CSE 6740 Homework 1

Anqi Wu, Spring 2025

Deadline: Feb. 04 Tuesday, 9:30 am

- There are 2 sections in gradescope: Homework 1 and Homework 1 Programming. Submit your answers as a PDF file to Homework 1 (including report for programming) and also submit your code in a zip file to Homework 1 Programming.
- All Homeworks are due by the beginning of class. Homework is penalized by 20% for each day that it is late (this applies additively, meaning that no credit is gained after 5 late days).
- We strongly encourage the use of LaTeX for your submission. Unreadable handwriting is subject to zero credit.
- Explicitly mention your collaborators if any.
- Recommended reading: PRML Section 9.1, 12.1
- Python and Matlab are allowed.

1 Probability [17 pts]

1. A company produces items in batches. Let the random variable D represent the number of defective items in a randomly chosen batch, with

$$P(D = n) = \frac{1}{2^n}, \quad n = 1, 2, 3, \ldots$$

Given that a batch has n defective items, each defective item is independently repaired with probability:

$$p_n = \frac{1}{n+1}.$$

What is the probability that at least one defective item is repaired in a randomly chosen batch? **[4 pts]**

2. In a hospital, 30% of patients are carriers of a certain contagious disease D. A diagnostic test is performed on all patients. Note that if a patient is a carrier of D, the test detects the disease as positive 92% of the time. If a patient is not a carrier of D, the test falsely detects the disease as positive 8% of the time. Moreover, patients who test positive are treated with a medication, which causes mild side effects in 10% of cases. Given that a patient chosen at random experiences side effects, what is the probability that the patient is a carrier of disease?**[4 pts]**

3. Victor has a choice to take a bus or walk to attend CSE6740 lecture. If he walks, he gets late with a probability of $\frac{1}{3}$. However, if he takes a bus, he gets late only with a probability of $\frac{1}{5}$. If he gets on time, he always keeps the same mode of travel the day after, whereas he always changes when he gets late. Let p be the probability that Charlie walks on the first day.

   (a) What is the probability that Charlie walks on the n^{th} day? **[4.5 pts]**

   (b) What is the probability that Charlie switches modes of travel on the n^{th} day? **[4.5 pts]**

2 Maximum Likelihood [10 pts]

1. Suppose we have n i.i.d (independent and identically distributed) data samples from the Pareto distribution, which has been used in economics for a density function with a slowly decaying tail:

$$f(x|x_0, \theta) = \theta {x_0}^{\theta} x^{-\theta-1}, \ x \geq x_0, \ \theta > 0$$

   assume that $x_0 > 0$ is given. Find the MLE of θ. **[4 pts]**

2. The lifetime X of a machine component follows a gamma distribution with a known shape parameter $k > 0$ and an unknown rate parameter $\theta > 0$. The probability density function is:

$$f(x \mid k, \theta) = \frac{\theta^k}{\Gamma(k)} x^{k-1} e^{-\theta x}, \quad x > 0.$$

   Here, $\Gamma(k)$ denotes the gamma function. The company observes n independent lifetimes: $x_1, x_2, \ldots, x_n$.

   (a) Find the MLE of θ. **[4 pts]**

   (b) Suppose the true value of θ is θ_0. How does the MLE $\hat{\theta}$ compare to θ_0 as the sample size n increases? Justify your answer intuitively or mathematically. **[2 pts]**

3 PCA [23 pts]

For PCA, suppose we use q directions, specified by q orthogonal length-one vectors $\vec{w_1}, ...\vec{w_q}$. We want to prove that minimizing the mean squared error is equivalent to maximizing the sum of the variances of the scores along these directions. Please prove the following sub-parts in order to arrive at this conclusion.

1. Write $\boldsymbol{w}$ for the matrix forms by stacking the $\vec{w_i}$. Prove that $\boldsymbol{w}^T\boldsymbol{w} = \boldsymbol{I}_q$. **[5 pts]**

2. Find the matrix of q-dimensional approximations based on these scores in terms of $\boldsymbol{x}$ and $\boldsymbol{w}$. Hint: your answer should reduce to $(\vec{x_i} \cdot \vec{w_1})\vec{w_1}$ when $q = 1$. **[5 pts]**

3. Using the conclusion from question 3.1, show that the MSE(mean squared error) of using the vectors $\vec{w_1}, ...\vec{w_q}$ is the sum of two terms, one of which depends only on $\boldsymbol{x}$ and not $\boldsymbol{w}$, and the other depends only on the scores along those directions (and not otherwise on what those directions are). Please show the necessary steps required to arrive at the answer (and not just the final solution) **[8 pts]**

4. Explain in what sense minimizing projection residuals is equivalent to maximizing the sum of variances along the different directions (using the previous conclusions). **[5 pts]**

4 Clustering [20 pts]

Given N data points $x^n (n = 1, ..., N)$, K-means clustering algorithm groups them into K clusters. With respect to K-means clustering answer the following question:

1. Consider the given single dimensional data with 4 data points $x_1 = 1, x_2 = 4, x_3 = 8, x_4 = 9$. Let's consider k = 3 for this situation. What is the optimal clustering for this data? **[4 pts]**
2. For the above part (1), show that by changing the center initialization we get a suboptimal cluster assignment that cannot be further improved.**[4 pts]**
3. Prove that the K-means algorithm converges to a local optimum in finite steps. **[8 pts]**
4. Original K-means algorithm uses Euclidian distance as the metric to compute the distance between data points. What is the disadvantage of using this distance function and suggest a solution to overcome this? **[4 pts]**

5 Programming: Handwritten Digit Recognition [Report 10 pts + Code 20 pts]

In this programming assignment, you will apply clustering algorithms for handwritten digit recognition. Before starting this assignment, we strongly recommend reading PRML Section 9.1.1, pages 428–430.

