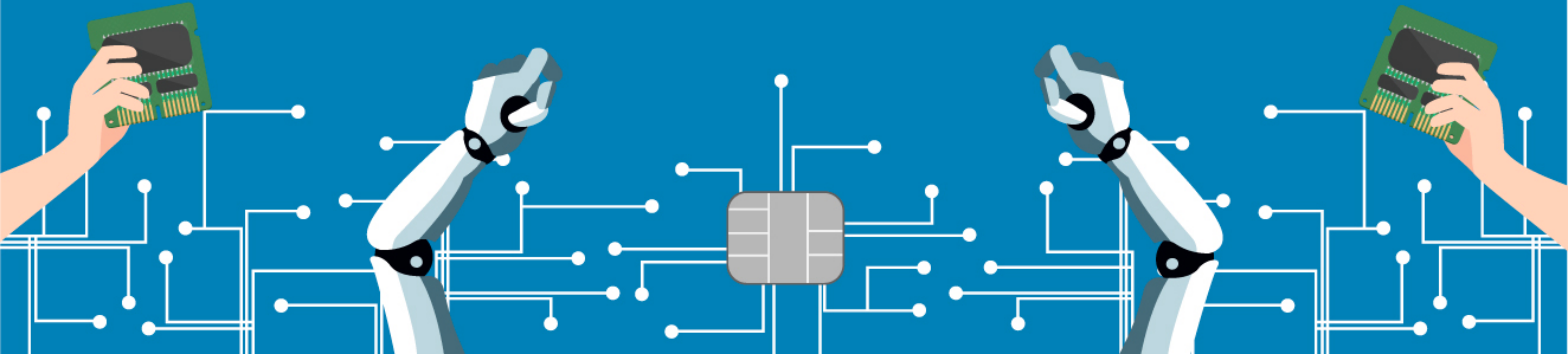
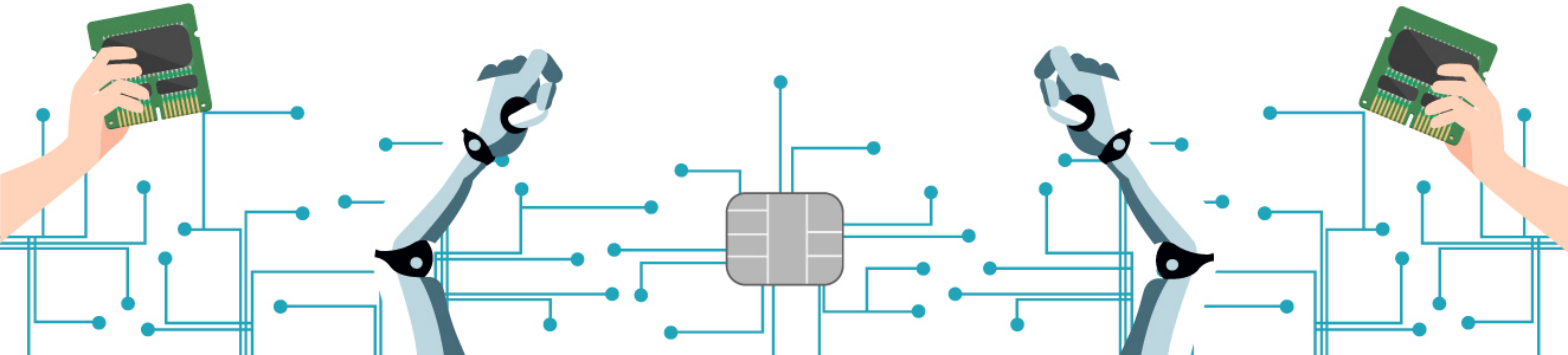
Kuggle

