Exercise 4

Prof. Kontorovich and Dr. Sabato

Submission guidelines, **please read and follow carefully**:

- You may submit the exercise in pairs.
- Submit using the submission system.
- The submission should be a zip file named "ex4.zip".
- The zip file should include **exactly the files enumerated below in the root — no subdirectories please**.
- The files in the zip file should be:
  1. A file called "answers.pdf" - The answers to the questions, including the graphs.
  2. A file called "kmeans.m" - The Matlab/Octave code for the requested function.
  3. A file called "spectral.m" - The Matlab/Octave code for the requested function.
- Note that you can put several auxiliary functions in this file after the definition of the main function. **Make sure that the single file works in Matlab/Octave before you submit it.**
- **Anywhere in the exercise where Matlab is mentioned, you can use Octave instead**.
- **Grading:** Question 1: 18 points. Question 2: 15 points. Question 3: 11 points. Question 4: 18 points. Question 5: 20 points. Question 6: 18 points.
- For questions use the course Forum, or email `inabd171@gmail.com`.

Question 1. For this question, use the data file `EX4q1_data.mat`, which contains data points $x_i \in \mathbb{R}^2$.

(a) Implement the k-means heuristic algorithm for Euclidean metric which we learned in class. The function should be implemented in the submitted file called "kmeans.m". The first line in the file (the signature of the function) should be:

```
function C = kmeans(X, k, t)
```

The input parameters are:

- k - the number of clusters
- t - the number of iterations to run
- X - a 2-D matrix of size $m \times d$. Row i in this matrix is a vector with d coordinates that describes example x_i from the training sample.

The function returns the variable C, which is a column vector of length m, where $C(i) \in \{1, \ldots, k\}$ is the identity of the cluster in which x_i has been assigned.

(b) Run your code on the data in the provided data file with k from 2 to 10, and provide a graph with the value of the k-means objective that your code found for each k. What is the trend that you observe? explain the trend.

(c) Plot the clustering you got for each k between 2 and 5, by drawing each point in the color that represents its cluster. Which clustering do you think is best and why? Is this the clustering with the smallest value of the k-means objective?

(d) Explain why the preferred k is not necessarily the one that gets the smallest value of the k-means objective.

(e) Run your k-means code on an **unlabeled** sample of size 1000 generated from all the digits in the MNIST data file `mnist_all.mat`, with $k = 10$. Use the labels to provide a table, showing: (1) for each cluster, which label is most common in it, and (2) for each cluster, what percentage of the points in the cluster have this label.

(f) Calculate the classification error on the sample, that would result if we classified all the points in each cluster using the cluster's most common label. Explain your calculation.

Question 2. For this question, use the data file `EX4q2_data.mat`, which contains data points $x_i \in \mathbb{R}^2$.

(a) Implement the unnormalized spectral clustering algorithm which we learned in class. The function should be implemented in the submitted file called "spectral.m". The first line in the file (the signature of the function) should be:

```
function C = spectral(W, k, t)
```

The input parameters are:

- k - the number of clusters
- t - the number of iterations to run in the internal k-means
- W - a 2-D matrix of size $m \times m$, the entries are W_{ij} the similarities between points i and j in the sample.

(b) Plot the points in the provided data file. Explain why standard k-means will not separate the points nicely. Assume $k = 2$.

(c) Run the spectral clustering algorithm on the given data file with, where the similarity measure is $W_{ij} := e^{-\|x_i - x_j\|^2}$. Plot the resulting clusters. Do the same for a run of k-means. Use $k = 2$ in both cases. Explain the differences in the results.

Question 3. Consider the following distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$: The distribution over examples $\mathcal{X}$ is uniform in $[0, 1]$, and the distribution of the label y for a given example x is $N(3x, 5x^2)$. Derive and prove the Bayes-optimal rule $h^* : \mathcal{X} \to \mathcal{Y}$ for regression with the squared loss for this distribution.

Note: we didn't do this in class for the case of continuous distributions, so you should derive it yourselves. Prove all your claims, including those that we have seen in some form in class.

Question 4. Recall the compressed sensing procedure, where we get the measurement vector $y = Wx$, where $y \in \mathbb{R}^k$, $x \in \mathbb{R}^d$, $W \in \mathbb{R}^{k \times d}$

(a) We showed that it is possible to recover x by finding $\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^d : y = Wx} \|x\|_1$. Write a linear program, in the format of linear programs studied in class, that finds $\hat{x}$. Prove its correctness.

(b) Prove that the minimization problem $\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^d : y = Wx} \|x\|_0$ is not convex.

Question 5. Suppose that you are given a collection of points $S = (X_1, \ldots, X_m)$ in $\mathbb{R}^d$ and told that there is some unknown basis B in which these points have a sparse representation.

(a) How would you go about discovering such a basis B? Write an optimization problem to solve for such a B. The problem need not be convex or efficiently solvable. (Hint: the "zero" pseudo-norm $|| \cdot ||_0$ might be useful.)
Note: There is more than one solution here. Explain why you think your specific choice is good, by giving an example of a possible real-life problem in which your solution makes the most sense.

(b) Suppose a basis B is known, so that the sample $S = (X_1, \ldots, X_m)$ is sparse with respect to this basis. Additionally we know that S resides in a low-dimensional subspace of $\mathbb{R}^d$ (or very nearly so).

We have a transmitter which is capable of linear operations, and a receiver which is a full computer. The transmitter has S, and needs to transmit it to the receiver using a small list of numbers. Write a short and clear description of the operations that the transmitter should perform, and the operations that the receiver should perform, so that we end up with a low-dimensional representation of S, and so that we transmit few numbers from the transmitter to the receiver. Prove that your solution gives the same distortion as an optimal solution in which the receiver knows the original S.

Question 6. (Solve this after we learn about Maximum Likelihood Estimators, in lecture 19) Consider the distribution density $f_\lambda(x)$ with parameter $\lambda > 0$ over the domain $[0, \infty)$, which is defined by $f_\lambda(x) := C_\lambda e^{-\lambda x}$, where C_λ is a constant depending only on λ.

(a) Determine the form of C_λ — that is, its exact dependence on λ. Prove your claim.

(b) Let $S = (X_1, \ldots, X_m)$ be an i.i.d. sample from the distribution with the density f_λ, where λ is unknown. Derive the Maximum Likelihood Estimator for the value of λ. Prove all your claims.