Curso: Fundamentos de Machine Learning

Aula 3

Tema da aula:

Aula 3 - Aprendizado não supervisionado

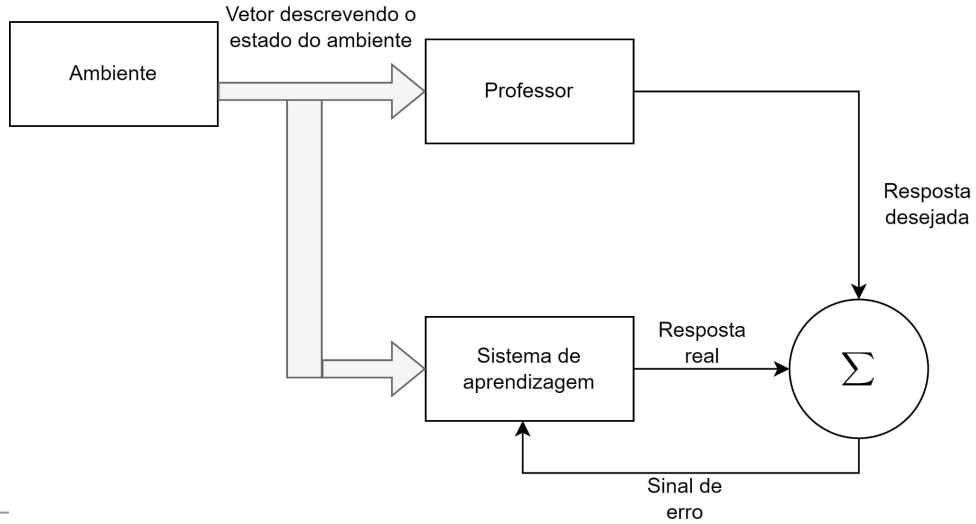
Conteúdo Programático

- Aprendizado não supervisionado;
- Técnicas de agrupamento;
- Conhecendo o K-Means;
- Técnicas de Redução de Dimensionalidade;
- Conhecendo o PCA.

© Copyright 2018-2019 www.infinityschool.com.br - All Rights Reserved

Aprendizado com professor

- Em termos conceituais, é possível considerar que o professor tem conhecimento sobre o ambiente. Com o conhecimento sendo representado como um conjunto de exemplos de entrada e saída (Haykin, S; 2008).



Aprendizado não supervisionado

- Na aprendizagem supervisionada, a aprendizagem acontece sob a supervisão de um professor. Entretanto, na aprendizagem não supervisionada ou sem professor, o aprendizado ocorre sem a presença de um professor, ou seja não há dados rotulados para que o algoritmo aprenda (Haykin, S; 2008).

Aprendizado não supervisionado

- O aprendizado não supervisionado consiste em treinar uma máquina a partir de dados que não estão rotulados e/ou classificados. Os algoritmos que fazem isso buscam descobrir padrões ocultos que agrupam as informações de acordo com semelhanças ou diferenças, por exemplo.

