PRINCIPAL COMPONET ANALYSIS (PCA)

Subject: Machine Learning

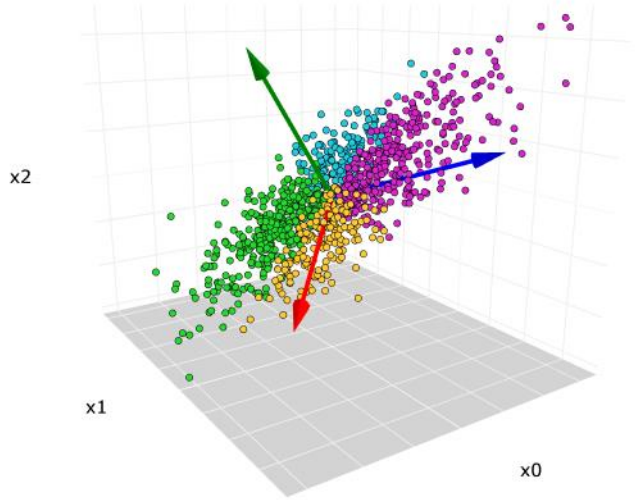
Topics:

- Definition of PCA
- How to apply
- Eigenvector and eigenvalues
- Covariance Matrix
- Find the PCA
- Example

"HOW CAN WE EXTRACT ESSENTIAL INFORMATION AND REDUCE THE COMPLEXITY OF OUR DATA IN AN INCREASINGLY INFORMATION-RICH WORLD WITH COMPLEX DATASETS?"

Definition

- Is a statistic method that is used to reduce the dimensionality of a dataset. This technique help us to simplify a dataset and chose the smaller number of predictors to predict a variable, objective or understand in a better way a BDD



https://www.hiberus.com/crecemos-contigo/analisis-de-componentes-principales/

How apply PCA

- Step 1: Standardization

The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

For example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1), which will lead to biased results. So, transforming the data to comparable scales can prevent this problem.

$$z = \frac{value - mean}{standard\ deviation}$$

How apply PCA

- Step 2: Covariance matrix computation

❑The first step is to calculate the covariance matrix, that show the relationship between the different characteristics and how they are different together.

❑Eigenvector represent the higher variability direction of the data and Eigenvalues indicates how much variability is associated with each eigenvector.

❑Sort the eigenvectors and eigenvalues.

- Step 3: Select the principal components
- Step 4: Data projection

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

So… What are the Eigenvectors and Eigenvalues?

First we need to understand how calcultate the covariance matrix:

Lecture	Humidity (x)	Temperature(y)	Light Level (z)
1	60	25	500
2	55	22	400
3	65	28	600
4	58	23	450
5	68	20	550

1. We need to calculate the mean for each data set: Temperature, Light Level…

$$Humidity\ Mean(HM) = \frac{60 + 55 + \ ... + 68}{5}$$

$$Temp.\ Mean(HM) = \frac{22 + 22 + \ ... + 20}{5}$$

$$Light.\ Mean(HM) = \frac{500 + 400 + \ ... + 550}{5}$$

2. Calculate the covariance matrix

$$Cov\ (X, Y, Z) = \frac{1}{n-1}\sum_{i=1}^{n}(x_1 - \bar{x})(y_1 - \bar{y})$$

$$Cov\ (X, Y, Z\ \ ...) = \frac{(60\ - HM)(25 - TM)(500 - LM) + (55 - HM)(22 - TM)(400 - LM)}{n(number\ of\ observations) - 1})$$

Covariance Matrix

$$\begin{vmatrix} Cov(Hum, Hum) & Cov(Hum, Temp) & Cov(Hum, Light) \\ Cov(Temp, Hum) & Cov(Temp, Temp) & Cov(Temp, Light) \\ Cov(Light, Hum) & Cov(Light, Temp) & Cov(Light, Light) \end{vmatrix}$$

What is next? Let's calculate the eigenvalues and eigenvectors.

- Using the Covariance Matrix, use the following formula to calculate the eigenvalue:

$$Det(Matriz\ de\ Covarianza - \lambda \times Identidad) = 0$$

Assuming that our covariance matrix is the following:

$$\begin{bmatrix} 31.75 & 12.75 & 262.5 \\ 12.75 & 6.44 & 132.0 \\ 262.5 & 132.0 & 273.0 \end{bmatrix}$$

The results of the equation are the following:

$$\lambda_1 \approx 2742.72$$
$$\lambda_2 \approx 25.98$$
$$\lambda_3 \approx 0.49$$

Calculate the Eigenvectors

The formula to calculate the eigenvector is:
$(\boldsymbol{Covariance\ Matrix} - \boldsymbol{\lambda} * \boldsymbol{Identity}) * \boldsymbol{Eigenvector} = \mathbf{0}$

The results are:

- For λ_1 the eigenvector is approximately:

$$v_1 = \begin{bmatrix} 0.096 \\ 0.046 \\ 0.995 \end{bmatrix}$$

- For λ_2 the eigenvector is approximately:

$$v_2 = \begin{bmatrix} 0.141 \\ 0.989 \\ -0.44 \end{bmatrix}$$

- For λ_3 the eigenvector is approximately:

$$v_1 = \begin{bmatrix} 0.986 \\ -0.137 \\ -0.092 \end{bmatrix}$$

1. We must order the Eigenvector depending on their eigenvalue, in this case:

$$\lambda_1 > \lambda_2 > \lambda_3$$

So, the λ_1 is the principal eigenvalue

2. Select what PCA are you going to use. Could be just 2 (K=2)
3. Data Projection:

To project the new data, it is necessary to multiply each observation by the eigenvector.

$$X_1\ projection = [T, H, L] * v_1$$
$$X_2\ projection = [T, H, L] * v_2$$

This projection process is applied to each observation in your data set to obtain the projection in principal component space.

