Aprendizagem de Máquina
Não Supervisionada

Redução de Dimensionalidade

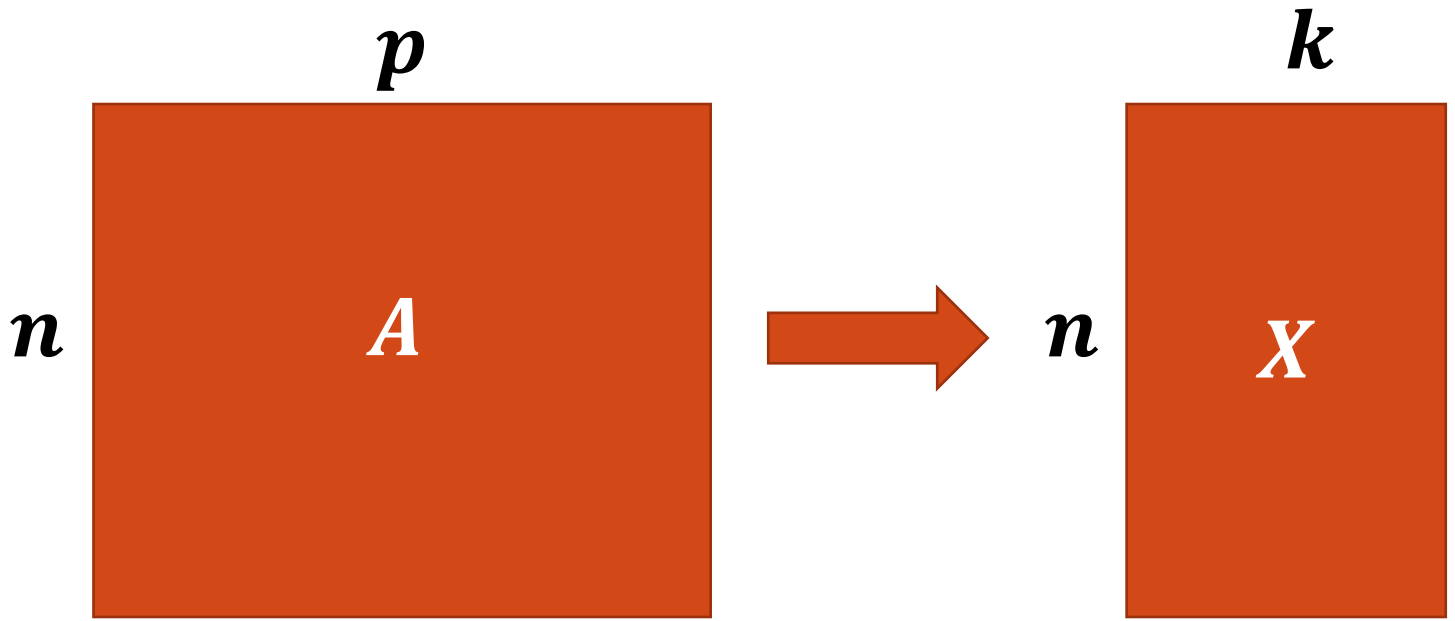
Prof. Leandro Balby Marinho
http://leandro.lsd.ufcg.edu.br

Redução de Dimensionalidade

- Alguns atributos podem não ser tão importantes quanto outros.
- Alguns atributos podem estar correlacionados entre si (redundância).
- Alta dimensionalidade leva a baixa performance de algoritmos de aprendizagem.
- Após 3 dimensões não conseguimos mais visualizer facilmente os dados e resultados das analises.
- **Redução de dimensionalidade comprime um grande conjunto de atributos em um subespaço de menor dimensão sem perderas informações importantes.**
- Principais tipos:
  - Seleção de atributos
  - **Métodos de Compressão**

