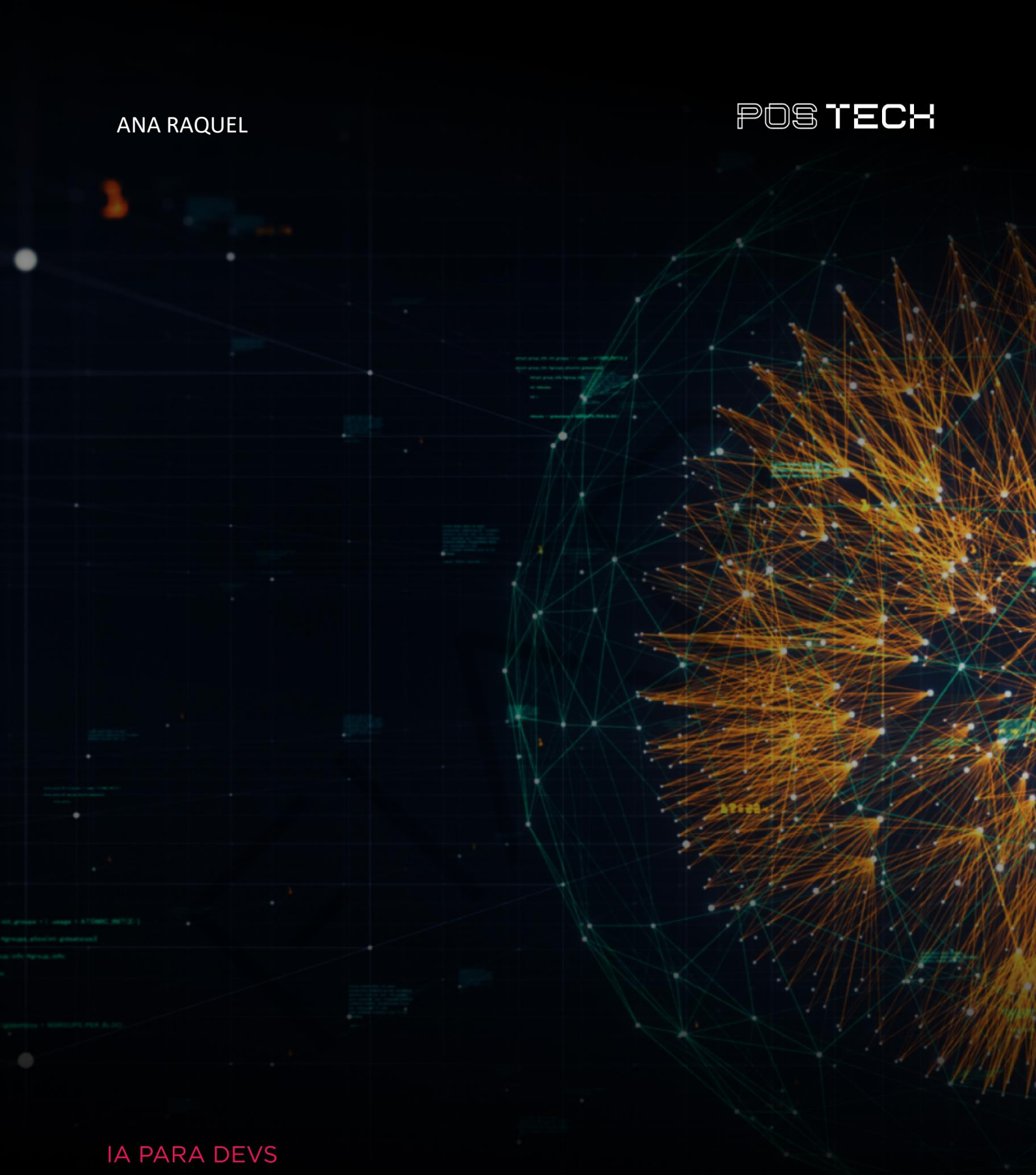
ANA RAQUEL

PDS TECH



IA PARA DEVS

MACHINE LEARNING

AULA 03

SUMÁRIO

O QUE VEM POR AÍ? 3
HANDS ON 4
SAIBA MAIS 5
O QUE VOCÊ VIU NESTA AULA? 13
REFERÊNCIAS 14

O QUE VEM POR AÍ?

