Data Mining

ENSIA 2023-2024

Lab sheet N°5 (Part 1): Dimensionality Reduction

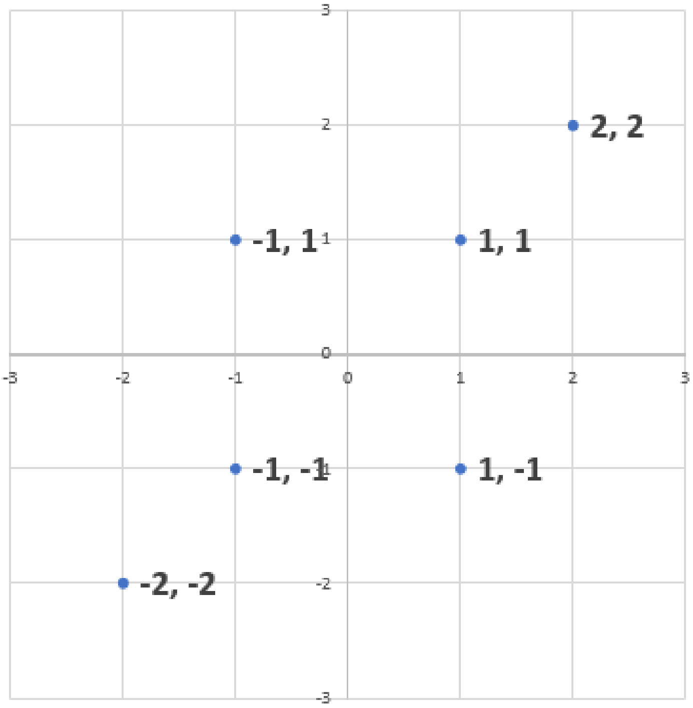
PCA & Feature selection

Exercise 1

1. Give the Steps of PCA.

1. Mean normalization
2. Compute the covariance matrix
3. Compute the eigenvectors/values with said matrix
4. Select the top K eigenvalues and their eigenvectors based on eigenvalues.
5. Create an orthogonal base with the eigenvectors
6. Transform data by multiplying with said base

2. Plot the centered dataset **X** with 5 data points in a 2D plane.
The points are (1,1), (2,2), (-1,-1), (-2,-2), (-1,1), (1,-1).



3. Calculate the covariance matrix for the previous dataset **X**.

X =

X_1	X_2
1	1
2	2
-1	-1
-2	-2
-1	1
1	-1
$\overline{X_1} = 0$	$\overline{X_2} = 0$

The data is centered because $\overline{X_1}$**=**$\overline{X_2} = 0 \Rightarrow$ $\Sigma = \frac{1}{n}\mathbf{X}^T\mathbf{X}$

The data X is centred $\Rightarrow \Sigma = \frac{1}{6} X^t X$

$$\Sigma = \frac{1}{6}\begin{bmatrix} 1 & 2 & -1 & -2 & -1 & 1 \\ 1 & 2 & -1 & -2 & 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 2 & 2 \\ -1 & -1 \\ -2 & -2 \\ -1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\Sigma = \frac{1}{6}\begin{bmatrix} 12 & 8 \\ 8 & 12 \end{bmatrix} = \begin{bmatrix} 2 & \frac{4}{3} \\ 4/3 & 2 \end{bmatrix}$$

4. Determine the eigenvalues and eigenvectors of the covariance matrix.

Eigenvalues: ⟷

$$\Sigma v = \lambda v \Rightarrow \mathrm{Det}(\Sigma - \lambda I) = 0$$

$$\mathrm{Det}\begin{pmatrix} 2-\lambda & \frac{4}{3} \\ \frac{4}{3} & 2-\lambda \end{pmatrix} = 0 \Rightarrow (\lambda-2)^2 - \frac{16}{9} = 0$$

$$\Rightarrow \lambda_1 = \frac{10}{3} \ , \ \lambda_2 = \frac{2}{3}$$

Eigenvectors:

$\lambda_1 = \frac{10}{3}$:

$$\Sigma v = \frac{10}{3} v = \Sigma \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{10}{3} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\begin{cases} 2x_1 + \frac{4}{3}x_2 = \frac{10}{3}x_1 \\ \frac{4}{3}x_1 + 2x_2 = \frac{10}{3}x_2 \end{cases} \Rightarrow \begin{cases} \frac{4}{3}x_2 = \frac{4}{3}x_1 \\ \frac{4}{3}x_1 = \frac{4}{3}x_2 \end{cases}$$

$$\Rightarrow x_1 = x_2 \quad \wedge \quad \|v_1\| = 1$$

$$\Rightarrow v_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$$

$\lambda_2 = \frac{2}{3}$:

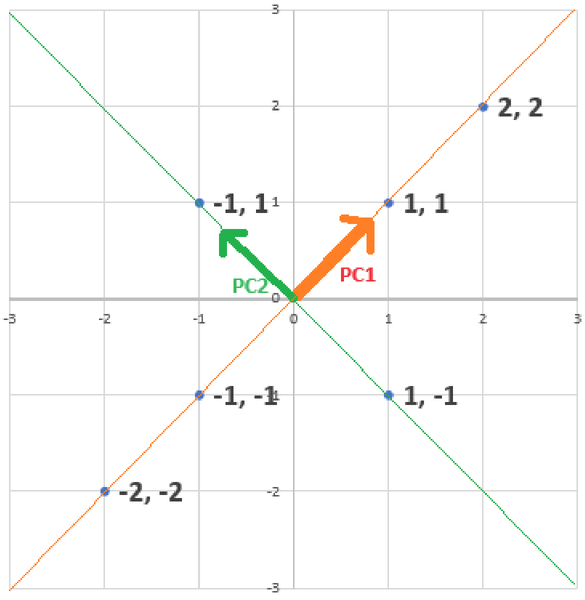
Same calculation

$$\begin{cases} 2x_1 + \frac{4}{3}x_2 = \frac{2}{3}x_1 \\ \frac{4}{3}x_1 + 2x_2 = \frac{2}{3}x_2 \end{cases} \Rightarrow \begin{cases} \frac{4}{3}x_2 = -\frac{4}{3}x_1 \\ \frac{4}{3}x_1 = -\frac{4}{3}x_2 \end{cases}$$

$$\Rightarrow x_2 = -x_1 \quad \wedge \quad \|v_2\| = 1$$

$$\Rightarrow v_2 \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ +\frac{\sqrt{2}}{2} \end{pmatrix}$$

5. Plot the two principal components (eigenvectors) on the same 2D plane as the original data points.



6. Deduce geometric insights or conclusions from the PCA decomposition.

The initial principal component, PC1, primarily accounts for the variations associated with four data points in our dataset, while the second component is orthogonal to the first one, capturing the variations caused by the remaining two data points.

7. Compute the reduced dataset, X_reduce, using a single principal component.

$$\mathbf{T_k} = \mathbf{XU}[:, \mathbf{1} : \mathbf{k}]$$

$$X_{reduced} = \begin{bmatrix} \sqrt{2} \\ 2\sqrt{2} \\ -\sqrt{2} \\ -2\sqrt{2} \\ 0 \\ 0 \end{bmatrix} = \sqrt{2} \begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \\ 0 \\ 0 \end{bmatrix}$$

8. Find the **X_approx** dataset, which represents the original data after reversing the reduction (**for students**).

Exercise 2

We have a binary classification problem with a dataset of 100 samples.

The distribution of the target variable ***Y*** is 60 samples with **Y=0** and 40 samples with **Y=1**.

1. Calculate the entropy of the initial dataset.

Entropy(Y) = -(0.6log2(0.6) + 0.4*log2(0.4)) = 0.97

Then, for two features (X1, X2), compute the entropy of Y given each feature. Here are the details for each feature:

- **Feature X1:**
  - When **X1=0**, there are 30 samples (**Y=0**: 15, **Y=1**: 15).
  - When **X1=1**, there are 70 samples (**Y=0**: 45, **Y=1**: 25).

