JOSÉ AHIRTON BATISTA LOPES FILHO

MACHINE LEARNING ENGINEERING

# APRENDIZADO NÃO SUPERVISIONADO

# AULA 03

**SUMÁRIO**

## O QUE VEM POR AÍ?

A redução da dimensionalidade é uma técnica fundamental em Aprendizado de Máquina e Análise de Dados que visa simplificar conjuntos de dados complexos, preservando suas características mais relevantes. Em muitos casos, os dados originais possuem um grande número de variáveis, o que pode tornar a análise e a visualização difíceis, além de aumentar o risco de overfitting (sobreajuste) nos modelos de Machine Learning.

Técnicas como Análise de Componentes Principais (PCA) são amplamente utilizadas para transformar dados de alta dimensão em representações de menor dimensão. A PCA, por exemplo, projeta os dados em um novo espaço de menor dimensão, maximizando a variação explicada pelas primeiras componentes principais. A redução da dimensionalidade não apenas melhora a eficiência computacional como também facilita a interpretação dos dados e a descoberta de padrões escondidos. Em resumo, essa técnica é essencial para lidar com a complexidade dos dados modernos e extrair insights valiosos de forma mais eficaz.

## HANDS ON

Neste Hands On, demonstraremos a capacidade da técnica de Análise de Componentes Principais (PCA) aplicando-a à base de dados breast_cancer do sklearn. O PCA é usado para reduzir a dimensionalidade dos dados, facilitando a visualização e a análise sem perder informações críticas. Inicialmente, carregamos e exploramos os dados, seguidos pela normalização para garantir que todas as características contribuam igualmente para nosso resultado.

Em seguida aplicamos o PCA, reduzindo as 30 características originais para suas duas componentes principais. Este exercício ilustra como o PCA pode simplificar conjuntos de dados complexos, preservando sua variabilidade essencial, e destaca sua utilidade em melhorar a interpretabilidade e a eficiência de modelos de Aprendizado de Máquina.

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
%matplotlib inline

from sklearn.datasets import load_breast_cancer

cancer = load_breast_cancer()

cancer.keys()

df =
pd.DataFrame(cancer['data'],columns=cancer['feature_names'])

# visualização do PCA

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaler.fit(df)

scaled_data = scaler.transform(df)

from sklearn.decomposition import PCA

pca = PCA(n_components=2)

pca.fit(scaled_data)
```

```
x_pca = pca.transform(scaled_data)

