Algoritmos Não Supervisionados – PCA

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Algoritmos Não Supervisionados

PCA

O objetivo deste módulo é apresentar conceitualmente os principais algoritmos de redução de dimensionalidade para que possamos simplificar os dados enquanto mantemos o máximo possível das informações importantes. E faremos um projeto explorando o primeiro destes algoritmos, que é o PCA, onde faremos o processo completo desde o EDA até a visualização dos resultados.

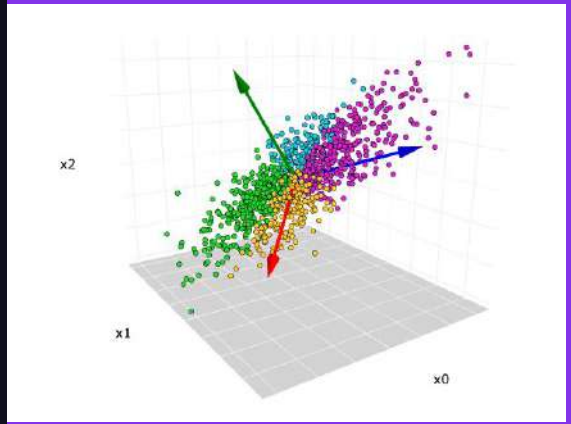
Agenda

- O que é redução de dimensionalidade
- Um passeio pelos algoritmos de redução de dimensionalidade
- O que é o algoritmo PCA
- SVD e PCA
- Métricas de algoritmos de redução de dimensionalidade
- Projeto – PCA

O que é redução de dimensionalidade

Os algoritmos de redução de dimensionalidade constituem uma classe de técnicas matemáticas aplicadas na análise de dados, cujo propósito fundamental é a simplificação dos dados ao reduzir o número de variáveis envolvidas. Essa redução é especialmente valiosa em contextos onde os dados apresentam alta complexidade dimensional, o que não apenas dificulta a análise visual e estatística, mas também aumenta o custo computacional e pode degradar o desempenho de algoritmos de aprendizado de máquina devido ao fenômeno conhecido como "maldição da dimensionalidade".

Na prática, esses algoritmos trabalham transformando um grande conjunto de variáveis em um menor, preservando tanto quanto possível as informações essenciais. Este processo é realizado através da identificação de padrões, correlações e estruturas fundamentais nos dados originais.



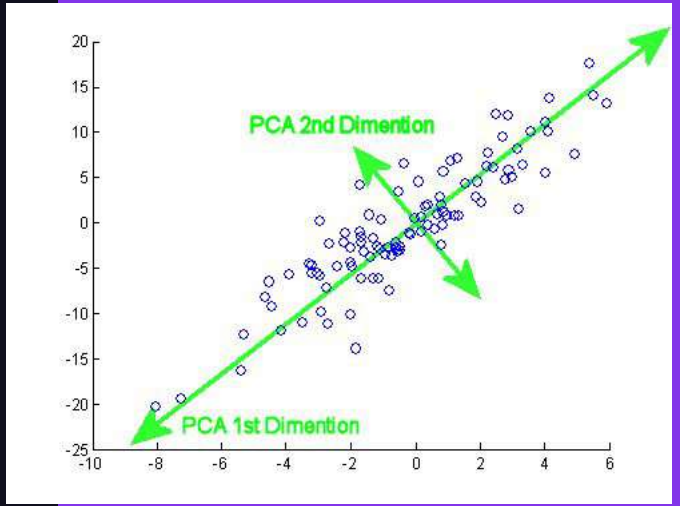
O que é redução de dimensionalidade

Principais objetivos e benefícios:

Redução de Complexidade: Dados de alta dimensão podem ser complexos de analisar e visualizar. A redução de dimensionalidade ajuda a simplificar esses dados para facilitar a interpretação.

Eliminação de Ruído: Ao focar em componentes principais ou características principais, a redução de dimensionalidade pode ajudar a eliminar o ruído, destacando apenas as características mais significativas dos dados.

Eficiência Computacional: Dados com menos dimensões requerem menos recursos computacionais para processamento. Isso é crucial em casos de aprendizado de máquina e análise de grandes volumes de dados, onde o tempo de processamento e a capacidade de memória podem ser limitantes.

