CSCI3320

FUNDAMENTALS OF MACHINE LEARNING TA NAMES: SHUO LIANG, YUE QIU

ASSIGNMENT 4

Deadline: 11:59 pm, April 14, 2024

Submit via Blackboard with VeriGuide receipt.

Please follow the course policy and the school's academic honesty policy.

1. Which of the following statements is true about Fisher's LDA for two classes? Only one choice is correct.

   [a] The goal of LDA is to minimize the distance between classes and to maximize the distance within classes.

   [b] The goal of LDA is to maximize the distance between classes and to minimize the distance within classes.

   [c] The goal of LDA is to minimize the distance between classes and within classes.

   [d] The goal of LDA is to maximize the distance between classes and within classes.

2. Which of the following statements is Wrong about the LDA for two classes? Only one choice is correct.

   [a] Fisher's linear discriminant attempts to find the vectors that maximizes the separation between classes of the projected data.

   [b] LDA takes into account the categories in the data.

   [c] LDA assumes linear decision boundary and variance-covariance homogeneity.

   [d] The desired LDA transformation is in the direction of the within-class covariance matrix.

3. Referring to Week9's Slides Page19, we have learned about Fisher's LDA for two classes. Prove that $\boldsymbol{w}$ reaches the maximum of $J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w}}$ when $\boldsymbol{w}$ satisfies $\mathbf{S}_B \boldsymbol{w} = \lambda \mathbf{S}_W \boldsymbol{w}$, where $\lambda = \frac{\boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w}}$.

   Hint: For clarity, we first define $f(\boldsymbol{w}) = \boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w}$ and $g(\boldsymbol{w}) = \boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w}$.
   Recall that $\frac{\partial}{\partial x} \frac{f(x)}{g(x)} = \frac{f'g - fg'}{g^2}$, where $f' = \frac{\partial}{\partial x} f(x)$ and $g' = \frac{\partial}{\partial x} g(x)$. Recall that $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \mathbf{A} \boldsymbol{x} = (\mathbf{A} + \mathbf{A}^T)\boldsymbol{x}$.
   $\mathbf{S}_B$ and $\mathbf{S}_W$ are symmetric, i.e., $\mathbf{S}_B^T = \mathbf{S}_B$ and $\mathbf{S}_W^T = \mathbf{S}_W$.

4. We have a dataset for 2 classes. For class 1, we have these samples {(1.4, 1.3, 0.8), (0.3, -0.4, -0.3), (0, -1.1, -2), (1.3, -0.5, -0.6)}. For class 2, we have these samples {(-1, -0.5, -1), (-0.5, -0.9, -0.2), (-1.4, 0.5, -1.2), (-0.8, -0.9, -1.3), (0.4, -0.1, 0.9), (1.1, -0.4, -0.3)}. Now we want to fit a Fisher's LDA model to this dataset. For this question, all your results should be rounded to 3 decimal places.

   (a) Calculate the mean of the two classes, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

   (b) Calculate the between-class scatter matrix $\mathbf{S}_B$.

   (c) Calculate the within-class scatter matrix $\mathbf{S}_W$.

   (d) Calculate the optimal $\boldsymbol{w}^*$ based on $\mathbf{S}_B$ and $\mathbf{S}_W$.

   Hint: You can use some software to solve the eigenvalue problem.

5. Python Programming Task: K-means implementation.

   (1) Plot the data points provided in "random_generated_data.csv".

   (2) Implement the K-means algorithm to fit the above data points with the cluster number k set to 5. Please plot the prediction results (i.e., the data points and their assigned cluster centers, please use different colors to represent different clusters) and report the number of iterations and the sum of squared distances of data points to their assigned cluster centers.

   (3) Repeat the fitting with different cluster numbers of 1,2,3,4,6,7,8,9,10,11,12. Plot the prediction results. Discuss how the cluster number affects the fitting results, and choose the optimal cluster number.

   Hint: You can use **pandas.read_csv** to load data points.

   Note: Please submit your Python implementation (e.g., Q5.py, Q5.ipynb) along with the report.

   Here is a submission example via Blackboard for your reference: report.pdf, Q5.py, veriguide_receipt.pdf.

   Please note that there are no restrict limitations regarding the submission format.

*** END ***

1. Which of the following statements is true about Fisher's LDA for two classes? Only one choice is correct.

   [a] The goal of LDA is to minimize the distance between classes and to maximize the distance within classes.

   [b] The goal of LDA is to maximize the distance between classes and to minimize the distance within classes.

   [c] The goal of LDA is to minimize the distance between classes and within classes.

