0 Instructions

Homework is due Tuesday, April 2, 2024 at 23:59pm Central Time. Please refer to `https://courses.grainger.illinois.edu/cs446/sp2024/homework/hw/index.html` for course policy on homeworks and submission instructions.

Reminder: Answers must be typeset. LaTeXand other methods of typesetting math are accepted.

1 PCA: 6pts

1. According to the definition of PCA, the first principal component of w is the one that makes the variance of the projected data Xw the largest. In this case, the direction of w is the line cross both data points (see Figuere 1), i.e. $w = (\frac{3}{5}, \frac{4}{5})^T$.

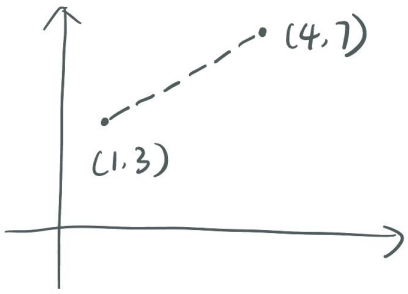

Figure 1: Q 1.1

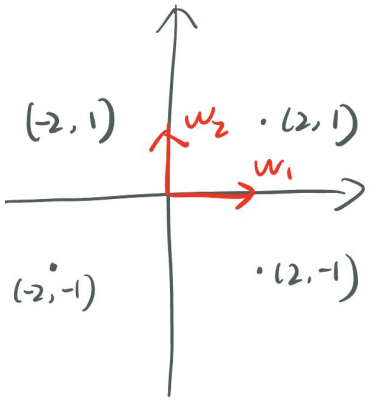

Figure 2: Q 1.2

2.

$$\mu = \frac{1}{4}\sum x_i = (1,4)^T$$

Then, we get the centralized data (as shown in Figure 2).

$$X = \begin{bmatrix} -2 & -2 & 2 & 2 \\ -1 & 1 & -1 & 1 \end{bmatrix}$$

The covariance matrix is,

$$\Sigma = \frac{1}{3}XX^T = \begin{bmatrix} \frac{16}{3} & 0 \\ 0 & \frac{4}{3} \end{bmatrix}$$

We know that the first and second principal components are corresponding to the first and second largest eigenvalues of Σ.

Because Σ is diagonal, the largest eigenvalue is $\frac{16}{3}$ and the corresponding eigenvector is $(1,0)^T$.

The second largest eigenvalue is $\frac{4}{3}$ and the corresponding eigenvector is $(0,1)^T$.

Therefore, the first and second principal components are $(1,0)^T$ and $(0,1)^T$ respectively.

3. $w^T\Sigma w$ is in a quadratic form. Because Σ is diagonal, the quadratic form only contains squared terms. Under the constraint $w^Tw = 1$, we have,

$$\begin{aligned} w^T\Sigma w &= 12w_1^2 + 6w_2^2 + 20w_3^2 + 10w_4^2 \\ &\leq 20 \end{aligned} \tag{1}$$

$w^T\Sigma w = 20$ only when $w = [0,0,1,0]^T$. Therefore, the optimal w is $[0,0,1,0]^T$.

2 Basics in Information Theory: 7pts

1.

$$\begin{aligned} Pr(X' = x) &= Pr(X' = x|B = 1)Pr(B = 1) + Pr(X' = x|B = 0)Pr(B = 0) \\ &= Pr(X = x)\lambda + Pr(X = x)(1 - \lambda) \end{aligned} \tag{2}$$

2.

$$\begin{aligned}
D_\lambda(P||Q) &= \lambda D_{KL}(P||\lambda P + (1-\lambda)Q) + (1-\lambda) D_{KL}(Q||\lambda P + (1-\lambda)Q) \\
&= \lambda \int p(x) \log \frac{p(x)}{\lambda p(x) + (1-\lambda)q(x)}\, dx + (1-\lambda) \int q(x) \log \frac{q(x)}{\lambda p(x) + (1-\lambda)q(x)}\, dx \\
&= \lambda \int p(x)(\log p(x) - \log(\lambda p(x) + (1-\lambda)q(x)))\, dx \\
&\quad + (1-\lambda) \int q(x)(\log q(x) - \log(\lambda p(x) + (1-\lambda)q(x)))\, dx \\
&= -\int (\lambda p(x) + (1-\lambda)q(x)) \log(\lambda p(x) + (1-\lambda)q(x))\, dx \\
&\quad + \lambda \int p(x) \log p(x)\, dx + (1-\lambda) \int q(x) \log q(x)\, dx \\
&= H(X') - \lambda H(X'|B=1) - (1-\lambda) H(X'|B=0) \\
&= H(X') - H(X'|B) \\
&= I(X'; B)
\end{aligned}$$

