CS 446 / ECE 449 — Homework 1

acard6

Version 1.0

Instructions.

- Homework is due **Tuesday, Feb 7th, at noon CDT**.
- Everyone must submit individually at gradescope under `hw1` and `hw1code`.
- The "written" submission at `hw1` **must be typed**, and submitted in any format gradescope accepts (to be safe, submit a PDF). You may use LaTeX, markdown, google docs, MS word, whatever you like; but it must be typed!
- When submitting at `hw1`, gradescope will ask you to mark out boxes around each of your answers; please do this precisely!
- Please make sure your NetID is clear and large on the first page of the homework.
- Your solution **must** be written in your own words. Please see the course webpage for full academic integrity information. Briefly, you may have high-level discussions with at most 3 classmates, whose NetIDs you should place on the first page of your solutions, and you should cite any external reference you use; despite all this, your solution must be written in your own words.
- We reserve the right to reduce the auto-graded score for `hw1code` if we detect funny business (e.g., your solution lacks any algorithm and hard-codes answers you obtained from someone else, or simply via trial-and-error with the autograder).
- When submitting to `hw1code`, only upload `hw1.py` and `hw1_utils.py`. Additional files will be ignored.

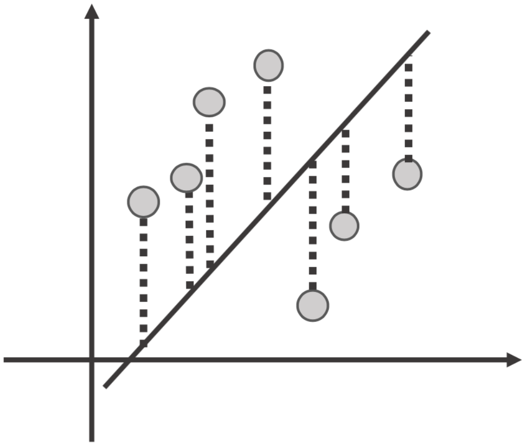
Version History.

1.0 Initial Version.

1. Principal Component Analysis

(a) For each of the following statements, specify whether the statement is true or false. If you think the statement is wrong, explain in 1 to 2 sentences why it is wrong.

- True or False: As shown in the figure below, PCA seeks a subspace such that the sum of all the vertical distance to the subspace (the dashed line) is minimized.



True

- True or False: PCA seeks a projection that best represents the data in a least-squares sense.
True
- True or False: PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables.
True
- True or False: The principal components are not necessarily orthogonal to each other.
False: PCA uses covariance matrix which are always symmetric and always have eigenvalues and vectors. The eigenvectors are always orthogonal to the symmetric matrix from which they stem from.

(b) Recall that PCA finds a direction w in which the projected data has highest variance by solving the following program:

$$\max_{w:||w||^2=1} w^T \Sigma w. \tag{1}$$

Here, Σ is a covariance matrix. You are given a dataset of two 2-dimensional points $(1, 3)$ and $(4, 7)$. Draw the two data points on the 2D plane. What is the first principal component w of this dataset?

$\mu_x = \frac{1+4}{2} = 2.5, \mu_y = \frac{3+3}{2} = 5$, Cov(x,y) $= \sum_{i=1}^{2} \frac{(x_i-\mu_x)*(y_i-\mu_y)}{2} = 3$, Var(x)=2.25, Var(y)=4
$\Sigma = \begin{bmatrix} 2.25 & 3 \\ 3 & 4 \end{bmatrix}$
Its eigenvalues are $\lambda = 0, \frac{25}{4}$, our largest eigenvalue greater than 0 and its corresponding eigenvector is $\begin{bmatrix} 2.25 & 3 \\ 3 & 4 \end{bmatrix}$ x = λx $\rightarrow \begin{bmatrix} -3 \\ 4 \end{bmatrix}$.

