CS 446 / ECE 449 — Homework 4

yuhangr2

Version 1.0

Instructions.

- Homework is due **Wednesday, November 3, at noon CST**; you have **3** late days in total for **all Homeworks**.
- Everyone must submit individually at gradescope under `hw4` and `hw4code`.
- The "written" submission at `hw4` **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 `hw4`, 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. You should cite any external reference you use.
- We reserve the right to reduce the auto-graded score for `hw4code` 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 `hw4code`, only upload `hw4.py` and `hw4_utils.py`. Additional files will be ignored.

1. Principal Component Analysis

(a) In the lecture, we mainly discuss the case where the data centers around the origin point (see slides 14/21). If the data does not center around the origin point, the PCA algorithm will first subtract the mean of all the data, then calculate the projection $\boldsymbol{w}$, and eventually add the mean back (see slides 15/21). In this question, we will prove that subtracting the mean is reasonable for PCA. However, since proving the general theorem is beyond the scope of this course, we will focus on the 2-dimensional case.

For a line on the xy-plane, we denote it as the set of roots/solutions to $f(\boldsymbol{p}) = 0$, where $f(\boldsymbol{p}) = \boldsymbol{v}^\top \boldsymbol{p} + b$. Provided N points on the xy plane: $\mathcal{P} = [\boldsymbol{p}^{(1)}, ..., \boldsymbol{p}^{(N)}]$, and $\boldsymbol{p}^{(i)} = (x^{(i)}, y^{(i)})$. We want to find the optimal line $\widetilde{f} = 0$ with parameters $\widetilde{\boldsymbol{v}}$ and $\widetilde{b}$ that minimizes the sum of the squared distances between the points and the line as the equation below.

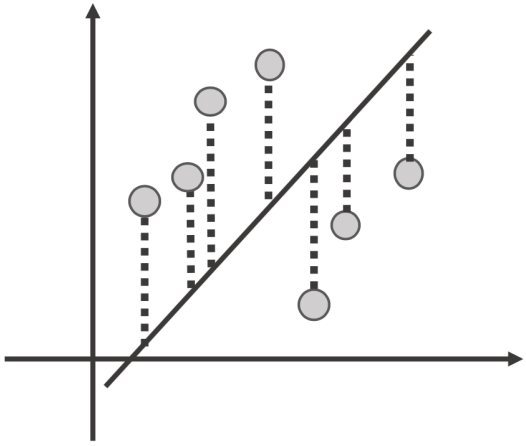
$$\widetilde{\boldsymbol{v}}, \widetilde{b} = \arg\min_{\boldsymbol{v}, b} \sum_{i=1}^{N} \left(\frac{\boldsymbol{v}^\top \boldsymbol{p}^{(i)} + b}{\|\boldsymbol{v}\|_2}\right)^2 \tag{1}$$

Prove that if the optimal line exists, then the mean of the N points is on the line.

Hint. Without loss of generality, consider the case where $\boldsymbol{v}$ is a unit vector. Note that $\boldsymbol{v}$ is the normal vector of the line.

(b) 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 line such that the sum of the vertical distances from the points to the line is minimized.



- True or False: PCA seeks a projection that best represents the data in a least-squares sense.
- True or False: The principal components are not orthogonal to each other.
- True or False: Solving PCA using SVD might result in a solution which corresponds to a local minimum.

(c) In PCA, assume we have

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 12 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}$$

Compute the largest eigenvalue of $\boldsymbol{X}^\top \boldsymbol{X}$ and its corresponding eigenvector.

Solution.

(a) Let the mean of the N points be $\bar{\boldsymbol{p}} = (\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$.
Without loss of generality, consider the case where $\boldsymbol{v}$ is a unit vector. So $\|\boldsymbol{v}\|_2^2 = 1$.
$f(\boldsymbol{v}, b) = \sum_{i=1}^{N} (\frac{\boldsymbol{v}^\top \boldsymbol{p}^{(i)} + b}{\|\boldsymbol{v}\|_2})^2 = \sum_{i=1}^{N} (\boldsymbol{v}^\top \boldsymbol{p}^{(i)} + b)^2$
$\frac{\partial f}{\partial b} = \sum_{i=1}^{N} (2b + 2v_1 x_i + 2v_2 y_i) = 0$ when f takes the minimum, because it is of the form $\|\boldsymbol{a}\boldsymbol{x} + \boldsymbol{b}\|^2$. Dividing the above equation by $2N$, we get $v_1\bar{x} + v_2\bar{y} + b = 0$, i.e. $\boldsymbol{v}^T\bar{\boldsymbol{p}} + b = 0$.
Therefore, if the optimal line exists, then the mean of the N points is on the line.

(b)
i. False. PCA seeks a line such that the sum of Euclidean distances of the points to the line is minimized.

ii. True.

iii. False. The principal components have to be orthonormal.

iv. True. SVDs may not be unique.