3 k-Means with Soft Assignments: 10pts

1.

$$\min_{\mu_1,\ldots,\mu_K} \min_{\substack{A \in [0,1]^{n\times K} \\ A \cdot \mathbf{1}_K = \mathbf{1}_n}} \sum_{i=1}^{n} \sum_{k=1}^{K} A_{ik} \|x^{(i)} - \mu_k\|_2^2 \leq \min_{\mu_1,\ldots,\mu_K} \min_{\substack{A \in \{0,1\}^{n\times K} \\ A \cdot \mathbf{1}_K = \mathbf{1}_n}} \sum_{i=1}^{n} \sum_{k=1}^{K} A_{ik} \|x^{(i)} - \mu_k\|_2^2$$

Because the right side is minimizing over a subset of feasible A of right side. Therefore, for every optimal $A*$ on the right side, we can always find the same $A*$ on the left side. Therefore, the left side is less than or equal to the right side.

2.

$$\min_{\mu_1,\ldots,\mu_K} \min_{\substack{A \in [0,1]^{n\times K} \\ A \cdot \mathbf{1}_K = \mathbf{1}_n}} \sum_{i=1}^{n} \sum_{k=1}^{K} A_{ik} \|x^{(i)} - \mu_k\|_2^2 \geq \min_{\mu_1,\ldots,\mu_K} \min_{\substack{A \in \{0,1\}^{n\times K} \\ A \cdot \mathbf{1}_K = \mathbf{1}_n}} \sum_{i=1}^{n} \sum_{k=1}^{K} A_{ik} \|x^{(i)} - \mu_k\|_2^2$$

Suppose mu_{l^*} is the solution of mu_l to $\|x^{(i)} - \mu_k\|_2^2 \geq \min_l \|x^{(i)} - \mu_l\|_2^2$. Then for any

$A \in [0,1]^{n \times K}$, we have,

$$\begin{aligned}
\sum_{k=1}^{K} A_{ik}||x^{(i)} - \mu_k||_2^2 &\geq \sum_{k=1}^{K} A_{ik}||x^{(i)} - \mu_{l^*}||_2^2 \\
&= (\sum_{k=1}^{K} A_{ik})||x^{(i)} - \mu_{l^*}||_2^2 \\
&= ||x^{(i)} - \mu_{l^*}||_2^2 \\
&= \min_{\substack{A \in \{0,1\}^{n \times K} \\ A \cdot \mathbf{1}_K = \mathbf{1}_n}} \sum_{k=1}^{K} A_{ik}||x^{(i)} - \mu_k||_2^2
\end{aligned}$$

Thus, for the optimal A that makes the left side minimum, the above inequality still holds, i.e.,

$$\min_{\substack{A \in [0,1]^{n \times K} \\ A \cdot \mathbf{1}_K = \mathbf{1}_n}} \sum_{k=1}^{K} A_{ik}||x^{(i)} - \mu_k||_2^2 \geq \min_{\substack{A \in \{0,1\}^{n \times K} \\ A \cdot \mathbf{1}_K = \mathbf{1}_n}} \sum_{k=1}^{K} A_{ik}||x^{(i)} - \mu_k||_2^2$$

Thus, the original inequality also holds.

3. From the above two question, we know that the left side and the right side are equal. In other words, the optimal solution to soft assignment problem is equal to the optimal solution to hard assignment problem.

4 Bernoulli Mixture Model: 18pts

Suppose the k-th element of z_i is 1, then the likelihood is,

1.

$$\begin{aligned}
\Pr(x^{(i)}, z_i | \pi, \mu) &= \Pr(x^{(i)} | z_i k = 1) \Pr(z_i k = 1) \\
&= \Pr(x^{(i)} | \mu_k) \pi_k \\
&= \pi_k \prod_{j=1}^{d} \mu_k^{x_j^{(i)}} (1 - \mu_k)^{1 - x_j^{(i)}}
\end{aligned} \tag{3}$$

Thus, the log-likelihood is,

$$\begin{aligned}\log \Pr(x^{(i)}, z_i|\pi, \mu) &= \log \pi_k + \sum_{j=1}^{d} \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}} \\ &= \log \pi_k + \sum_{j=1}^{d} x_j^{(i)} \log \mu_k + \sum_{j=1}^{d} (1-x_j^{(i)}) \log(1-\mu_k)\end{aligned}$$

2.