(c) Now you are given a dataset of four points $(2, 0)$, $(2, 2)$, $(6, 0)$ and $(6, 2)$. Draw the four data points on the 2D plane. Given this dataset, what is the dimension of the covariance matrix Σ in Eq. (1)? Also, explicitly write down the values of Σ given the dataset.
the mean and variance of x and y are $\mu_x = 4, \mu_y = 1, Var(x) = 4, Var(y) = 1, withtheCov(x, y) = 0$. The dimensions of the covariance matrix Σ is 2,

$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

(d) What is the optimal w and the optimal value of the program in Eq. (1) given

$$\Sigma = \begin{bmatrix} 12 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}.$$

Knowing the optimal w is found by taking the eigenvector of the largest eigenvalue of the covariance matrix, which after being computed leaves the eigenvalues to be $\lambda = 10, 12, 6, 20$, taking the largest value and finding its vector is is $\lambda = 20 \rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ and plugginh this back into Eq. 1 we get $\max_{w:||w||^2=1} w^T \Sigma w = 20$

2. K-Means 1

(a) Mention if K-Means is a supervised or an un-supervised method and state the reason.
Asnwer un-supervised, since with kmean there is no requirement to label data of anysort, rather the algorithm is rather simply just clustering data together and dealing with distortion.

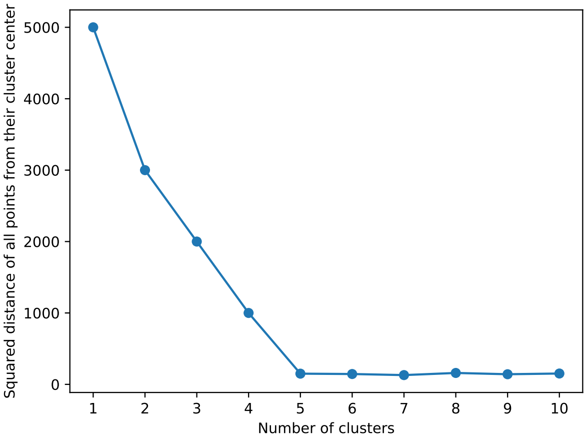
(b) Assume that you are trying to cluster data points x_i for $i \in \{1, 2, \ldots, D\}$ into K clusters each with center μ_k where $k \in \{1, 2, \ldots, K\}$. The objective function for doing this clustering involves minimization of the Euclidean distance between the points and the cluster centers. It is given by

$$\min_{\mu} \min_{r} \sum_{i \in D} \sum_{k=1}^{K} \frac{1}{2} r_{ik} \|x_i - \mu_k\|_2^2.$$

How do you ensure hard assignment of one data point to one and only one cluster at a given time?
Hint: By hard assignment we mean that you are 100 % sure that a point either belongs or doesn't belong to a cluster.
Answer: to best ensure that we are 100% certain about what cluster a point belongs to we look to see what cluster minimizes its distance to the point and if it just so happens that a point is equdistance from multiple cluster then we simply just take the first cluster presented

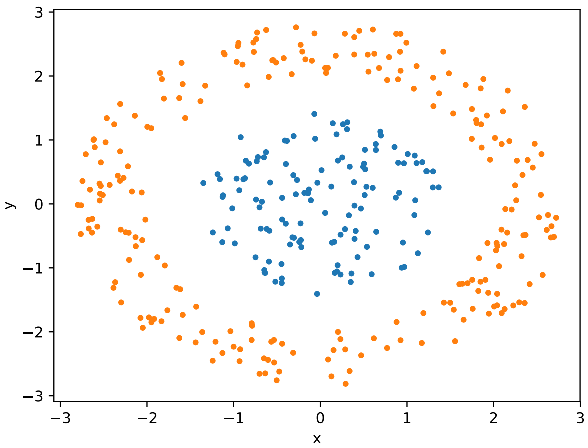
(c) How does your answer to part b change if we want to obtain a soft assignment instead?
Hint: By soft assignment we mean that a point belongs to a cluster with some probability.
Answer: If we were to use soft clustering then we could use percentages and probability where a point would stores a the likelihood that its near a certain cluster since we no longer have to worry about beign set to one main cluster, but rather the likelyhood of being in a cluster. This leaves us with less ambiguity as to what to do with points that have clusters with equal probability since they can belong to multiple.