(c) $A - \lambda I = 0$. Solving the equation we get the largest eigenvalue is 20. And the corresponding eigenvector is $\boldsymbol{v} = [0, 0, 1, 0]^T$.

2. K-Means

We are given a dataset $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$ of d-dimensional points $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ which we are interested in partitioning them into K clusters, each having a cluster center $\boldsymbol{\mu}_k \in \mathbb{R}^d$ ($k \in [K]$) via the K-Means algorithm. Note that we denote $[K] = \{1, \ldots, K\}$ and similarly for $[N]$. Recall (from our lecture notes) that the algorithm optimizes the following cost function:

$$\min_{\boldsymbol{\mu}_k \in \mathbb{R}^d, k \in [K]} \quad \min_{r_{ik} \in \{0,1\}, i \in [N], k \in [K]} \quad \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{1}{2} r_{ik} \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k\|_2^2 \quad \text{s.t.} \quad \sum_{k=1}^{K} r_{ik} = 1 \quad \forall k \in [K]. \tag{2}$$

(a) Given fixed cluster centers $\boldsymbol{\mu}_k$, $\forall k \in [K]$, what is the optimal assignments r_{ik} for the optimization in (2)? (assume that there is no $\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_p\|_2^2 = \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_q\|_2^2$ for $\forall p \neq q$)

(b) Given fixed assignments r_{ik}, $\forall i \in [N], k \in [K]$, we assume each cluster center has at least one data point assigned to it, i.e., $\sum_{i=1}^N r_{ik} > 0$. What are the optimal cluster centers $\boldsymbol{\mu}_k$ for the optimization (2)?

(c) Implement the K-Means (with $K = 2$) algorithm for a 2-dimensional dataset and submit your code to `hw4code` on gradescope. For the given dataset, visualize the clusters at the first three steps and attach the plots in your written (typed) report. Please see `hw4.py` for details.

Solution.

(a) The optimal assignment is the nearest neighbor cluster centers for $\boldsymbol{x}^{(i)}$. In other words, $r_{ik} = 1$ if $k = \text{argmin}_{\text{k} \in 1,\ldots,\text{K}} \|\mathbf{x}^{(\text{i})} - \mu_{\text{k}}\|_2^2$, or 0 otherwise.

(b) $\boldsymbol{\mu}_k = \frac{\sum_{i=1}^N r_{ik} \boldsymbol{x}^{(i)}}{\sum_{i=1}^N r_{ik}}$.

(c) Figures are included in the following.

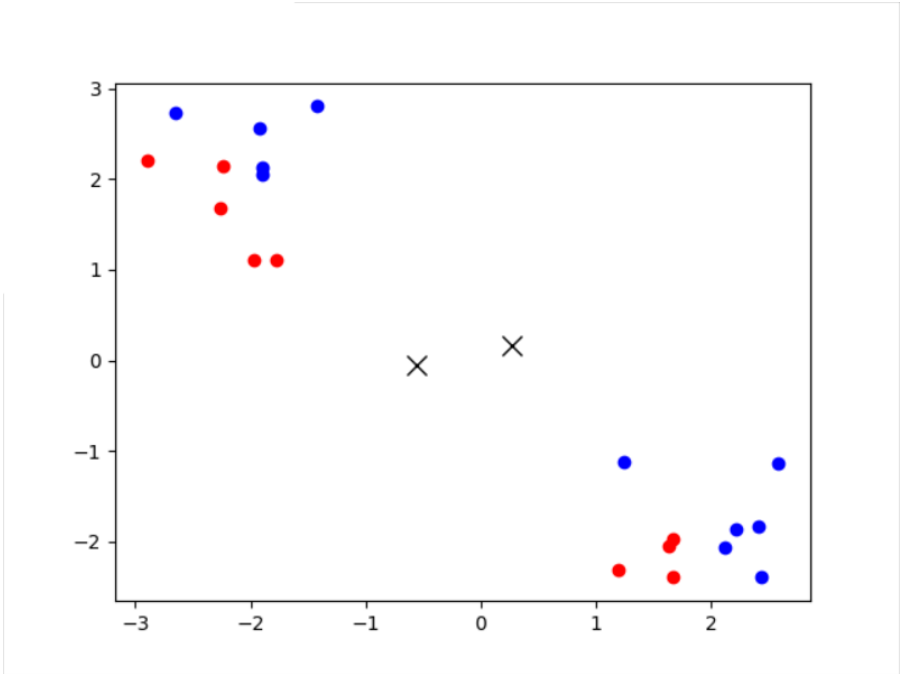


Figure 1: Q2(C): First iteration.