2020_02 Kuggle 정규세션_W6

2020.10.27
발표자: 정성준



CONTENTS



01 Linear Discriminant Analysis(LDA) 개요

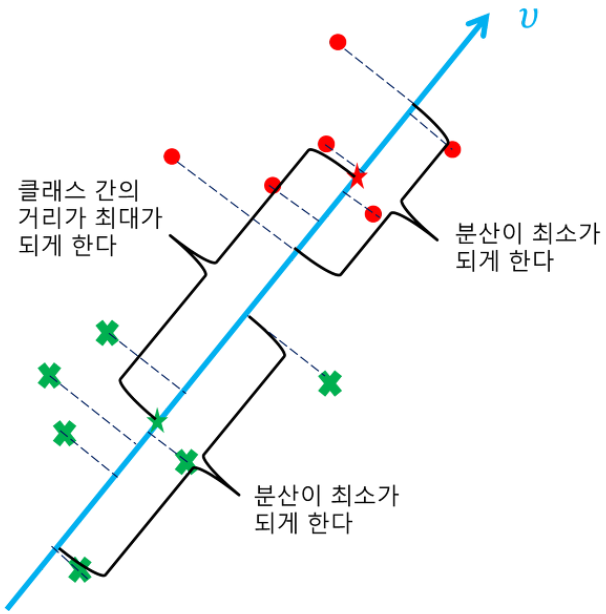
- **LDA, Linear Discriminant Analysis**

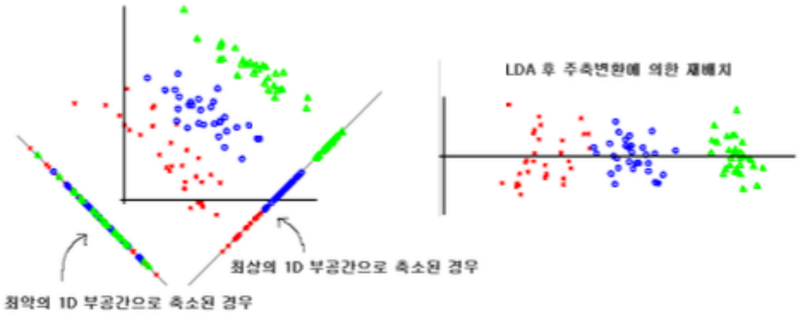
:클래스간 분산(between-class scatter)과 클래스내 분산(within-class scatter)의 비율을 최대화하는 방식으로 데이터에 대한 특징 벡터의 차원을 축소하는 방법

- LDA는 아래와 같은 두 가지 성질을 가진 벡터를 찾는다.

1) 데이터 포인트들을 투영시켰을 때 각 클래스에 속하는 투영들의 평균간의 거리의 합이 최대가 되게 하는 벡터. 다른 말로, **클래스 간(between-class)의 거리가 최대가 되게 하는 벡터**.

2) 데이터 포인트들을 투영시켰을 때 클래스 내의 투영들의 분산이 최소가 되게 하는 벡터. 다른 말로, **클래스 내(within-class)의 분산이 최소가 되게 하는 벡터**.