P(y=0|x1=0) = 15/30 = 0.5
P(y=1|x1=0) = 15/30 = 0.5
P(y=0|x1=1) = 45/70 = 0.642
P(y=1|x1=1) = 25/70 = 0.375
Entropy(Y|X1=0) =- [0.5 log2(0.5)+0.5log2(0.5)] = 1
Entropy(Y|X1=1) = - [0.642*log2(0.642)+0.375*log2(0.375)] = 0.940
P(X1=0) = 30/100 = 0.3
P(X1=1) = 70/100 = 0.7

Entropy(Y|X1) = 0.3*1 +0.7*0.940 = 0.958

- **Feature X2:**
  - When **X2=0**, there are 40 samples (**Y=0**: 25, **Y=1**: 15).
  - When **X2=1**, there are 60 samples (**Y=0**: 35, **Y=1**: 25).

P(y=0|X2=0) = 25/40 = 0.625
P(y=1|X2=0) = 15/40 = 0.375
P(y=0|X2=1) = 45/70 = 0.583
P(y=1|X2=1) = 25/70 = 0.416
Entropy(Y|X2=0) =- [0.625 log2(0.625)+0.375log2(0.375)] = 0.954
Entropy(Y|X2=1) = - [0.583*log2(0.583)+0.416*log2(0.416)] = 0.980
P(X2=0) = 30/100 = 0.4
P(X2=1) = 70/100 = 0.6

Entropy(Y|X2) = 0.4*0.954 +0.6*0.980 = 0.969

2. Using the information gain formula, calculate the information gain for each feature and rank them in descending order.

IG(X1) = Entropy(Y) - Entropy(Y|X1) = 0.97 - 0.958 = 0.012
IG(X2) = Entropy(Y) - Entropy(Y|X2) = 0.97 - 0.969 = 0.001

The ranking is: X1 is better than X2 because it has a bigger information gain.

3. Prove that entropy is always positive.

$$\text{Entropy}(\text{Set}) = -\sum P_i \log_2(P_i)$$

$$0 < P_i \leq 1 \Rightarrow \log_2(P_i) < 0$$

$$\Rightarrow P_i \cdot \log_2(P_i) < 0$$

$$\Rightarrow \sum P_i \log_2(P_i) < 0$$

$$\Rightarrow -\sum P_i \cdot \log_2(P_i) > 0$$

$$\Rightarrow \text{entropy}(\text{Set}) \geq 0$$

4. What is the highest achievable Information Gain when the target variable can take on **n** different categories? (**For student**)

Exercise 3 (20 minutes)

1. Calculate the maximum number of models trained in forward selection with ***100*** features, and similarly for backward elimination.

Forward selection:

The total number of models trained, NB, is calculated as the sum of 100 + 99 + 98 + ... + 1, which equals 100*101 / 2 = 5050

Backward elimination:

The total number of models trained, NB, is determined by summing 100 + 99 + 98 + ... + 1, resulting in the same count of models, 100*101 / 2 = 5050

Note: ***When evaluating the quality of a subset using the average of k models, such as in k-fold cross-validation, the previous count should be multiplied by k.***

2. Analyze the computational demands for both methods in scenarios with a high number of initial features with the goal of selecting only a small subset.

If the number of initial features is ***N*** and we want to select ***M*** features (***M***<<***N***) then:

Forward Selection:

The number of trained models is N +(N-1)+ . . . + (N-M+1)=M(2N-M+1)/2 = O(N)
M<<N
Example: if N= 1000 and M = 3
The number of trained models = 1000+999+998 = 2997

Backward elimination:

The number of trained models is N+(N-1) + ... + (M+1)= (N-M+1)(N+M+1)/2=O(N*N)
Example: if N= 1000 and M = 3
The number of trained models = 1000+999+998+.....+4 = 997*(4+1000)/2 = 500494

Backward elimination is more expensive (quadratic vs. linear) if we want to select a few features out of a large number of features.

3. Discuss the possibility of combining the two methods to develop an improved method (**for students**).