scaled_data.shape

x_pca.shape

# plotando nossas componentes principais

plt.figure(figsize=(8,6))
plt.scatter(x_pca[:,0],x_pca[:,1],c=cancer['target'],cmap='pl
asma')
plt.xlabel('Primeiro componente principal')
plt.ylabel('Segundo componente principal')
```

Código-fonte 1 – Demonstração 4 – Redução de Dimensionalidade com Principal Component Analysis (PCA)
Fonte: Elaborado pelo autor (2024)

## SAIBA MAIS

A Análise de Dados moderna frequentemente lida com conjuntos de dados de alta dimensionalidade, ou seja, nos quais cada instância é caracterizada por um grande número de atributos. No entanto, essa riqueza de informações pode ser acompanhada por desafios significativos, como a chamada **maldição da dimensionalidade**, que pode levar a problemas de desempenho computacional, dificuldades de interpretação e overfitting em modelos de Aprendizado de Máquina.

A Redução de Dimensionalidade surge como uma abordagem fundamental para lidar com esses desafios, buscando extrair informações relevantes de conjuntos de dados de alta dimensionalidade e representá-los de forma mais compacta, preservando assim as características essenciais. Nesta aula exploraremos conceitos, técnicas e aplicações da redução de dimensionalidade, com foco especial no método do Principal Component Analysis (PCA), uma das técnicas mais amplamente utilizadas nesse contexto.

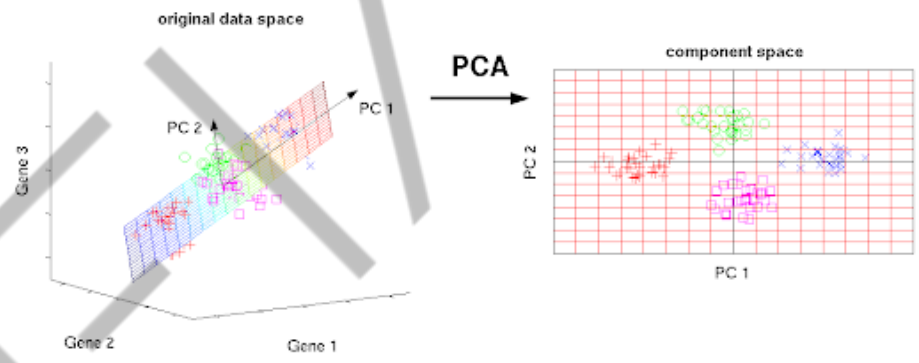


Figura 1 – Visão da transformação dimensional de dados via Análise de Componentes Principais (PCA)
Fonte: Shojaeifard; Yazdani (2021)

Ao compreender os princípios subjacentes da redução de dimensionalidade e as estratégias práticas para sua implementação, profissionais de Análise de Dados e Aprendizado de Máquina estarão melhor equipados(as) para lidar com conjuntos de dados complexos e extrair insights significativos deles. Vamos mergulhar na jornada de explorar a redução de dimensionalidade e seus benefícios na Análise de Dados e Modelagem Estatística.

**Maldição da Dimensionalidade**

A Maldição da Dimensionalidade refere-se aos desafios específicos enfrentados quando lidamos com conjuntos de dados de alta dimensionalidade. À medida que o número de dimensões (ou atributos) aumenta, a densidade dos dados tende a diminuir drasticamente. Isso significa que, em espaços de alta dimensão, os dados se tornam cada vez mais dispersos e as relações entre as instâncias individuais se tornam mais difíceis de discernir.

Como resultado, muitas técnicas de Análise de Dados e Aprendizado de Máquina podem encontrar dificuldades para generalizar padrões e inferir relações significativas entre os dados.

A alta dimensionalidade pode ter vários impactos negativos nos algoritmos de Aprendizado de Máquina. Além do aumento da complexidade computacional, a presença de muitos atributos pode levar à superposição de exemplos nos espaços de alta dimensão, dificultando a distinção entre diferentes classes ou clusters. Além disso, a presença de atributos irrelevantes ou redundantes pode introduzir ruído nos modelos, levando a um desempenho subótimo e a uma maior probabilidade de overfitting.

Diante desses desafios, os objetivos da redução de dimensionalidade são claros: simplificar a representação dos dados, preservando ao máximo as informações relevantes e reduzindo a complexidade do problema. Ao reduzirmos o número de dimensões, buscamos então extrair um conjunto de características mais compacto que **capture as principais variações nos dados** e **elimine a redundância e o ruído desnecessário**. Com uma representação mais eficiente, os modelos de Aprendizado de Máquina podem se beneficiar de uma melhor capacidade de generalização, interpretabilidade e desempenho preditivo.

Além do Principal Component Analysis (PCA), que será a mais utilizada no decorrer desse material, existem várias outras técnicas de redução de dimensionalidade amplamente empregads na Análise de Dados e Aprendizado de Máquina. Algumas dessas técnicas incluem t-Distributed Stochastic Neighbor Embedding (t-SNE), Linear Discriminant Analysis (LDA) e também os autoencoders.

O t-SNE é uma técnica não linear de redução de dimensionalidade que se destaca na visualização de dados de alta dimensionalidade. Ele mapeia os pontos de

dados de alta dimensão para um espaço de baixa dimensão de forma que pontos semelhantes no espaço original permaneçam próximos no espaço reduzido. O t-SNE é frequentemente usado para explorar a estrutura intrínseca dos dados e identificar agrupamentos ou padrões ocultos.