Redução de Dimensionalidade

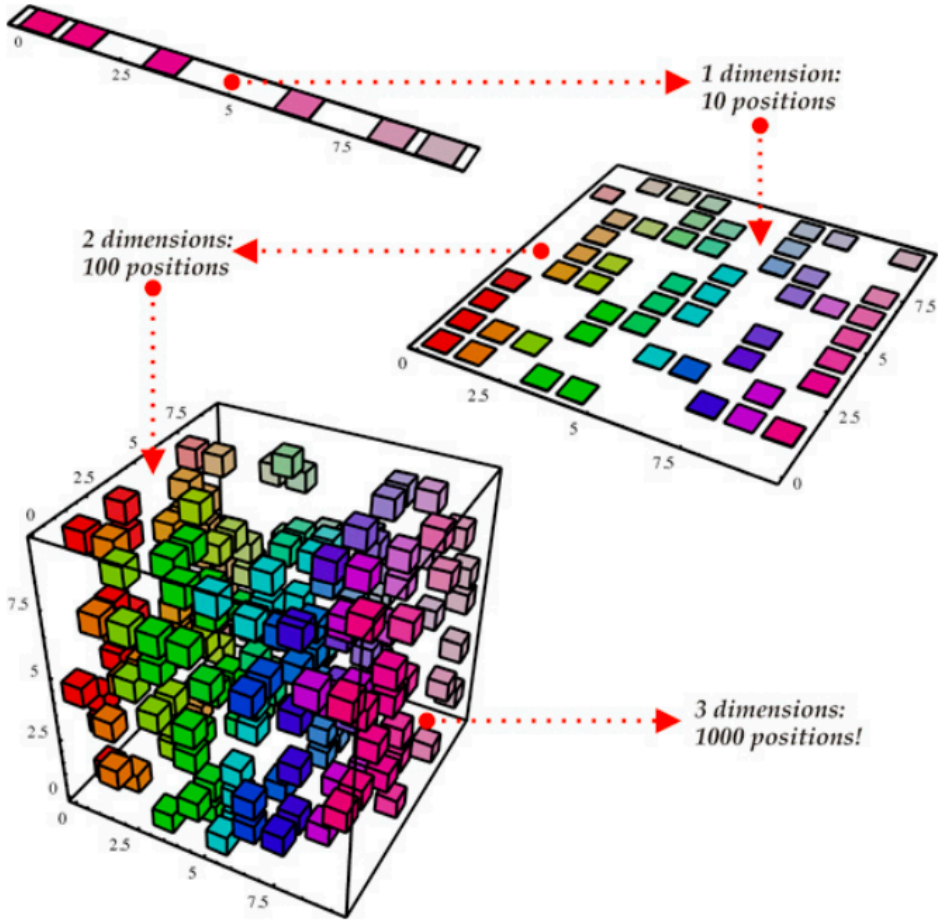
Sumarização de dados com (p) variáveis por um subconjunto menor de (k) variáveis derivadas.



Redução de Dimensionalidade

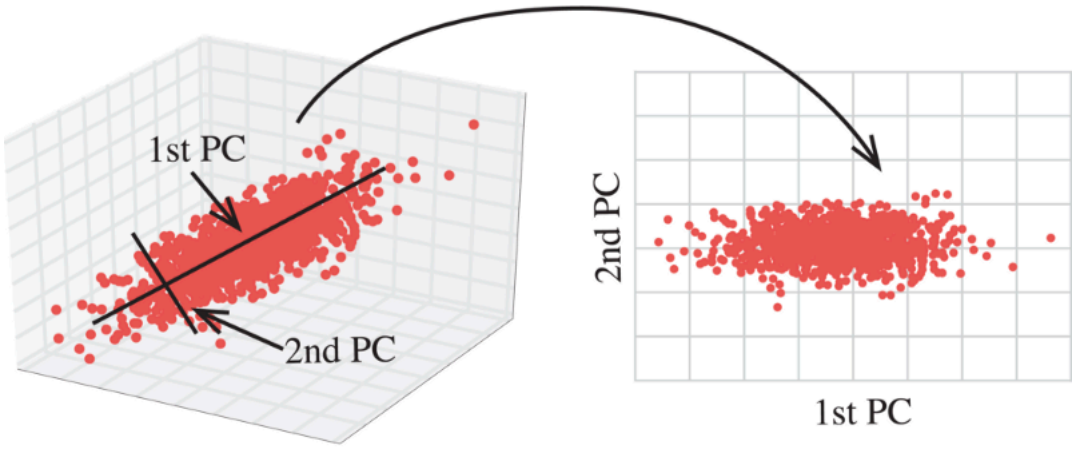
Trade-off entre:

- Clareza da representação, fácil entendimento
- Super simplificação: perda de informação relevante.

