CSE6740 CDA Homework

Name:
GTID:

Deadline: Sep 2nd 11:59 pm ET

1 Principal Component Analysis (PCA) [15 pts]

1.1 Reconstruction error and variance maximization [5 pts]

Assume the centered data matrix $X \in \mathbb{R}^{n \times d}$ has rows $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$. Let $q \leq d$ and choose q unit-length, mutually orthogonal directions $\mathbf{w}_1, \ldots, \mathbf{w}_q \in \mathbb{R}^d$. Define $w = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_q] \in \mathbb{R}^{d \times q}$.

(a) Write w as the matrix formed by stacking the $\mathbf{w}_i$. Prove that $w^\top w = I_q$.

(b) **Scores:** Find the $n \times q$ matrix of q-dimensional scores in terms of x and w. *Hint: your answer should reduce to* $\mathbf{x}_i^\top \mathbf{w}_1$ *when* $q = 1$.

(c) **Reconstructions:** Find the $n \times d$ matrix of d-dimensional approximations (reconstructions) based on these scores, in terms of x and w. *Hint: your answer should reduce to* $(\mathbf{x}_i^\top \mathbf{w}_1)\mathbf{w}_1^\top$ *when* $q = 1$.

(d) Show that the mean-squared reconstruction error obtained using $\mathbf{w}_1, \ldots, \mathbf{w}_q$ is the sum of two terms: one that depends only on x (and not on w), and another that depends only on the scores along those directions (and not otherwise on what those directions are).

(e) Explain in what sense minimizing projection residuals is equivalent to maximizing the sum of the variances of the scores along the different directions.

Solution:

(a) Proof that $w^\top w = I_q$

Since $w = [\mathbf{w}_1 \ \mathbf{w}_2 \ \cdots \ \mathbf{w}_q] \in \mathbb{R}^{d \times q}$, we have:

$$w^\top w = \begin{bmatrix} \mathbf{w}_1^\top \\ \mathbf{w}_2^\top \\ \vdots \\ \mathbf{w}_q^\top \end{bmatrix} [\mathbf{w}_1 \ \mathbf{w}_2 \ \cdots \ \mathbf{w}_q]$$

The (i, j)-th entry of $w^\top w$ is:

$$(w^\top w)_{ij} = \mathbf{w}_i^\top \mathbf{w}_j$$

Since the directions are mutually orthogonal and unit-length:

$$\mathbf{w}_i^\top \mathbf{w}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Therefore:

$$w^\top w = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_q$$

(b) Scores Matrix

The $n \times q$ matrix of scores is:

$$\boxed{S = xw}$$

where $x \in \mathbb{R}^{n \times d}$ and $w \in \mathbb{R}^{d \times q}$.
Verification: The i-th row of S is:

$$S_i = \mathbf{x}_i^\top w = [\mathbf{x}_i^\top \mathbf{w}_1 \quad \mathbf{x}_i^\top \mathbf{w}_2 \quad \cdots \quad \mathbf{x}_i^\top \mathbf{w}_q]$$

When $q = 1$, this reduces to $\mathbf{x}_i^\top \mathbf{w}_1$

(c) Reconstructions Matrix

The $n \times d$ matrix of reconstructions is:

$$\boxed{\hat{x} = xww^\top}$$

Verification: The i-th row of $\hat{x}$ is:

$$\hat{\mathbf{x}}_i^\top = \mathbf{x}_i^\top ww^\top = \sum_{j=1}^{q} (\mathbf{x}_i^\top \mathbf{w}_j)\mathbf{w}_j^\top$$

When $q = 1$, this reduces to $(\mathbf{x}_i^\top \mathbf{w}_1)\mathbf{w}_1^\top$