Já o LDA é uma técnica de redução de dimensionalidade supervisionada que visa encontrar as projeções de maior discriminação entre as classes em um conjunto de dados. Ao contrário do PCA, que é uma técnica não supervisionada, o LDA leva em consideração a informação de classe durante a redução de dimensionalidade. Ele procura maximizar a separação entre as classes enquanto minimiza a dispersão dentro de cada classe, resultando em uma representação de baixa dimensionalidade que preserva a estrutura discriminativa dos dados.

Por sua vez, os autoencoders são uma classe de modelos de Aprendizado Profundo usados para aprender representações eficientes de dados de alta dimensionalidade. Eles consistem em uma arquitetura de rede neural, a qual mapeia os dados de entrada para uma representação latente de dimensionalidade inferior e, em seguida, tenta reconstruir os dados de entrada a partir dessa representação latente. Os autoencoders podem aprender automaticamente características importantes dos dados e são altamente flexíveis e adaptáveis a uma variedade de problemas de redução de dimensionalidade.

Cada técnica de redução de dimensionalidade tem suas próprias características e é adequada para diferentes tipos de conjuntos de dados e objetivos de análise. Enquanto o PCA é eficaz para reduzir a dimensionalidade mantendo a maior parte da variância dos dados, o t-SNE é mais adequado para visualização e exploração de estruturas complexas. O LDA, por sua vez, é ideal para problemas de classificação em que a separação entre classes é importante. Já os autoencoders oferecem uma abordagem comumente mais flexível e adaptável, especialmente em contextos nos quais os dados têm uma estrutura não linear ou complexa.

A redução de dimensionalidade é então uma técnica bastante versátil, que encontra aplicação em uma ampla variedade de domínios. A seguir, são apresentados exemplos de casos de uso em diferentes áreas, destacando como a redução de dimensionalidade pode melhorar a eficiência computacional e a qualidade dos modelos de Aprendizado de Máquina.

Por exemplo, na Visão Computacional, imagens muitas vezes são representadas por vetores de alta dimensionalidade, em que cada pixel corresponde a uma dimensão. A redução de dimensionalidade é frequentemente usada para extrair características significativas das imagens, permitindo análise e classificação mais eficientes. Por exemplo: técnicas como PCA podem ser aplicadas para extrair as principais características de uma imagem, reduzindo a quantidade de dados necessária para processamento e melhorando a precisão dos modelos de reconhecimento de imagem.

Já em Processamento de Linguagem Natural (PLN), textos podem ser representados como vetores de alta dimensionalidade usando técnicas como a contagem de palavras ou embeddings de palavras. A redução de dimensionalidade pode ser útil para encontrar representações mais compactas e informativas dos textos, facilitando tarefas como classificação de documentos, análise de sentimentos e sumarização de textos. Por exemplo, o LDA pode ser usado para extrair tópicos latentes de grandes coleções de documentos, permitindo uma análise mais eficiente e interpretação dos textos.

Na bioinformática, dados biológicos, como sequências de DNA ou expressões genéticas, podem ser extremamente dimensionais. A redução de dimensionalidade é crucial para a análise e interpretação desses dados, ajudando pesquisadores(as) a identificar padrões e relações significativas. Técnicas de redução de dimensionalidade como o t-SNE podem ser usadas para visualizar a similaridade entre diferentes amostras biológicas, ajudando a identificar grupos de genes ou proteínas associados a fenótipos específicos.

Em finanças, conjuntos de dados financeiros podem ser altamente dimensionais, com uma grande quantidade de variáveis e características. A redução de dimensionalidade pode ser empregada para identificar fatores latentes subjacentes nos dados financeiros, facilitando a modelagem e previsão de séries temporais, análise de risco e detecção de anomalias. Por exemplo: técnicas como PCA podem ser usadas para identificar os principais fatores de risco em carteiras de investimento e otimizar a alocação de ativos.

Em todos esses casos de uso, a redução de dimensionalidade oferece benefícios significativos, incluindo uma redução na complexidade computacional, uma representação mais compacta e interpretável dos dados e uma melhoria na eficiência

e precisão dos modelos de aprendizado de máquina. Ao extrair características relevantes e descartar informações redundantes, a redução de dimensionalidade permite uma análise mais eficaz e eficiente dos dados em uma variedade de diferentes domínios.

Ao se aplicar técnicas de redução de dimensionalidade, é importante estar ciente dos desafios comuns e das considerações práticas para garantir resultados eficazes. A seguir, são discutidos alguns desses desafios e as estratégias para lidar com eles de maneira eficaz:

- **Perda de Informação**: ao reduzir a dimensionalidade, há o risco de se perder informações importantes contidas nos dados originais. É essencial garantir que a redução de dimensionalidade preserve as características relevantes dos dados e minimize a perda de informação.
- **Overfitting**: em alguns casos, técnicas de redução de dimensionalidade podem levar ao overfitting, em que o modelo se ajusta demasiadamente aos dados de treinamento e falha em generalizar para novos dados. É importante monitorar o desempenho do modelo em conjuntos de dados de teste e ajustar os parâmetros conforme necessário para evitar overfitting.
- **Escolha de Técnica Adequada**: existem várias técnicas de redução de dimensionalidade disponíveis e a escolha da adequada depende do tipo de dados, do objetivo da análise e das características específicas do problema. É importante entender as vantagens e limitações de cada técnica e selecionar aquela mais adequada ao contexto do problema.

**Considerações Práticas**

- **Pré-processamento de Dados**: antes de aplicar técnicas de redução de dimensionalidade, é importante realizar um pré-processamento adequado dos dados, incluindo a remoção de valores ausentes, normalização de atributos e tratamento de outliers. Um pré-processamento cuidadoso pode melhorar a eficácia das técnicas de redução de dimensionalidade e evitar problemas potenciais durante a análise.
- **Avaliação de Desempenho**: é essencial avaliar o desempenho das técnicas de redução de dimensionalidade em conjunto com os modelos de

Aprendizado de Máquina subsequentes. Isso pode ser feito por meio de validação cruzada, análise de componentes principais (PCA) ou outras técnicas de avaliação de desempenho. A seleção da técnica mais adequada deve ser baseada não apenas na eficácia da redução de dimensionalidade, mas também no desempenho global do modelo.

- **Experimentação Iterativa**: a redução de dimensionalidade é um processo iterativo que requer experimentação e ajustes contínuos. É importante explorar diferentes técnicas, parâmetros e configurações para encontrar a combinação ideal que atenda aos requisitos específicos do problema.

**Estratégias para Lidar com Dados de Alta Dimensionalidade**

- **Seleção de Atributos**: uma abordagem comum para lidar com dados de alta dimensionalidade é a seleção de atributos, em que apenas os atributos mais relevantes são mantidos para análise. Isso pode ser feito manualmente com base no conhecimento do domínio ou usando métodos automatizados, como a análise de importância de atributos.
- **Agrupamento Hierárquico**: o clustering hierárquico é uma técnica eficaz para agrupar atributos semelhantes em clusters menores, reduzindo assim a dimensionalidade dos dados. Isso pode simplificar a análise e melhorar a interpretabilidade dos resultados.
- **Incremental PCA**: para conjuntos de dados muito grandes que não cabem na memória, o Incremental PCA pode ser uma opção viável. Esta técnica permite calcular os componentes principais em lotes menores de dados e depois combinar os resultados para obter uma redução de dimensionalidade global.

Ao longo do material da aula, discutiremos a motivação por trás da redução de dimensionalidade e apresentaremos o PCA. Exploraremos casos de uso em diferentes domínios, como Visão Computacional, demonstrando como a redução de dimensionalidade pode melhorar a eficiência computacional e a qualidade dos modelos de Aprendizado de Máquina.

Além disso, abordaremos os desafios comuns e considerações práticas ao aplicar técnicas de redução de dimensionalidade, fornecendo estratégias para lidar

com dados de alta dimensionalidade de forma eficaz. Destacamos a importância da avaliação de desempenho e da experimentação iterativa para encontrar a combinação ideal de técnicas para um determinado problema.

Olhando para o futuro, há várias direções de pesquisa promissoras na área de redução de dimensionalidade. Avanços em técnicas de Aprendizado Profundo, como autoencoders e redes adversariais generativas (GANs), têm o potencial de revolucionar a forma como lidamos com dados de alta dimensionalidade. Além disso, o desenvolvimento de métodos escaláveis e eficientes para lidar com conjuntos de dados cada vez maiores continuará a ser uma área de interesse.

Em suma, a redução de dimensionalidade desempenha um papel fundamental na Análise de Dados e no Aprendizado de Máquina, permitindo uma representação mais compacta e interpretação dos dados, bem como uma melhoria na eficiência e precisão dos modelos. Com a rápida evolução da tecnologia e a crescente complexidade dos conjuntos de dados, a pesquisa contínua nesta área é essencial para impulsionar a inovação e avançar no campo do Aprendizado de Máquina.