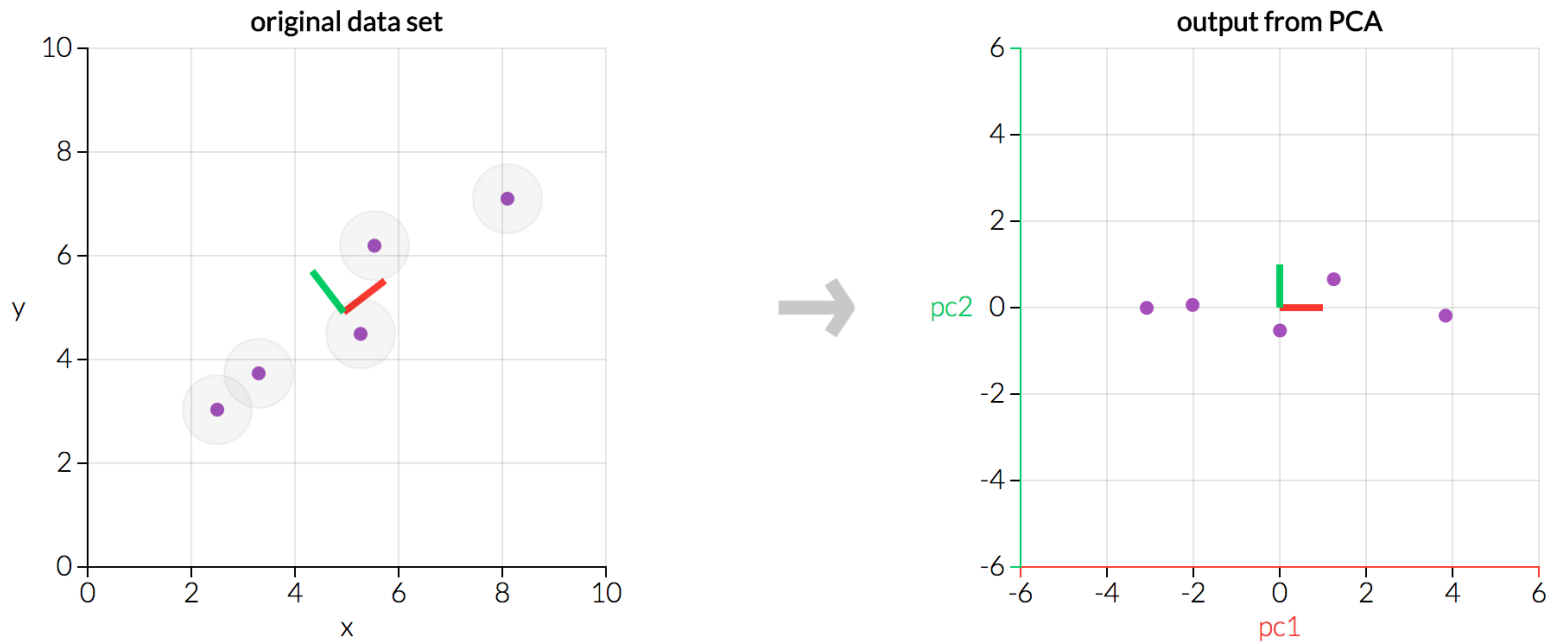


Análise de Componentes Principais (PCA)

- Criado por Pearson (1901) e Hotelling (1933).
- Recebe uma matriz de n objetos por p variáveis e a sumariza por variáveis não correlacionadas (componentes principais) que são combinações lineares das p variáveis originais.
- Os primeiros k componentes incorporam tanto quanto possível a variação entre os objetos.