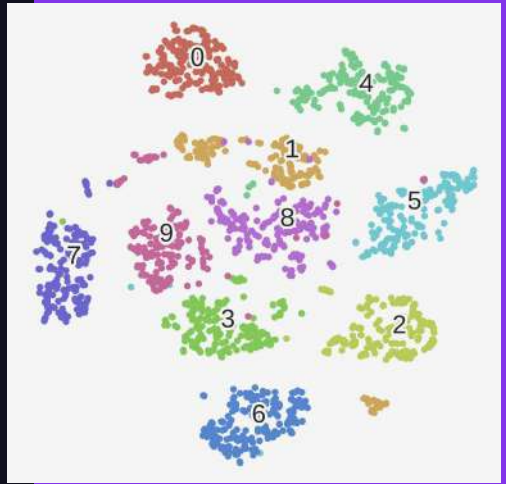


O que é redução de dimensionalidade

Melhoria no Desempenho de Algoritmos: Muitos algoritmos de aprendizado de máquina têm seu desempenho prejudicado pela "maldição da dimensionalidade", que é quando o aumento no número de dimensões leva a um espaçamento maior entre os pontos de dados. Reduzindo a dimensionalidade, pode-se melhorar a acurácia e eficácia desses algoritmos.

Visualização de Dados: É difícil visualizar dados com muitas dimensões diretamente. A redução para duas ou três dimensões permite o uso de gráficos bidimensionais ou tridimensionais para explorar e comunicar os dados de forma efetiva.

Descoberta de Estruturas Subjacentes: A redução de dimensionalidade pode revelar estruturas ocultas nos dados, que não são aparentes em uma análise de alta dimensão. Isso pode incluir agrupamentos ou padrões correlacionados.

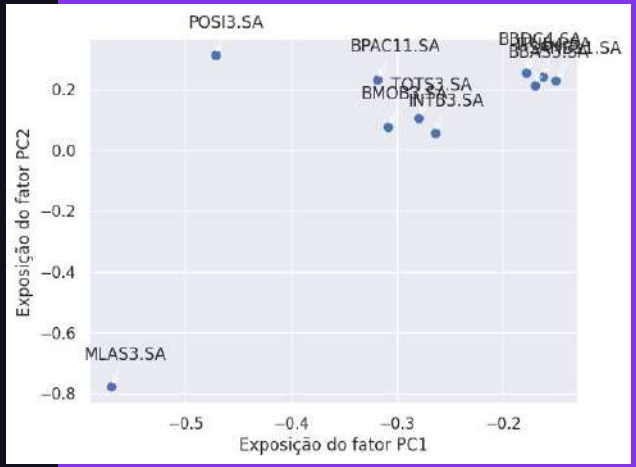


O que é redução de dimensionalidade

Alguns usos de algoritmos de redução de dimensionalidade:

Finanças e Investimentos

Análise de Risco e Portfólio: No setor financeiro, a redução de dimensionalidade é frequentemente usada para simplificar os dados de mercado e identificar os principais fatores que afetam os retornos de investimento. Por exemplo, o PCA (Análise de Componentes Principais) pode ser usado para identificar as principais variáveis que influenciam a movimentação de um conjunto diversificado de ativos, permitindo aos analistas de risco e gerentes de portfólio entender melhor as correlações e volatilidades dos ativos.

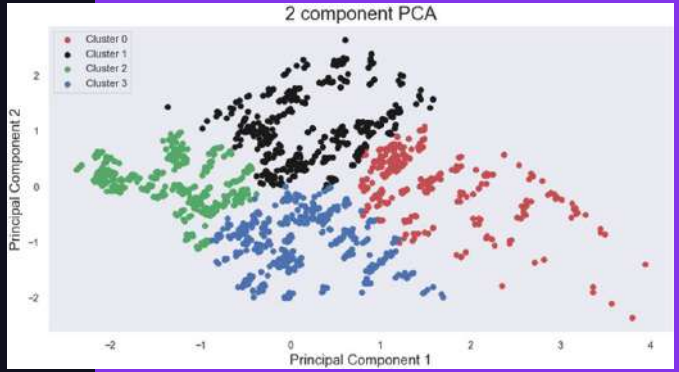


O que é redução de dimensionalidade

Alguns usos de algoritmos de redução de dimensionalidade:

Marketing Digital

Segmentação de Clientes: Algoritmos de redução de dimensionalidade, como PCA e t-SNE, podem ser aplicados para entender melhor as características dos clientes com base em grandes conjuntos de dados de comportamento online. Isso permite que as empresas segmentem seus clientes de maneira mais eficaz, personalizando o marketing para grupos específicos com base em seus hábitos e preferências de compra.

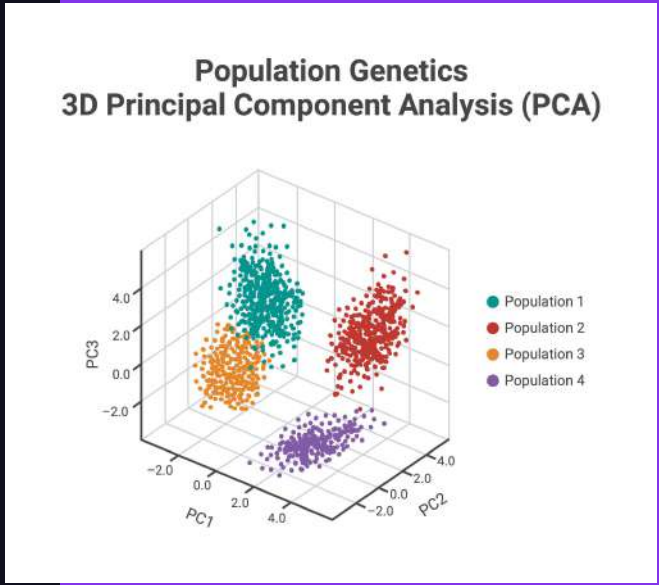


O que é redução de dimensionalidade

Alguns usos de algoritmos de redução de dimensionalidade:

Saúde e Bioinformática

Genômica e Pesquisa de Doenças: Na bioinformática, a redução de dimensionalidade é usada para analisar dados genômicos de alta dimensionalidade, como a expressão gênica ou sequências de DNA/RNA. Isso ajuda os pesquisadores a identificar padrões ou marcadores biológicos relevantes para doenças específicas, facilitando o desenvolvimento de tratamentos personalizados e a medicina de precisão.

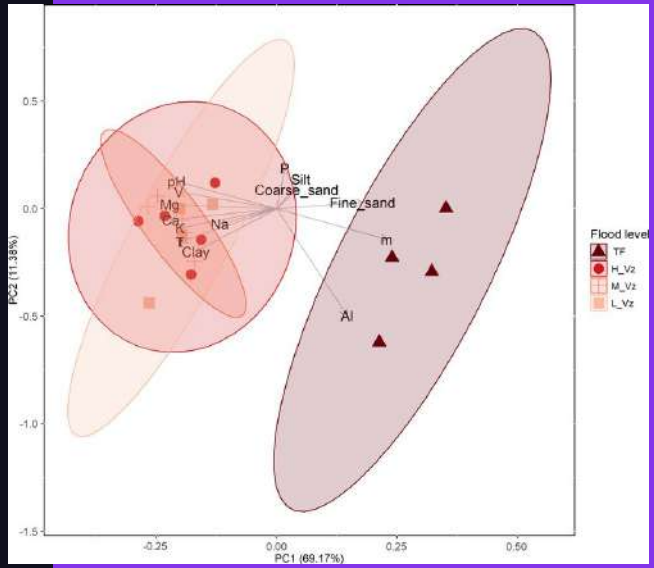


O que é redução de dimensionalidade

Alguns usos de algoritmos de redução de dimensionalidade:

Varejo e Gestão de Inventário

Otimização de Inventário: No varejo, a redução de dimensionalidade pode ser usada para analisar padrões de compra e comportamento do consumidor, ajudando na previsão de demanda e na otimização de estoque. Algoritmos como PCA podem reduzir a complexidade dos dados de transações, destacando os principais fatores que influenciam as decisões de compra dos consumidores.



