Yinhong Qin
yq2021@nyu.edu

Homework 3

ECE 6143A-Machine Learning
October 21, 2021

Problem 1

The mean of all teapot images and the top 3 eigenvectors of its covariance matrix are listed below:

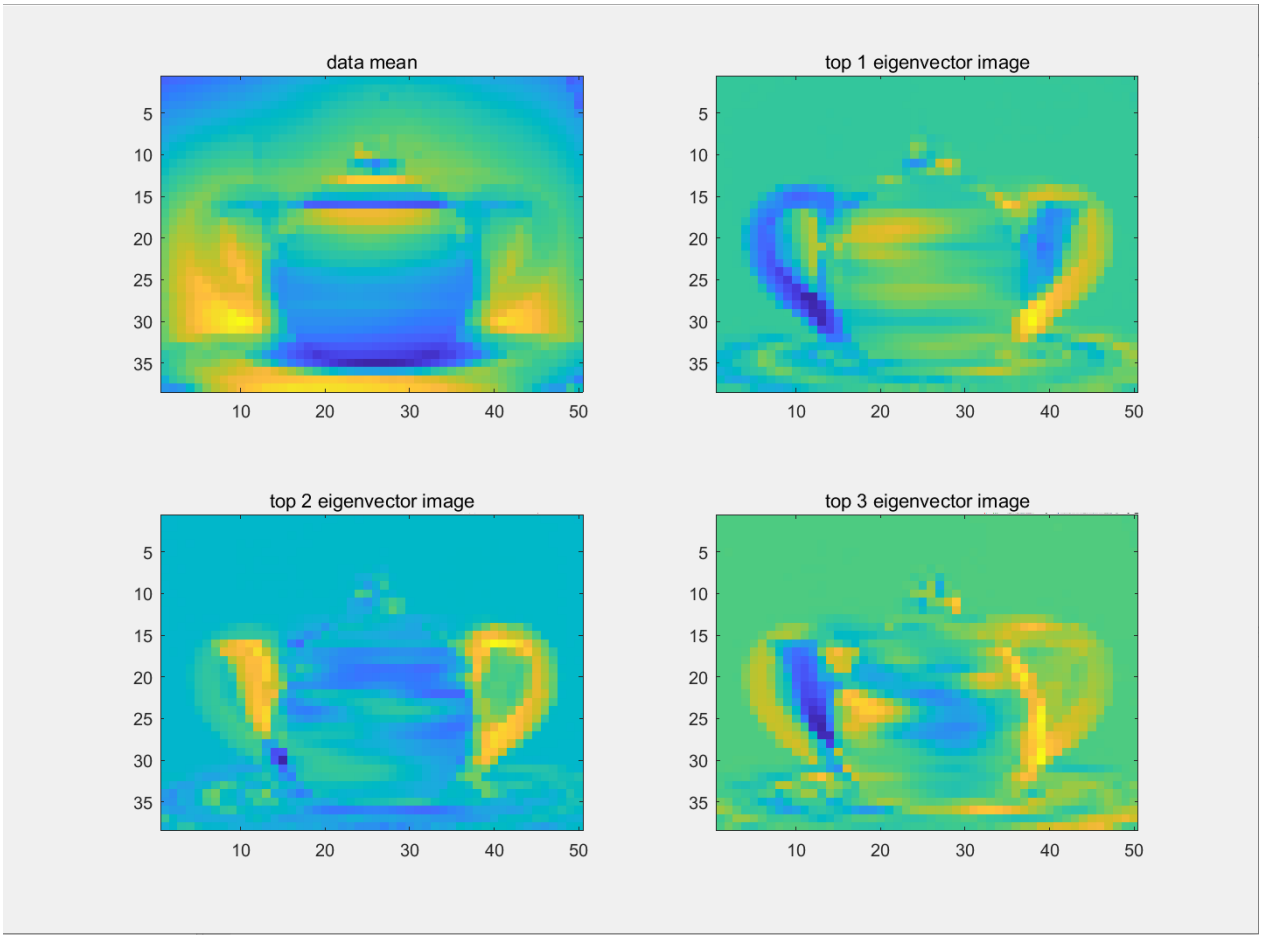

Figure 1: Mean and top 3 eigenvectors images.