Figura 1: Aprendizado não supervisionado

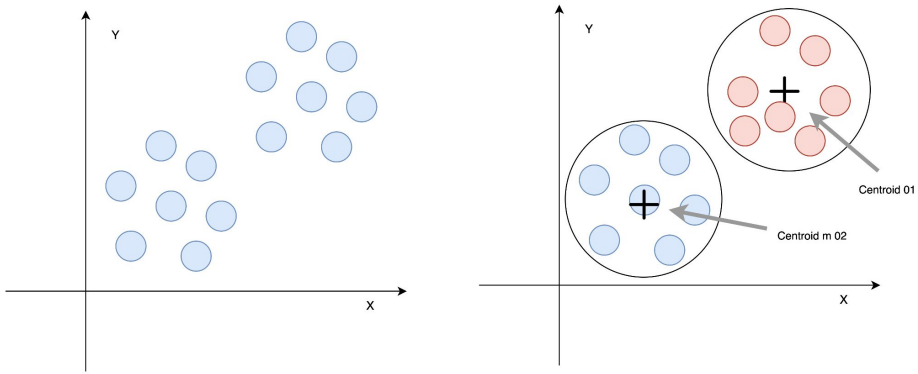
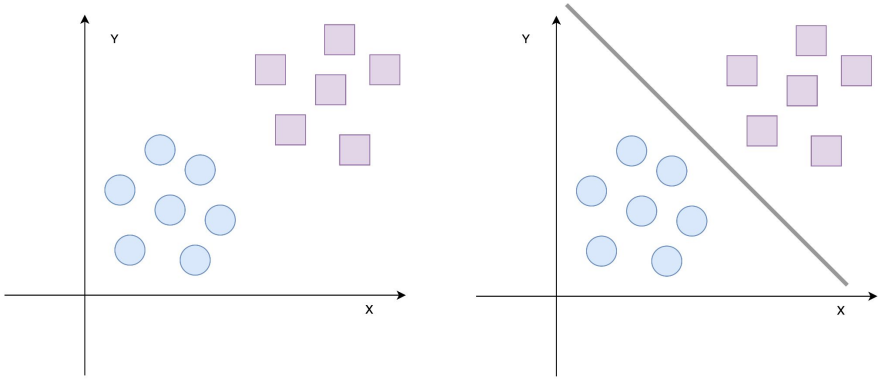
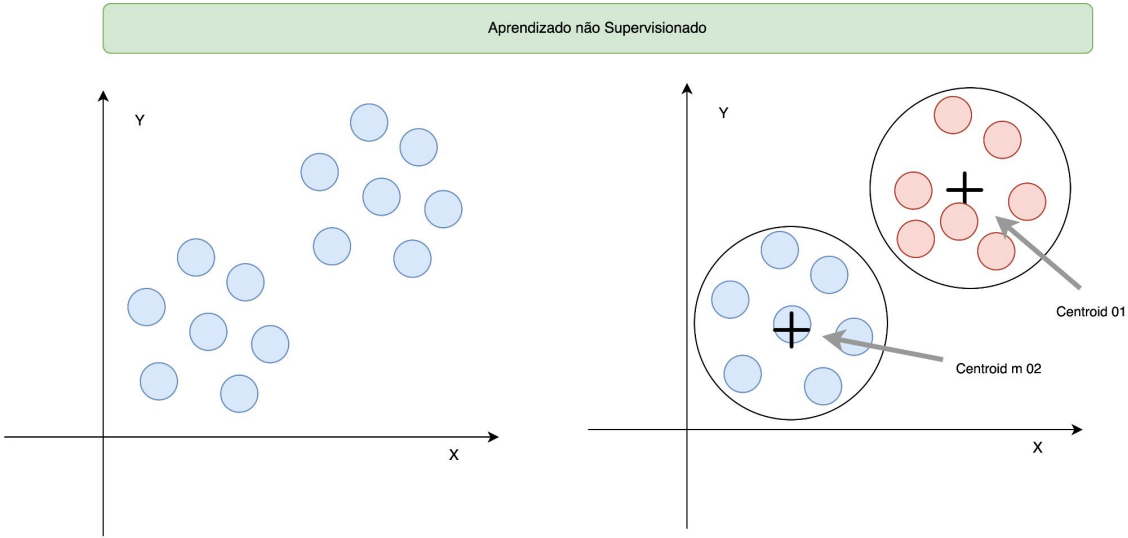