(d) Looking at the following plot, what is the best choice for the number of clusters?



Answer: since after 5 clusters and onward the spread from the cluster center seems to level out the best choice is 5 clusters.

(e) Would K-Means be an efficient algorithm to cluster the following data? Explain your answer in a couple of lines.



Answer: No, since trying to classify the data into two distinct groups would require multiple groups. The overall mean of all the datapoints is the center of the and trying to tether the outter ring to a single cluster would be hard for that reason it is unreasonable to use k-mean clustering for this data

3. K-Means 2

We are given a dataset $\mathcal{D} = \{(x)\}$ of 2d points $x \in \mathbb{R}^2$ which we are interested in partitioning into K clusters, each having a cluster center μ_k ($k \in \{1, \ldots, K\}$) via the k-Means algorithm. This algorithm optimizes the following cost function:

$$\min_{\mu_k, r} \sum_{x \in \mathcal{D}, k \in \{1,\ldots,K\}} \frac{1}{2} r_{x,k} \|x - \mu_k\|_2^2 \qquad \text{s.t.} \quad \begin{cases} r_{x,k} \in \{0,1\} & \forall x \in \mathcal{D}, k \in \{1,\ldots,K\} \\ \sum_{k \in \{1,\ldots,K\}} r_{x,k} = 1 & \forall x \in \mathcal{D} \end{cases} \tag{2}$$

(a) What is the domain for μ_k?

(b) Given fixed cluster centers μ_k $\forall k \in \{1, \ldots, K\}$, what is the optimal $r_{x,k}$ for the program in Eq. 2? Provide a reason?

(c) Given fixed $r_{x,k}$ $\forall x \in \mathcal{D}, k \in \{1, \ldots, K\}$, what are the optimal cluster centers μ_k $\forall k \in \{1, \ldots, K\}$ for the program in Eq. 2?

Hint: Reason by first computing the derivative w.r.t μ_k.

(d) Using Pseudo-code, sketch the algorithm which alternates the aforementioned two steps. Is this algorithm guaranteed to converge and why? Is this algorithm guaranteed to find the global optimum? What is the reason?

Hint: you can provide a counter-example to invalidate a statement.

(e) Please implement the aforementioned two steps. For the given dataset, after how many updates does the algorithm converge, what cost function value does it converge to and what are the obtained cluster centers? Visualize clusters at each step and attach the plots here. Please at least report numbers with one decimal point.

Remark: how we count updates: when computing a set of new centroids from initialization, we call this one update.

Hint: You may find `hw1_utils.vis_cluster` useful.

Answer A) If the data is in the 2D-plane then the domain of μ_k is also 2D as a real value

Answer B) The optimal $r_{x,k}$ is

$$r_{x,k} = \begin{cases} 1 & \text{for } \arg\min_k |x^i - \mu_k|^2 \\ 0 & \text{else} \end{cases} \tag{3}$$

Answer C) The optimal cluster centers $\mu_k \forall$ is

$$\mu_k = \frac{\sum_{i \in D} r_{ik} x^{(i)}}{\sum_{i \in D} r_{ik}}$$

Answer D) The following algorithm may not always be guaranteed to find global optimal since there is a possibility that when initializing the random starting centroids it can get stuck in local minima and not reach the correct optimum. To overcome this issue it is best to run it many times with different initial cluster centroid to ensure that optimum can be found.