(d) Mean-Squared Reconstruction Error Decomposition

The reconstruction error matrix is:

$$E = x - \hat{x} = x - xww^\top = x(I_d - ww^\top)$$

The mean-squared reconstruction error is:

$$\text{MSE} = \frac{1}{n}\|E\|_F^2 = \frac{1}{n}\text{tr}(E^\top E) \tag{1}$$

$$= \frac{1}{n}\text{tr}((x(I_d - ww^\top))^\top (x(I_d - ww^\top))) \tag{2}$$

$$= \frac{1}{n}\text{tr}((I_d - ww^\top)x^\top x(I_d - ww^\top)) \tag{3}$$

Using the fact that $(I_d - ww^\top)$ is a projection matrix (idempotent and symmetric):

$$(I_d - ww^\top)^2 = I_d - ww^\top$$

Therefore:

$$\text{MSE} = \frac{1}{n}\text{tr}((I_d - ww^\top)x^\top x) \tag{4}$$

$$= \frac{1}{n}\text{tr}(x^\top x) - \frac{1}{n}\text{tr}(ww^\top x^\top x) \tag{5}$$

$$= \frac{1}{n}\text{tr}(x^\top x) - \frac{1}{n}\text{tr}(w^\top x^\top xw) \tag{6}$$

Since $w^\top x^\top xw = (xw)^\top (xw) = S^\top S$ where $S = xw$ are the scores:

$$\text{MSE} = \underbrace{\frac{1}{n}\text{tr}(x^\top x)}_{\text{depends only on } x} - \underbrace{\frac{1}{n}\text{tr}(S^\top S)}_{\text{depends only on scores}}$$

(e) Equivalence of Minimizing Error and Maximizing Variance

From part (d), we have:

$$\text{MSE} = \frac{1}{n}\text{tr}(x^\top x) - \frac{1}{n}\text{tr}(S^\top S)$$

The first term $\frac{1}{n}\text{tr}(x^\top x) = \sum_{i=1}^n \|\mathbf{x}_i\|^2/n$ is the total variance of the original data (since data is centered), which is constant regardless of the choice of w.

The second term $\frac{1}{n}\text{tr}(S^\top S)$ represents the sum of variances along the chosen directions:

$$\frac{1}{n}\text{tr}(S^\top S) = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{w}_j)^2 = \sum_{j=1}^{q}\text{Var(scores along } \mathbf{w}_j)$$

Therefore:

$$\text{Minimize MSE} \iff \text{Maximize } \frac{1}{n}\text{tr}(S^\top S) \iff \text{Maximize sum of score variances}$$

Interpretation: Since the total variance is fixed, minimizing the reconstruction error is equivalent to maximizing the variance captured by the projections onto the chosen directions. This is the fundamental principle of PCA: find directions that capture maximum variance, which equivalently minimize reconstruction error.

1.2 First Principal Component via Projection-Error Minimization [10pts]

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be centered and standardized so each coordinate has unit variance. For a unit vector $\mathbf{v}$, define $f_{\mathbf{v}}(\mathbf{x})$ to be the projection of $\mathbf{x}$ onto the line $\mathcal{V} = \{\alpha\mathbf{v} : \alpha \in \mathbb{R}\}$, i.e.,

$$f_{\mathbf{v}}(\mathbf{x}) = \arg\min_{\mathbf{u}\in\mathcal{V}} \|\mathbf{x} - \mathbf{u}\|^2$$

Show that the unit vector minimizing the mean squared projection error,

$$\arg\min_{\|\mathbf{v}\|_2=1} \sum_{i=1}^{n} \left\|\mathbf{x}_i - f_{\mathbf{v}}(\mathbf{x}_i)\right\|_2^2,$$

is the first principal component of the data.