**Livros**

"Pattern Recognition and Machine Learning" de Christopher M. Bishop: este livro aborda uma ampla gama de tópicos em Reconhecimento de Padrões e Aprendizado de Máquina, incluindo técnicas de Aprendizado Não Supervisionado, como clustering e redução de dimensionalidade.

"The Elements of Statistical Learning: Data Mining, Inference, and Prediction" de Trevor Hastie, Robert Tibshirani e Jerome Friedman: este livro é uma referência fundamental em aprendizado de máquina e mineração de dados, cobrindo uma ampla gama de tópicos, incluindo técnicas de aprendizado supervisionado e não supervisionado, como clustering, análise de componentes principais e métodos de regressão.

**Artigos Científicos:**

"Principal Component Analysis" de I. T. Jolliffe: este artigo clássico apresenta uma explicação detalhada do método de redução de dimensionalidade conhecido como Análise de Componentes Principais (PCA), que é amplamente utilizado em diversas áreas para simplificar conjuntos de dados complexos enquanto preserva sua variabilidade essencial.

"Stochastic Neighbor Embedding" de Geoffrey Hinton e Sam Roweis: este artigo introduz a técnica de Stochastic Neighbor Embedding (SNE), que é uma abordagem inovadora para a redução de dimensionalidade que preserva a estrutura local dos dados em espaços de menor dimensão, sendo especialmente útil para visualização de dados complexos.

"T-SNE: Stochastic Neighbor Embedding", de Laurens van der Maaten e Geoffrey Hinton: este artigo apresenta o t-SNE, uma melhoria do SNE, que resolve problemas de otimização do método original, oferecendo uma técnica poderosa e eficaz para a redução de dimensionalidade e visualização de dados de alta dimensão.

Estas obras demonstram um pouco dos desafios ao se manejar grandes conjuntos de dados, em que a alta dimensionalidade pode dificultar a visualização e a extração de informações relevantes, oferecendo soluções robustas e inovadoras para esses problemas. A capacidade de lidar eficazmente com alta dimensionalidade é fundamental para o desenvolvimento de modelos de Aprendizado de Máquina eficientes, robustos e interpretáveis. Isso permite a pesquisadores(as) e praticantes extrair insights mais significativos dos dados e construir soluções mais eficazes para problemas complexos.

**Análise de Componentes Principais (PCA)**

A Análise de Componentes Principais (PCA) é uma técnica amplamente utilizada para redução de dimensionalidade e extração de características em conjuntos de dados de alta dimensionalidade. Nesta seção, exploraremos os princípios matemáticos por trás do PCA, as etapas para sua implementação e suas aplicações em Aprendizado de Máquina.

O PCA busca encontrar uma nova base de coordenadas que maximize a variância dos dados. Isso é alcançado por meio da decomposição dos dados originais em componentes principais ortogonais, que são vetores que capturam a direção de máxima variância nos dados. Matematicamente, o PCA envolve a diagonalização da matriz de covariância dos dados ou a realização de uma decomposição de valores singulares (SVD) na matriz de dados.

A decomposição da matriz de covariância em análise de componentes principais (PCA) é uma etapa crucial para identificar os principais componentes de variância nos dados. Aqui está como a decomposição pode ser expressa:

$$\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T$$

. Em que:

- **V** é a matriz de autovetores, sendo que cada coluna representa um autovetor.
- **D** é uma matriz diagonal contendo os autovalores correspondentes aos autovetores de **V**.

Os autovetores representam as direções dos eixos principais dos dados, enquanto os autovalores indicam a quantidade de variância explicada por cada componente principal. A ordem dos autovalores (e, portanto, dos autovetores) é tal que o primeiro autovalor é o maior e o último é o menor, representando a quantidade decrescente de variância explicada por cada componente principal.

Muitas vezes acontece que há muitos recursos no conjunto de dados e uma pequena fração de informações está presente em cada recurso ou variável. Por exemplo: suponha que temos um conjunto de dados que consiste em 50 colunas

(recursos); então será quase impossível visualizar esses 50 recursos no mesmo gráfico e procurar por insights nos dados.

Então, o que fazemos é encontrar uma representação de baixa dimensão dos dados que capture o máximo de informações possível. Se pudermos obter uma representação bidimensional dos dados que capture a maior parte da informação, então poderemos representar graficamente as observações neste espaço de baixa dimensão. O PCA fornece uma ferramenta para fazer exatamente isso!

