ISYE/CSE 6740 Homework 1

August 30, 2019

- Submit your answers as an electronic copy on Canvas.
- No unapproved extension of deadline is allowed. Zero credit will be assigned for late submissions. Email request for late submission may not be replied.
- For typed answers with LaTeX (recommended) or word processors, extra credits (10 pts) will be given. If you handwrite, try to be clear as much as possible. No credit may be given to unreadable handwriting.
- Explicitly mention your collaborators if any.

1 Probability [10 pts]

(a) Stores A, B, and C have 50, 75, and 100 employees and, respectively, 50, 60, and 70 percent of these are women. Resignations are equally likely among all employees, regardless of stores and sex. Suppose an employee resigned, and this was a woman. What is the probability that she has worked in store C? [5 pts]

(b) A laboratory blood test is 95 percent effective in detecting a certain disease when it is, in fact, present. The test also yields a false positive result for 1 percent of the healthy persons tested. That is, if a healthy person is tested then with probability 0.01 the test result will imply he has the disease. If 0.5 percent of the population actually has the disease, what is the probability a person has the disease given that his test result is positive? [5 pts]

2 Linear algebra [15 pts]

(a) Let the scalar α be defined by $\alpha = y^T x$, where y is $n \times 1$, and both y and x are functions of vector z. Prove that [7 pts]

$$\tfrac{\partial \alpha}{\partial z} = x^T \tfrac{\partial y}{\partial z} + y^T \tfrac{\partial x}{\partial z}.$$

(b) Let A be a nonsingular, $m \times m$ matrix whose elements are functions of the scalar parameter α. Prove that [8 pts]

$$\tfrac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \tfrac{\partial A}{\partial \alpha} A^{-1}.$$

3 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) Poisson distribution [5 pts]

The Poisson distribution is defined as

$$P(x_i = k) = \frac{\lambda^k e^{-\lambda}}{k!} (k = 0, 1, 2, ...).$$

What is the maximum likelihood estimator of λ?

(b) Multinomial distribution [5 pts]

The probability density function of Multinomial distribution is given by

$$f(x_1, x_2, \ldots, x_k; n, \theta_1, \theta_2, \ldots, \theta_k) = \frac{n!}{x_1! x_2! \cdots x_k!} \prod_{j=1}^{k} \theta_j^{x_j},$$

where $\sum_{j=1}^{k} \theta_j = 1, \sum_{j=1}^{k} x_j = n$. What is the maximum likelihood estimator of $\theta_j, j = 1, \ldots k$?

(c) Gaussian normal distribution [5 pts]

Suppose we have n i.i.d (Independent and Identically Distributed) data samples from a univariate Gaussian normal distribution $\mathcal{N}(\mu, \sigma^2)$, which is given by

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

What is the maximum likelihood estimator of μ and σ^2?

4 Clustering [20 pts]

[a-b] Given N data points $\mathrm{x}^n (n = 1, \ldots, N)$, K-means clustering algorithm groups them into K clusters by minimizing the distortion function over $\{r^{nk}, \mu^k\}$

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r^{nk} \|\mathrm{x}^n - \mu^k\|^2,$$

where $r^{nk} = 1$ if x^n belongs to the k-th cluster and $r^{nk} = 0$ otherwise.

(a) Prove that using the squared Euclidean distance $\|\mathbf{x}^n - \mu^k\|^2$ as the dissimilarity function and minimizing the distortion function, we will have

$$\mu^k = \frac{\sum_n r^{nk}\mathbf{x}^n}{\sum_n r^{nk}}.$$

That is, μ^k is the center of k-th cluster. [5 pts]

(b) Prove that K-means algorithm converges to a local optimum in finite steps. [5 pts]

[c-d] In class, we discussed bottom-up hierarchical clustering. For each iteration, we need to find two clusters $\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_m\}$ and $\{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_p\}$ with the minimum distance to merge. Some of the most commonly used distance metrics between two clusters are:

- Single linkage: the minimum distance between any pairs of points from the two clusters, i.e.

$$\min_{\substack{i=1,\ldots,m \\ j=1,\ldots,p}} \|\mathrm{x}_i - \mathrm{y}_j\|$$

- Complete linkage: the maximum distance between any parts of points from the two clusters, i.e.

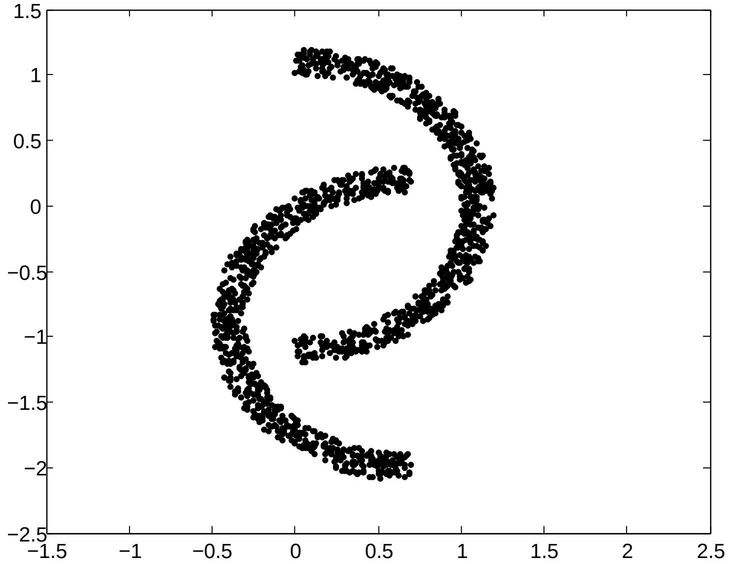
$$\max_{\substack{i=1,\ldots,m \\ j=1,\ldots,p}} \|\mathrm{x}_i - \mathrm{y}_j\|$$

- Average linkage: the average distance between all pair of points from the two clusters, i.e.

$$\frac{1}{mp}\sum_{i=1}^{m}\sum_{j=1}^{p} \|\mathrm{x}_i - \mathrm{y}_j\|$$

(c) When we use the bottom up hierarchical clustering to realize the partition of data, which of the three cluster distance metrics described above would most likely result in clusters most similar to those given by K-means? (Suppose K is a power of 2 in this case). [5 pts]

(d) For the following data (two moons), which of these three distance metrics (if any) would successfully separate the two moons? [5 pts]



5 Programming: Image compression [30 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 $k = \arg\min_j D(\mathrm{x}^n, \mu^j)$, 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 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.
- `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. Upload `mykmeans.m` and `mykmedoids.m` files with your implementation. 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 2 and 3 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.

6 Spectral clustering [10 points]

Consider an undirected graph with non-negative edge weights w_{ij} and graph Laplacian L. Suppose there are m connected compoenents $A_1, A_2, \ldots, A_m$ in the graph. Show that there are m eigenvectors of L corresponding to eigenvalue zero, and the indicator vectors of these components $I_{A_1}, \ldots, I_{A_m}$ span the zero eigenspace.