01 Linear Discriminant Analysis(LDA) 개요

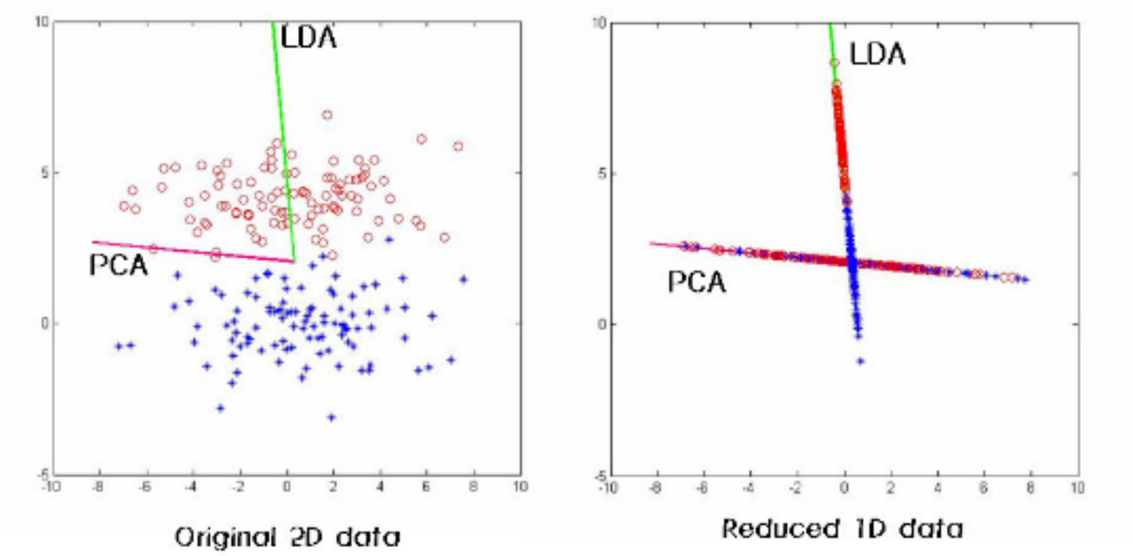
- **LDA vs PCA**

1. 선형판별분석(LDA)

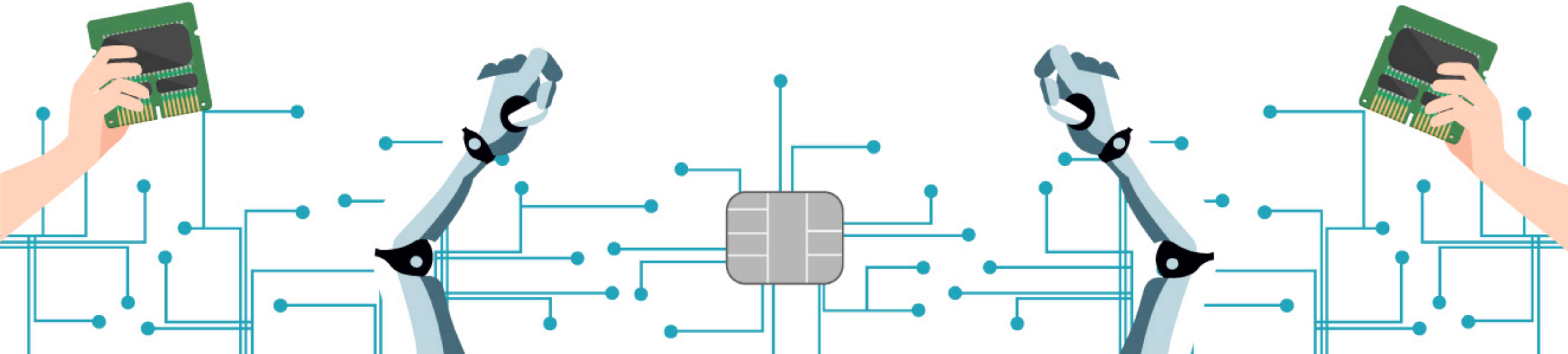
: 데이터의 최적 분류(best discrimination between classes)의 관점에서 차원을 축소하는 방법

2. 주성분분석(PCA)

: 데이터의 최적 표현(best description of the data in its entirety)의 관점에서 데이 터를 축소하는 방법



CONTENTS



02 LDA의 이해

- **Fisher's linear discriminant**

:클래스 내 분산(within-class scatter)으로 평균 간의 차이를 정규화하여 목적 함수로 표 현하고, 이를 최대화하는 방법

:Fisher의 선형 판별식은 목적 함수를 최대화하는 선형 변환행렬 w를 찾아내는 것 → w라는 변환 행렬은 같은 클래스의 표본들은 인접하게 사영을 취하고, 다른 클래스간의 사영은 가능한 중심이 멀리 떨어지도록 변환시키는 행렬을 의미

