CSE/ISYE 6740 Homework 1

Anqi Wu, Fall 2022

Deadline: Sep. 22 Thursday, 12:30 pm

- 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 [15 pts]

1. We select a positive integer I with $P\{I = n\} = \frac{1}{2^n}$. If $I = n$, we toss a coin with probability of heads $p = e^{-n}$. What is the probability that the result is heads? **[3.5 pts]**
2. In a network of computers, 15% of the computers are infected by a virus V. An anti-virus scan has the property that if a computer is infected with V, the scan will detect the infection to be positive 95% of the time. However, if the computer is not infected, the scan will still detect the infection to be positive 10% of the time. All the computers which are detected to be infected, are applied with a corrective software-patch, which causes corruption of computers' files 20% of the time. Given that a computer picked at random has corrupted files, what is the probability that it was actually infected with the virus V to begin with? **[3.5 pts]**
3. Charlie has a choice to take a bus or walk to attend CSE6740 lecture. If he walks, he gets late with a probability of $\frac{1}{2}$. However, if he takes a bus, he gets late only with a probability of $\frac{1}{6}$. Further, 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 pts]**
   (b) What is the probability that Charlie gets late on the n^{th} day? **[4 pts]**

2 Maximum Likelihood [15 pts]

Suppose we have n i.i.d (independent and identically distributed) data samples from the following probability distribution. This problem asks you to build a log-likelihood function, and find the maximum likelihood estimator of the parameter(s).

(a) Exponential distribution [5 pts]

The exponential distribution is defined as

$$P(x|\beta) = \frac{1}{\beta}e^{-\frac{x}{\beta}}, \text{ with } 0 \leq x < \infty$$

Please find the MLE of the parameter β

(b) Pareto distribution [5 pts]

The Pareto distribution has been used in economics for a density function with a slowly decaying tail:

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

assume that $x_0 > 0$ is given. Find the MLE of θ.

(c) Normal linear regression model [5 pts]

The regression equations can be written in matrix form as

$$y = X\beta + \varepsilon$$

where y is the $N \times 1$ vector of observations of the dependent variable, X is the $N \times K$ matrix of regressors, and ε is the $N \times 1$ error terms. With the i.i.d assumption, multivariate normal distribution of ε on X, and full rank X, we can construct that the likelihood function of the linear regression model is

$$L(\beta, \sigma^2; y, X) = \left(2\pi\sigma^2\right)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2\right)$$

Show that the MLE of the regression coefficients β and the variance of the error terms σ^2 are

$$\begin{aligned} \hat{\beta}_N &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \\ \hat{\sigma}_N^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - x_i\hat{\beta}_N\right) \end{aligned}$$

3 PCA [20 pts]

Suppose that we use q directions, given by q orthogonal length-one vectors $\vec{w_1}, ...\vec{w_q}$. Please prove that minimizing the mean squared error is equivalent to maximizing the sum of the variances of the scores along these directions.

1. Write $\boldsymbol{w}$ for the matrix forms by stacking the $\vec{w_i}$. Prove that $\boldsymbol{w}^T\boldsymbol{w} = \boldsymbol{I}_q$. [4 pts]
2. Find the matrix of p-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 $q = 1$. [4 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). [10 pts]
4. Explain in what sense minimizing projection residuals is equivalent to maximizing the sum of variances along the different directions. [2 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 $x_1 = 1, x_2 = 3, x_3 = 6, x_4 = 7$. 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: Image Compression [Report 10 pts + Code 20 pts]

In this programming assignment, you are going to apply clustering algorithms for image compression. Before starting this assignment, we strongly recommend reading PRML Section 9.1.1, page 428 – 430.

To ease your implementation, we provide a skeleton code containing image processing part. `homework1.m` is designed to read an RGB bitmap image file, then cluster pixels with the given number of clusters K. It shows converted image only using K colors, each of them with the representative color of centroid. To see what it looks like, you are encouraged to run `homework1('beach.bmp', 3)` or `homework1('football.bmp', 2)`, for example.

Your task is implementing the clustering parts with two algorithms: *K-means* and *K-medoids*. We learned and demonstrated K-means in class, so you may start from the sample code we distributed.

The file you need to edit is `mykmeans.m` and `mykmedoids.m`, provided with this homework. In the files, you can see it calls Matlab function `kmeans` initially. Comment this line out, and implement your own in the files. You would expect to see similar result with your implementation of K-means, instead of `kmeans` function in Matlab.

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 $k =_j D(\mathrm{x}^n, \mu^j)$, 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 Euclidean distance, and also you 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 as well as way of choosing representatives.

Formatting instruction

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

Input

- `pixels`: the input image representation. Each row contains one data point (pixel). For image dataset, it contains 3 columns, each column corresponding to Red, Green, and Blue component. Each component has an integer value between 0 and 255.
- `K`: the number of desired clusters. Too high value of K may result in empty cluster error. Then, you need to reduce it.

Output

- `class`: cluster assignment of each data point in pixels. The assignment should be 1, 2, 3, etc. For $K = 5$, for example, each cell of class should be either 1, 2, 3, 4, or 5. The output should be a column vector with `size(pixels, 1)` elements. Start from 0 if you are using python.
- `centroid`: location of K centroids (or representatives) in your result. With images, each centroid corresponds to the representative color of each cluster. The output should be a matrix with K rows and 3 columns. The range of values should be [0, 255], possibly floating point numbers.

Hand-in

Both of your code and report will be evaluated. Submit `mykmeans.m` and `mykmedoids.m` files as a zip to Homework 1 Programming (submit `homework1.py` if you are using python). In your report, answer to 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.
2. Attach a picture of your own. We recommend size of 320 × 240 or smaller.
3. Run your K-medoids implementation with the picture you chose above, with several different K. (e.g, small values like 2 or 3, large values like 16 or 32) What did you observe with different K? How long does it take to converge for each K?
4. Run your K-medoids implementation with different initial centroids/representatives. Does it affect final result? Do you see same or different result for each trial with different initial assignments? (We usually randomize initial location of centroids in general. To answer this question, an intentional poor assignment may be useful.)
5. Repeat question 3 and 4 with K-means. Do you see significant difference between K-medoids and K-means, in terms of output quality, robustness, or running time?

Note

- You may see some error message about empty clusters even with Matlab implementation, when you use too large K. Your implementation should treat this exception as well. That is, do not terminate even if you have an empty cluster, but use smaller number of clusters in that case.
- We will grade using test pictures which are not provided. We recommend you to test your code with several different pictures so that you can detect some problems that might happen occasionally.
- 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. Also, directly calling Matlab function `kmeans` or other clustering functions is not allowed.