$$\begin{aligned}Pr(z_{ik}=1|x^{(i)}) &= \frac{Pr(x^{(i)}|z_{ik}=1)Pr(z_{ik}=1)}{\sum_{k=1}^{K} Pr(x^{(i)}|z_i)Pr(z_{ik}=1)} \\ &= \frac{\pi_k \prod_{j=1}^{d} \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}}{\sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}}\end{aligned}$$

3.

$$\begin{aligned}\min -\log \prod_{i \in D} p(x^i|\mu, \pi) &= \min -\sum_{i \in D} \log p(x^i|\mu, \pi) \\ &= \min -\sum_{i \in D} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}\end{aligned}$$

Take the derivative with respect to μ_k and π_k , we have,

$$\begin{aligned}
\frac{d}{d\mu_k} &= -\sum_{i\in D} \frac{(\pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}} (1-\mu_k)^{(1-x_j^{(i)})})'}{\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}} \\
&= -\sum_{i\in D} \frac{(\pi_k \mu_k^{\sum_{j=1}^d x_j^{(i)}} (1-\mu_k)^{\sum_{j=1}^d (1-x_j^{(i)})})'}{\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}} \\
&= -\sum_{i\in D} \frac{}{\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}} \\
&= -\sum_{i\in D} \frac{\pi_k (\sum_{j=1}^d x_j^{(i)}) \mu_k^{(\sum_{j=1}^d x_j^{(i)})-1} (1-\mu_k)^{\sum_{j=1}^d (1-x_j^{(i)})}}{\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}} \\
&\quad -\sum_{i\in D} \frac{\pi_k \mu_k^{\sum_{j=1}^d x_j^{(i)}} (\sum_{j=1}^d (1-x_j^{(i)})) (1-\mu_k)^{\sum_{j=1}^d (1-x_j^{(i)})-1}}{\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}} \\
&= -\sum_{i\in D} \frac{\pi_k \mu_k^{(\sum_{j=1}^d x_j^{(i)})-1} (1-\mu_k)^{\sum_{j=1}^d (1-x_j^{(i)})-1}}{\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}} (1-\mu_k)^{1-x_j^{(i)}}} ((1-\mu_k) \sum_{j=1}^d x_j^{(i)} + \mu_k \sum_{j=1}^d (1-x_j^{(i)})) \\
&= -\sum_{i\in D} \frac{r_{ik}}{\mu_k(1-\mu_k)} ((1-\mu_k) \sum_{j=1}^d x_j^{(i)} + \mu_k \sum_{j=1}^d (1-x_j^{(i)}))
\end{aligned}$$

set the derivative to 0, we have,

$$\begin{aligned}
&-\sum_{i\in D} r_{ik} ((1-\mu_k) \sum_{j=1}^d x_j^{(i)} + \mu_k \sum_{j=1}^d (1-x_j^{(i)})) = 0 \\
&-\sum_{i\in D} r_{ik} (\sum_{j=1}^d (1-2x_j)\mu_k + \sum_{j=1}^d x_j) = 0 \\
&\mu_k = \frac{\sum_{i\in D} r_{ik} \sum_{j=1}^d x_j}{\sum_{i\in D} r_{ik} \sum_{j=1}^d (2x_j - 1)}
\end{aligned}$$

For π_k, we add a Lagrange multiplier to the objective function, and take derivative

with respect to π_k, we have,

$$\begin{aligned}
&\sum_{i\in D} \frac{\mu_k^{\sum_{j=1}^d x_j^{(i)}}(1-\mu_k)^{\sum_{j=1}^d (1-x_j^{(i)})}}{\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}}(1-\mu_k)^{1-x_j^{(i)}}} + \lambda = 0 \\
&\sum_{i\in D} \frac{\pi_k\mu_k^{\sum_{j=1}^d x_j^{(i)}}(1-\mu_k)^{\sum_{j=1}^d (1-x_j^{(i)})}}{\sum_{k=1}^K \pi_k \prod_{j=1}^d \mu_k^{x_j^{(i)}}(1-\mu_k)^{1-x_j^{(i)}}} + \lambda\pi_k = 0 \\
\Rightarrow &\sum_{i\in D} r_{ik} + \lambda\pi_k = 0 \qquad (4) \\
\Rightarrow &\sum_{k=1}^K \sum_{i\in D} r_{ik} + \lambda = 0 \\
\Rightarrow &\lambda = -\sum_{k=1}^K \sum_{i\in D} r_{ik} = -N \qquad (5)
\end{aligned}$$

Take $r_{ik} = -N$ back to $\sum_{i\in D} r_{ik} + \lambda\pi_k = 0$, we get $\pi_k = \frac{\sum_{i\in D} r_{ik}}{N}$.