– 사영 표본의 각 클래스에 대한 분산

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

– 사영 표본의 클래스내 분산

$$(\tilde{s}_1^2 + \tilde{s}_2^2)$$

– 목적 함수

$$J(W) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{(\tilde{s}_1^2 + \tilde{s}_2^2)}$$

02 LDA의 이해

- **Fisher's linear discriminant**

: 수학적 이해

L : 클래스 수
N_i : 클래스 i 내 샘플의 개수
N : 전체 샘플의 개수
$x_j^{(i)}$: 클래스 i 내 j 번째 샘플

$$N = N_1 + N_2 + \cdots + N_L$$

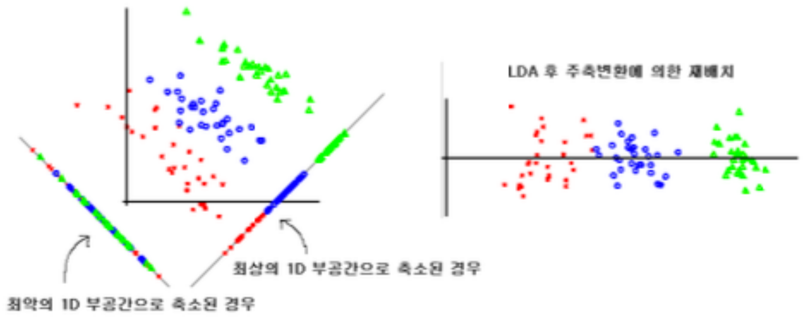
클래스 1부터 클래스L 까지의 데이터포인트들을 나열하면 아래와 같다

$$x_1^{(1)},\ x_2^{(1)},\ldots,\ x_{N_1}^{(1)},\ x_1^{(2)},\ x_2^{(2)},\ldots,\ x_{N_2}^{(2)},\ldots,\ x_1^{(L)},\ x_2^{(L)},\ldots,\ x_{N_L}^{(L)}$$

클래스 1부터 클래스L 까지의 데이터포인트들을 벡터v에 사상(linear map)하면 다음과 같다

$$v^Tx_1^{(1)},\ v^Tx_2^{(1)},\ldots,\ v^Tx_{N_1}^{(1)},\ v^Tx_1^{(2)},\ v^Tx_2^{(2)},\ldots,\ v^Tx_{N_2}^{(2)},\ldots,\ v^Tx_1^{(L)},\ v^Tx_2^{(L)},\ldots,\ v^Tx_{N_L}^{(L)}$$

사상한 벡터들을 가지고 두 가지를 구한다
1) 클래스 간의 거리의 합
2) 각 클래스 내의 분산의 합



02 LDA의 이해

L : 클래스 수
N_i : 클래스 i 내 샘플의 개수
N : 전체 샘플의 개수
$x_j^{(i)}$: 클래스 i 내 j 번째 샘플

- Fisher's linear discriminant

1) 클래스 간의 거리의 합

클래스 i 에 속하는 데이터 포인트들의 사상들의 평균

$$\overline{m_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} v^T x_j^{(i)} = v^T\left(\frac{1}{N_i}\sum_{j=1}^{N_i} x_j^{(i)}\right) = v^T m_i \qquad m_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_j^{(i)}$$