Algorithm 1 k-means clustering

```
Input $\mathcal{D} \leftarrow$ set of 2D points
Input K $\leftarrow$ number of cluster centroid $\mu_k$
randomly assign centroids to points in data
**repeat**
  **for** iteration $\leftarrow$ 1 to max_iter **do**
    **for all** $x_i \in \mathcal{D}$ **do**
      $r_{ik} \leftarrow \arg\min_k ||x_i - \mu_k||^2$
    **end for**
    **for** $k \leftarrow 1toK$ **do**
      $\mu_k \leftarrow$ ($\sum x^i$ that are a part of that $\mu_k$) / (the number of points in the $\mu_k$)
    **end for**
  **end for**
**until** max iteration or all cluster $\mu_k$ reach equalibrium
```

4. Gaussian Mixture Models

Consider a Gaussian mixture model with K components ($k \in \{1, \dots, K\}$), each having mean μ_k, variance σ_k^2, and mixture weight π_k. Further, we are given a dataset $\mathcal{D} = \{x_i\}$, where $x_i \in \mathbb{R}$. We use $z_i = \{z_{ik}\}$ to denote the latent variables.

(a) What is the log-likelihood of the data according to the Gaussian Mixture Model (use μ_k, σ_k, π_k, K, x_i, and $\mathcal{D}$)?

(b) Assume $K = 1$, find the maximum likelihood estimate for the parameters (μ_1, σ_1^2, π_1).

(c) What is the probability distribution on the latent variables, i.e., what is the distribution $p(z_{i,1}, z_{i,2}, \cdots, z_{i,K})$ underlying Gaussian mixture models. Also give its name.

(d) For general K, what is the posterior probability $p(z_{ik} = 1|x_i)$? To simplify, wherever possible, use $\mathcal{N}(x_i|\mu_k, \sigma_k)$, a Gaussian distribution over $x_i \in \mathbb{R}$ having mean μ_k and variance σ_k^2.

(e) How are k-Means and Gaussian Mixture Model related? (There are three conditions)
Hint: Think of variance, π_k, and hard/soft assignment.

(f) Show that:

$$\lim_{\epsilon \to 0} -\epsilon \log \sum_{k=1}^{K} \exp\left(-F_k/\epsilon\right) = \min_k F_k, \quad \epsilon \in \mathbb{R}^+$$

Hint: Use l'Hopital's rule.

(g) Consider the modified Gaussian Mixture Model objective:

$$\min_{\mu} - \sum_{x_i \in \mathcal{D}} \epsilon \log \sum_{k=1}^{K} \exp\left(-(x_i - \mu_k)^2/\epsilon\right).$$

Conclude that the objective for k-Means is the 0-temperature limit of Gaussian Mixture Model.
Hint: Let $F_k = (x - \mu_k)^2$ and apply the equation you proved in (f).

Answer A) The log likelihood of of $p(x_i|\pi, \mu, \sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)$ is $log(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \sigma_k))$

Answer B) Assuming that $K = 1$ we can find the MLE by maximizing the log-likelihood function, which in this case is $log(N(x_i|\mu_k, \sigma_k) = log(p(x_i)|\mu, \sigma) = \sum_(i \in \mathcal{D}) \frac{(x_i - \mu)^2}{2\sigma^2} + \frac{N}{2} log(2\pi\sigma^2)$, where N is the size of the dataset. since K=1 there is no need for summation and $\pi_1 = 1$ since it can be no other value. Thus maximizing the MLE for μ and σ, we get the sample mean and variance $\mu = \frac{1}{N} \sum_{i \in \mathcal{D}} x_i$, $\sigma^2 = \frac{1}{N} \sum_{i \in \mathcal{D}} (x_i - \mu)^2$

Answer C) the probability of the auxiliary variable is $p(z_{ik} = 1) = \pi_k, \, \Pi_{k=1}^{K} \pi_k^{z_{ik}}, z_i = [z_{i1}, ..., z_{iK}]^T$

Answer D) for a general K the posterior probability is $p(z_{ik}|x_i) = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x_i|\mu_{\hat{k}}, \sigma_{\hat{k}})}$

Answer E) while both kmean and gmm are unsupervised learning techniques that use pre-determined clusters numbers, they relate in how they label the datas relation to one another using distance to determine how related they are to the rest of the data in some shape. As well as their use of the variance of the data to better form relations between each iteration to better fit a mold for the data, and determing how to assign data to a cluster whether the use of probability π or not.