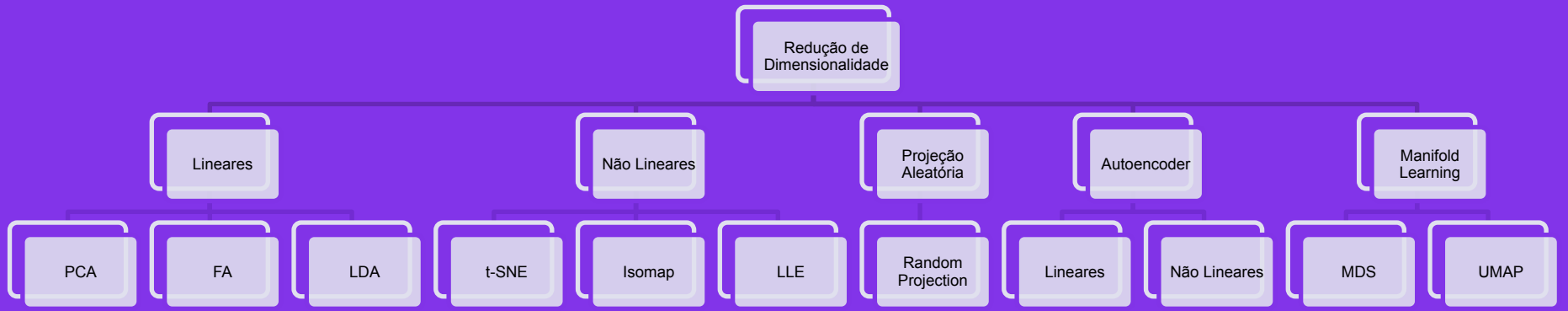
O que é redução de dimensionalidade

Alguns usos de algoritmos de redução de dimensionalidade:

Telecomunicações

Gerenciamento de Tráfego e Planejamento de Rede: Empresas de telecomunicações usam redução de dimensionalidade para analisar e gerenciar grandes volumes de dados de tráfego de rede. Isso permite a identificação de padrões e tendências no uso da rede, ajudando no planejamento e otimização da infraestrutura de rede para melhorar a qualidade do serviço e a eficiência operacional.

Um passeio pelos algoritmos de redução de dimensionalidade



Um passeio pelos algoritmos de redução de dimensionalidade

PCA (Principal Component Analysis)

Transforma os dados para um novo sistema de coordenadas, reduzindo a dimensionalidade ao escolher os primeiros componentes principais que capturam a maior variância possível.

FA (Factor Analysis)

Modela as variáveis observadas em termos de um número menor de variáveis não observadas (fatores), assumindo que as variáveis observadas são combinações lineares dos fatores mais o ruído.

LDA (Linear Discriminant Analysis)

Busca eixos que maximizem a separação entre múltiplas classes. O objetivo é achar uma combinação linear das características que melhor separa duas ou mais classes de objetos.

t-SNE (t-Distributed Stochastic Neighbor Embedding)

Reduz a dimensionalidade ao converter distâncias entre pontos em probabilidades condicionais e tentando minimizar a divergência entre essas probabilidades no espaço de alta e baixa dimensão.

Um passeio pelos algoritmos de redução de dimensionalidade

LLE (Local Linear Embedding)

Reduz a dimensão ao preservar distâncias locais entre os pontos. Assume que cada ponto e seus vizinhos mais próximos estão em um plano ou subespaço linear local.

Random Projection

Projeta os dados em um espaço de menor dimensionalidade de forma aleatória, mas preserva as distâncias entre os pontos de acordo com a teoria de Johnson-Lindenstrauss.

MDS (Multidimensional Scaling)

Visa colocar cada objeto em um espaço de baixa dimensão de forma que as distâncias entre os pontos sejam preservadas o máximo possível.

