20171195014 정승표

#1. PCA 코딩

In [2]:
```
import numpy as np
import pandas as pd

from sklearn.preprocessing import StandardScaler
```

In [4]:
```
df = pd.read_csv('https://raw.githubusercontent.com/uiuc-cse/data-fa14/gh-pages/data/iris.csv')
df.head()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In [5]:
```
X = df.iloc[:,:-1]
label = df.iloc[:,-1]
X.head()
```

Out[5]:

	sepal_length	sepal_width	petal_length	petal_width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2

In [6]:
```
# Step 1. Center Data
X_scaled = StandardScaler().fit_transform(X)
X_scaled[:5]
```

Out[6]:
```
array([[-0.90068117,  1.03205722, -1.3412724 , -1.31297673],
       [-1.14301691, -0.1249576 , -1.3412724 , -1.31297673],
       [-1.38535265,  0.33784833, -1.39813811, -1.31297673],
       [-1.50652052,  0.10644536, -1.2844067 , -1.31297673],
       [-1.02184904,  1.26346019, -1.3412724 , -1.31297673]])
```

In [7]:
```
# Step 2. Compute Covariance Matrix
cov_matrix = X_scaled.T @ X_scaled / (X_scaled.shape[0]-1)
cov_matrix
```

$Var(X) = \frac{X^T X}{n-1}$

Out[7]:
```
array([[ 1.00671141, -0.11010327,  0.87760486,  0.82344326],
       [-0.11010327,  1.00671141, -0.42333835, -0.358937  ],
       [ 0.87760486, -0.42333835,  1.00671141,  0.96921855],
       [ 0.82344326, -0.358937  ,  0.96921855,  1.00671141]])
```

In [8]:
```
# Step 3. Eigenvalue decomposition
eigvals, eigvecs = np.linalg.eig(cov_matrix) #TODO
eigvals
```

eigenvalue와 eigenvector 생성

Out[8]:
```
array([2.93035378, 0.92740362, 0.14834223, 0.02074601])
```

In [9]:
```
eigvecs
```

Out[9]:
```
array([[ 0.52237162, -0.37231836, -0.72101681,  0.26199559],
       [-0.26335492, -0.92555649,  0.24203288, -0.12413481],
       [ 0.58125401, -0.02109478,  0.14089226, -0.80115427],
       [ 0.56561105, -0.06541577,  0.6338014 ,  0.52354627]])
```

```
# Ratio of explained variance per PC
explained_variances = []
for i in range(len(eigvals)):
    explained_variances.append(eigvals[i] / np.sum(eigvals))

print(np.sum(explained_variances), '₩n', explained_variances)
```

```
1.0000000000000002
[0.7277045209380133, 0.2303052326768066, 0.03683831957627389, 0.005151926808906326]
```

⇒ PC1, PC2가 설명하는 분산의 정도(설명력)을 확인!

```
# Visualization (Embedding)
pc1 = np.dot(X_scaled, eigvecs[:,0]) #TODO
pc2 = np.dot(X_scaled, eigvecs[:,1]) #TODO
res = pd.DataFrame(pc1, columns=['PC1'])
res['PC2'] = pc2
res['label'] = label
res.head()
```

	PC1	PC2	label
0	-2.264542	-0.505704	setosa
1	-2.086426	0.655405	setosa
2	-2.367950	0.318477	setosa
3	-2.304197	0.575368	setosa
4	-2.388777	-0.674767	setosa

#2 PCA의 고유값 고유벡터의 관계성증명

선형결합의 variance를 최대화하는 단위벡터 δ 찾기

$$\max_{\delta} \mathrm{var}(\delta^T X) = \delta^T \mathrm{var}(X)\,\delta = \delta^T \Sigma \delta \quad \text{s.t. } \|\delta\|=1 \Leftrightarrow \delta^T\delta = 1$$

라그랑지안

$$L_0 = \delta^T \Sigma \delta - \lambda(\delta^T\delta - 1)$$

$$\frac{\partial L_0}{\partial \delta} = 2\Sigma\delta - 2\lambda\delta = 0 \Leftrightarrow \Sigma\delta = \lambda\delta \Rightarrow \lambda = \Sigma$$

Σ의 eigenvalue

Σ의 eigen vector

$$\therefore \delta^T\Sigma\delta = \delta^T\lambda\delta = \lambda\underbrace{\delta^T\delta}_{1} = \lambda$$

⇒ maximum variance = λ 이다.