Solution:
Note that:

$$\begin{aligned}
f_{\mathbf{v}}(\mathbf{x}) &= \arg\min_{\mathbf{u}\in V} \|\mathbf{x} - \mathbf{u}\|^2 \\
&= \arg\min_{\mathbf{u}\in V} (\mathbf{x} - \mathbf{u})^\top (\mathbf{x} - \mathbf{u}) \\
&= \mathbf{v} \cdot \arg\min_{\alpha\in\mathbb{R}} (\mathbf{x} - \alpha\mathbf{v})^\top (\mathbf{x} - \alpha\mathbf{v}) \\
&= \mathbf{v} \cdot \arg\min_{\alpha\in\mathbb{R}} \left(\mathbf{x}^\top\mathbf{x} - 2\alpha\,\mathbf{x}^\top\mathbf{v} + \alpha^2\mathbf{v}^\top\mathbf{v}\right) \\
&= \mathbf{v} \cdot \frac{2\,\mathbf{x}^\top\mathbf{v}}{2\,\mathbf{v}^\top\mathbf{v}} \\
&= \mathbf{v}\,\mathbf{x}^\top\mathbf{v}
\end{aligned}$$

where the second last step is by minimizing a quadratic convex function and using the fact that $\|v\|^2 = 1$. Now,

$$\begin{aligned}
&= \arg\min_{\|v\|=1} \sum_{i=1}^{n} \|x_i - f_v(x_i)\|^2 \\
&= \arg\min_{\|v\|=1} \sum_{i=1}^{n} \left(x_i - v x_i^\top v\right)^\top \left(x_i - v x_i^\top v\right) \\
&= \arg\min_{\|v\|=1} \sum_{i=1}^{n} \left(x_i^\top x_i - 2\left(x_i^\top v\right)^2 + \left(x_i^\top v\right)^2\right) \\
&= \arg\min_{\|v\|=1} \sum_{i=1}^{n} -\left(x_i^\top v\right)^2 \\
&= \arg\max_{\|v\|=1} \sum_{i=1}^{n} v^\top x_i x_i^\top v \\
&= \arg\max_{\|v\|=1} v^\top \left(\sum_{i=1}^{n} x_i x_i^\top\right) v
\end{aligned}$$

The last equation gives the result of the eigenvector corresponding to the largest eigenvalue of

the covariance matrix, and when projecting the original matrix onto this eigenvector, we obtain the first principal component.

2 Density Estimation

2.1 Call Center Counts and Likelihood-based Estimation [5pts]

A call center records the number of customer calls per hour. Over one week you collect data for 70 hours. Typical hourly counts are: 6, 9, 7, 11, ...

1. Suggest an appropriate probability distribution for this data and justify your choice.
2. Find the parameter estimate that makes the observed data most probable under your chosen distribution.

Solution:
Since the data are nonnegative integer counts measured per fixed time unit (hour), a **Poisson distribution** is suitable. It models independent event counts with a constant rate λ.

The likelihood for $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ is

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!},$$

so the log-likelihood is

$$\ell(\lambda) = -n\lambda + \Big(\sum_{i=1}^{n} x_i\Big) \log \lambda - \sum_{i=1}^{n} \log(x_i!).$$

Differentiating and setting to zero:

$$\frac{d\ell}{d\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \quad \Longrightarrow \quad \hat{\lambda} = \bar{X}.$$

Thus, the MLE of the Poisson rate is the sample mean of observed calls per hour.

—

2.2 Pareto Distribution Parameter Estimation [5pts]

The Pareto distribution has been used in economics to model income and wealth distributions with slowly decaying tails. Its density is given by

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

where x_0 is known. Based on n i.i.d. samples $x_1, \ldots, x_n$, find the estimator of θ that maximizes the likelihood function.

Solution:
The log-likelihood is

$$\ell(\theta) = n \log \theta + n\theta \log x_0 - (\theta + 1) \sum_{i=1}^{n} \log x_i.$$

Differentiating:

$$\frac{d\ell}{d\theta} = \frac{n}{\theta} + n \log x_0 - \sum_{i=1}^{n} \log x_i = 0.$$

Solving gives

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log \left(\frac{x_i}{x_0} \right)}.$$

Thus the maximum likelihood estimator of θ depends on the ratio of each sample to x_0.

—

2.3 Parametric vs Nonparametric Density Choice [5pts]

Consider the following two figures showing empirical density plots.