UMAP (Uniform Manifold Approximation and Projection)

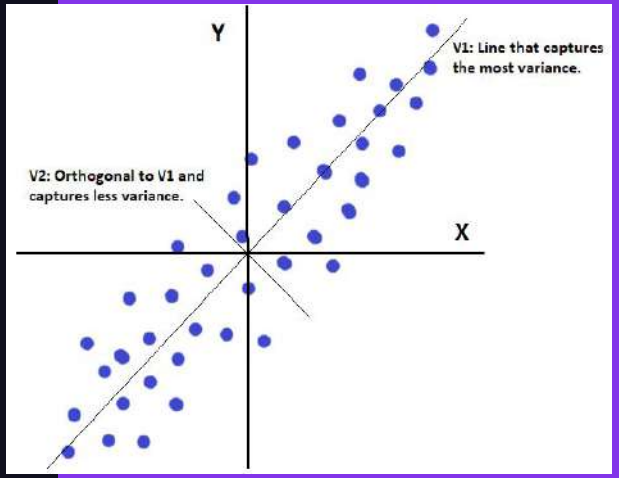
Similar ao t-SNE em seus objetivos, mas baseia-se em uma abordagem matemática diferente e, muitas vezes, é mais rápido e melhor para preservar a estrutura global dos dados.

O que é o algoritmo PCA

O algoritmo PCA (Principal Component Analysis) é uma técnica estatística utilizada para reduzir a dimensionalidade dos dados enquanto preserva o máximo possível de sua variância. Segue os passos que este algoritmo segue:

1) Normalização dos Dados: Antes de aplicar o PCA, é comum normalizar os dados para que cada característica (variável) tenha média zero e variância unitária. Isso é importante para evitar que variáveis com maior magnitude dominem o resultado.

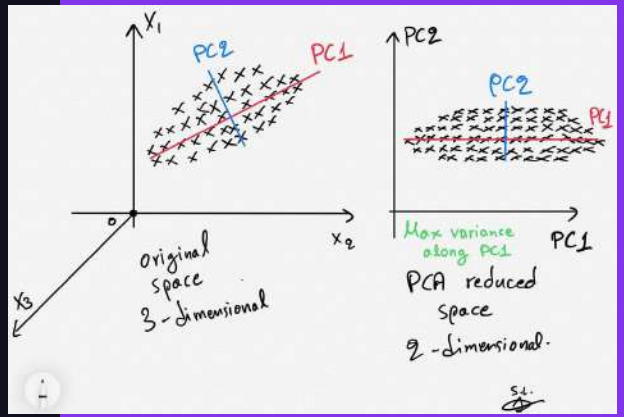
2) Cálculo da Matriz de Covariância: A matriz de covariância é calculada a partir dos dados normalizados. Essa matriz ajuda a entender como as variáveis nos dados estão variando juntas.



O que é o algoritmo PCA

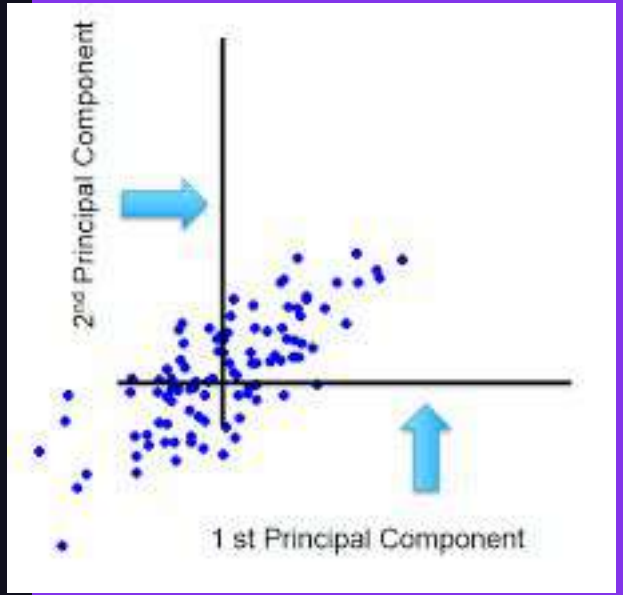
3) Cálculo dos Autovalores e Autovetores: Os autovalores e autovetores da matriz de covariância são calculados. Os autovetores representam as direções dos novos eixos no espaço de dados original, onde os dados têm maior variância. Os autovalores associados a cada autovetor indicam a quantidade de variância que cada um dos novos eixos captura.