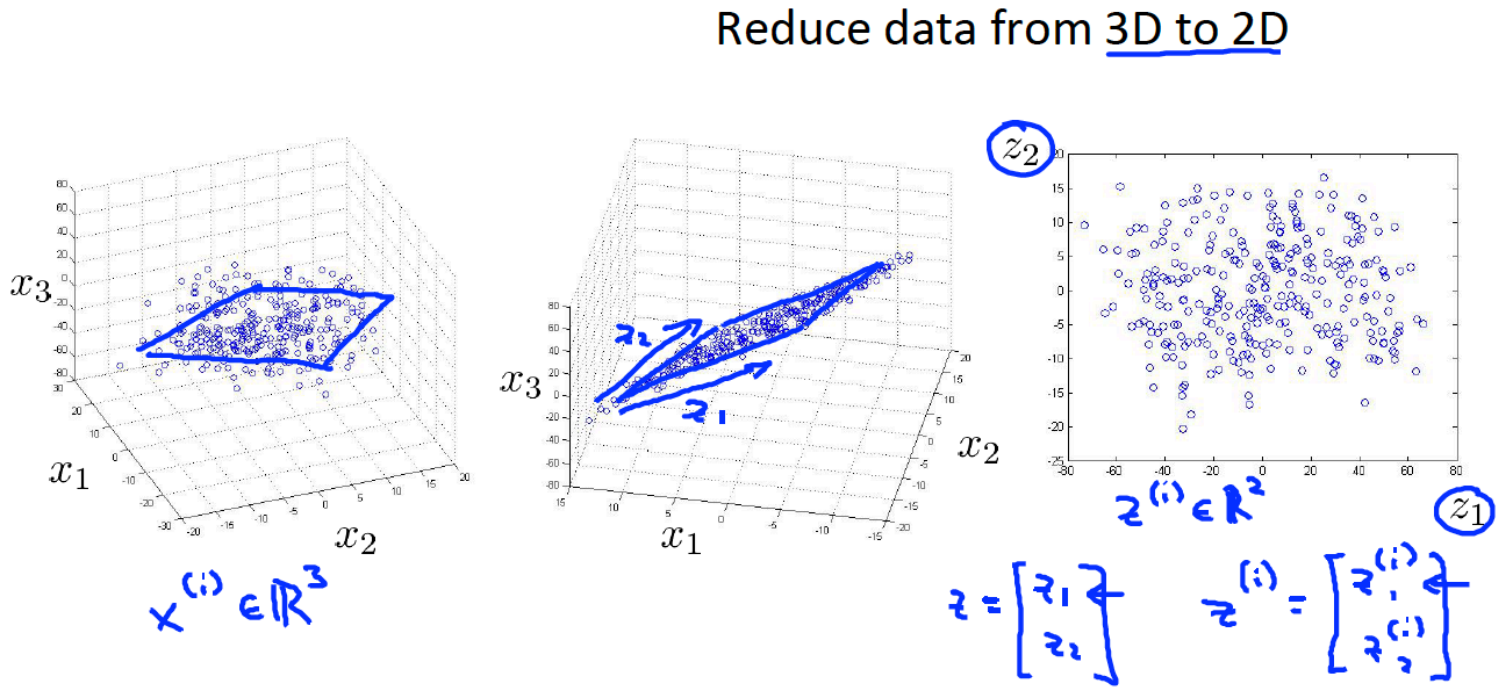


Definição do Problema



- Reduzir de 2-dimensões para 1-dimensão: Encontrar a direção (um vetor $u^{(1)} \in \mathbb{R}^n$) para onde projetar os dados de forma a minimizar o erro de projeção.