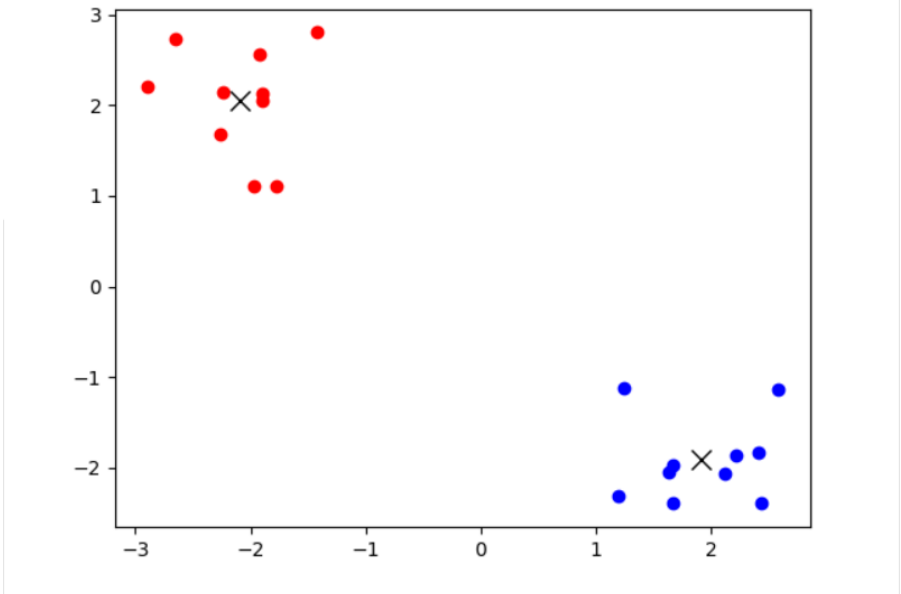


Figure 2: Q2(C): Second iteration.

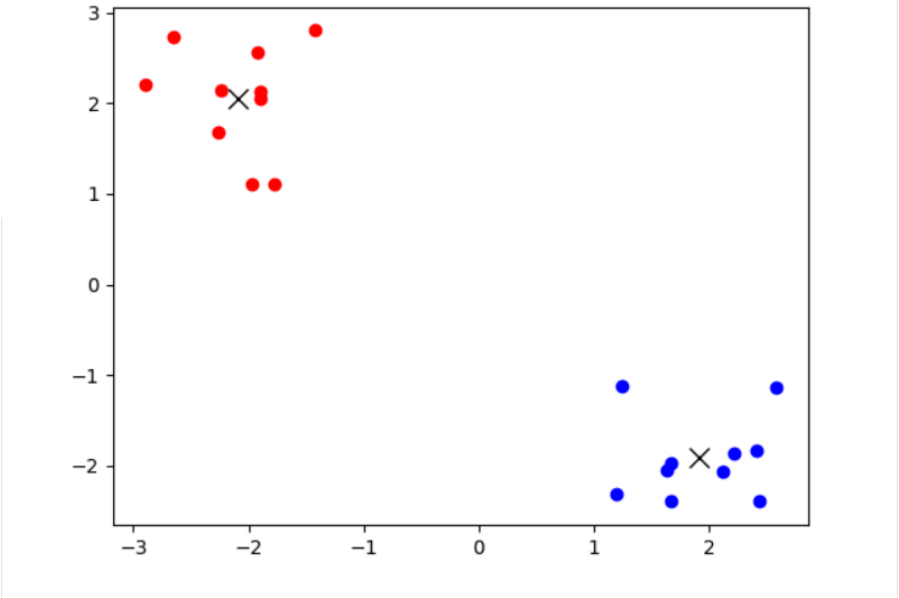


Figure 3: Q2(C): Second iteration.

3. Gaussian Mixture Models

Consider a one-dimensional Gaussian mixture model with K components ($k \in \{1, \ldots, K\}$), each having mean μ_k, variance σ_k^2, and mixture weight π_k. Further, we are given a dataset $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$, where $x^{(i)} \in \mathbb{R}$.

(a) What is the log-likelihood of the data, i.e., $\log p(\mathcal{D} \mid \mu_k, \sigma_k, \pi_k, 1 \le k \le K)$, according to the Gaussian Mixture Model, assuming the data samples $x^{(i)}$ are i.i.d.?

(b) Recall the Expectation-Maximization procedure for Gaussian mixture models from our lecture where $r_{ik} = p_{\psi^{(t)}}(\boldsymbol{z}^{(i)} = \boldsymbol{e}_k \mid x^{(i)})$. Prove that $\sum_{k=1}^K r_{ik} = 1$.

(c) Recall the Expectation-Maximization procedure for Gaussian mixture models from our lecture where Lagrangian is

$$L = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik}[\tfrac{1}{2\sigma_k^2}(x^{(i)} - \mu_k)^2 + \tfrac{1}{2}\log(2\pi\sigma_k^2) - \log \pi_k] + \lambda(\sum_{k=1}^{K} \pi_k - 1).$$

Show how to derive $\pi_k = \frac{1}{N}\sum_{i=1}^N r_{ik}$ from setting $\frac{\partial L}{\partial \lambda} = 0$ and $\frac{\partial L}{\partial \pi_k} = 0$ $(1 \le k \le K)$.