클래스 간의 거리구하기

$$\frac{N_1}{N}\frac{N_2}{N}(\overline{m_1}-\overline{m_2})^2+\frac{N_1}{N}\frac{N_3}{N}(\overline{m_1}-\overline{m_3})^2+\cdots+\frac{N_1}{N}\frac{N_L}{N}(\overline{m_1}-\overline{m_L})^2$$
$$+\frac{N_2}{N}\frac{N_3}{N}(\overline{m_2}-\overline{m_3})^2+\frac{N_2}{N}\frac{N_4}{N}(\overline{m_2}-\overline{m_4})^2+\cdots+\frac{N_2}{N}\frac{N_L}{N}(\overline{m_2}-\overline{m_L})^2$$
$$\vdots$$
$$+\frac{N_?}{N}\frac{N_?}{N}(\overline{m_?}-\overline{m_?})^2+\frac{N_?}{N}\frac{N_?}{N}(\overline{m_?}-\overline{m_?})^2+\cdots+\frac{N_{L-1}}{N}\frac{N_L}{N}(\overline{m_{L-1}}-\overline{m_L})^2$$

$$\rightarrow \sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(\overline{m_i}-\overline{m_j})^2$$

02 LDA의 이해

L : 클래스 수
N_i : 클래스 i 내 샘플의 개수
N : 전체 샘플의 개수
$x_j^{(i)}$: 클래스 i 내 j 번째 샘플

- Fisher's linear discriminant

1) 클래스 간의 거리의 합

클래스 간의 거리구하기

$$\frac{N_1}{N}\frac{N_2}{N}(\overline{m_1}-\overline{m_2})^2+\frac{N_1}{N}\frac{N_3}{N}(\overline{m_1}-\overline{m_3})^2+\cdots+\frac{N_1}{N}\frac{N_L}{N}(\overline{m_1}-\overline{m_L})^2$$
$$+\frac{N_2}{N}\frac{N_3}{N}(\overline{m_2}-\overline{m_3})^2+\frac{N_2}{N}\frac{N_4}{N}(\overline{m_2}-\overline{m_4})^2+\cdots+\frac{N_2}{N}\frac{N_L}{N}(\overline{m_2}-\overline{m_L})^2$$
$$\vdots$$
$$+\frac{N_?}{N}\frac{N_?}{N}(\overline{m_?}-\overline{m_?})^2+\frac{N_?}{N}\frac{N_?}{N}(\overline{m_?}-\overline{m_?})^2+\cdots+\frac{N_{L-1}}{N}\frac{N_L}{N}(\overline{m_{L-1}}-\overline{m_L})^2$$

$$\longrightarrow \sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(\overline{m_i}-\overline{m_j})^2$$

$$\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(\overline{m_i}-\overline{m_j})^2$$
$$=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(\overline{m_i}-\overline{m_j})(\overline{m_i}-\overline{m_j})^T$$
$$=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(v^Tm_i-v^Tm_j)(v^Tm_i-v^Tm_j)^T \qquad (\because\ \overline{m_i}=v^Tm_i)$$
$$=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}v^T(m_i-m_j)(m_i-m_j)^Tv$$
$$=v^T\left(\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(m_i-m_j)(m_i-m_j)^T\right)v$$
$$=v^TS_b^{LDA}v \longrightarrow$$ Between-class scatter matrix

02 LDA의 이해

- **Fisher’s linear discriminant**

1) 클래스 간의 거리의 합

클래스 간의 거리구하기

$$
\begin{aligned}
&\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(\overline{m_i}-\overline{m_j})^2\\
&=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(\overline{m_i}-\overline{m_j})(\overline{m_i}-\overline{m_j})^T\\
&=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(v^Tm_i-v^Tm_j)(v^Tm_i-v^Tm_j)^T \qquad (\because\ \overline{m_i}=v^Tm_i)\\
&=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}v^T(m_i-m_j)(m_i-m_j)^Tv\\
&=v^T\left(\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(m_i-m_j)(m_i-m_j)^T\right)v\\
&=v^TS_b^{LDA}v
\end{aligned}
$$

→ Between-class scatter matrix

$$
\begin{aligned}
S_b^{LDA}&=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(m_i-m_j)(m_i-m_j)^T\\
&=\sum_{i=1}^{L}\frac{N_i}{N}(m_i-m_0)(m_i-m_0)^T
\end{aligned}
$$