The random picked 10 images before and after reconstruction are listed below:

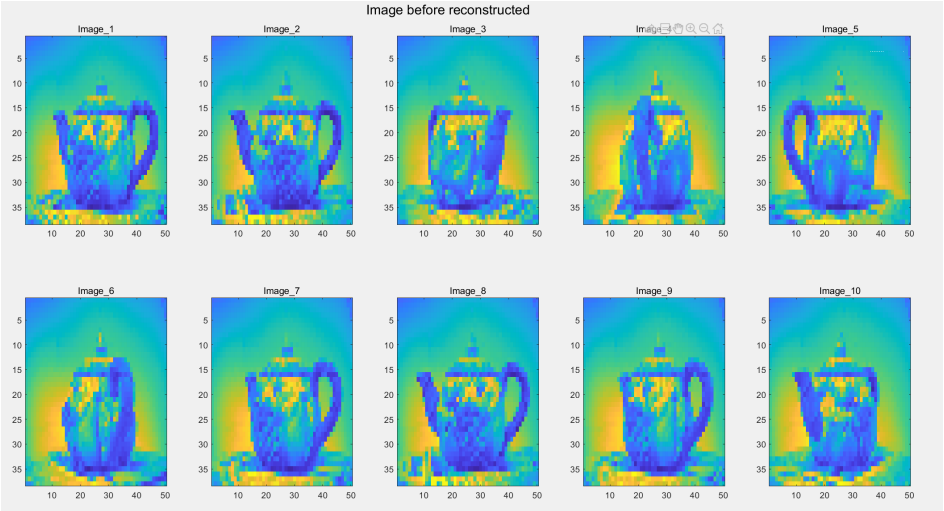

Figure 2: 10 images before reconstruction.

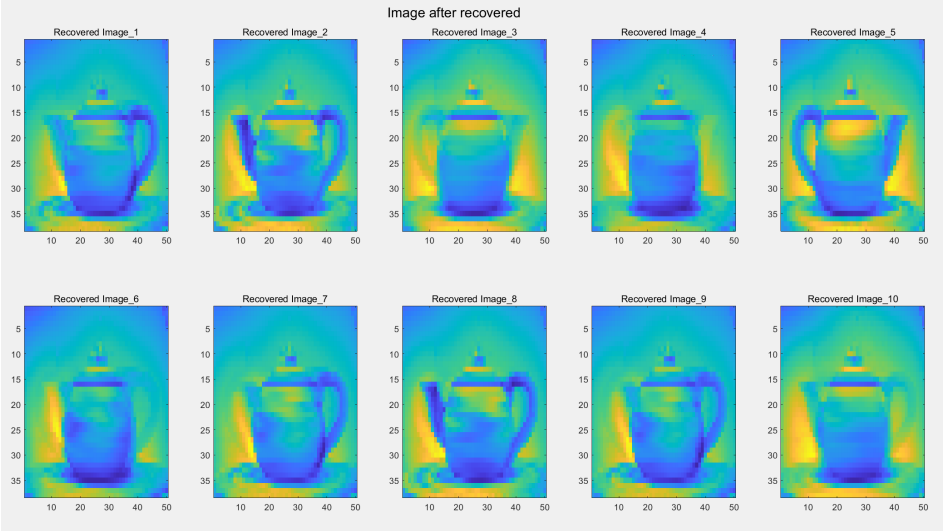

Figure 3: 10 images before reconstruction.

From results listed above, we can see that the mean of the dataset gives us a general view of the teapot the top 3 eigenvectors provide the most important and fundamental features of the teapot like the handle and the spout. We can also notice that the reconstruction with the data mean and top 3 eigenvectors is roughly similar to the original image. They are not completely the same because the number of eigenvectors is limited so the reconstruction cannot reproduce all the details.

Problem 2

Let the box containing 8 apples and 4 oranges be "box A". Thus another box containing 10 apples and 2 oranges becomes "box B".