To ease your implementation, we provide skeleton code that loads the images and labels and calls the functions you need to implement. We also provide a function that computes the Euclidean distance between two cluster centers, so you don't need to worry about implementing it.

Your task is to implement the clustering components using two algorithms: *K-means* and *K-medoids*. We covered K-means in class, so you can begin with the sample code we distributed. The file you need to edit is `HW1_Programming.ipynb`, which is provided with this homework. In the data folder, you will find two files: `images.idx3-ubyte` and `labels.idx3-ubyte`. These files contain images of five handwritten digits $[0, 1, 2, 3, 4]$ and their corresponding labels. Your task is to cluster the images and return the cluster assignments to the driver function.

We will assess the cluster representations by visualizing the cluster centers and evaluating accuracy. To compute accuracy, we find the most common label in each cluster and assume it is the correct label for that cluster. We then calculate the percentage of correctly assigned labels across all clusters.

Next, we add 100 "contaminated" data points to the dataset. These are images containing only Gaussian noise, with labels set to -1. We will test your clustering functions again on this modified dataset to observe how the algorithms handle these outliers. Note that these outliers will not be included in the accuracy calculation. The code for adding these outliers has already been provided.

You are required to complete the Python functions 'mykmeans' and 'mykmedoids' with your own implementation. **Using built-in clustering functions is not allowed.** Your code will be evaluated based on the correctness of your implementation rather than the accuracy values. However, you should strive to achieve the best accuracy possible.

K-medoids

In class, we learned that the basic K-means works in Euclidean space for computing distance between data points as well as for updating centroids by arithmetic mean. Sometimes, however, the dataset may work better with other distance measures. It is sometimes even impossible to compute arithmetic mean if a feature is categorical, e.g, gender or nationality of a person. With K-medoids, you choose a representative data point for each cluster instead of computing their average.

Given N data points $\mathrm{x}^n (n = 1, ..., N)$, K-medoids clustering algorithm groups them into K clusters by minimizing the distortion function $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r^{nk} D(\mathrm{x}^n, \mu^k)$, where $D(\mathrm{x}, \mathrm{y})$ is a distance measure between two vectors x and y in same size (in case of K-means, $D(x, y) = \|\mathrm{x} - \mathrm{y}\|^2$), μ^k is the center of k-th cluster; and $r^{nk} = 1$ if x^n belongs to the k-th cluster and $r^{nk} = 0$ otherwise. In this exercise, we will use the following iterative procedure.

- Initialize the cluster center μ^k, $k = 1, ..., K$.
- Iterate until convergence:
  - Update the cluster assignments for every data point x^n: $r^{nk} = 1$ if $k = \arg\min_j D(\mathrm{x}^n, \mu^j)$, and $r^{nk} = 0$ otherwise.
  - Update the center for each cluster k: choosing another representative if necessary.

There can be many options to implement the procedure; for example, you can try many distance measures in addition to the Euclidean distance, and can be creative for deciding a better representative of each cluster. We will not restrict these choices in this assignment. You are encouraged to try many distance measures and ways of choosing representatives.

Programming [20 pts]

Both the `mykmeans` and `mykmedoids` functions take input and output format as follows. You should not alter this definition; or your submission will print an error, which leads to zero credit.

Input

- `digits`: N data points, each of dimension D = 784. The shape being (N, D). Each row contains a flattened image with 784 pixels (28x28). Each pixel value is a grayscale integer between 0 and 255
- `K`: the number of desired clusters. Higher values of K may result in empty cluster error. Then, you need to reduce it.

Output

- `cluster_assignments`: cluster assignment of each data point. For $K = 5$, for example, each image should be assigned to either 0, 1, 2, 3, or 4. The output should be a row vector with N elements.
- `centers`: location of K centers (or representatives) in your result. With images, each centroid corresponds to the representative image of each cluster. The output should be of the shape (K, 768). In case of medoids it will be a subset of the original points.

Hand-in [10 Points]

Both your code and report will be evaluated. Submit `HW1_Programming.ipynb` files as a .zip to Homework 1 Programming. In your report, answer the following questions:

1. Within the K-medoids framework, you have several choices for detailed implementation. Explain how you designed and implemented details of your K-medoids algorithm, including (but not limited to) how you chose representatives of each cluster, what distance measures you tried and chose one, or when you stopped iteration. [2pts]

2. Report your accuracies for the non-contaminated dataset using both k-means and k-medoids for the following values of K: 5, 8 [2pts]

3. Attach an image of the cluster centers for both k-means and k-medoids for $K = 5$ [4pts]

4. Compare the accuracies of k-means and k-medoids before and after the contamination for $K = 5$. Which method is more resilient to the contamination and why? [2pts]

Note

- Your K-means clusters are likely to be more accurate than K-medoids. This is normal. Your code will be judged on the correctness of your implementation, rather than the accuracy values that you get after clustering.

- Your accuracy is likely to vary significantly with different values of K

- If we detect copy from any other student's code or from the web, you will not be eligible for any credit for the entire homework, not just for the programming part. Directly calling the built-in function `kmeans` or other clustering functions is not allowed.