L : 클래스 수

N_i : 클래스 i 내 샘플의 개수

N : 전체 샘플의 개수

$x_j^{(i)}$: 클래스 i 내 j 번째 샘플

$$
\begin{aligned}
S_b^{LDA}&=\sum_{i=1}^{L-1}\sum_{j=i+1}^{L}\frac{N_i}{N}\frac{N_j}{N}(m_i-m_j)(m_i-m_j)^T\\
&=\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i}{N}\frac{N_j}{N}(m_i-m_j)(m_i-m_j)^T\\
&=\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i}{N}\frac{N_j}{N}(m_i-m_j)(m_i^T-m_j^T)\\
&=\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i}{N}\frac{N_j}{N}(m_im_i^T-m_im_j^T-m_jm_i^T+m_jm_j^T)\\
&=\frac{1}{2}\Big(\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i}{N}\frac{N_j}{N}m_im_i^T-\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i}{N}\frac{N_j}{N}m_im_j^T\\
&\qquad-\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i}{N}\frac{N_j}{N}m_jm_i^T+\sum_{i=1}^{L}\sum_{j=1}^{L}\frac{N_i}{N}\frac{N_j}{N}m_jm_j^T\Big)\\
&=\frac{1}{2}\Big(\sum_{i=1}^{L}\frac{N_i}{N}m_im_i^T\sum_{j=1}^{L}\frac{N_j}{N}-\sum_{i=1}^{L}\frac{N_i}{N}m_i\sum_{j=1}^{L}\frac{N_j}{N}m_j^T\\
&\qquad-\sum_{i=1}^{L}\frac{N_i}{N}m_i^T\sum_{j=1}^{L}\frac{N_j}{N}m_j+\sum_{i=1}^{L}\frac{N_i}{N}\sum_{j=1}^{L}\frac{N_j}{N}m_jm_j^T\Big)\\
&=\frac{1}{2}\Big(\sum_{i=1}^{L}\frac{N_i}{N}m_im_i^T-\sum_{i=1}^{L}\frac{N_i}{N}m_i\sum_{j=1}^{L}\frac{N_j}{N}m_j^T\\
&\qquad-\sum_{i=1}^{L}\frac{N_i}{N}m_i^T\sum_{j=1}^{L}\frac{N_j}{N}m_j+\sum_{j=1}^{L}\frac{N_j}{N}m_jm_j^T\Big) \qquad \Big(\because\ \sum_{j=1}^{L}\frac{N_j}{N}
\end{aligned}
$$

