Name: Jianning Chen
NUID: 002952943
E-mail: chen.jiann@northeastern.edu

Homework 3
EECE 5644

2022-11-17

1 Question 1

The pdf of a 2-dimensional real-valued pdf is given by:

$$p(\mathbf{x}) = p(\mathbf{x}|L=0)P(L=0) + p(\mathbf{x}|L=1)P(L=1)$$

Where $p(L=0) = 0.6$ and $p(L=1) = 0.4$ are priori probabilities and it is known that:

$$p(\mathbf{x}|L=0) = w_1 g(\mathbf{x}|\mathbf{m}_{01}, \mathbf{C}_{01}) + w_2 g(\mathbf{x}|\mathbf{m}_{02}, \mathbf{C}_{02})$$

$$p(\mathbf{x}|L=1) = g(\mathbf{x}|\mathbf{m}_1, \mathbf{C}_1)$$

It is also given that $w_1 = w_2 = 0.5$, and

$$\mathbf{m}_{01} = \begin{bmatrix} 5 \\ 0 \end{bmatrix}, \mathbf{C}_{01} = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}, \mathbf{m}_{02} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \mathbf{C}_{02} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}, \mathbf{m}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \mathbf{C}_{01} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Using the information known, 4 datasets can be generated with 100 samples, 1000 samples, 10000 samples and 20000 samples:

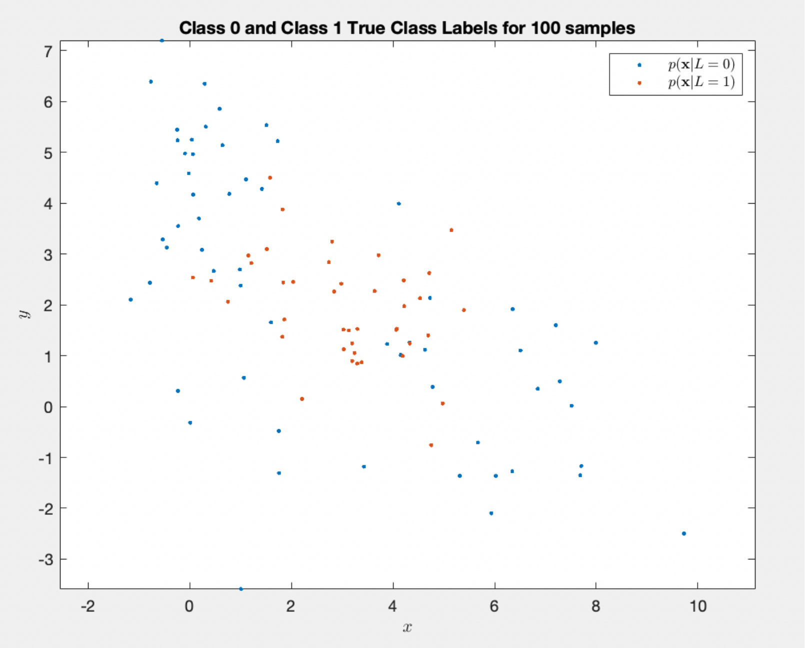

Figure 1: 100 random samples of the pdf ($\mathcal{D}^{100}_{train}$)

1.1 Part 1

As shown in *Figure 4*, the samples cannot be separated with a linear classifier, therefore, the most optimal classifier would be a non-linear classifier. Since class 2(the red dots) are surrounded by class 1(the blue dots), therefore, using dimensionality reduction and kernel PCA is the optimal approach for this question.

Referring to Figure 5, since the sample data is not linearly separable, the kernel trick I pick is rbf kernel which has a math formula below:

$$\mathcal{K}(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(\frac{-||\mathbf{x}_1 - \mathbf{x}_2||^2}{2\sigma^2}\right)$$

$$\left(\frac{1}{n}\mathbf{I} - \frac{1}{n^2}\mathbf{1}_{n\times n}\right)\mathcal{K}\tilde{\mathbf{u}}_j = \tilde{\lambda}_j \tilde{\mathbf{u}}_j$$