Figura 2: Aprendizado supervisionado



Aprendizado não supervisionado

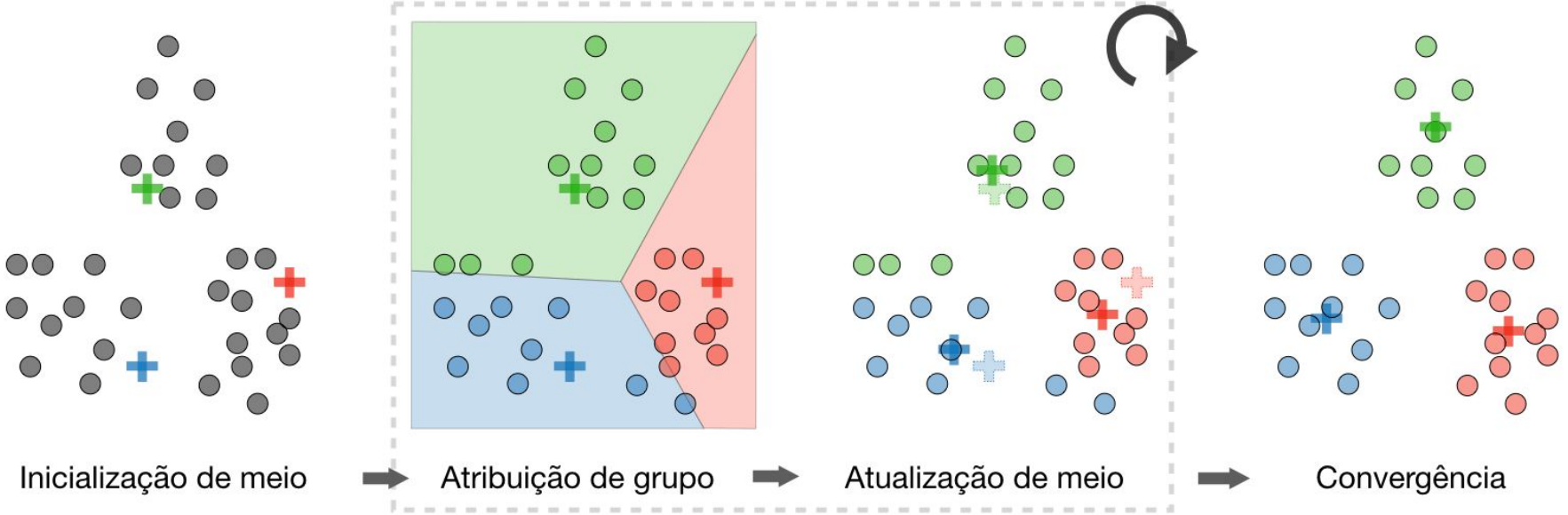
- **Agrupamento:** A técnica de agrupamento consiste em agrupar dados não rotulados com base em suas semelhanças ou diferenças..
- **Regras de Associação:** Ao usar as regras de associação, buscamos descobrir relações que descrevem grandes porções dos dados. A associação é muito utilizada em análises de cestas de compras, no qual a empresa pode tentar entender relações de preferências de compras entre os produtos.



Conhecendo o K-Means

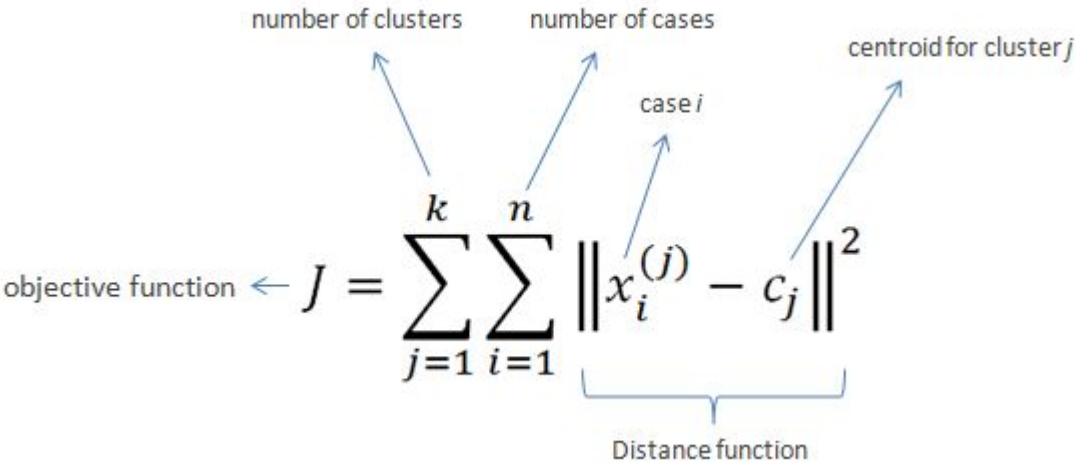
$$c^{(i)} = \arg\min_j ||x^{(i)} - \mu_j||^2 \quad \text{e} \quad \mu_j = \frac{\sum_{i=1}^{m} 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^{m} 1_{\{c^{(i)}=j\}}}$$

...



Conhecendo o K-Means

...



Conhecendo o K-Means

Algoritmo:

1. Agrupa os dados em k grupos onde k é predefinido;
2. Selecione k pontos aleatoriamente como centros de cluster;
3. Atribua objetos ao centro de cluster mais próximo de acordo com a função de distância euclidiana;
4. Calcule o centróide ou a média de todos os objetos em cada cluster;
5. Repita as etapas 3 e 4 até que os mesmos pontos sejam atribuídos a cada cluster em rodadas consecutivas.