Você já parou para pensar na dimensão das bases de dados que são trabalhadas em modelos de aprendizado de máquina? Parece algo simples, mas não é somente ao volume de instâncias (linhas) da base que devemos nos atentar, mas também no número de features, que também pode ser um ponto importante de ser analisado no aprendizado de máquina. Nesta aula, você aprenderá sobre um problema muito comum no mundo da ciência de dados: o problema da alta dimensionalidade.

HANDS ON

Você vai aprender na prática como empregar uma das técnicas mais utilizadas da estatística para reduzir a dimensionalidade dos dados: o PCA (Análise de Componentes Principais). Vamos lá?!

SAIBA MAIS

O problema da alta dimensionalidade

Você conhece o problema da alta dimensionalidade? Também conhecido como a maldição da alta dimensionalidade, este tipo de problema ocorre quando a base de dados utilizada em um aprendizado de máquina possui um número muito grande de features (colunas). Esse alto volume de variáveis pode ser um grande problema, pois quanto mais dimensões (linhas e colunas) possui o modelo, mais esparsos são os dados, e maior é o risco de overfitting (nome dado quando o algoritmo decora os dados de treinamento e não sabe generalizar novos casos). A redução de dimensionalidade entra na parte de técnicas não supervisionadas, na qual não queremos rotular os dados dado uma variável alvo, mas sim agrupar esses dados a fim de criar uma representação compacta desse grande volume de features.

Normalmente, quando estamos criando um modelo preditivo, na etapa de feature engineering selecionamos as melhores variáveis para inserir no modelo que melhor responde o problema alvo. Nesses casos, mesmo analisando as features, podemos enfrentar uma situação em que vamos precisar utilizar todas as colunas da base de dados no aprendizado de máquina. Por exemplo, imagine que você tem uma base de dados com 30 colunas para criar um modelo de classificação. Vamos supor que você estudou todas as variáveis e observou que será preciso utilizar todas no seu modelo de Machine Learning. Se criar um modelo preditivo supervisionado com esse volume de features, você pode cair em overfitting. Para solucionar este cenário, poderíamos pensar aqui em técnicas de redução de dimensionalidade.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	Target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

Queremos criar esssa tabela de representação compactada

	PC1	PC2	target
0	-2.264703	0.480027	0
1	-2.080961	-0.674134	0
2	-2.364229	-0.341908	0
3	-2.299384	-0.597395	0
4	-2.389842	0.646835	0

Figura 1 - Exemplo de redução de dimensionalidade
Fonte: elaborado pela autora (2024)

Existem muitas técnicas para solucionar esse tipo de problema, tais como:

- ✓ Análise de Componentes Principais (PCA).
- ✓ T-SNE (t-Distributed Stochastic Neighbor Embedding).
- ✓ Autoencoders.
- ✓ Kernel PCA (KPCA).
- ✓ Linear Discriminant Analysis (LDA).
- ✓ LLE (Locally Linear Embedding).
- ✓ Isomap.

Você deve estar se perguntando: mas como essas técnicas criam uma representação compacta desses dados? Como é possível criar essas representações e manter a originalidade dos dados? Bem, nesta aula será apresentada uma das técnicas mais importantes e mais utilizadas para reduzir dimensionalidade: a Análise de Componentes Principais (PCA). Acompanhe comigo os próximos passos para entender o funcionamento desta técnica tão importante na ciência e análise de dados.

Análise de Componentes Principais (PCA)

Na área de análise de dados, frequentemente nos deparamos com o desafio de lidar com uma grande quantidade de informações. Quando iniciamos um projeto de análise, especialmente na fase de exploração, muitas vezes nos deparamos com um grande número de características interessantes que podem ser utilizadas para treinar modelos, criar dashboards ou identificar padrões. Surge a pergunta: quais características são realmente relevantes para a análise?

É nesse contexto que entra o PCA (Análise de Componentes Principais), uma técnica introduzida por Karl Pearson, em 1909. O principal objetivo do PCA é reduzir a complexidade de conjuntos de dados com diversas variáveis, ajudando a identificar as características mais essenciais, ou seja, aquelas que capturam a maior parte da variação nos dados (ou o quão diferentes os dados estão da média).

Em resumo, o PCA possibilita uma **representação mais concisa dos dados, eliminando informações menos cruciais e mantendo as que têm maior importância para a análise e interpretação**. É importante notar que o PCA não exclui

diretamente as características que estamos analisando, mas sim nos auxilia a focar nas mais relevantes para os nossos objetivos.