Ele encontra uma representação de baixa dimensão de um conjunto de dados que contém o máximo de variação possível. A ideia é que sobre cada uma das observações que vivem em algum espaço p-dimensional, nem todas essas dimensões sejam igualmente interessantes. Logo, o PCA busca um pequeno número de dimensões que sejam tão interessantes quanto possível, em que o conceito de interessante é medido pela quantidade que as observações variam ao longo de cada dimensão.

Vejamos um pequeno exemplo: suponha que temos dados de publicidade que consistem em duas características: tamanho da população (pop) dada em dezenas de milhares de pessoas e gastos com publicidade para uma determinada empresa (anúncio) em milhares de dólares para 100 cidades (nº de observações).

Portanto, para os dois recursos teremos uma visão bidimensional dos dados. Quando procuramos a visão dimensional inferior, estamos preparados para encontrar os componentes principais usando os dados ou determinado recurso que melhor explique a variação nos dados (que normalmente são menores que o número de recursos presentes nos dados).

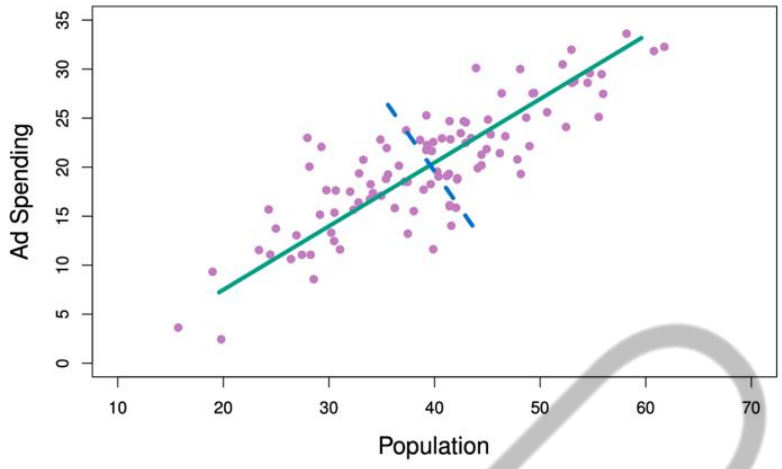


Figura 2 – Verificando nossos componentes principais para um problema de compra de mídia para propaganda
Fonte: James et al. (2023)

A linha sólida verde na figura 2 representa a primeira direção do componente principal (Z1) dos dados. Podemos ver a olho nu que esta é a direção ao longo da qual há maior variabilidade nos dados (em geral, cidades mais populosas terão maiores gastos com publicidade).

Isto é, se projetássemos as 100 observações nesta reta, então as observações projetadas resultantes teriam a maior variância possível; projetar as observações em qualquer outra linha produziria observações projetadas com menor variância. Projetar um ponto em uma linha envolve simplesmente encontrar o local na linha que está mais próximo do ponto, como ilustrado na figura a seguir.

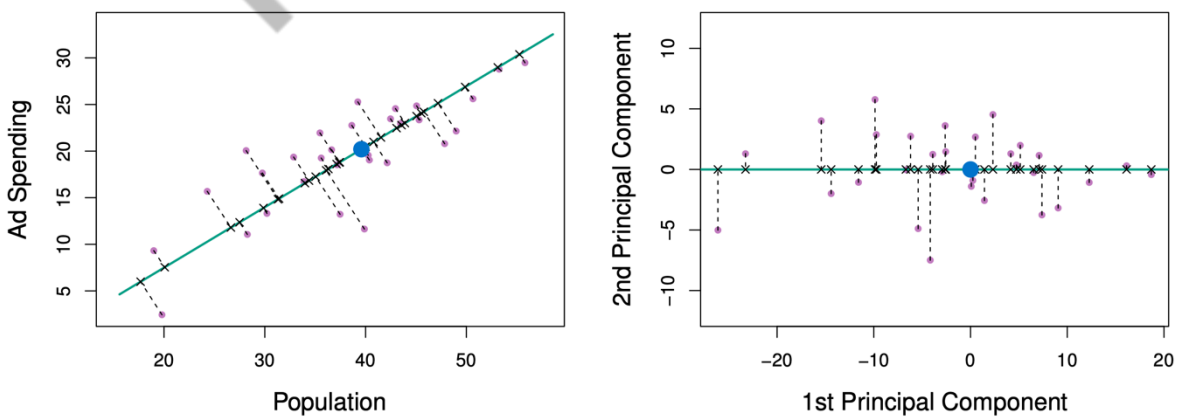

Figura 3 – Verificando nossos componentes principais para um problema de compra de mídia para propaganda
Fonte: James et al. (2023)

**Implementação do PCA**

As etapas básicas para implementação do PCA são as seguintes:

- **Padronização dos Dados**: os dados são padronizados para garantir que todas as variáveis tenham média zero e variância unitária.
- **Cálculo da Matriz de Covariância ou SVD**: em seguida, é calculada a matriz de covariância dos dados ou realizada a decomposição de valores singulares na matriz de dados.
- **Cálculo dos Autovetores e Autovalores**: os autovetores e autovalores da matriz de covariância (ou da matriz de covariância empírica) são computados.
- **Seleção de Componentes Principais**: os autovetores correspondentes aos maiores autovalores são selecionados como os componentes principais.
- **Projeção dos Dados**: os dados originais são projetados nos componentes principais para obter uma representação de dimensionalidade reduzida.

A projeção dos dados nos componentes principais pode ser expressa como:

$$\text{Projeção}_{\text{PCA}}(X) = X \cdot \mathbf{V}_k$$

Em que:

- X é a matriz de dados original com dimensão $n \times m$n×m.
- Vk é a matriz de autovetores dos k principais componentes, contendo os k autovetores mais relevantes de V, resultando em uma matriz de dimensão m×k.
- O resultado da multiplicação X·Vk é a matriz de dados projetada no espaço dos componentes principais, com dimensão n×k. Cada linha dessa matriz

representa as coordenadas dos dados originais no novo espaço de menor dimensionalidade definido pelos principais componentes.

**Exemplo Prático**

Temos agora a disposição um conjunto de dados bidimensional $X$ com duas características $x_1$ e $x_2$ e desejamos aplicar o PCA de forma a reduzir a dimensionalidade do conjunto de dados para uma dimensão. Nesse caso, você pode seguir os seguintes passos:

- **Padronize os dados**: é importante padronizar os dados para garantir que cada característica tenha média zero e variância unitária.
- **Calcule a matriz de covariância**: determine a matriz de covariância dos dados padronizados.
- **Calcule os autovetores e autovalores**: compute os autovetores e autovalores da matriz de covariância.
- **Selecione o autovetor correspondente ao maior autovalor**: este autovetor representará a direção do componente principal.
- **Projete os dados no novo espaço**: projetar os dados originais no espaço unidimensional definido pelo autovetor selecionado.

A seguir, temos um exemplo de como fazer isso em Python se utilizando de numpy e scikit-learn:

```
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA


# Exemplo de dados bidimensionais
X = np.array([[1, 2], [2, 3], [3, 4], [4, 5], [5, 6]])
```

```
# Padronizar os dados
X_std = StandardScaler().fit_transform(X)

# Aplicar PCA para redução de dimensionalidade para 1 componente principal
pca = PCA(n_components=1)
X_pca = pca.fit_transform(X_std)

# Componente principal (autovetor)
principal_component = pca.components_[0]

# Variância explicada pelo componente principal
explained_variance = pca.explained_variance_ratio_[0]

print("Componente Principal:", principal_component)
print("Variância explicada:", explained_variance)

# Projeção dos dados no novo espaço unidimensional
print("Dados projetados:")
print(X_pca)
```

Código-fonte 2 – Demonstração 5 – Redução de Dimensionalidade com Principal Component Analysis (PCA)
Fonte: Elaborado pelo autor (2024)

Neste exemplo, os dados bidimensionais são primeiro padronizados. Em seguida, o PCA é aplicado para reduzir a dimensionalidade para um componente principal. O autovetor correspondente ao maior autovalor (que representa a direção do componente principal) é armazenado em principal_component, e a variância explicada por esse componente é impressa. Finalmente, os dados são projetados no novo espaço unidimensional definido pelo componente principal e são impressos.

**Aplicações**

O PCA tem uma ampla gama de aplicações em aprendizado de máquina e análise de dados, incluindo redução de dimensionalidade, visualização de dados e extração de características. Ao aplicar o PCA, é importante considerar o número de componentes principais a serem retidos, a interpretabilidade dos resultados e as limitações do método, como a sensibilidade a outliers.

Pudemos observar, portanto, que o PCA é uma técnica poderosa para redução de dimensionalidade e extração de características em conjuntos de dados de alta dimensionalidade. Compreender os princípios matemáticos por trás do PCA e sua implementação prática é fundamental para seu uso eficaz em aplicações do mundo real. Uma das consequências da dimensionalidade é que os dados se tornam esparsos à medida que o número de dimensões aumenta. Em espaços de alta dimensão, a maioria dos pontos de dados tende a se concentrar nas bordas da região de amostragem, deixando vastas regiões do espaço praticamente vazias. Isso dificulta a estimativa de densidades e a generalização de modelos.

À medida que o número de dimensões aumenta, torna-se cada vez mais difícil visualizar os dados de forma eficaz. Enquanto podemos facilmente visualizar dados bidimensionais ou tridimensionais em gráficos, a visualização de espaços de alta dimensão é impraticável. Isso torna difícil entender a estrutura dos dados e identificar padrões visualmente. Algoritmos que funcionam bem em espaços de baixa dimensão podem se tornar computacionalmente inviáveis em espaços de alta dimensão.

Isso ocorre porque o número de cálculos necessários aumenta exponencialmente com o número de dimensões. Muitos algoritmos de Aprendizado de Máquina sofrem de baixo desempenho ou até mesmo falham em espaços de alta dimensão devido a esse aumento na complexidade computacional, exigindo então o uso de técnicas como o PCA.

A maldição da dimensionalidade continua a ser um desafio significativo em Ciência de Dados e Aprendizado de Máquina, com implicações importantes para a análise e modelagem de dados em espaços de alta dimensão. Compreender os efeitos da dimensionalidade e adotar as estratégias adequadas para mitigar seus efeitos é essencial para o desenvolvimento de modelos eficazes em ambientes de alta dimensionalidade.

# MERCADO, CASES E TENDÊNCIAS

### Matéria EN-US – Empresa de Streaming e Produção de Conteúdo (Netflix)

A Netflix emprega redes neurais autoencoder para aprimorar seu sistema de recomendação de conteúdo. Autoencoders são um tipo de rede neural artificial treinada para reconstruir dados de entrada, aprendendo efetivamente uma representação comprimida dos dados originais.

