MACHINE LEARNING

Redução de dimensionalidade

cognitiza

Your Data & AI school

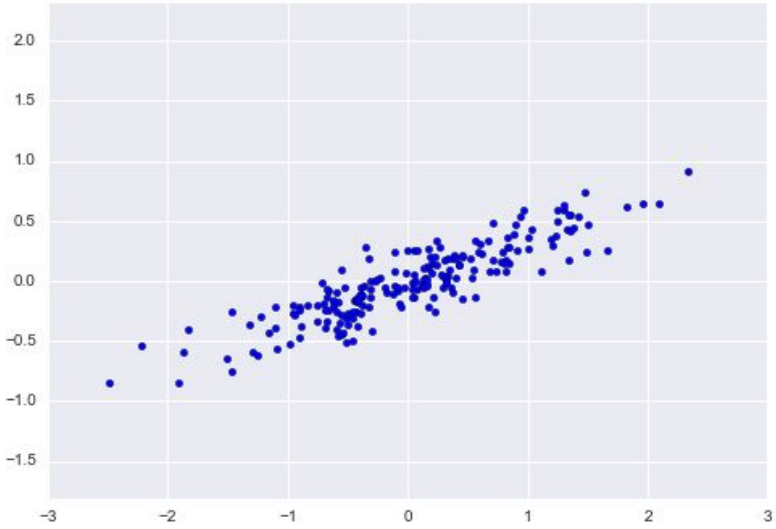
Prof. Altino Dantas

Redução de dimensionalidade | **Características**

- Existem vários algoritmos, tais como: PCA, LDA, FA;
- Método de aprendizado não-supervisionado;
- Utilizado para facilitar a visualização de dados quando o dataset possui muitas dimensões;
- Útil para melhorar o desempenho de algoritmos de aprendizado supervisionado;
- Remoção de ruído e extração de características são outras aplicações do PCA.

Redução de dimensionalidade | **PCA**

- Comportamento fácil de se entendido para duas dimensões;
- Aprendizado não-supervisionado está interessado encontrar a relação entre variáveis e não predizer uma com base em outras;
- No PCA, isso é feito encontrando-se uma lista de eixos principais que expliquem os dados.



Redução de dimensionalidade | **PCA**

- Na mplemetação do sklearn, o PCA pode ser executado da seguinte forma:

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
pca.fit(X)
```

- E produz como resultado os componentes principais e a variância de cada eixo.