Funcionamento do PCA

A ideia do PCA é combinar múltiplas variáveis preditoras numéricas em um conjunto menor de variáveis, que são combinações lineares ponderadas do conjunto original. O menor conjunto de variáveis, os componentes principais, “explica” a maior parte da variabilidade do conjunto completo de variáveis, reduzindo assim a dimensão dos dados.

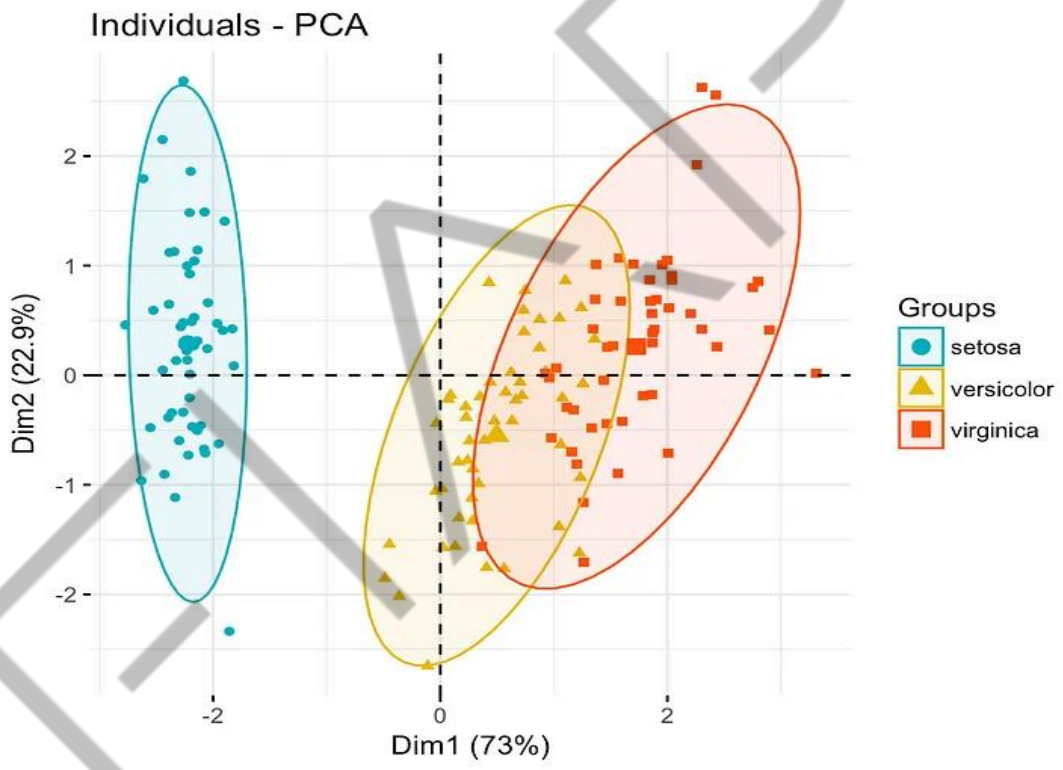


Figura 2 - Exemplo de componentes principais
Fonte: Medium (2020)

Padronização de variáveis no PCA

A padronização de variáveis é um processo crucial ao aplicar a Análise de Componentes Principais (PCA), uma técnica sensível à escala. Se as unidades de medida das variáveis diferem, aquelas com magnitudes mais elevadas podem distorcer a análise.

Por exemplo, em um conjunto de dados com distâncias em metros e valores em milhares, a segunda variável, devido à sua amplitude maior, pode ter uma influência desproporcional na variabilidade total durante o PCA. A padronização, frequentemente realizada por meio da normalização z, é um procedimento que ajusta as variáveis de um conjunto de dados para que todas possuam uma média de zero e um desvio padrão de um. A normalização z transforma cada variável de forma que sua distribuição seja centrada na média zero e escalada pelo desvio padrão.

Métrica estatística sobre qualidade da minha redução de componentes

A técnica do PCA parece muito interessante! No entanto, como posso ter certeza de que o PCA não está omitindo informações essenciais para o meu modelo?

Para esclarecer essa preocupação, utilizamos uma técnica chamada "Razão de Variância Explicada" (ou Explained Variance Ratio). Essa métrica estatística quantifica o quanto de variação em um conjunto de dados pode ser atribuído a cada um dos componentes principais (que são criados pelo PCA) em relação à média. Isso é fundamental, pois permite classificar os componentes por ordem de importância e concentrar a atenção dos mais significativos ao interpretar os resultados da análise.