   [d] The goal of LDA is to maximize the distance between classes and within classes.

[b]

since Fisher's LDA maximizes $J(w) = \frac{w^T S_B w}{w^T S_W w}$

so the goal is to "maximize" between-class scatter matrix S_B and "minimize" within-class scatter matrix S_W

since $S_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$, to maximize, difference between μ_1, μ_2 should be big

since $S_W = \sum_{n:y_n=1} (x_n - \mu_1)(x_n - \mu_1)^T + \sum_{n:y_n=2} (x_n - \mu_2)(x_n - \mu_2)^T$, the difference between point and center should be small

so the goal is to maximize the distance between class and minimize distance within class

2. Which of the following statements is Wrong about the LDA for two classes?// Only one choice is correct.

   [a] Fisher's linear discriminant attempts to find the vectors that maximizes the separation between classes of the projected data.

   [b] LDA takes into account the categories in the data.

   [c] LDA assumes linear decision boundary and variance-covariance homogeneity.

   [d] The desired LDA transformation is in the direction of the within-class covariance matrix.

[d].

the result of the decision boundary tries to maximize the separation after projection, so besides within class information it should also take account of the between class information like the relative position of 2 classes,

so the desired transformation is in the direction that maximizes the separation.

the direction is not necessarlly the same with the within-class covariance matrix.

3. Referring to Week9's Slides Page19, we have learned about Fisher's LDA for two classes. Prove that $\boldsymbol{w}$ reaches the maximum of $J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w}}$ when $\boldsymbol{w}$ satisfies $\mathbf{S}_B \boldsymbol{w} = \lambda \mathbf{S}_W \boldsymbol{w}$, where $\lambda = \frac{\boldsymbol{w}^T \mathbf{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \mathbf{S}_W \boldsymbol{w}}$.

proof: define $f(x) = w^T S_B w$, $g(w) = w^T S_W w$

$$\Rightarrow \frac{\partial}{\partial w} J(w) = \frac{\partial}{\partial w} \frac{f(w)}{g(w)} = \frac{\frac{\partial}{\partial w} f(w) \cdot g(w) - f(w) \cdot \frac{\partial}{\partial w} g(w)}{g(w) \cdot g(w)}$$

since $\frac{\partial}{\partial w} f(w) = \frac{\partial}{\partial w} w^T S_B w = (S_B + S_B^T) w = 2 S_B w$ (since S_B is symmetric)

$\frac{\partial}{\partial w} g(w) = \frac{\partial}{\partial w} w^T S_W w = (S_W + S_W^T) w = 2 S_W w$ (since S_W is symmetric)

$$\Rightarrow \frac{\partial}{\partial w} J(w) = \frac{2 S_B W \cdot W^T S_W W - 2 W^T S_B W S_W W}{W^T S_W W W^T S_W W}$$

when $J(w)$ reaches maximum:

$$0 = \frac{\partial}{\partial w} J(w) = \frac{2 S_B W \cdot W^T S_W W - 2 W^T S_B W S_W W}{W^T S_W W W^T S_W W}$$

$$\Rightarrow 0 = 2 S_B W \cdot W^T S_W W - 2 W^T S_B W S_W W$$

$$\Rightarrow S_B W W^T S_W W = W^T S_B W S_W W$$

$$\Rightarrow S_B W = \frac{W_T S_B W}{W^T S_W W} S_W W$$

$$\Rightarrow S_B W = \lambda S_W W, \text{ where } \lambda = \frac{W_T S_B W}{W^T S_W W}$$

4. We have a dataset for 2 classes. For class 1, we have these samples {(1.4, 1.3, 0.8), (0.3, -0.4, -0.3), (0, -1.1, -2), (1.3, -0.5, -0.6)}. For class 2, we have these samples {(-1, -0.5, -1), (-0.5, -0.9, -0.2), (-1.4, 0.5, -1.2), (-0.8, -0.9, -1.3), (0.4, -0.1, 0.9), (1.1, -0.4, -0.3)}. Now we want to fit a Fisher's LDA model to this dataset. For this question, all your results should be rounded to 3 decimal places.

   (a) Calculate the mean of the two classes, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

   (b) Calculate the between-class scatter matrix $\mathbf{S}_B$.

   (c) Calculate the within-class scatter matrix $\mathbf{S}_W$.

   (d) Calculate the optimal $\boldsymbol{w}^*$ based on $\mathbf{S}_B$ and $\mathbf{S}_W$.