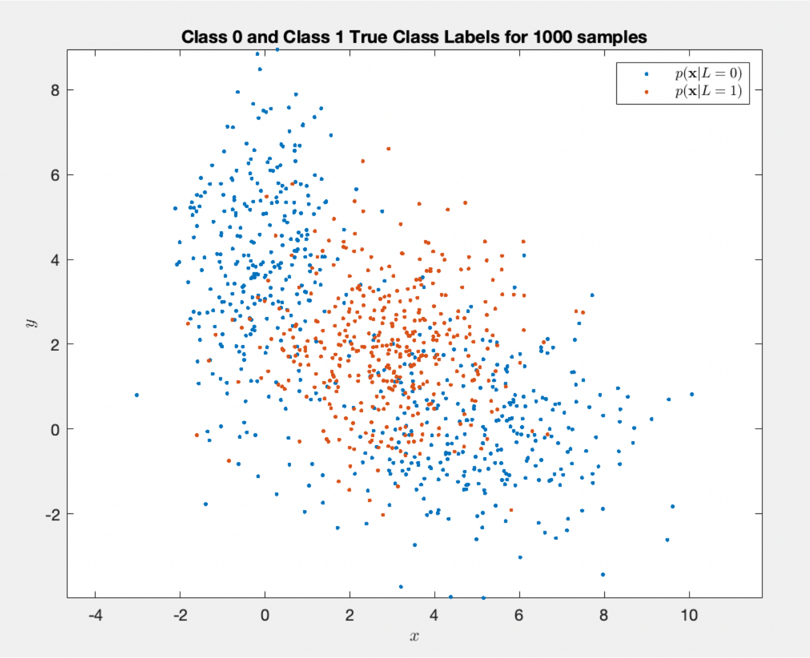

Figure 2: 1000 random samples of the pdf ($\mathcal{D}^{1000}_{train}$)

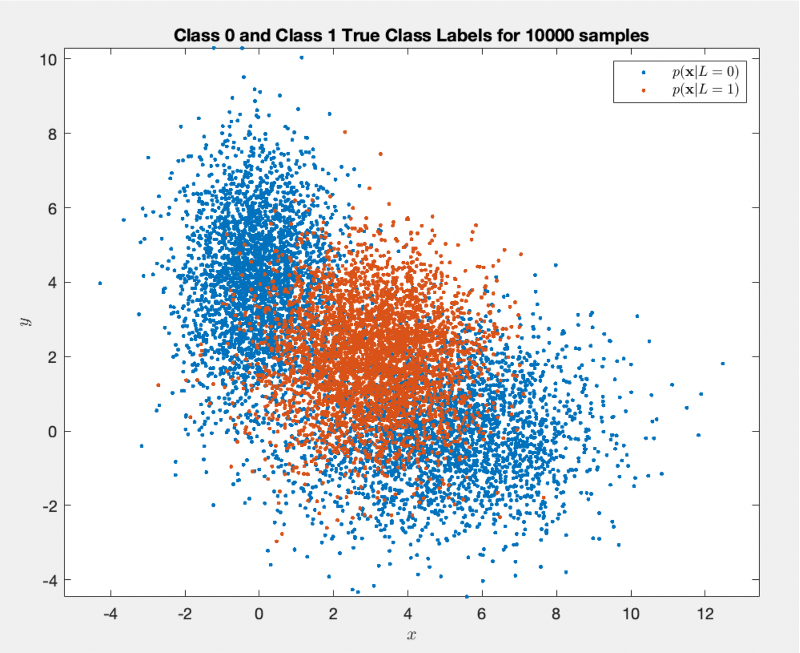

Figure 3: 10000 random samples of the pdf ($\mathcal{D}^{10000}_{train}$)

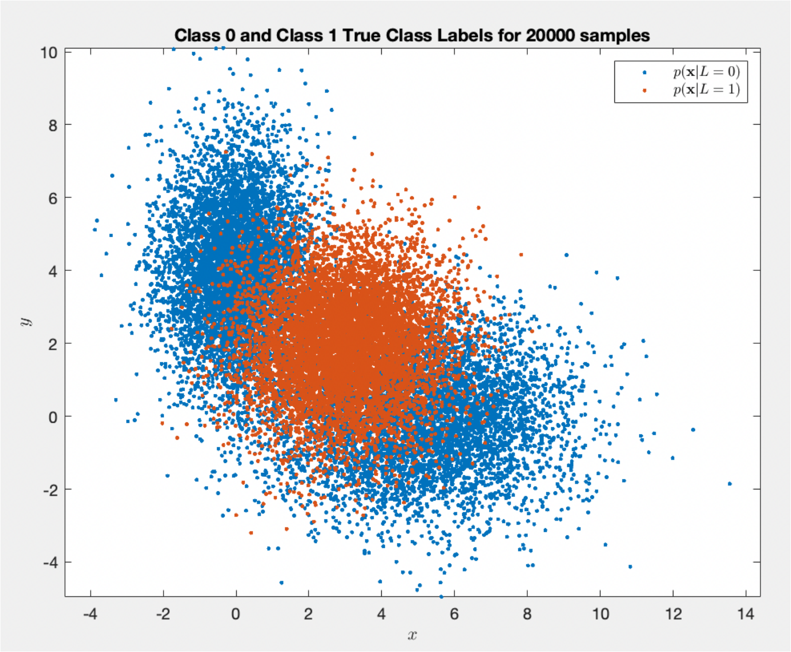

Figure 4: 20000 random samples of the pdf ($\mathcal{D}^{20000}_{validate}$)

$$\tilde{\mathbf{e}}_j = \frac{1}{\sqrt{\tilde{\mathbf{u}}_j^T \mathcal{K} \tilde{\mathbf{u}}_j}} \mathbf{X}^T \tilde{\mathbf{u}}_j$$

$$\tilde{\mathbf{A}}_p = [\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \tilde{\mathbf{e}}_3, ...] = \mathbf{X}^T \mathbf{U}_p$$

Here we have to find $\mathbf{U}_1$ and $\mathbf{U}_2$ in order to find the kernel PCA.