Example

Step 1: We have the following dataset

Punto	Distancia (cm)	Temperatura (°C)	Luz (lux)	Presión (Pa)	Humedad Relativa (%)
1	30	25	300	1000	40
2	25	23	350	980	45
...	...	...	...	...	...
1000	35	27	400	1020	35

Step 2: Normalize de data

1. We must calculate the mean (average) for each feature:

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

Where:

- μ_j is the mean of feature j
- n is the number of points (rows)
- x_{ij} is the value in row i and column j

Step 3: Calculate the standard deviation for each feature

1. We calculate the standard deviation for each feature

$$\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \mu_j)^2}$$

Where:

- σ_j is the standard deviation of feature j
- n is the number of points (rows)
- x_{ij} is the value in row i and column j
- μ_j is the mean of feature j

Example

Step 4: Data normalization

We use the calculate mean and standard deviation to normalize each value

$$\text{Normalized value}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Where:

- $\text{Normalized value}_{ij}$ is the normalized value in row i and column j
- x_{ij} is the value in row i and column j
- μ_j is the mean of feature j
- σ_j is the standard deviation of feature j

• Covariance matrix

Let's represent the matrix X with dimensions n (observations) x m (variables). The covariance matrix is calculated as follows:

1. Calculated the mean for each variable

   $$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

   Where:

   - μ_j is the mean of variable j.
   - x_{ij} is the value of variable j in observation i.
   - n is the number of observations.

2. Center the data

We subtract the mean of each variable form all the values of that variable in the dataset . This give us a Matrix Xcentered of the same dimension as X but with a mean of zero for each variable

$$X_{\text{centered}} = X - \mu$$

Example

Step 6: Calculated the covariance matrix

The covariance matrix S is calculated using the following formula:

The covariance matrix S is calculated using the following formula:

$$S = \frac{1}{n-1}(X^T_{\text{centered}} \cdot X_{\text{centered}})$$

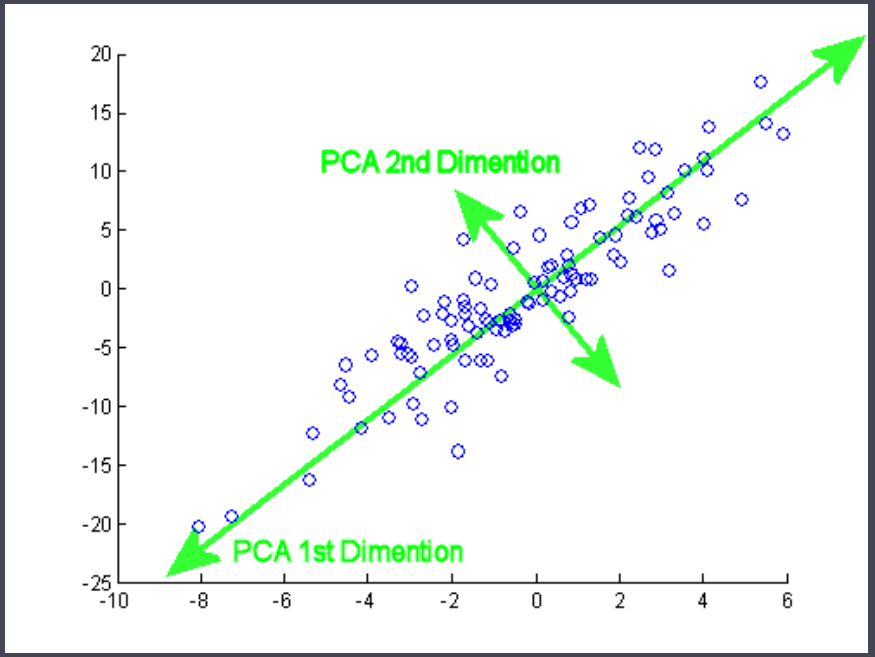
Where:

- n is the number of observations.
- X_{centered} is the centered data matrix.
- X^T_{centered} is the transpose of X_{centered}.
- "·" represents matrix multiplication.

The resulting matrix S is a symmetric matrix of dimensions m × m that contains the covariances between all combinations of variables.

Step 7: Select the principal components

1. Once we have calculated the covariance matrix and its corresponding eigenvectors and eigenvalues, the next step is to select the principal components. We sort the eigenvectors according to their corresponding eigenvalues in decreasing order.



Example

Step 8: Data Transformation

2. Once we have selected the N eigenvectors, we multiply the normalized data matrix by these N eigenvectors to obtain the new representation of the data in the principal component space.

$$X_{\text{transformed}} = X_{\text{normalized}} \times V$$

Where:

- $X_{\text{transformed}}$ is the transformed data in the principal component space.
- $X_{\text{normalized}}$ is the normalized data matrix.
- V is the matrix of selected eigenvectors.

3. This new representation in the principal component space can be utilized for planning and navigation of the robot in the environment. By reducing the dimensionality of the data while retaining the relevant information for the robot's decisions,