4) Seleção dos Componentes Principais: Os autovetores são ordenados por seus autovalores correspondentes em ordem decrescente. Os primeiros autovetores (aqueles com os maiores autovalores) são selecionados como os componentes principais. O número de componentes principais selecionados depende do quanto de variância queremos reter nos dados reduzidos.

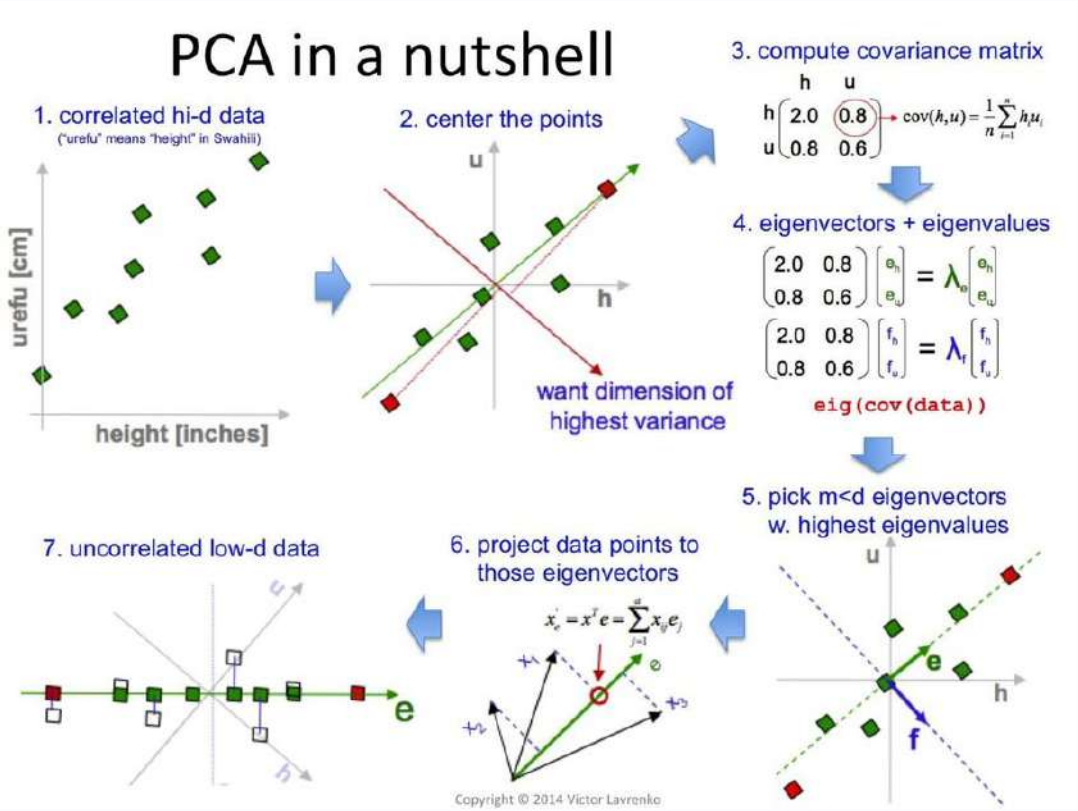


O que é o algoritmo PCA

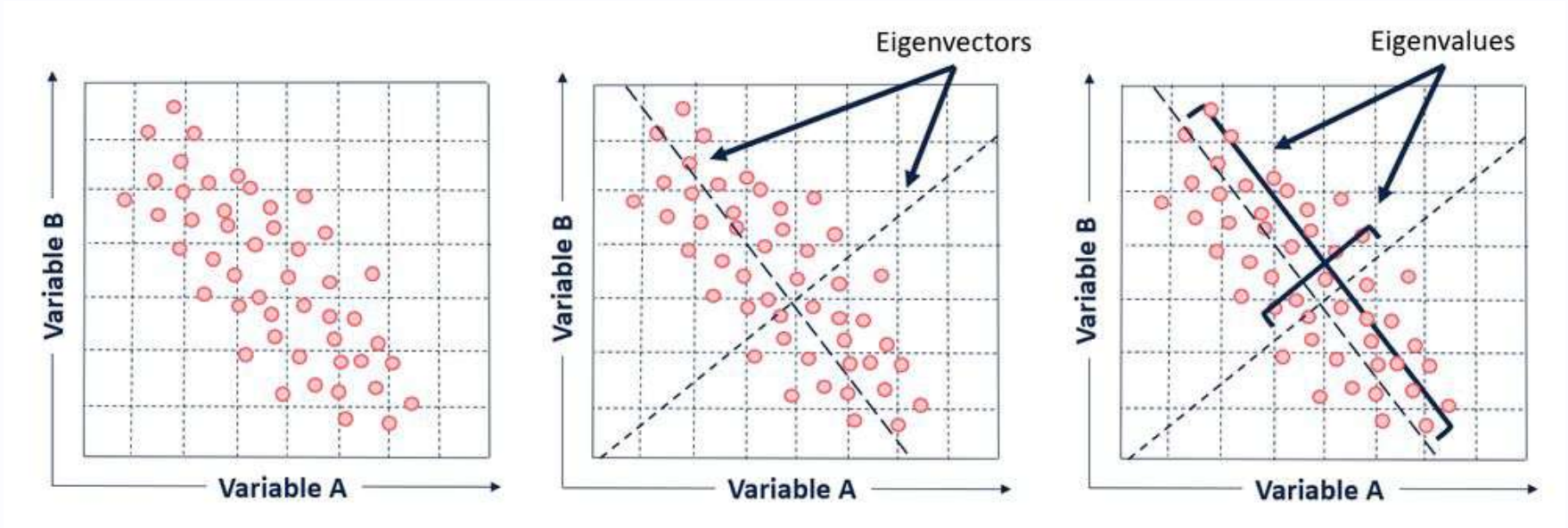
5) Transformação dos Dados: Finalmente, os dados originais são projetados nos autovetores selecionados. Isso transforma os dados do espaço original para um novo espaço com menos dimensões (os componentes principais), onde os dados ainda retêm a maior parte de sua informação original, mas com menos ruído e redundância.



O que é o algoritmo PCA



O que é o algoritmo PCA



SVD e PCA

SVD (Singular Value Decomposition) e PCA estão diretamente relacionados porque a SVD fornece uma maneira computacional de realizar o PCA.

Usar SVD para fazer PCA é como usar uma ferramenta sofisticada para automaticamente encontrar os melhores ângulos para capturar as características mais informativas dos dados.

SVD e PCA

SVD: Desmontando a Matriz

Imagine que você tem uma pilha de livros que representam diferentes aspectos dos dados, como um livro para cada variável (como idade, peso, altura, etc.). A SVD ajuda a "desmontar" essa pilha de livros em três pilhas menores:

- Uma pilha mostra os tipos de informações (as direções dos dados, como padrões ou tendências).
- A segunda pilha diz quão importante é cada tipo de informação (os valores, que mostram a força ou impacto de cada padrão).
- A terceira pilha mostra como esses padrões estão relacionados aos livros originais.

SVD e PCA

PCA: Encontrando o Melhor Ângulo

Agora, imagine que você está tentando fotografar sua pilha original de livros, mas você quer que a foto capture o máximo de informações com o mínimo de fotos possíveis.

O PCA ajuda a encontrar o melhor ângulo para tirar essas fotos. Esses "melhores ângulos" são basicamente os componentes principais que capturam a maior parte da informação sobre como os livros variam uns dos outros.

SVD e PCA

Relação entre SVD e PCA

Quando você aplica a SVD aos seus dados (pilha de livros), uma das pilhas menores que você obtém mostra os melhores ângulos para olhar seus dados — esses são os componentes principais que o PCA busca. Então, usando a SVD nos dados, você pode encontrar diretamente esses componentes principais.

SVD e PCA

Relação entre SVD e PCA

Dados Centrados: Para usar o PCA, você geralmente subtrai a média de cada variável para que seus dados estejam centrados ao redor do zero. Isso é como garantir que sua câmera está focada corretamente antes de tirar a foto.