Definição do Problema

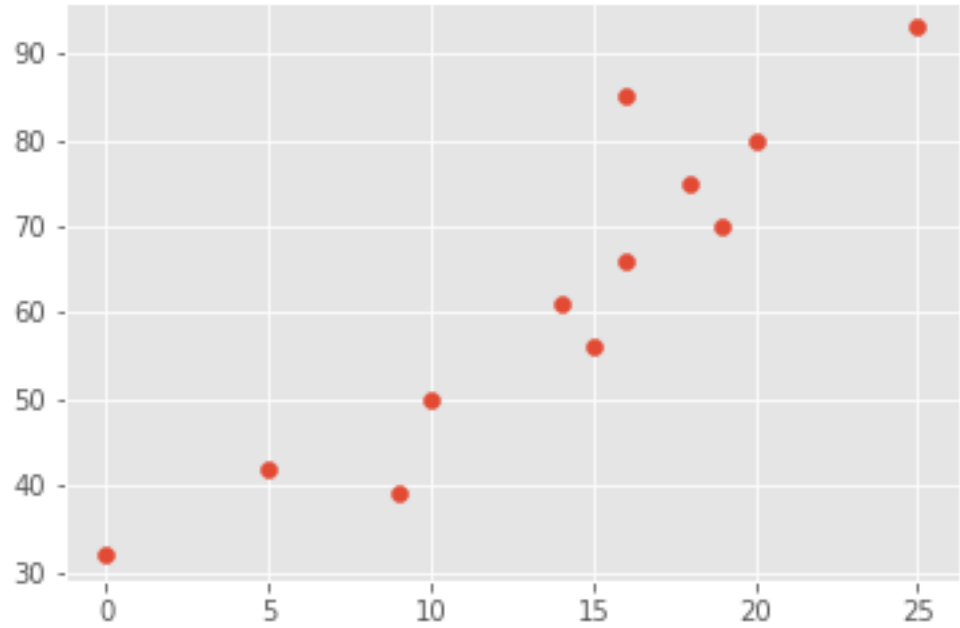


- Reduzir de n-dimensões para k-dimensões: Encontrar k vetores $u^{(1)}, u^{(2)}, \cdots, u^{(k)}$ para onde projetar os dados de forma a minimizar o erro de projeção.

Passos do Algoritmo

1. Entrada: Matriz de dados X:

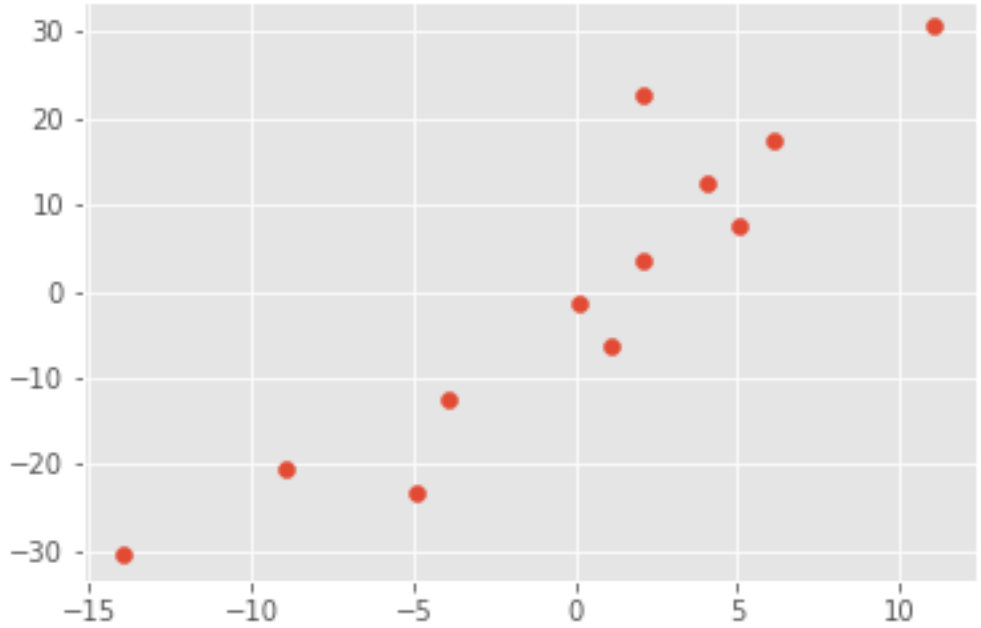
	Hours(H)	*Mark(M)*
Data	9	39
	15	56
	25	93
	14	61
	10	50
	18	75
	0	32
	16	85
	5	42
	19	70
	16	66
	20	80
Totals	167	749
Averages	13.92	62.42



Passos do Algoritmo

2. Para cada coluna, substitua cada entrada pelo valor da entrada menos a média da coluna. Seja essa nova matriz Z.

H	M	$(H_i - \bar{H})$	$(M_i - \bar{M})$
9	39	-4.92	-23.42
15	56	1.08	-6.42
25	93	11.08	30.58
14	61	0.08	-1.42
10	50	-3.92	-12.42
18	75	4.08	12.58
0	32	-13.92	-30.42
16	85	2.08	22.58
5	42	-8.92	-20.42
19	70	5.08	7.58
16	66	2.08	3.58
20	80	6.08	17.58
Total			
Average			