```
print(pca.components_)
```

```
[[ 0.94446029  0.32862557]
 [ 0.32862557 -0.94446029]]
```

```
print(pca.explained_variance_)
```

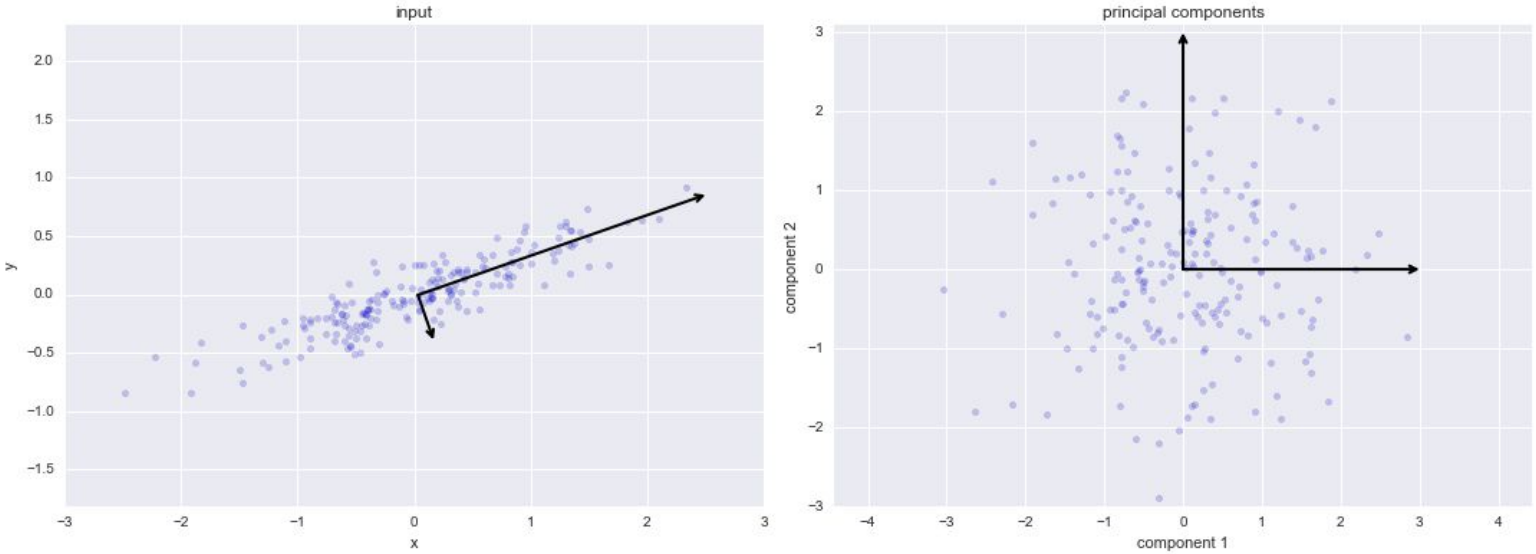
```
[ 0.75871884  0.01838551]
```

Redução de dimensionalidade | **PCA**

- Os vetores representam os eixos principais dos dados e o comprimento é a indicação do quão importante o eixo é na descrição da distribuição dos dados, ou seja, é a variância dos dados naquele eixo;
- A projeção de cada ponto de dados nos eixos principais são os "componentes principais" dos dados;

Redução de dimensionalidade | **PCA**

Os dados projetados nas dimensões originais ao lado dos dados projetados nos componentes principais (através de uma transformação);



Redução de dimensionalidade | **PCA**

- Vejamos o exemplo das espécies no dataset Iris;
- Este dataset possui quatro características e uma variáveis dependente;

	sepal length	sepal width	petal length	petal width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

Redução de dimensionalidade | **PCA**

- Para a tarefa de redução de dimensionalidade, dispensa-se a variável dependente;
- Como o PCA é muito sensível à escala dos dados entre os eixos, os dados precisam ser normalizados:

	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
4	5.0	3.6	1.4	0.2

Standardization

	sepal length	sepal width	petal length	petal width
0	-0.900681	1.032057	-1.341272	-1.312977
1	-1.143017	-0.124958	-1.341272	-1.312977
2	-1.385353	0.337848	-1.398138	-1.312977
3	-1.506521	0.106445	-1.284407	-1.312977
4	-1.021849	1.263460	-1.341272	-1.312977

Redução de dimensionalidade | **PCA**

- Depois de inferir os componentes principais, os dados originais podem ser transformados;

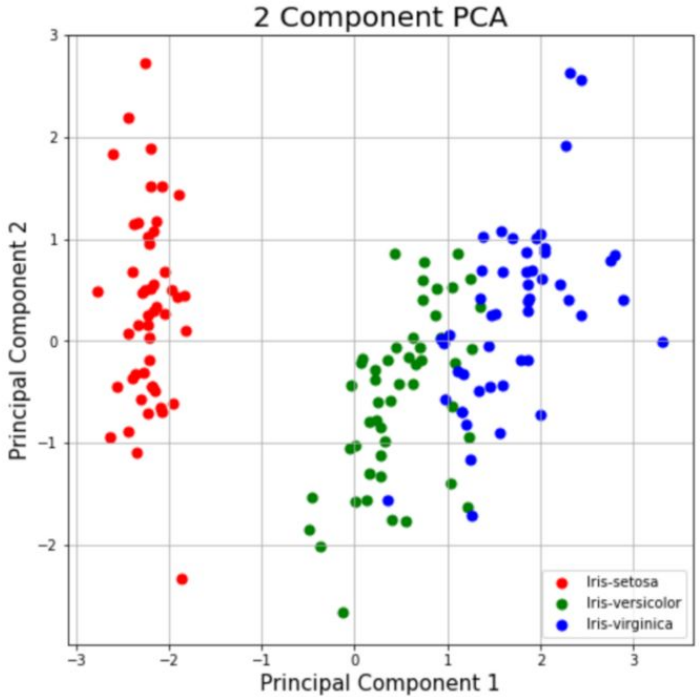
	sepal length	sepal width	petal length	petal width
0	-0.900681	1.032057	-1.341272	-1.312977
1	-1.143017	-0.124958	-1.341272	-1.312977
2	-1.385353	0.337848	-1.398138	-1.312977
3	-1.506521	0.106445	-1.284407	-1.312977
4	-1.021849	1.263460	-1.341272	-1.312977

PCA
(2 components)

	principal component 1	princial component 2
0	-2.264542	0.505704
1	-2.086426	-0.655405
2	-2.367950	-0.318477
3	-2.304197	-0.575368
4	-2.388777	0.674767

Redução de dimensionalidade | **PCA**

- No espaço 2-dimensional obtido pelo PCA, é possível projetar as espécies;
- Percebe-se que a espécie setosa pode ser facilmente distinguida das demais;



Redução de dimensionalidade | **PCA**

- A variância de cada componente explica o quão importante ele é para a distribuição dos dados originais;
- No processo de redução de dimensões, é esperado que haja perdas, logo, a soma das variâncias explicadas pelos componentes tende a ser menor do 100%;
- Costumamos usar a taxa de variância como critério para a escolha da quantidade de componentes a considerar.

Redução de dimensionalidade | **PCA**

O link abaixo apresenta outras aplicações de PCA

https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

Obrigado.