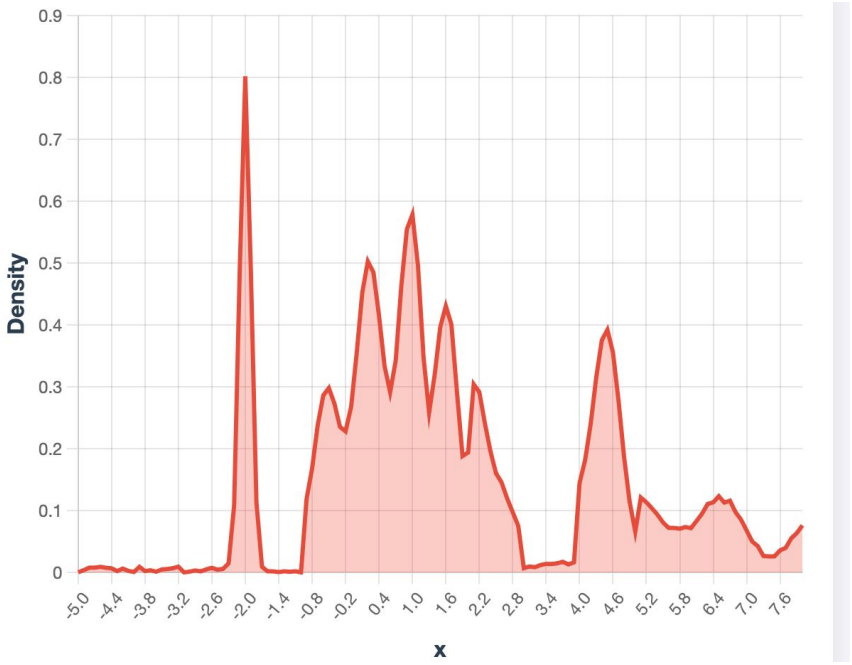

Figure A

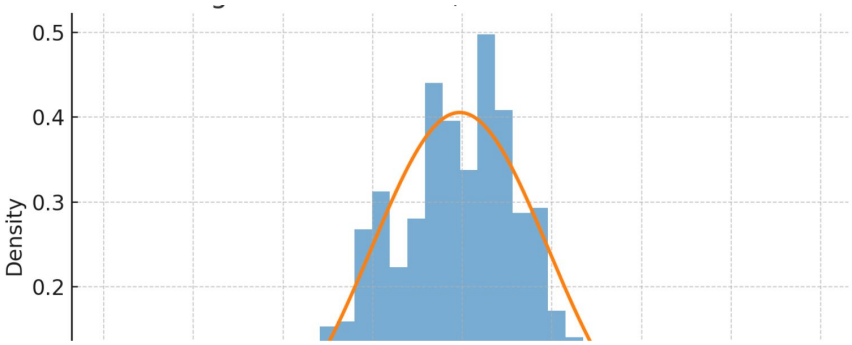

Figure B

1. For Figure A, decide whether a **parametric** or **nonparametric** density estimation method is more appropriate. Justify your answer and suggest a suitable method.

2. Repeat the same question for Figure B.

Solution:
(a) Figure A: The density is multimodal and highly irregular. A simple parametric distribution cannot capture the multiple peaks. Hence, a **nonparametric method** such as Kernel Density Estimation (KDE) with an appropriate bandwidth is more suitable.

(b) Figure B: The density is unimodal, symmetric, and bell-shaped, resembling a Gaussian. Thus, a **parametric model**, specifically the Normal distribution with parameters estimated by MLE, is the appropriate choice.