Ao codificar preferências do usuário e histórico de visualização em um espaço de dimensão inferior usando autoencoders, a Netflix pode capturar padrões complexos e relacionamentos no comportamento do usuário. Isso permite recomendações de conteúdo mais precisas e personalizadas, melhorando assim o engajamento e a satisfação do usuário. Saiba mais aqui.

## O QUE VOCÊ VIU NESTA AULA?

Nesta aula sobre Redução de Dimensionalidade e Análise de Componentes Principais (PCA), exploramos uma técnica fundamental para redução de dimensionalidade e extração de características em conjuntos de dados multivariados. Aprendemos que o PCA busca identificar os principais modos de variabilidade nos dados, representados por autovetores, enquanto os autovalores indicam a magnitude dessa variabilidade em cada direção.

Ao aplicar o PCA, a matriz de covariância dos dados é decomposta em autovetores e autovalores, permitindo a projeção dos dados em um novo espaço dimensional com menor dimensionalidade, mantendo a maior parte da informação original.

Além disso, compreendemos o processo prático de implementação do PCA utilizando bibliotecas como scikit-learn em Python, indo desde a padronização dos dados até a projeção no espaço de menor dimensionalidade, passando pela seleção dos componentes principais com maior variância explicada. Essa técnica não apenas simplifica a representação dos dados, mas também facilita a visualização e a análise, ajudando a identificar padrões e estruturas subjacentes nos conjuntos de dados complexos.

A aplicação do PCA é amplamente utilizada em diversas áreas, incluindo reconhecimento de padrões, processamento de imagem (como visto no código da Demonstração 5) e biologia, entre outros, sendo uma ferramenta valiosa para análise exploratória e modelagem de dados.

# REFERÊNCIAS

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006. Disponível em: <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>. Acesso em: 12 jul. 2024.

HINTON, G.; ROWEIS, S. **Stochastic Neighbor Embedding**. In: Advances In Neural Information Processing Systems 15, NIPS 2002, Vancouver. Proceedings... Vancouver: MIT Press, 2003. p. 833-840. Disponível em: <https://papers.nips.cc/paper_files/paper/2002/hash/6150ccc6069bea6b5716254057a194ef-Abstract.html>. Acesso em: 12 jul. 2024.

JAMES, G. et al. **An Introduction to Statistical Learning with Applications in Python**. [s.l.]: Springer, 2023. Disponível em: <https://www.statlearning.com/>. Acesso em: 12 jul. 2024.

JOLLIFFE, I. T. **Principal Component Analysis**. New York: Springer, 2002. Disponível em: <http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf>. Acesso em: 12 jul. 2024.

MAATEN, L. V. D.; HINTON, G. T-SNE: Stochastic Neighbor Embedding. **Journal of Machine Learning Research**, 9, 2008, 2579-2605. Disponível em: <https://www.cs.toronto.edu/~fritz/absps/sne.pdf>. Acesso em: 12 jul. 2024.

SHOJAEIFARD, A. R.; YAZDANI, H. R. **Hybrid Completely Positive Tensorial PCA (HCPT-PCA) method for face recognition (FR)**. Em: The First Conference on Mathematics and its Applications, 11-12 agosto 2021, Shahid Chamran University of Ahvaz, Ahvaz, Irã. Disponível em: <https://www.researchgate.net/publication/353902821_Hybrid_Completely_Positive_Tensorial_PCA_HCPT-PCA_method_for_face_recognition_FR>. Acesso em: 15 jul. 2024.

TAN, P. et al. **Introduction to Data Mining**. Essex: Pearson, 2014. Disponível em: <https://www.ceom.ou.edu/media/docs/upload/Pang-Ning_Tan_Michael_Steinbach_Vipin_Kumar_-_Introduction_to_Data_Mining-Pe_NRDK4fi.pdf >. Acesso em: 12 jul. 2024.

## PALAVRAS-CHAVE

Inteligência Artificial. Aprendizado de Máquina. Aprendizado Não Supervisionado.