(a):

$$\mu_1 = \frac{1}{4}\left[\begin{pmatrix}1.4\\1.3\\0.8\end{pmatrix} + \begin{pmatrix}0.3\\-0.4\\-0.3\end{pmatrix} + \begin{pmatrix}0\\-1.1\\-2\end{pmatrix} + \begin{pmatrix}1.3\\-0.5\\-0.6\end{pmatrix}\right] = \begin{pmatrix}0.750\\-0.175\\-0.525\end{pmatrix}$$

$$\mu_2 = \frac{1}{6}\left[\begin{pmatrix}-1\\-0.5\\-1\end{pmatrix} + \begin{pmatrix}-0.5\\-0.9\\-0.2\end{pmatrix} + \begin{pmatrix}-1.4\\0.5\\-1.2\end{pmatrix} + \begin{pmatrix}-0.8\\-0.9\\-1.3\end{pmatrix} + \begin{pmatrix}0.4\\-0.1\\0.9\end{pmatrix} + \begin{pmatrix}1.1\\-0.4\\-0.3\end{pmatrix}\right] = \begin{pmatrix}-\frac{11}{30}\\-\frac{23}{60}\\-\frac{31}{60}\end{pmatrix} \approx \begin{pmatrix}-0.367\\-0.383\\-0.517\end{pmatrix}$$

(b)

$$\mu_2 - \mu_1 = \begin{pmatrix}-\frac{11}{30}\\-\frac{23}{60}\\-\frac{31}{60}\end{pmatrix} - \begin{pmatrix}0.75\\-0.175\\-0.525\end{pmatrix} = \begin{pmatrix}-\frac{67}{60}\\-\frac{5}{24}\\\frac{1}{120}\end{pmatrix}$$

$$S_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T = \begin{pmatrix}-\frac{67}{60}\\-\frac{5}{24}\\\frac{1}{120}\end{pmatrix}\begin{pmatrix}-\frac{67}{60} & -\frac{5}{24} & \frac{1}{120}\end{pmatrix} = \begin{pmatrix}\frac{4489}{3600} & \frac{67}{288} & -\frac{67}{7200}\\ \frac{67}{288} & \frac{25}{576} & -\frac{1}{576}\\ -\frac{67}{7200} & -\frac{1}{576} & \frac{1}{14400}\end{pmatrix} \approx \begin{pmatrix}1.247 & 0.233 & -0.009\\ 0.233 & 0.043 & -0.002\\ -0.009 & -0.002 & \underline{0.00007}\end{pmatrix}$$

↑ note: if keep 3 decimal places, it's 0, but according to Piazza, I kept a smaller value

(c)

$$S_1 = \sum_{x \in class 1} (x-\mu_1)(x-\mu_1)^T = \left[\begin{pmatrix}1.4\\1.3\\0.8\end{pmatrix} - \begin{pmatrix}0.75\\-0.175\\-0.525\end{pmatrix}\right]^2 + \left[\begin{pmatrix}0.3\\-0.4\\-0.3\end{pmatrix} - \begin{pmatrix}0.75\\-0.175\\-0.525\end{pmatrix}\right]^2 + \left[\begin{pmatrix}0\\-1.1\\-2\end{pmatrix} - \begin{pmatrix}0.75\\-0.175\\-0.525\end{pmatrix}\right]^2 + \left[\begin{pmatrix}1.3\\-0.5\\-0.6\end{pmatrix} - \begin{pmatrix}0.75\\-0.175\\-0.525\end{pmatrix}\right]^2$$

$$= \begin{pmatrix}1.490 & 1.575 & 1.825\\ 1.575 & 3.188 & 3.293\\ 1.825 & 3.293 & 3.988\end{pmatrix}$$

$$S_2 = \sum_{x \in class 2} (x-\mu_2)(x-\mu_2)^T = \left[\begin{pmatrix}-1\\-0.5\\-1\end{pmatrix} - \begin{pmatrix}-\frac{11}{30}\\-\frac{23}{60}\\-\frac{31}{60}\end{pmatrix}\right]^2 + \left[\begin{pmatrix}-0.5\\-0.9\\-0.2\end{pmatrix} - \begin{pmatrix}-\frac{11}{30}\\-\frac{23}{60}\\-\frac{31}{60}\end{pmatrix}\right]^2 + \left[\begin{pmatrix}-1.4\\0.5\\-1.2\end{pmatrix} - \begin{pmatrix}-\frac{11}{30}\\-\frac{23}{60}\\-\frac{31}{60}\end{pmatrix}\right]^2 + \left[\begin{pmatrix}-0.8\\-0.9\\-1.3\end{pmatrix} - \begin{pmatrix}-\frac{11}{30}\\-\frac{23}{60}\\-\frac{31}{60}\end{pmatrix}\right]^2$$