2 Question 2

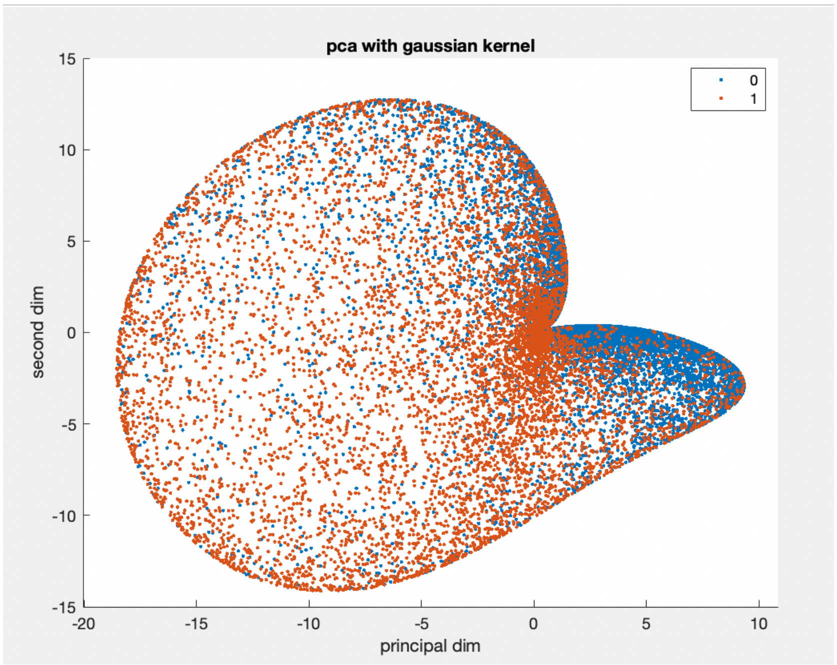

Figure 5: The sample distribution after applying kernel PCA of $\mathcal{D}^{20000}_{validate}$

3 Question 3

From the question statement we know for loss function $\lambda(\alpha_i|\omega_j)$ is:

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \qquad i, j = 1, ..., c \\ \lambda_r & i = c + 1 \\ \lambda_s & \text{otherwise} \end{cases}$$

The conditional risk function is given by:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$$

Using the zero-one loss function, $\lambda(\alpha_i|\omega_j) = 0$ when $i = j$ and $\lambda(\alpha_i|\omega_j) = 1$ when $i \neq j$, we get:

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \\ &= \sum_{j \neq i} P(\omega_j|\mathbf{x}) \\ &= 1 - P(\omega_i|\mathbf{x}) \end{aligned}$$

Using the same calculation method, we have $R(\alpha_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$. In order to accept ω_i, $R(\alpha_i|\mathbf{x}) \leq R(\alpha_j|\mathbf{x})$, and that gives equation:

$$\begin{aligned} 1 - P(\omega_i|\mathbf{x}) &\leq 1 - P(\omega_j|\mathbf{x}) \\ P(\omega_i|\mathbf{x}) &\geq P(\omega_i|\mathbf{x}) \end{aligned}$$

For $j = 1, 2, ..., c$, $R(\alpha_i|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$, and for $i = 1, 2, ..., c$, the loss function $\lambda(\alpha_i|\omega_j) = \lambda_s$ when $i \neq j$. Which means that $R(\alpha_i|\mathbf{x}) = \lambda_s(1 - P(\omega_i|\mathbf{x}))$. For $i = c + 1$, $R(\alpha_i|\mathbf{x}) = R(\alpha_{c+1}|\mathbf{x}) = \lambda_r$. In order to decide ω_i, $R(\alpha_i|\mathbf{x})$ has to be the smallest in all circumstances, so we have $R(\alpha_i|\mathbf{x}) \leq R(\alpha_{c+1}|\mathbf{x})$, that gives equation:

$$\begin{aligned} \lambda_s(1 - P(\omega_i|\mathbf{x})) &\leq \lambda_r \\ 1 - P(\omega_i|\mathbf{x}) &\leq \frac{\lambda_r}{\lambda_s} \\ P(\omega_i|\mathbf{x}) &\geq 1 - \frac{\lambda_r}{\lambda_s} \end{aligned}$$

When $\lambda_r = 0$, $P(\omega_i|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$ will always be false unless $P(\omega_i|\mathbf{x}) = 1$, and in this case ω_i will always be rejected except when $P(\omega_j|\mathbf{x}) = 0$. When $\lambda_r > \lambda_s$, $P(\omega_i|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$ will always be true, that means ω_i will always be accepted.