Passos do Algoritmo

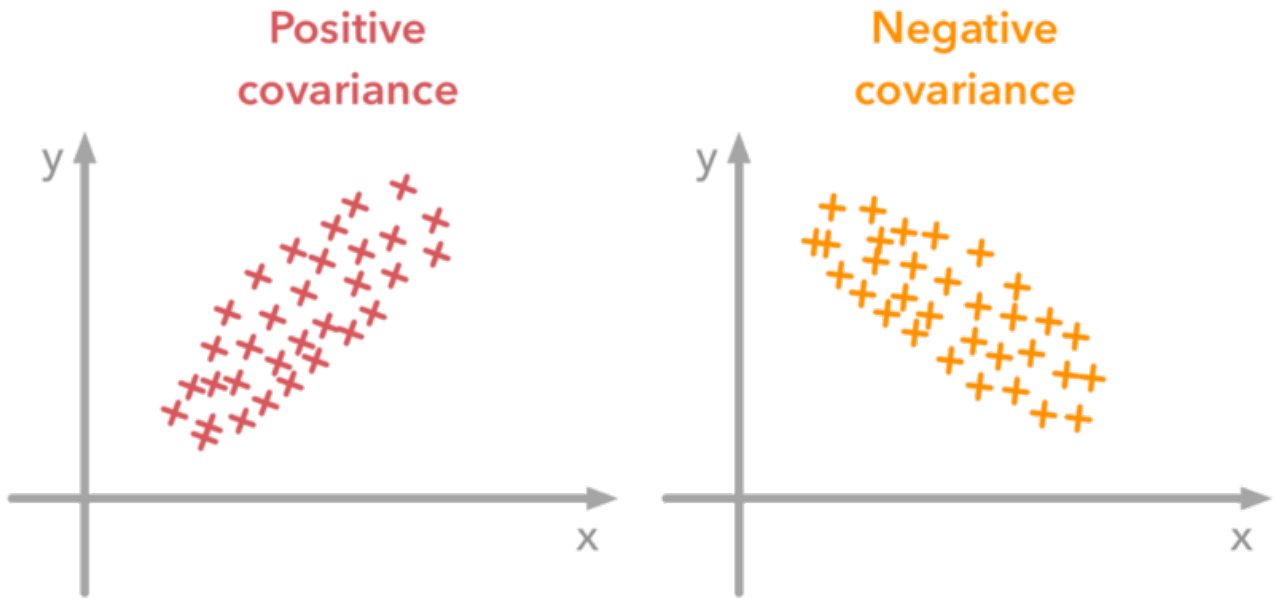
3. Calcule a matriz de covariância a partir de Z.

$$\begin{pmatrix} 47.7 & 122.9 \\ 122.9 & 370 \end{pmatrix}$$

Co-Variância

Calcula como duas variáveis variam em conjunto:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$



Co-Variância

Calcula a varição conjunto de duas variáveis: $cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{(n-1)}$

H	M	$(H_i - \bar{H})$	$(M_i - \bar{M})$	$(H_i - \bar{H})(M_i - \bar{M})$
9	39	-4.92	-23.42	115.23
15	56	1.08	-6.42	-6.93
25	93	11.08	30.58	338.83
14	61	0.08	-1.42	-0.11
10	50	-3.92	-12.42	48.69
18	75	4.08	12.58	51.33
0	32	-13.92	-30.42	423.45
16	85	2.08	22.58	46.97
5	42	-8.92	-20.42	182.15
19	70	5.08	7.58	38.51
16	66	2.08	3.58	7.45
20	80	6.08	17.58	106.89
Total				1149.89
Average				104.54

Matriz de Co-Variância

Calcula todas as co-variâncias possíveis entre todas as variáveis. Por exemplo, para três variáveis:

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

Passos do Algoritmo

4. Calcule os autovetores e autovalores da matriz de co-variância.

- A intuição é o que os autovetores mais importantes (com os maiores autovalores) capturam a maior parte da variância.
- Normalmente ussa-se o método Decomposição de Valores Singulares (Singular Value Decomposition-SVD)
  - SVD(X) = UΣV* onde X (mxn), U(mxm), V*(nxn) e Σ (mxn) é uma matrix retangula contendo os autovalores na diagonal principal.

Passos do Algoritmo

5. Ordene os autovalores de acordo com sua importância e chame de matriz P.

6. Calcule os novos dados projetados: $Z^* = ZP$

Referências

- Data Mining: Concepts and Techniques: Jiawei Han, Jian Pei, Micheline Kamber. 2011
- Introduction to Data Mining: Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Anuj Karpatne. 2019
- Especialização online em Machine Learning: Universidade de Washington.