$$
\sum_{i=1}^{L}\frac{N_i}{N}m_i = m_0
$$

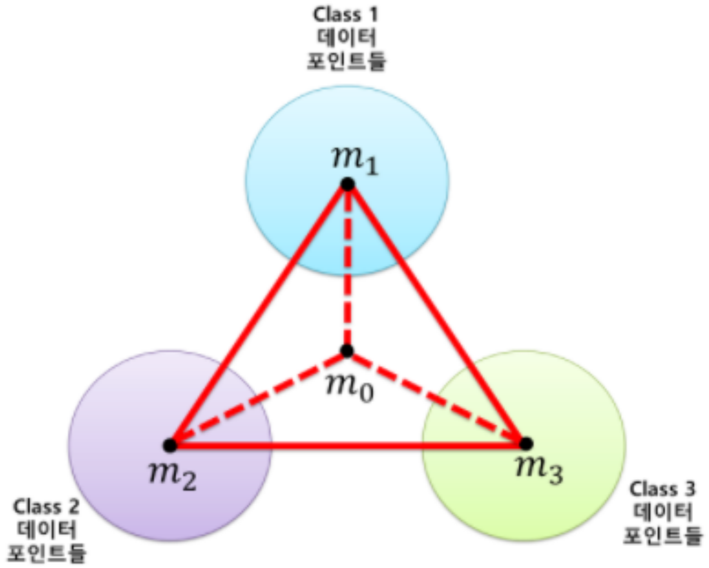
$$
\begin{aligned}
S_b^{LDA}&=\frac{1}{2}\Big(\sum_{i=1}^{L}\frac{N_i}{N}m_im_i^T-\sum_{i=1}^{L}\frac{N_i}{N}m_i\sum_{j=1}^{L}\frac{N_j}{N}m_j^T\\
&\qquad-\sum_{i=1}^{L}\frac{N_i}{N}m_i^T\sum_{j=1}^{L}\frac{N_j}{N}m_j+\sum_{j=1}^{L}\frac{N_j}{N}m_jm_j^T\Big)\\
&=\frac{1}{2}\Big(\sum_{i=1}^{L}\frac{N_i}{N}m_im_i^T-m_0m_0^T-m_0m_0^T+\sum_{j=1}^{L}\frac{N_j}{N}m_jm_j^T\Big)\\
&=\sum_{i=1}^{L}\frac{N_i}{N}m_im_i^T-m_0m_0^T\\
&=\sum_{i=1}^{L}\frac{N_i}{N}m_im_i^T-m_0m_0^T-m_0m_0^T+m_0m_0^T\\
&=\sum_{i=1}^{L}\frac{N_i}{N}m_im_i^T-\sum_{i=1}^{L}\frac{N_i}{N}m_im_0^T-m_0\sum_{i=1}^{L}\frac{N_i}{N}m_i^T+\sum_{i=1}^{L}\frac{N_i}{N}m_0m_0^T\\
&=\sum_{i=1}^{L}\frac{N_i}{N}(m_im_i^T-m_im_0^T-m_0m_i^T+m_0m_0^T)\\
&=\sum_{i=1}^{L}\frac{N_i}{N}(m_i-m_0)(m_i^T-m_0^T)\\
&=\sum_{i=1}^{L}\frac{N_i}{N}(m_i-m_0)(m_i-m_0)^T
\end{aligned}
$$

02 LDA의 이해

- Fisher's linear discriminant

1) 클래스 간의 거리의 합

$$S_b^{LDA} = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \frac{N_i}{N} \frac{N_j}{N} (m_i - m_j)(m_i - m_j)^T$$
$$= \sum_{i=1}^{L} \frac{N_i}{N} (m_i - m_0)(m_i - m_0)^T$$



Beween-class scatter matrix : 각 클래스 평균 사이의 거리들의 합은 각 클래스 평균과 전체 평균간의 거리들의 합과 같다.

02 LDA의 이해

L : 클래스 수
N_i : 클래스 i 내 샘플의 개수
N : 전체 샘플의 개수
$x_j^{(i)}$: 클래스 i 내 j 번째 샘플

- **Fisher's linear discriminant**

1) 각 클래스 내의 분산의 합

클래스 i 에 속하는 데이터 포인트들의 사상들의 분산

$$\sum_{j=1}^{N_i} \frac{1}{N} (v^T x_j^{(i)} - \overline{m_i})^2$$

모든 클래스의 사상의 분산의 합

$$\sum_{i=1}^{L} \sum_{j=1}^{N_i} \frac{1}{N} (v^T x_j^{(i)} - \overline{m_i})^2$$

$$= \sum_{i=1}^{L} \sum_{j=1}^{N_i} \frac{1}{N} (v^T x_j^{(i)} - v^T m_i)(v^T x_j^{(i)} - v^T m_i)^T \qquad (\because \ \overline{m_i} = v^T m_i)$$