3 Expectation-Maximization (EM) Algorithm [20 pts]

3.1 Relation to K-means [8pts]

Consider a Gaussian mixture model in $\mathbb{R}^d$ with fixed covariance matrices $\epsilon\mathbf{I}$ for all the mixture components, where $\mathbf{I}$ is the identity matrix and $\epsilon > 0$ is a given constant. The class-conditional density is

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{d/2}} \exp\left(-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right).$$

We now consider the EM algorithm for a mixture of K Gaussians of this form while treating ϵ as a fixed constant instead of a parameter to estimate. Let $\{\pi_k\}_{k=1}^K$ be the mixing proportions and $\{\mathbf{x}^i\}_{i=1}^n$ the data points. The latent variable z_i indicates the Gaussian component to which the data point $\mathbf{x}^i$ belongs.

(a) Write down the posterior probability $\tau_k^i = p(z_i = k|\mathbf{x}^i)$ in this setting.

Solution:

$$\tau_k^i = p(z_i = k|\mathbf{x}^i) = \frac{\pi_k \exp\left(-\frac{1}{2\epsilon}\|\mathbf{x}^i - \boldsymbol{\mu}_k\|^2\right)}{\sum_{k'=1}^K \pi_{k'} \exp\left(-\frac{1}{2\epsilon}\|\mathbf{x}^i - \boldsymbol{\mu}_{k'}\|^2\right)}$$

(b) Show that in the limit $\epsilon \to 0$, maximizing the expected complete-data log likelihood

$$\mathbb{E}_{z_1 \sim p(z_1|\mathbf{x}^1), \ldots, z_n \sim p(z_n|\mathbf{x}^n)} \left[\log \prod_{i=1}^n p(\mathbf{x}^i, z_i|\boldsymbol{\pi}, \boldsymbol{\mu})\right]$$

is equivalent to minimizing the K-means objective $\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}^i - \boldsymbol{\mu}_{c(i)}\right\|^2$, where $c(i) = \arg\min_k \left\|\mathbf{x}^i - \boldsymbol{\mu}_k\right\|^2$.

Solution:
As $\epsilon \to 0$,

$$\tau_k^i = \text{softmax}_k\left(\log \pi_k - \frac{1}{2\epsilon}\|\mathbf{x}^i - \boldsymbol{\mu}_k\|^2\right) \to \mathbf{1}\{k = c(i)\}$$

Therefore,

$$\mathbb{E}_{z_1, z_2, \ldots, z_n}\left[\log \prod_{i=1}^n p(\mathbf{x}^i, z_i|\boldsymbol{\pi}, \boldsymbol{\mu})\right] \to \sum_{i=1}^n \log p\left(\mathbf{x}^i, c(i)|\boldsymbol{\pi}, \boldsymbol{\mu}\right)$$

We have

$$\sum_{i=1}^n \log p\left(\mathbf{x}^i, c(i)|\boldsymbol{\pi}, \boldsymbol{\mu}\right) = \sum_{i=1}^n \log \pi_{c(i)} - \frac{1}{2\epsilon}\sum_{i=1}^n \|\mathbf{x}^i - \boldsymbol{\mu}_{c(i)}\|^2 + \text{const}$$

When $\epsilon \to 0$, the $-\frac{1}{2\epsilon}\sum_{i=1}^n \|\mathbf{x}^i - \boldsymbol{\mu}_{c(i)}\|^2$ term dominates, so maximizing the expected complete-data log-likelihood is equivalent to minimizing the K-means objective $\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}^i - \boldsymbol{\mu}_{c(i)}\right\|^2$.

3.2 Mixture of Exponential Distributions [12pts]

In this question, you will extend the EM algorithm to the mixture of exponential distributions. Suppose we have N i.i.d samples $x_1, x_2, ..., x_N \in [0, \infty)$ drawn from a mixture of K exponential components. Let mixture weights $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K)$ satisfy $\pi_k \geq 0$ and $\sum_{k=1}^K \pi_k = 1$, and component rates $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_K)$ satisfy $\lambda_k > 0$. Given the component k, the likelihood of observing an instance x is $P(x|k) = \lambda_k e^{-\lambda_k x}$.

(a) Write down the log-likelihood $L(\boldsymbol{\pi}, \boldsymbol{\lambda})$ for N observations $x_1, x_2, ..., x_N$.

