Machine Learning, 2021 Fall
Assignment 4

Notice

Due 23:59 (CST), Dec. 28, 2021

Plagiarizer will get 0 points.
LaTeX is highly recommended. Otherwise you should write as legibly as possible.

1 Matrix Factorization

Suppose you are invited to design a course recommendation system for ShanghaiTech University. You are given an incomplete matrix $M \in R^{m \times n}$, $M_{i,j}$ is the i-th student's rating for the j-th course. Ω is a real subset of $\{1, 2, \cdots, m\} \times \{1, 2, \cdots, n\}$, $\forall (i, j) \in \Omega$, $M_{i,j}$ is given.($m < n$)

(a)We wish the completion of M has a low rank, please explain why.

(b)You decided to use SVD introduced in the course to solve this problem, we allow negative ratings, assume there are k hidden factors and $k < m$. Give the corresponding optimization problem using the notation M and Ω.

(c)Although we allow negative ratings, extremely high or low ratings are sometimes confusing, how to adjust the optimization problem to avoid extremely high or low ratings?

Solution:
(a)Similar students likes similar courses, and many students are similar, so the rating matrix should be low rank.

(b)

$$\begin{aligned} \min \quad & \sum_{(i,j)\in\Omega} (M_{i,j} - \hat{M}_{i,j})^2 \\ \text{s.t.} \quad & \hat{M} = UV^\top \\ & U \in R^{m\times k} \\ & V \in R^{n\times k} \end{aligned}$$

(c)

$$\begin{aligned} \min \quad & \sum_{(i,j)\in\Omega} (M_{i,j} - \hat{M}_{i,j})^2 + \frac{1}{2}||\hat{M}||_F^2 \\ \text{s.t.} \quad & \hat{M} = UV^\top \\ & U \in R^{m\times k} \\ & V \in R^{n\times k} \end{aligned}$$

2 Dimensionality Reduction

You can run the following code to see the difference of PCA and LDA on the Iris dataset.

```
from sklearn import datasets
from sklearn.decomposition import PCA
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd

iris = datasets.load_iris()
iris_X = iris.data#data
iris_y = iris.target#label

model_pca = PCA(n_components=2)
X_pca = model_pca.fit(iris_X).transform(iris_X)
fig = plt.figure(figsize=(10,8))
plt.scatter(X_pca[:, 0], X_pca[:, 1],marker='o',c=iris_y)
plt.title('PCA to 2D')
plt.show()

model_pca = PCA(n_components=1)
X_pca = model_pca.fit(iris_X).transform(iris_X)
fig = plt.figure(figsize=(10,8))
y = [0]*X_pca.shape[0]
plt.scatter(X_pca[:, 0], y,marker='o',c=iris_y)
plt.title('PCA to 1D')
plt.show()

model_lda = LinearDiscriminantAnalysis(n_components=2)
X_lda = model_lda.fit(iris_X, iris_y).transform(iris_X)
fig = plt.figure(figsize=(10,8))
plt.scatter(X_lda[:, 0], X_lda[:, 1], marker='o',c=iris_y)
plt.title('LDA to 2D')
plt.show()

model_lda = LinearDiscriminantAnalysis(n_components=1)
X_lda = model_lda.fit(iris_X, iris_y).transform(iris_X)
fig = plt.figure(figsize=(10,8))
y = [0]*X_lda.shape[0]
plt.scatter(X_lda[:, 0], y, marker='o',c=iris_y)
plt.title('LDA to 1D')
plt.show()
```

(a)Compare the result of PCA and LDA, tell the difference between PCA and LDA and explain why.
(b)Given the data in the following picture, draw the PCA direction and LDA direction.

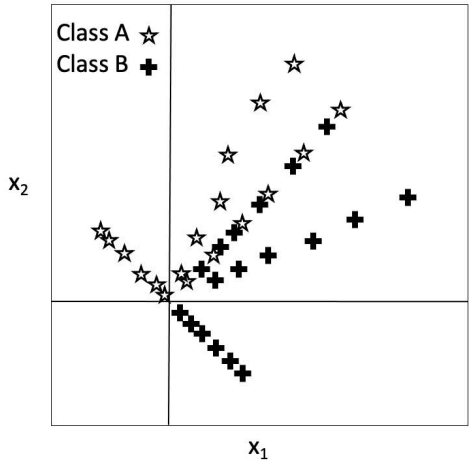

Figure 1: Question(b)

(c)Suppose the origin data has D dimensions,C classes , after applying LDA, what is the range of the dimension of new data? Why?

Solution:
(a):
PCA has more overlapping between classes.
PCA selects the direction on which the projected data have largest variance. Since it is unsupervised, PCA assumes that the larger the variance is, the more information it contains, and using principal components to represent the original data can remove the redundant dimensions and achieve dimensionality reduction. LDA selects the direction of small intra-class variance and large inter-class variance after projection. The category labeling information is used to find the discriminative dimensions in the data, so that the original data are projected in these directions and the different categories are distinguished as much as possible.
(b):

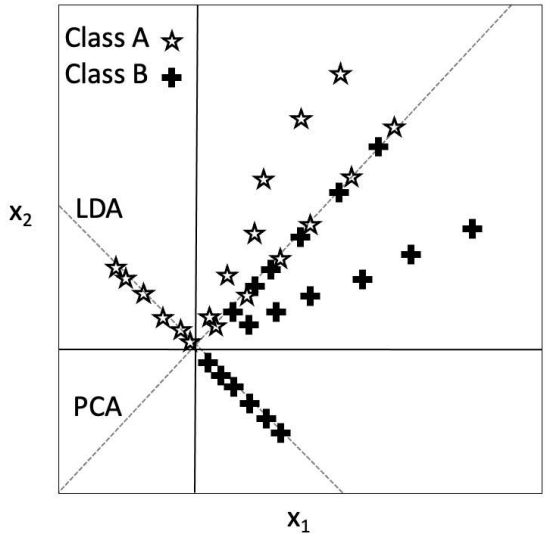

Figure 2: Question(b)

(c):
$\{1, 2, \cdots, C-1\}$
Since $\mu = \frac{1}{n}\sum_{i=1}^{c} n_i \mu_i$ is a linear combination of $\mu_1, \mu_2, \cdots, \mu_C$
$\mu_C - \mu$ can be write as a linear combination of $\mu_1 - \mu, \mu_2 - \mu, \cdots, \mu_{C-1} - \mu$
The rank of $S_B = \sum_{i=1}^{C} n_i(\mu_i - \mu)(\mu_i - \mu)^\top$ is at most $C-1$, thus $S_w^{-1}S_B$ has at most $C-1$ eigen vectors.
So the new data has at most $C-1$ dimensions.