the MNIST dataset which is a database of handwritten digits.

(a) Load the data `mnist_067` available in the HW4 folder. The included data is a subset of the MNIST data consisting of the digits $0, 6$ and 7. Project the data points onto the first two principal components. Run K-means with K = 3 clusters and 100 replicates on the latent representations. Plot your optimal clustering result, i.e. show a scatter plot of the data points in different colors corresponding to their K-means label obtained from the replicate with the lowest error.

(b) Recall the results of problem 6 in HW3. Given those results, what could you conclude about your result in (a) ?

[**Remark**: For this problem, you can use an inbuilt PCA function and K-means function from a solver of your choice.]