Solution:
By definition,

$$L(\boldsymbol{\pi}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \lambda_k e^{-\lambda_k x_n} \right).$$

(b) Given $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$, derive the lower bound of the log-likelihood $L(\boldsymbol{\pi}, \boldsymbol{\lambda})$ using Jensen's inequality, and write down the update rule for E-step.

Solution:
Let $z_n \in \{1, 2, ..., K\}$ be the latent variable indicating the component to which x_n belongs.

Using Jensen's inequality,

$$\begin{aligned} L(\boldsymbol{\pi}, \boldsymbol{\lambda}) &= \sum_{n=1}^{N} \log \left(\sum_{z_n=1}^{K} q(z_n) \frac{\pi_{z_n} \lambda_{z_n} e^{-\lambda_{z_n} x_n}}{q(z_n)} \right) \\ &\geq \sum_{n=1}^{N} \sum_{z_n=1}^{K} q(z_n) \log \frac{\pi_{z_n} \lambda_{z_n} e^{-\lambda_{z_n} x_n}}{q(z_n)} \end{aligned}$$

E-step:

To maximize the lower bound above, we should have

$$q(z_n) = P(z_n | x_n; \boldsymbol{\pi}, \boldsymbol{\lambda})$$

Therefore,

$$q(z_n) = \frac{P(x_n, z_n | \boldsymbol{\pi}, \boldsymbol{\lambda})}{P(x_n | \boldsymbol{\pi}, \boldsymbol{\lambda})} = \frac{\pi_{z_n} \lambda_{z_n} e^{-\lambda_{z_n} x_n}}{\sum_{k=1}^{K} \pi_k \lambda_k e^{-\lambda_k x_n}}$$

(c) Write down the M-step which maximizes your lower bound written above. To receive full credits, you should provide the answer step by step.

Solution:

The optimization formulation of the M-step is:

$$\max_{\boldsymbol{\pi},\boldsymbol{\lambda}} \sum_{n=1}^{N}\sum_{z_n=1}^{K} q(z_n)\log\frac{\pi_{z_n}\lambda_{z_n}e^{-\lambda_{z_n}x_n}}{q(z_n)}$$

$$\text{subject to} \quad \sum_{k=1}^{K}\pi_k = 1$$

Dropping terms independent of $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$, and introducing a Lagrange multiplier μ, the Lagrangian is

$$L = \sum_{n=1}^{N}\sum_{z_n=1}^{K} q(z_n)\left(\log\pi_{z_n} + \log\lambda_{z_n} - \lambda_{z_n}x_n\right) + \mu\left(1 - \sum_{k=1}^{K}\pi_k\right)$$

Taking the partial derivative with respect to $\boldsymbol{\pi}$ and set it to zero, we get

$$\pi_j^{\text{new}} = \frac{\sum_{n=1}^{N} q(z_n = j)}{\mu}, j = 1, 2, ...K$$

Since $\sum_{k=1}^{K}\pi_k = 1$, we can get $\mu = N$, and

$$\pi_j^{\text{new}} = \frac{\sum_{n=1}^{N} q(z_n = j)}{N}, j = 1, 2, ..., K$$

Taking the partial derivative with respect to $\boldsymbol{\lambda}$ and set it to zero, we obtain,

$$\lambda_j^{\text{new}} = \frac{\sum_{n=1}^{N} q(z_n = j)}{\sum_{n=1}^{N} q(z_n = j)x_n}, j = 1, 2, ..., K$$

4 Programming

Please use this link to download all the required files. This homework contains only a ipynb, which you can make a copy and run on Google Colab

Deliverables

For the programming part, please submit your `.ipynb` file to the programming autograder. Then, use `File` (top-left corner) → `Print` to generate and submit a PDF.

Expected files

- `HW1.pdf`
- `HW1.ipynb`

- `hw1.ipynb - Colab.pdf`