Aplicando SVD: Depois de centrar seus dados, você aplica a SVD. A matriz que você obtém (a pilha que mostra como os padrões estão relacionados aos livros originais) contém os mesmos componentes principais que o PCA busca.

Capturando Informações: As "fotos" (componentes principais) tiradas pelo PCA são aquelas que retêm a maior quantidade de informações variadas dos dados originais, o que nos ajuda a entender melhor os dados com menos "fotos" ou menos componentes.

Métricas de Algoritmos de Redução de Dimensionalidade

Erro de reconstrução dos dados. Este é a diferença entre os dados originais e os dados reconstruídos, que são obtidos aplicando a transformação inversa do algoritmo de redução de dimensionalidade. Quanto menor o erro de reconstrução dos dados, melhor o algoritmo é em reter as características essenciais dos dados. Podemos usar várias métricas para quantificar o erro de reconstrução dos dados, como MSE, RMSE ou MAE.

Coeficiente de Correlação de Distâncias: Para algoritmos como t-SNE e MDS, que visam preservar as relações de distância entre pontos, uma métrica útil é o coeficiente de correlação entre as distâncias no espaço original e no espaço de dimensionalidade reduzida. Um coeficiente próximo de 1 indica que as distâncias relativas entre pontos são bem preservadas.

Métricas de Algoritmos de Redução de Dimensionalidade

Taxa de compressão dos dados. Essa é a razão entre o tamanho dos dados originais e o tamanho dos dados reduzidos. Quanto maior a taxa de compressão dos dados, mais eficiente é o algoritmo na redução da dimensionalidade dos dados. No entanto, uma alta taxa de compressão de dados não significa necessariamente uma alta qualidade de reconstrução dos dados. Precisamos equilibrar o compromisso entre a compressão de dados e a reconstrução de dados ao escolher um algoritmo de redução de dimensionalidade.

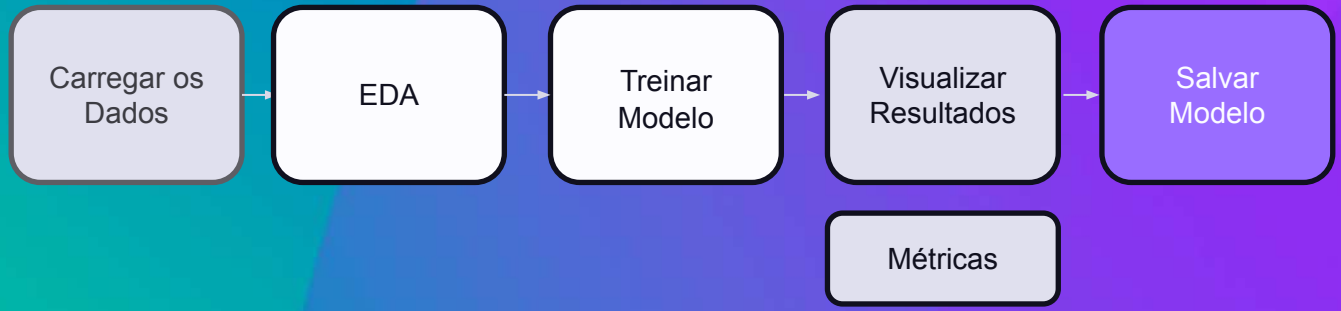
Precisão na classificação dos dados em problemas de aprendizado supervisionado: A precisão é medida pela proporção de rótulos corretamente previstos. Métodos como regressão logística e KNN podem ser usados para esse teste. Algoritmos que mantêm alta precisão de classificação são considerados eficazes em preservar as características discriminativas dos dados.

Projeto – PCA

Uma empresa de fast-food deseja abrir novas lojas ao redor do mundo e precisa apresentar de forma simples, como os países estão organizados em termos de variáveis ou indicadores macroeconômicos como inflação, arrecadação, expectativa de vida, dentre outros.

Desta forma, para que seja possível representar os países num gráfico de apenas 3 dimensões, iremos construir um algoritmo de redução de dimensionalidade que reduza a quantidade de variáveis, permitindo assim a visualização necessária para suportar a tomada de decisão para a Diretoria.

Estrutura do Projeto



Code Time ...

rocketseat.com.br