A variância explicada é uma ferramenta que ajuda a escolher quantas dimensões devem ser mantidas em um conjunto de dados reduzido. Além disso, ela também pode ser usada para avaliar a qualidade de um modelo de aprendizado de máquina. Em geral, um modelo com uma alta variância explicada tende a ter um bom poder de previsão, enquanto um modelo com uma baixa variância explicada pode não ser tão preciso. No entanto, é importante lembrar que a interpretação depende da pergunta que estamos tentando responder com os dados.

Imagine que você possui um conjunto de dados distribuídos no espaço. A Análise de Componentes Principais (PCA) equivale a encontrar a melhor linha reta onde você pode projetar esses pontos de forma que fiquem o mais próximos possível dessa linha. Essa linha é o que chamamos de primeiro componente principal. Se desejar adicionar outra dimensão, é possível encontrar a melhor linha reta que seja perpendicular à primeira, onde você pode projetar os pontos. Isso corresponderia ao segundo componente principal, e o processo continua para mais dimensões. Cada

uma dessas "linhas" captura uma parte da "informação" contida nos seus dados, e isso é mensurado pela variância ao longo dessa linha.

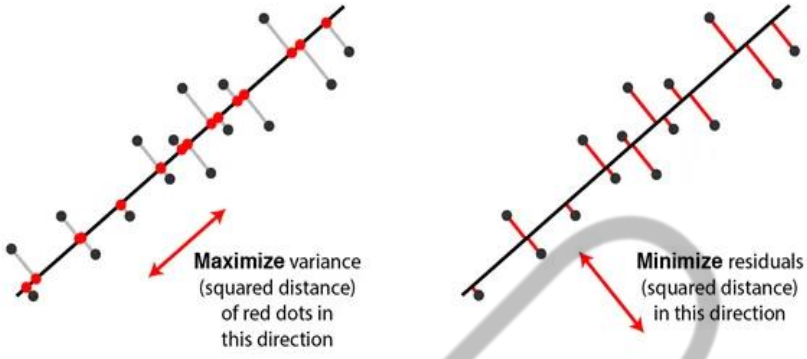


Figura 3 - Autovetor e Autovalor
Fonte: Medium (2019)

Essas linhas que representam as informações são denominadas autovetores. Cada autovetor possui um autovalor correspondente, que indica a quantidade de informação capturada por essa linha, ou seja, os pontos que ela atravessa.

Portanto, ao aplicar o PCA, nosso objetivo é identificar os melhores autovetores, onde cada autovalor representa uma determinada parcela de informação contida no nosso conjunto de dados.

PCA na prática com Python

Mas como podemos aplicar essa técnica utilizando a linguagem do Python? Para compreender a aplicação, acompanhe um exemplo que será realizado utilizando o dataset clássico da flor Íris. O conjunto de dados contém 3 classes de 50 instâncias cada, onde cada classe se refere a um tipo de planta de Íris.

```
# Vamos então trabalhar com um exemplo clássico para entendermos melhor
o conceito na prática!
# Vamos importar o dataset íris
! pip install scikit-learn
from sklearn import datasets
import pandas as pd

iris = datasets.load_iris()
df = pd.DataFrame(iris.data, columns=iris.feature_names)
df['Target'] = iris.get('target')
df.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	Target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

Figura 4 - Base de dados Iris
Fonte: elaborado pela autora (2024)

```
# A seguir, separaremos todas as colunas na lista de 'recursos' para
uma variável 'X' e a variável 'destino' para 'y'.
features = ['sepal length (cm)', 'sepal width (cm)', 'petal length
(cm)', 'petal width (cm)']
X = df[features].values
y = df['Target'].values
```

Depois de separar nossos dados em X para as variáveis características e Y para a target, vamos escalonar os dados. Como a Análise de Componentes Principais (PCA), é uma técnica sensível à escala, esse passo se torna fundamental!