Consider $P(A)$ is the probability of the event "Choose the box A". It is obvious that

$$P(A) = \frac{1}{2}$$

$P(B)$ is the probability of the event "Choose a apple from either box A or B". Thus, we can have:

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \\ &= \frac{2}{3} \times \frac{1}{2} + \frac{5}{6} \times \frac{1}{2} \\ &= \frac{9}{12} = \frac{3}{4} \end{aligned}$$

The probability of the event "Choose an apple from box A" is $P(A|B)$. To compute this, we can use the Bayes' theorem:

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{\frac{2}{3} \times \frac{1}{2}}{\frac{3}{4}} \\ &= \frac{4}{9} \end{aligned}$$

Problem 3

Consider n data examples $X = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^d, y \in \{0, 1\}$. Then we can have,

$$\begin{aligned} p(\mathbf{x}_i, y_i; \boldsymbol{\theta}) &= p(y_i; \theta) p(\mathbf{x}_i; y_i, \boldsymbol{\theta}) \\ &= (\alpha^{y_i}(1-\alpha)^{1-y_i}) N(\mathbf{x}_i; \boldsymbol{\mu}_{y_i}, \boldsymbol{\Sigma}_{y_i}) \end{aligned}$$

To recover all the parameters $\theta = \{\alpha, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\}$ over the dataset $X = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$ using Maximum Likelihood Estimation, we can have:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log\, p(X; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log\, p(\mathbf{x}_i, y_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log\, [p(y_i; \boldsymbol{\theta}) p(\mathbf{x}_i; y_i, \boldsymbol{\theta})] \\ &= \sum_{i=1}^{n} \log\, p(y_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \log\, p(\mathbf{x}_i; y_i, \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log\, p(y_i; \alpha) + \sum_{y_i \in 0} \log\, p(\mathbf{x}_i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \sum_{y_i \in 1} \log\, p(\mathbf{x}_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \end{aligned}$$

Then we can have:

$$\begin{aligned} \frac{\partial \sum_{i=1}^{n} \log\, p(y_i; \alpha)}{\partial \alpha} &= \sum_{i=1}^{n} \frac{\partial \log\, p(y_i; \alpha)}{\partial \alpha} \\ &= \sum_{i=1}^{n} \frac{\partial \log\, \alpha^{y_i}(1-\alpha)^{1-y_i}}{\partial \alpha} \\ &= \sum_{i=1}^{n} (\frac{y_i}{\alpha} - \frac{1-y_i}{1-\alpha}) = 0 \end{aligned}$$

Let N_0 be the number of data examples whose $y_i = 0$, and N_1 be the number of data examples whose $y_i = 1$. Therefore, $\alpha = \frac{N_1}{N_0 + N_1}$.

For the remained part in $l(\theta)$, we can first compute how to use MLE to recover the parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate Gaussian $N(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Since

$$N(x; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp(\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

thus the log likelihood function is

$$\begin{aligned} \log N(x; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \\ &= -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + Const \end{aligned}$$

First we compute the estimate of $\boldsymbol{\mu}$. Consider

$$\begin{aligned} \frac{\partial \log N(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} &= \frac{\partial -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} \\ &= \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{0} \end{aligned}$$

Therefore, for the given dataset X, we can have

$$\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{0}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

Then, to compute the estimate of $\boldsymbol{\Sigma}$ we need first provide some pre-requisites.

Lemma 1. *For a matrix $A_{n\times n}$,*

$$\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial A} = \mathbf{x}\mathbf{x}^T$$

Proof. $\mathbf{x}^T A \mathbf{x}$ is a scalar that

$$\mathbf{x}^T A \mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} x_i x_j$$

Consider the element-wise partial derivative

$$\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial A_{ij}} = x_i x_j$$

Therefore,

$$\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial A} = \mathbf{x}\mathbf{x}^T$$

□

Lemma 2. *For a invertible square matrix $A_{n\times n}$,*

$$\frac{\partial \log(\det(A))}{\partial A} = A^{-T}$$

Proof.

$$\frac{\partial \log(\det(A))}{\partial A} = \frac{1}{\det(A)}\frac{\partial \det(A)}{\partial A}$$