Answer F) so consider $f(\epsilon) = -epsilon, g(\epsilon) = \sum_{k=1}^{K} exp(-F_k/\epsilon), h(\epsilon) = log(g(\epsilon))$ the limit as f approaches 0 is 0 and the limit as h approaches 0 from the right is $log \sum_{k=1}^{K} exp(-F_k/\epsilon) = log \sum_{k=1}^{K} exp(-\inf) = log(k * 0) = log(0) \to \lim_{x \to 0^+} log(x) = -\inf$ so we get $\lim_{\epsilon \to 0} f(\epsilon)h(\epsilon) = 0 * -\inf$ by l'Hopitals we get $f'(\epsilon) = -1, h'(\epsilon) = \frac{1}{g(\epsilon)} g'(\epsilon) = \frac{\sum_{k=1}^{k} F_k * exp(-F_k/\epsilon)}{\epsilon^2 \sum_{k=1}^{K} exp(-F_k/\epsilon)} = \frac{1}{\epsilon^2} \sum_{k=1}^{K} \frac{F_k * exp(-F_k/\epsilon)}{exp(-F_k/\epsilon)} = \sum \frac{F_k}{\epsilon^2}$

$$\lim_{\epsilon \to 0} \sum_{k=1}^{K} F_k/\epsilon^2 = \lim_{\epsilon \to 0} \frac{d^2}{d\epsilon^2} \sum_{k=1}^{K} F_k/\epsilon^2 = \min_k F_k/1 = \min_k F_k$$

Answer G) if $F_k = (x-\mu_k)^2$ then

$$\lim_{\epsilon\to 0^+}\min_{\mu} - \sum_{x_i\in(D)} \epsilon log \sum exp(F_k/\epsilon) = \min_{\mu} \sum_{x_i\in(D)} min_k F_k = \min_{\mu}\min_{k} \sum_{x_i\in\mathcal{D}} (x_i-\mu)^2$$

is the cost function of k-means clustering, therefore the k-mean is the 0-temp limit of the gaussian mixture model

Appendix

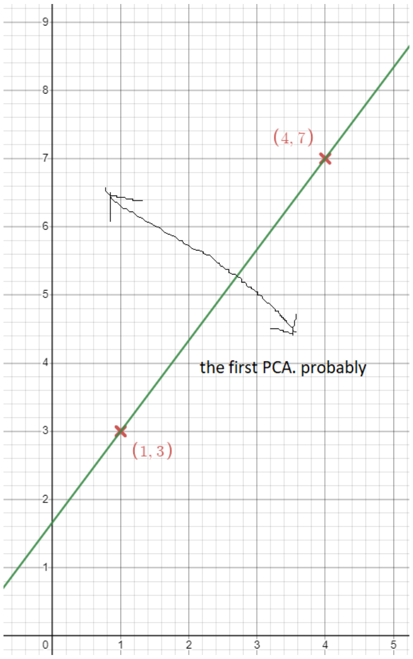


figure 1: the drawing for Q1 part b

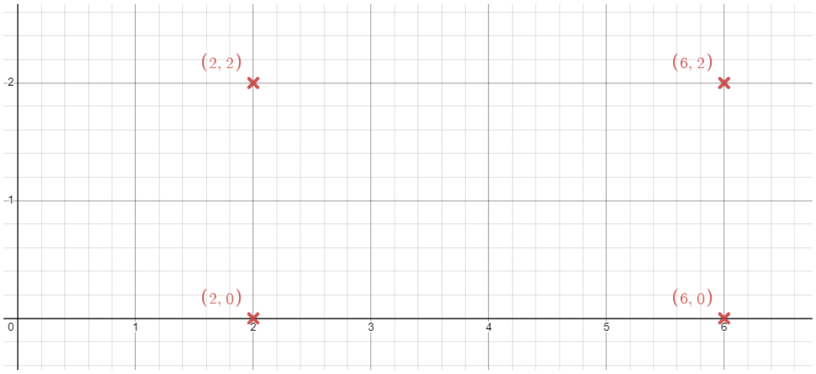


figure 2: the drawing for Q1 part c