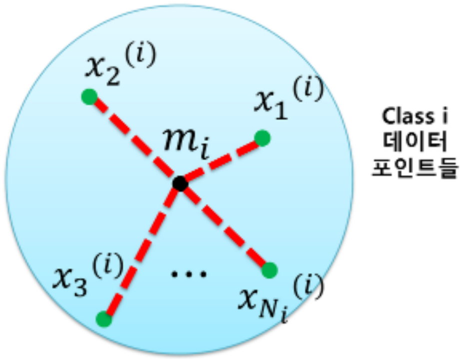
$$= \sum_{i=1}^{L} \sum_{j=1}^{N_i} \frac{1}{N} v^T (x_j^{(i)} - m_i)(x_j^{(i)} - m_i)^T v$$

$$= v^T \left(\sum_{i=1}^{L} \sum_{j=1}^{N_i} \frac{1}{N} (x_j^{(i)} - m_i)(x_j^{(i)} - m_i)^T \right) v$$

$$= v^T S_w^{LDA} v$$

Within-class scatter matrix

$$S_w^{LDA} = \sum_{i=1}^{L} \sum_{j=1}^{N_i} \frac{1}{N} (x_j^{(i)} - m_i)(x_j^{(i)} - m_i)^T$$



within-class scatter matrix: 각 클래스 내의 데이터포인트들과 각 클래스의 평균 사이의 거리를 모두 더한다.

02 LDA의 이해

- **Fisher's linear discriminant**

목적함수를 v에 대해 미분하고 0을 가지는 v를 찾는다!!

$$v = \underset{v \in R^d}{arg\ max} \frac{v^T S_b^{LDA} v}{v^T S_w^{LDA} v}$$

$$= \underset{v^T S_w^{LDA} v = 1}{arg\ max}\ v^T S_b^{LDA} v$$

$$S_b^{LDA} v = \lambda S_w^{LDA} v$$

$$(S_w^{LDA})^{-1} S_b^{LDA} v = \lambda v$$

$(S_w^{LDA})^{-1} S_b^{LDA}$ 의 고유 값(eigen value)과 고유벡터(eigen vector)를 찾아준다.

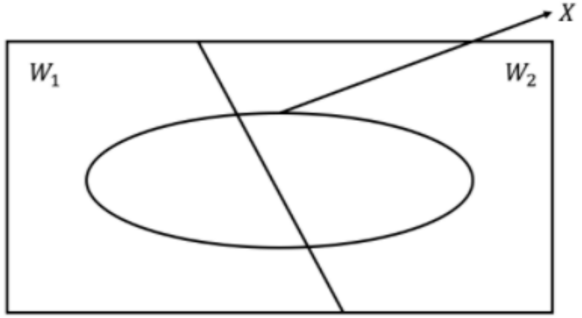
$$S_W^T S_B = [e_1 \ \cdots \ e_n] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} e_1^T \\ \cdots \\ e_n^T \end{bmatrix}$$

가장 높은 고유 값을 가지는 고유벡터가 사상 벡터가 된다.

02 LDA의 이해

- Bayes's rule

LDA 가정 : 데이터 분포가 **다변량 정규분포(multivariate normal distribution)**



$$P(W_i|x) = \frac{P(x|W_i)P(W_i)}{P(x)} = \frac{P(x|W_i)P(W_i)}{P(x|W_1)P(W_1) + P(x|W_2)P(W_2)}$$

- P(Wi|x)는 새로운 데이터 x가 주어졌을 때(=정답 범주를 모를 때=예측할 때) WiWi일 확률, 즉 검증데이터가 특정 범주에 할당하기 위한 스코어를 의미합니다. 범주 분류를 위한 확률(스코어)를 내어주는 함수를 **판별함수(discriminant function)**라고 합니다
- 판별함수는 다변량 정규분포를 따른다고 가정한다.
- 두 범주의 판별함수가 같아지는 선형식을 구하고 선형식의 coefficeint를 구하면 위에서 구한 사상벡터가 나온다.

01 Linear Discriminant Analysis(LDA) 개요

- 선형판별분석
- LDA vs PCA

02 LDA의 이해

- Fisher's linear discriminant
- Basysian rule

03 LDA 실습

- Sklearn, iris dataset을 이용한 실습

감사합니다