(d) A nice property of Gaussian mixture is the universal approximation property. Loosely speaking, given any distribution, there exist a Gaussian mixture that can approximate it up to arbitrary accuracy.

Let us verify this powerful property on discrete distributions. Consider n points $z_1, z_2, \ldots, z_n \in \mathbb{R}$, and a discrete distribution characterized by a probability mass function q on $\mathbb{R}$:

$$q(x) = \begin{cases} q_i & \text{if } x = z_i, \\ 0 & \text{otherwise,} \end{cases}$$

where $q_i > 0$ and $\sum_{i=1}^n q_i = 1$.

Construct a Gaussian mixture on $\mathbb{R}$ that approximates the distribution characterized by the probability mass function q . It is sufficient to describe the construction and show intuitively why it is a good approximation. You don't need to rigorously prove your results.

(e) We have seen the Gaussian mixture with finitely many components (K components), where π_k ($k \in \{1, \ldots, K\}$) are the mixture weights. Noting that $\sum_{k=1}^K \pi_k = 1$ and $\pi_k \ge 0$, we can see that the mixture weights represent a probability mass function. Therefore, we can generalize the Gaussian mixture to combine infinitely many components by using a probability density function.

Formally, let $\mu : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}_+$ be two functions representing the means and variances, respectively. Let $\pi : \mathbb{R} \to \mathbb{R}_+$ be a probability density function on $\mathbb{R}$ representing the mixture weights. Denote the density of the generalized Gaussian mixture as

$$p(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma(t)^2}} \exp\Big(-\frac{(x-\mu(t))^2}{2\sigma(t)^2}\Big)\pi(t)\ dt. \tag{3}$$

Prove that p is a valid density function, i.e., $\forall x \in \mathbb{R} : p(x) \ge 0$ and $\int_{-\infty}^{\infty} p(x)dx = 1$ (assuming $p(x)$ and $\int_{-\infty}^{\infty} p(x)dx$ exist).

Hint: See the Fubini's Theorem, but you don't need to worry about the mathematical details.

Solution.

(a) $\sum_{i=1}^N \log \sum_{k=1}^K \pi_k N(x^{(i)} | \mu_k, \sigma_k)$

(b) $r_{ik} = p_{\psi^{(t)}}(\boldsymbol{z}^{(i)} = \boldsymbol{e}_k \mid x^{(i)}) = \frac{p_{\psi^{(t)}}(\boldsymbol{z}^{(i)}=\boldsymbol{e}_k) p_{\psi^{(t)}}(x^{(i)}|\boldsymbol{z}^{(i)}=\boldsymbol{e}_k)}{\sum_{j=1}^{K} p_{\psi^{(t)}}(\boldsymbol{z}^{(i)}=\boldsymbol{e}_j) p_{\psi^{(t)}}(x^{(i)}|\boldsymbol{z}^{(i)}=\boldsymbol{e}_j)}$

The denominator is just the sum of the numerator for different k. So $\sum_{k=1}^{K} r_{ik} = 1$.

(c) From $\frac{\partial L}{\partial \lambda} = 0$ and $\frac{\partial L}{\partial \pi_k} = 0$ we have:
$\sum_{k=1}^{K} \pi_k = 1$ and $\sum_{i=1}^{N} r_{ik} = \lambda \pi_k$.
$\sum_{k=1}^{K} \lambda \pi_k = \sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} = \sum_{i=1}^{N} 1 = N$.
Since $\sum_{k=1}^{K} \pi_k = 1$, $\lambda = N$.
Therefore, $\pi_k = \frac{1}{N} \sum_{i=1}^{N} r_{ik}$.

(d) $q(x) = \sum_{i=1}^{n} q_i N(x_i|z_i, 0)$. The intuition is that we treat each q_i as a Gaussian distribution with mean z_i and 0 deviation.

(e) $p(x) \geq 0$: The probability density is always greater or equal to zero no matter what value t has. Hence the integration $p(x)$ is also always greater or equal to zero.
$\int_{-\infty}^{\infty} p(x)dx = 1$:
$\int_{-\infty}^{\infty} p(x)dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma(t)^2}} \exp\Big(- \frac{(x-\mu(t))^2}{2\sigma(t)^2} \Big) \pi(t)\ dt\ dx$
$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma(t)^2}} \exp\Big(- \frac{(x-\mu(t))^2}{2\sigma(t)^2} \Big) \pi(t)\ dx\ dt$
$= \int_{-\infty}^{\infty} \pi(t)\ dt$
$= 1$