Suponha que queremos agrupar os visitantes de um site usando apenas a idade (espaço unidimensional) da seguinte forma:

n= 19

15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61.65

Clusters iniciais (centróide aleatório ou média):

k = 2

c 1 = 16

c 2 = 22

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Iteração 1

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	16	22	1	7	1	
15	16	22	1	7	1	**15.33**
16	16	22	0	6	1	
19	16	22	9	3	2	
19	16	22	9	3	2	
20	16	22	16	2	2	
20	16	22	16	2	2	
21	16	22	25	1	2	
22	16	22	36	0	2	
28	16	22	12	6	2	
35	16	22	19	13	2	
40	16	22	24	18	2	**36.25**
41	16	22	25	19	2	
42	16	22	26	20	2	
43	16	22	27	21	2	
44	16	22	28	22	2	
60	16	22	44	38	2	
61	16	22	45	39	2	
65	16	22	49	43	2	

Exemplo extraído de: https://www.saedsayad.com/clustering_kmeans.htm

Iteração 2

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	15.33	36.25	0.33	21.25	1	
15	15.33	36.25	0.33	21.25	1	
16	15.33	36.25	0.67	20.25	1	
19	15.33	36.25	3.67	17.25	1	
19	15.33	36.25	3.67	17.25	1	**18.56**
20	15.33	36.25	4.67	16.25	1	
20	15.33	36.25	4.67	16.25	1	
21	15.33	36.25	5.67	15.25	1	
22	15.33	36.25	6.67	14.25	1	
28	15.33	36.25	12.67	8.25	2	
35	15.33	36.25	19.67	1.25	2	
40	15.33	36.25	24.67	3.75	2	
41	15.33	36.25	25.67	4.75	2	
42	15.33	36.25	26.67	5.75	2	
43	15.33	36.25	27.67	6.75	2	**45.9**
44	15.33	36.25	28.67	7.75	2	
60	15.33	36.25	44.67	23.75	2	
61	15.33	36.25	45.67	24.75	2	
65	15.33	36.25	49.67	28.75	2	

Exemplo extraído de: https://www.saedsayad.com/clustering_kmeans.htm

Iteração 3

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	18.56	45.9	3.56	30.9	1	
15	18.56	45.9	3.56	30.9	1	
16	18.56	45.9	2.56	29.9	1	
19	18.56	45.9	0.44	26.9	1	
19	18.56	45.9	0.44	26.9	1	
20	18.56	45.9	1.44	25.9	1	**19.50**
20	18.56	45.9	1.44	25.9	1	
21	18.56	45.9	2.44	24.9	1	
22	18.56	45.9	3.44	23.9	1	
28	18.56	45.9	9.44	17.9	1	
35	18.56	45.9	16.44	10.9	2	
40	18.56	45.9	21.44	5.9	2	
41	18.56	45.9	22.44	4.9	2	
42	18.56	45.9	23.44	3.9	2	
43	18.56	45.9	24.44	2.9	2	**47.89**
44	18.56	45.9	25.44	1.9	2	
60	18.56	45.9	41.44	14.1	2	
61	18.56	45.9	42.44	15.1	2	
65	18.56	45.9	46.44	19.1	2	

Exemplo extraído de: https://www.saedsayad.com/clustering_kmeans.htm

Iteração 4

x_i	c_1	c_2	Distance 1	Distance 2	Nearest Cluster	New Centroid
15	19.5	47.89	4.50	32.89	1	
15	19.5	47.89	4.50	32.89	1	
16	19.5	47.89	3.50	31.89	1	
19	19.5	47.89	0.50	28.89	1	
19	19.5	47.89	0.50	28.89	1	
20	19.5	47.89	0.50	27.89	1	**19.50**
20	19.5	47.89	0.50	27.89	1	
21	19.5	47.89	1.50	26.89	1	
22	19.5	47.89	2.50	25.89	1	
28	19.5	47.89	8.50	19.89	1	
35	19.5	47.89	15.50	12.89	2	
40	19.5	47.89	20.50	7.89	2	
41	19.5	47.89	21.50	6.89	2	
42	19.5	47.89	22.50	5.89	2	
43	19.5	47.89	23.50	4.89	2	**47.89**
44	19.5	47.89	24.50	3.89	2	
60	19.5	47.89	40.50	12.11	2	
61	19.5	47.89	41.50	13.11	2	
65	19.5	47.89	45.50	17.11	2	