Since

$$adj(A) = \det(A)A^{-1}$$

and we also consider the element-wise partial derivative

$$\frac{\partial \det(A)}{\partial A_{ij}} = adj(A)_{ji}$$

This is because the $adj(A)_{ij}$ has no relation with A_{ij} by the definition of adjugate matrix. Finally, we can have

$$\begin{aligned}\frac{\partial \log(\det(A))}{\partial A_{ij}} &= \frac{1}{\det(A)}\frac{\partial \det(A)}{\partial A} \\ &= \frac{1}{\det(A)}adj(A)_{ji} \\ &= A^{-1}_{ji}\end{aligned}$$

Therefore,

$$\frac{\partial \log(\det(A))}{\partial A} = A^{-T}$$

□

Consider

$$\frac{\partial \log N(x;\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}} = \frac{\partial(-\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))}{\partial \boldsymbol{\Sigma}}$$

With lemmas proved above, we can know that

$$\frac{\partial(-\frac{1}{2}\log|\boldsymbol{\Sigma}|)}{\partial \boldsymbol{\Sigma}} = \frac{\partial(\frac{1}{2}|\boldsymbol{\Sigma}^{-1}|)}{\partial \boldsymbol{\Sigma}} = \frac{1}{2}\boldsymbol{\Sigma}^T = \frac{1}{2}\boldsymbol{\Sigma}$$

and

$$\frac{\partial(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T$$

Thus, over the whole dataset we can have

$$\sum_{i=1}^{n}(\frac{1}{2}\mathbf{\Sigma} - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T) = 0$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T)$$

Finally, we can have that

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{N_0}\sum_{y_i=0}\mathbf{x}_i,\ \hat{\mathbf{\Sigma}}_1 = \frac{1}{N_0}\sum_{y_i=0}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)^T)$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{1}{N_1}\sum_{y_i=1}\mathbf{x}_i,\ \hat{\mathbf{\Sigma}}_2 = \frac{1}{N_1}\sum_{y_i=1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_2)^T)$$

For the decision boundary, we can consider

$$q(\mathbf{x}) = \frac{p(y=1;\mathbf{x})}{p(y=0;\mathbf{x})}$$

when $q(\mathbf{x}) > 1$, it indicates that sample $\mathbf{x}$ has more likely to be class 1.
Use Bayes theorem and take the log

$$\begin{aligned} q(\mathbf{x}) &= \frac{p(y=1;\mathbf{x})}{p(y=0;\mathbf{x})} \\ &= \frac{\frac{p(\mathbf{x};y=1)p(y=1)}{p(\mathbf{x})}}{\frac{p(\mathbf{x};y=0)p(y=0)}{p(\mathbf{x})}} \\ &= \frac{p(\mathbf{x};y=1)p(y=1)}{p(\mathbf{x};y=0)p(y=0)} \end{aligned}$$

$$\log q(\mathbf{x}) = \log p(\mathbf{x};y=1) + \log p(y=1) - \log p(\mathbf{x};y=0 - \log p(y=0)$$

$$\log q(\mathbf{x}) \propto -(\mathbf{x}-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + (\mathbf{x}-\boldsymbol{\mu}_2)^T\mathbf{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + Const$$

$$\log q(\mathbf{x}) \propto -(\mathbf{x}^T\mathbf{\Sigma}_1^{-1}\mathbf{x} - 2\boldsymbol{\mu}_1^T\mathbf{\Sigma}_1^{-1}\mathbf{x} + \boldsymbol{\mu}_1^T\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1) + (\mathbf{x}^T\mathbf{\Sigma}_2^{-1}\mathbf{x} - 2\boldsymbol{\mu}_2^T\mathbf{\Sigma}_2^{-1}\mathbf{x} + \boldsymbol{\mu}_2^T\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2) + Const$$

Clearly, this form above is quadratic when $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, and is linear when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

Consider the quadratic term above

$$-\mathbf{x}^T\boldsymbol{\Sigma}_1^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}_2^{-1}\mathbf{x} = \mathbf{x}^T(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})\mathbf{x}$$

Thus $(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})$ determines the direction of the quadratic line.

Since $\det(\boldsymbol{\Sigma})$ measures the spread of multivariate Gaussian distribution, it is trivial to see that the decision boundary will directed the class that is more concentrated around its mean.