```
from sklearn.preprocessing import StandardScaler
# Normalizando os dados utilizando o standardScaler
# (Padroniza as features removendo a média e escala a variância a uma
unidade.
# Isso significa que para cada feature, a média seria 0, e o Desvio
Padrão seria 1)
X = StandardScaler().fit_transform(X)
#Visualizando nossos dados padronizados
df_padronizado = pd.DataFrame(data=X, columns=features)
display(df_padronizado.head())
```

Agora, importaremos o PCA usando Sklearn e projetaremos nossos dados originais, que possuem 4 (quatro) dimensões, em 2 (duas) dimensões (ou também chamados de componentes principais).

```
# importando PCA da biblioteca sklearn
from sklearn.decomposition import PCA
# Instanciando o pca e a quantidade de componentes que desejamos obter
pca = PCA(n_components=2)
# Aplicando PCA nas nossas features
principalComponents = pca.fit_transform(X)

# Criando um novo dataframe para visualizarmos como ficou nossos dados
reduzidos com o PCA
df_pca = pd.DataFrame(data = principalComponents,
                  columns = ['PC1', 'PC2'])

target = pd.Series(iris['target'], name='target')
result_df = pd.concat([df_pca, target], axis=1)
result_df
```

	PC1	PC2	target
0	-2.264703	0.480027	0
1	-2.080961	-0.674134	0
2	-2.364229	-0.341908	0
3	-2.299384	-0.597395	0
4	-2.389842	0.646835	0

Figura 5 - Resultado dos componentes principais
Fonte: elaborado pela autora (2024)

Vamos olhar agora como ficou a variabilidade de cada componente principal, ou seja, o quanto cada coluna gerada a partir do PCA está próxima/distante da média dos nossos dados.

```
print('Variance of each component:', pca.explained_variance_ratio_)
print('\n Total Variance Explained:',
round(sum(list(pca.explained_variance_ratio_))*100, 2))
```

```
Variance of each component: [0.72962445 0.22850762]

 Total Variance Explained: 95.81
```

O que significam esses valores? Se você tem um PC1 de 0.72962445, isso significa que o primeiro componente principal da sua análise é responsável por

aproximadamente 72.96% da variância nos seus dados. Em outras palavras, esse único componente principal captura cerca de 72.96% da informação ou "estrutura" dos seus dados em termos de variância. Os restantes da variância são capturados pelos outros componentes principais (PC2, PC3 etc.).

Por que a soma não chega a 100%?

A razão pela qual não soma a 100% é devido à existência de mais componentes principais que explicam o restante da variância. No entanto, esses componentes adicionais são frequentemente desconsiderados na prática, pois contribuem muito pouco para a interpretação dos dados.

A ideia subjacente à PCA é reduzir a dimensionalidade dos seus dados, mantendo a maior parte da informação (variância). Portanto, geralmente é aceitável ignorar esses componentes adicionais se eles não contribuírem de forma significativa para a variância total.

Quantos componentes escolher para meu conjunto de dados?

Ao utilizar o PCA, você decide quantos componentes principais manter dos seus dados originais. A escolha depende de equilibrar a simplificação dos dados com a retenção da informação. Geralmente, você opta por um número que capture uma grande parte da variabilidade dos dados, como 95% ou 99%, evitando incluir componentes desnecessários. A decisão final depende das necessidades do seu projeto e dos objetivos específicos de análise.

O QUE VOCÊ VIU NESTA AULA?

Nessa aula, você aprendeu o que é a redução de dimensionalidade dos dados e como podemos utilizar a técnica de PCA (Análise de Componentes Principais) para explicar a maior parte da variabilidade do conjunto de variáveis, reduzindo, assim, a dimensão dos dados.

REFERÊNCIAS

GÉRON, **Aurélien. Estatística prática para cientistas de dados: 50 conceitos essenciais**. O'Reilly Media, Inc., 2019

GOONEWARDANA, H. **PCA: Application in Machine Learning. 2019**. Disponível em: <https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db>. Acesso em: 06 fev. 2024.

JAADI, Z. **A Step-by-Step Explanation of Principal Component Analysis (PCA)**. 2023. Disponível em: https://builtin.com/data-science/step-step-explanation-principal-component-analysis. Acesso em: 06 fev. 2024.

MACHADO, F. **Aprendizado Não Supervisionado | Redução de Dimensionalidade**. 2020. Disponível em: https://medium.com/turing-talks/aprendizado-n%C3%A3o-supervisionado-redu%C3%A7%C3%A3o-de-dimensionalidade-479ecfc464ea. Acesso em: 06 fev. 2024.

PALAVRAS-CHAVE

PCA, Redução de Dimensionalidade, Covariância.