5 Variational Autoencoder (VAE): 19pts

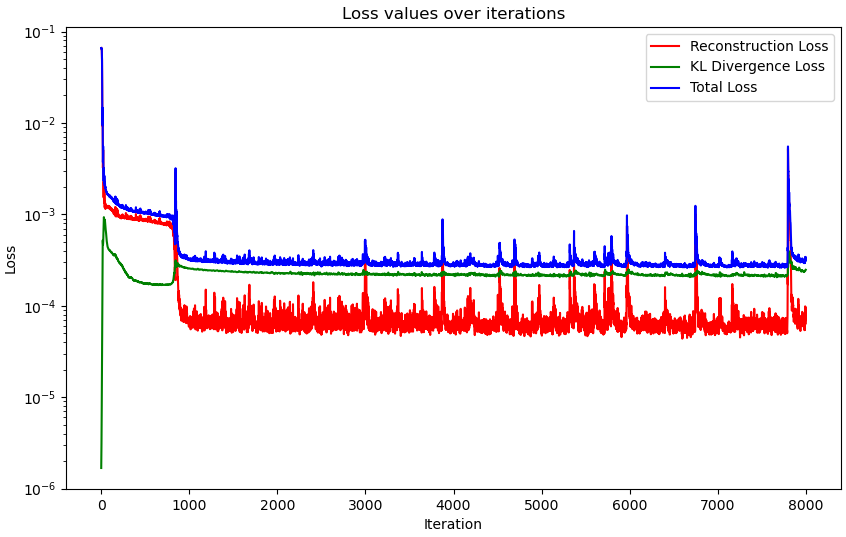

Figure 3: Empirical risk

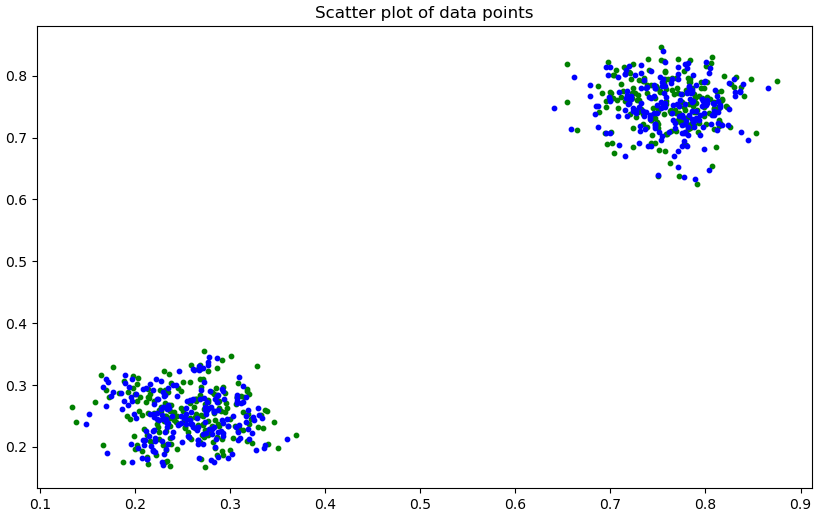

Figure 4: Training samples and its reconstruction

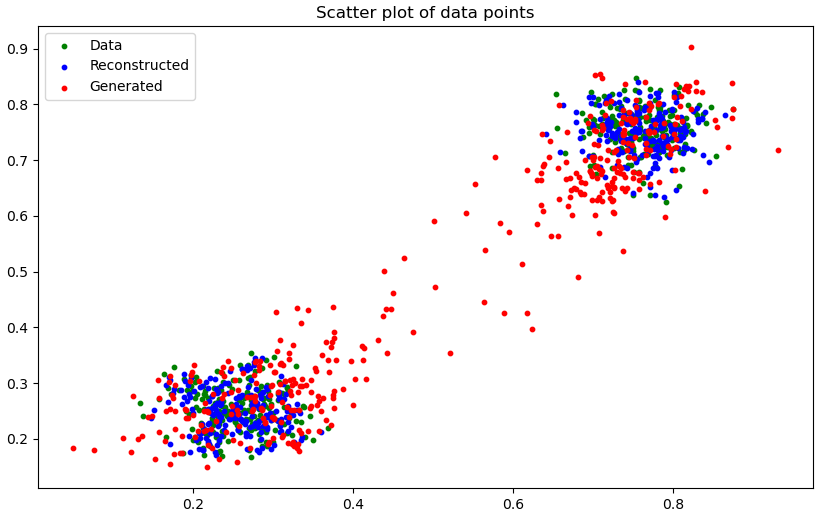

Figure 5: Training, reconstructed and generated samples