$$+ \left[\begin{pmatrix}0.4\\-0.1\\0.9\end{pmatrix} - \begin{pmatrix}-\frac{11}{30}\\-\frac{23}{60}\\-\frac{31}{60}\end{pmatrix}\right]^2 + \left[\begin{pmatrix}1.1\\-0.4\\-0.3\end{pmatrix} - \begin{pmatrix}-\frac{11}{30}\\-\frac{23}{60}\\-\frac{31}{60}\end{pmatrix}\right]^2$$

$$\approx \begin{pmatrix}4.413 & -0.353 & 2.713\\ -0.353 & 1.408 & 0.092\\ 2.713 & 0.092 & 3.468\end{pmatrix}$$

$$\Rightarrow S_W = S_1 + S_2 \approx \begin{pmatrix}5.903 & 1.222 & 4.538\\ 1.222 & 4.596 & 3.384\\ 4.538 & 3.384 & 7.456\end{pmatrix}$$

(d)

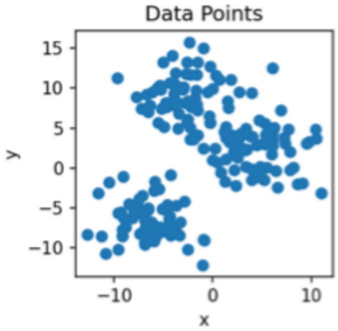
$$w^* \propto S_w^{-1}(\mu_2-\mu_1) = \begin{pmatrix} 5.903 & 1.222 & 4.538 \\ 1.222 & 4.596 & 3.384 \\ 4.538 & 3.384 & 7.456 \end{pmatrix}^{-1} \begin{pmatrix} -\frac{67}{60} \\ -\frac{5}{24} \\ \frac{1}{120} \end{pmatrix}$$

$$\approx \begin{pmatrix} 0.343 & 0.094 & -0.252 \\ 0.094 & 0.353 & -0.217 \\ -0.252 & -0.217 & 0.386 \end{pmatrix} \begin{pmatrix} -\frac{67}{60} \\ -\frac{5}{24} \\ \frac{1}{120} \end{pmatrix}$$

$$\approx \begin{pmatrix} -0.405 \\ -0.180 \\ 0.330 \end{pmatrix}$$

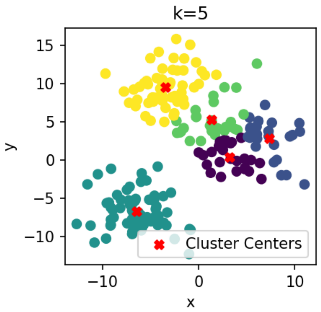
5. (1).

the data are plotted as below:

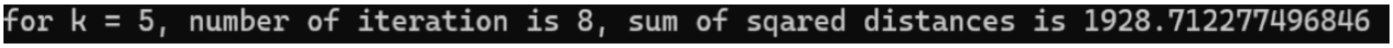


(2)

prediction resut:



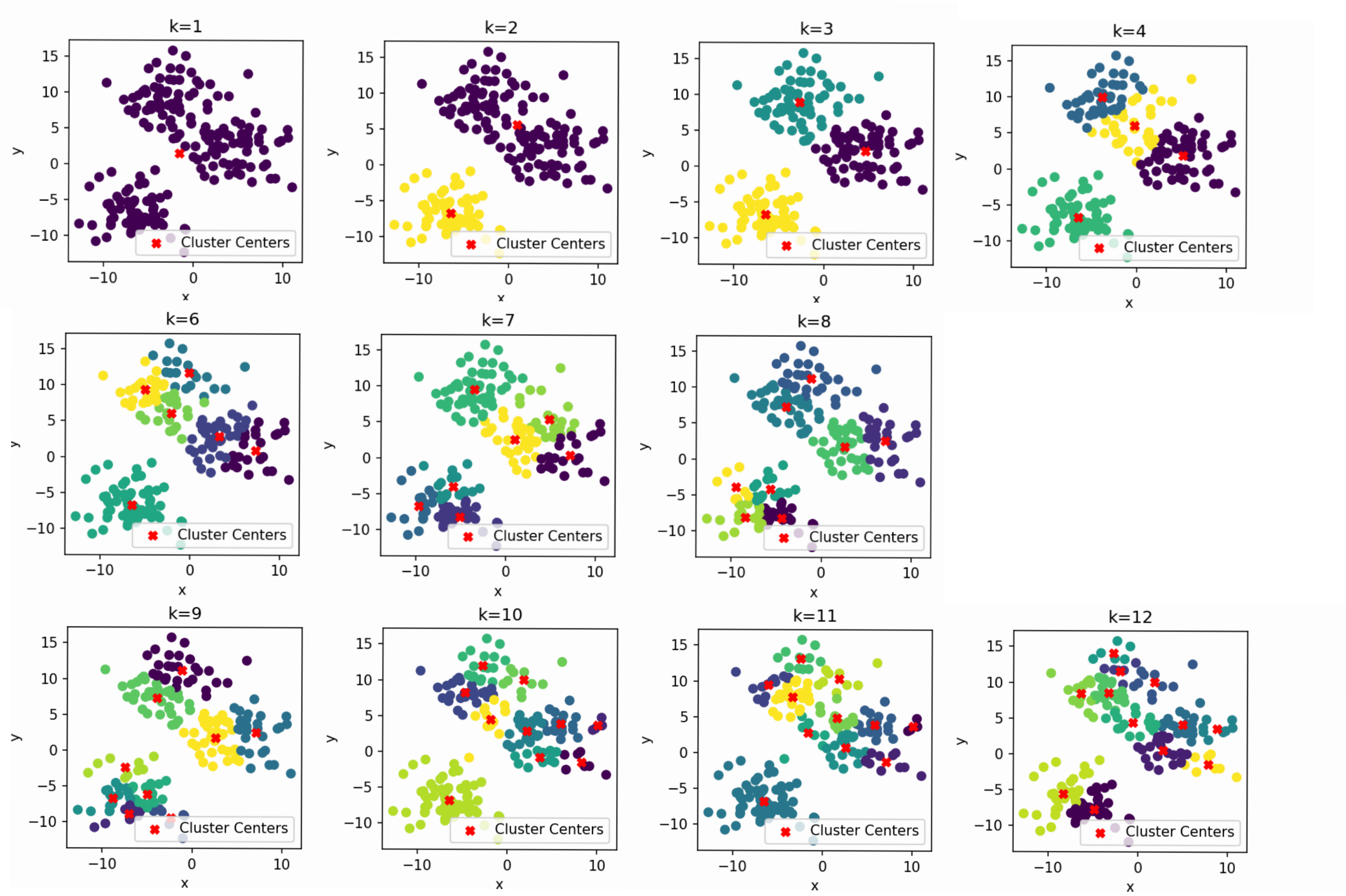
number of iteration and sum of squared · distances:



```
for k = 5, number of iteration is 8, sum of sqared distances is 1928.712277496846
```

(3)

prediction result:

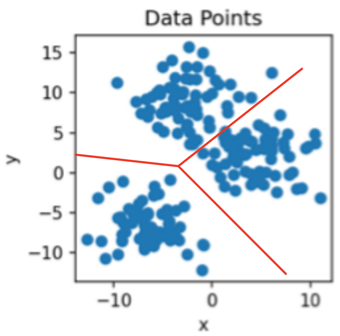


when the cluster number increaces, the number of clusters generated at the end increases, so that in the final data, more clusters will be generated, and each with smaller size.

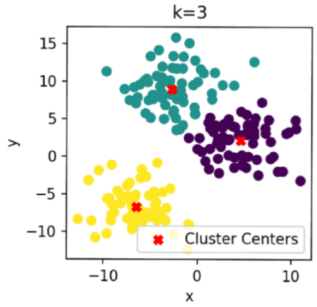
Also, when number of cluster increases, the squared distance sum to centers gets smaller.

when cluster number is too small (eg. k=1) then too few clusters are generated and different clusters are grouped to one. when cluster number is too big (eg. k=12) too many clusters are generated and points that originally belong to one cluster are separated into multiple clusters

from the data it's quite clear that there should be 3 clusters:

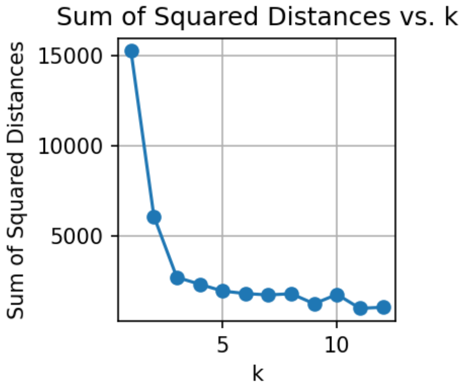


and it's also the case that the result with k=3 looks best for clustering



So that the optimal cluster number is k=3

alternative: Elbow method



we can see that when k=3, the sum of square distances' decrease becomes slow, i.e. meets the elbow

so k should be 3.