Exemplo extraído de: https://www.saedsayad.com/clustering_kmeans.htm

Vamos a prática

1. Baixe (ou faça a leitura online com pandas) o dataset iris UCI (Disponível em https://archive.ics.uci.edu/ml/datasets/iris);
2. Divida o *dataset* em duas variáveis (x e y), **x** deve conter os dados de características (sepal length in cm, sepal width in cm, petal length in cm, petal width in cm) e na variável **y** deve conter a classe (ou target) que vai indicar qual das três classes (Iris Setosa, Iris Versicolour, Iris Virginica) cada dado representa;
3. Para simplificar o problema, redefina a variável **x**, pegando apenas as 100 primeiras linhas e mantendo apenas a primeira e terceira coluna (sepal length in cm, petal width in cm);
4. Implemente uma função para calcular a distância euclidiana, dado uma entrada de x,y;
5. Implementar uma função para calcular os centróides dados uma lista de grupos.

https://colab.research.google.com/drive/17S31iDkgl_yE1nI3xALYFOsmInx1VZWy?usp=sharing

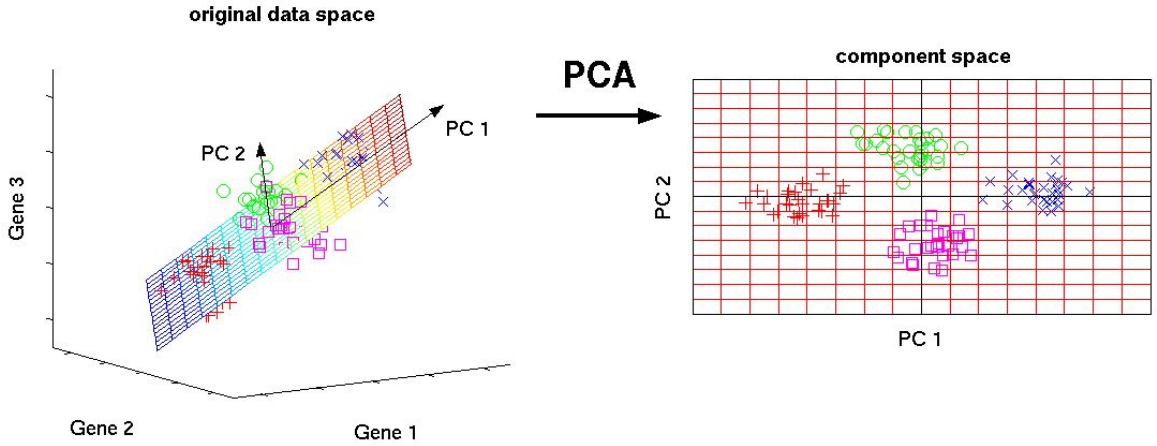
Vamos a prática

1. Usando a mesma divisão dos dados anteriores, acesse a documentação do sklearn e implemente o algoritmo de agrupamentos K-means.

O que é redução de dimensionalidade ?

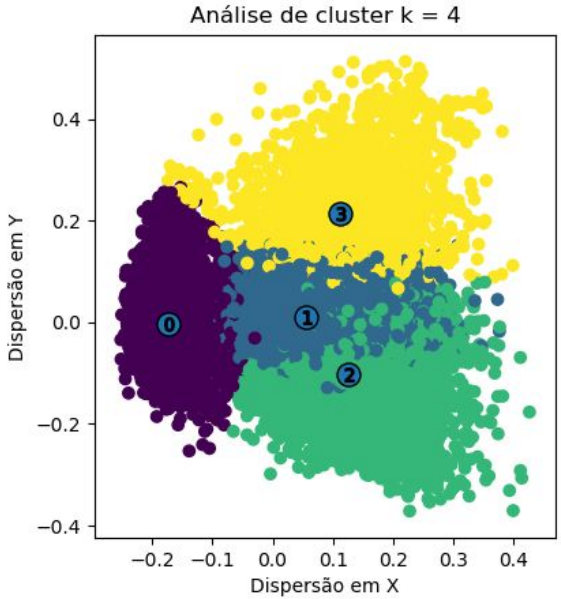
A redução da dimensionalidade (features) é o processo de reduzir o número de variáveis aleatórias que serão inseridas em um modelo para treino.

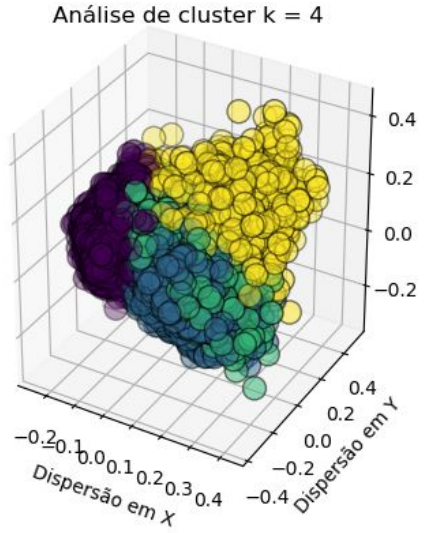
Imagine que você possua um dataset com centenas de colunas(features), a redução de dimensionalidade traria as dimensões para um número mais fácil de se trabalhar, algumas poucas dezenas por exemplo. Seria como converter uma esfera de três dimensões para um círculo de duas dimensões.



O que é redução de dimensionalidade ?

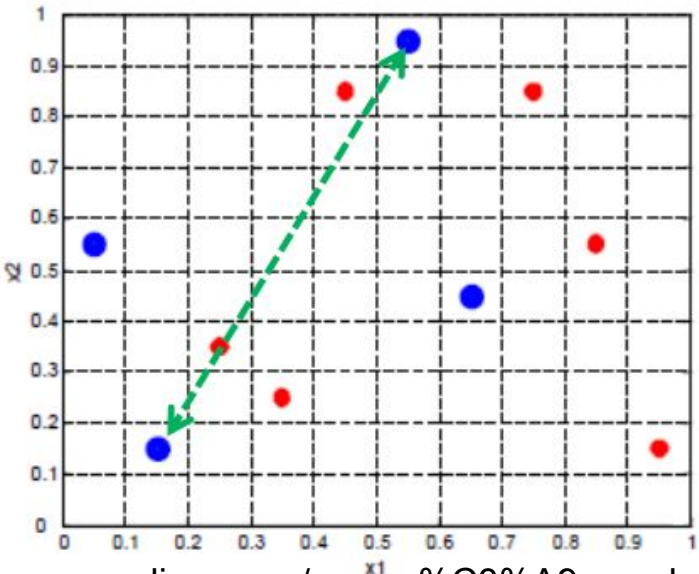
Para fazer isso, precisamos identificar quais são as variáveis principais (ou seja, mais importantes). Chamamos de Principal Components o conjunto de variáveis que não são linearmente correlacionadas.

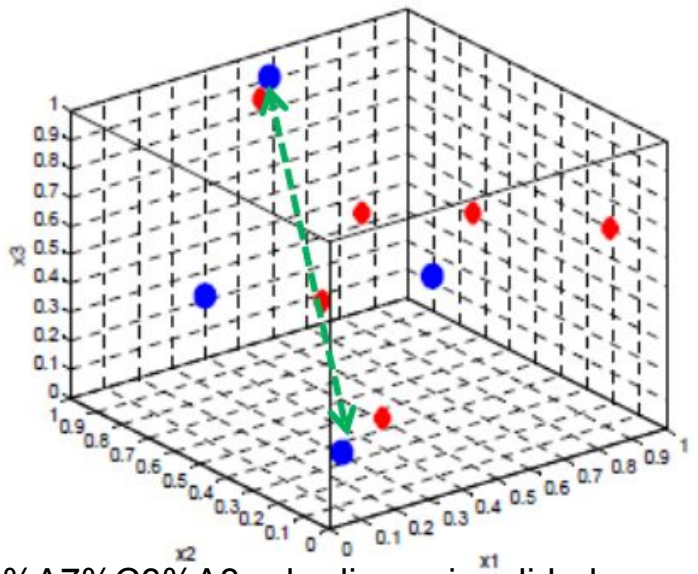




Por que realizar esse procedimento?

Quando o **número** de features aumenta, o **número de amostras** precisa aumentar também para que o número de combinações entre features e classes seja **satisfatório**. Isso faz com que o modelo fique cada vez mais **complexo.**





https://ealexbarros.medium.com/o-que-%C3%A9-a-redu%C3%A7%C3%A3o-de-dimensionalidade-em-machine-learning-cc2a89e3cdec

Vamos a prática

1. Usando a biblioteca sklearn, faça a implementação do método de análise de componentes principais, usado para redução de dimensionalidade usando o dataset iris completo. Ou seja, transforme o dataset que contém quatro variáveis em um que contém duas.

71 3901 1052 | 71 9 9204 0134
@infinity.school
www.infinityschool.com.br
Salvador Shopping Business | Torre Europa Sala 310 Caminho das Árvores, Salvador - BA CEP: 40301-155