Algoritmos Não Supervisionados – t-SNE

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Algoritmos Não Supervisionados

t-SNE

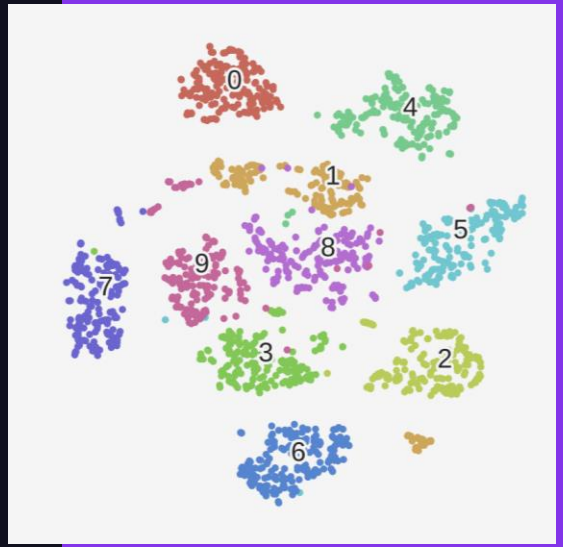
O objetivo deste módulo é apresentar o algoritmo de t-SNE que pertence a classe de algoritmos de redução de dimensionalidade e trabalharemos num projeto para um empresa de cosméticos onde faremos o processo completo desde o EDA até a visualização dos resultados.

Agenda

- O que é t-SNE
- Etapas do t-SNE
- A mecânica do t-SNE
- Desafios e Limitações do t-SNE
- Aplicações do t-SNE
- t-SNE vs PCA
- O hiperparâmetro Perplexity
- Projeto – t-SNE

O que é t-SNE

O t-Distributed Stochastic Neighbor Embedding (t-SNE) é um algoritmo poderoso para redução de dimensionalidade, amplamente utilizado para a visualização de conjuntos de dados de alta dimensão. Desenvolvido por Laurens van der Maaten e Geoffrey Hinton em 2008, o t-SNE é particularmente bem-sucedido em capturar a estrutura local de dados complexos e revelar agrupamentos em um espaço de baixa dimensão (geralmente 2D ou 3D).



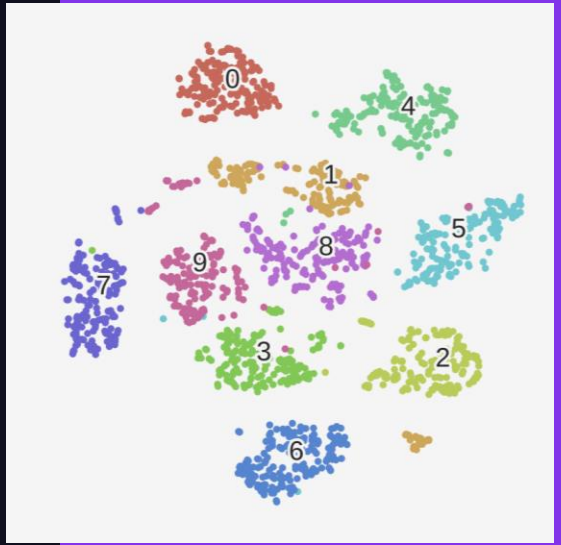
Como o t-SNE Funciona?

O t-SNE começa com a compreensão de que em um espaço de alta dimensão, cada ponto de dados pode ser visto como um ponto em um espaço multidimensional. O algoritmo transforma as distâncias entre os pontos em probabilidades condicionais que representam similaridades. A ideia é que pontos que são próximos uns dos outros têm uma alta probabilidade de serem "vizinhos" e aqueles que estão distantes têm uma probabilidade baixa.

O que é t-SNE

Por que usar t-SNE?

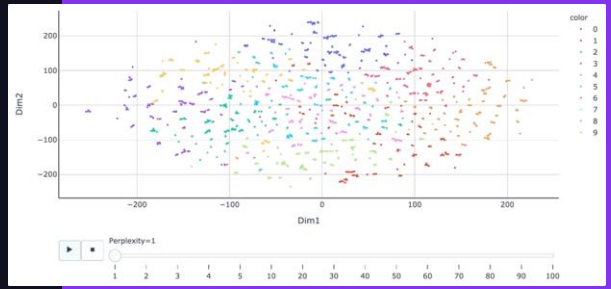
Utilizamos o t-SNE para simplificar a complexidade dos dados, permitindo-nos observar padrões e grupos que são difíceis de visualizar em múltiplas dimensões. Por exemplo, na análise de dados genéticos, o t-SNE pode ajudar a identificar subgrupos de células com padrões de expressão gênica similares, apesar da complexidade e da grande dimensão dos dados.



Etapas do t-SNE

1) Medindo Similaridades

Imagine que cada ponto de dados em um espaço de alta dimensão seja uma pessoa em uma festa. O t-SNE começa calculando o quanto cada pessoa (ponto de dados) prefere ficar perto de cada outra pessoa. Essa "preferência" é medida usando a distância entre eles: pessoas que estão mais próximas são mais propensas a serem "amigas". Isso é convertido em probabilidades, com amigos próximos tendo alta probabilidade de ficarem juntos e conhecidos distantes tendo baixa probabilidade.



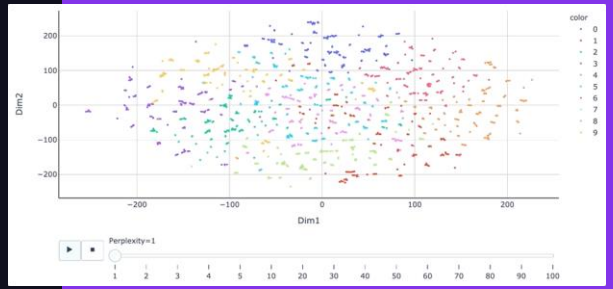
2) Criando o Mapa

Depois, o t-SNE cria um mapa, que é uma representação mais simples onde as pessoas (pontos de dados) também estão presentes, mas desta vez, o espaço é muito menor (como mudar de um grande salão para uma sala pequena). Aqui, ele tenta colocar todos no espaço pequeno mantendo as amizades (similaridades) tanto quanto possível.

Etapas do t-SNE

3) Comparação de Similaridades

Agora, o t-SNE olha para as probabilidades no espaço original (o grande salão) e no espaço pequeno (a sala) e verifica quão bem as amizades foram preservadas. Se duas pessoas que eram próximas no espaço original não estiverem próximas no espaço reduzido, há uma penalidade. O t-SNE ajusta as posições no espaço pequeno tentando minimizar essas penalidades.



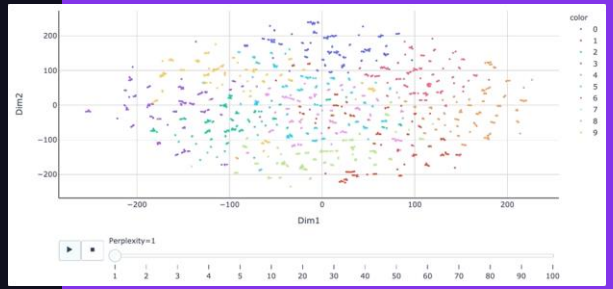
4) Otimização

Este processo de ajustar as posições é um pouco como um jogo de quebra-cabeça onde você tenta fazer todos felizes com suas posições relativas. O t-SNE faz muitos pequenos ajustes, sempre tentando manter amigos próximos, até que encontre uma configuração no espaço menor onde as amizades sejam bem representadas, de acordo com as probabilidades originais.

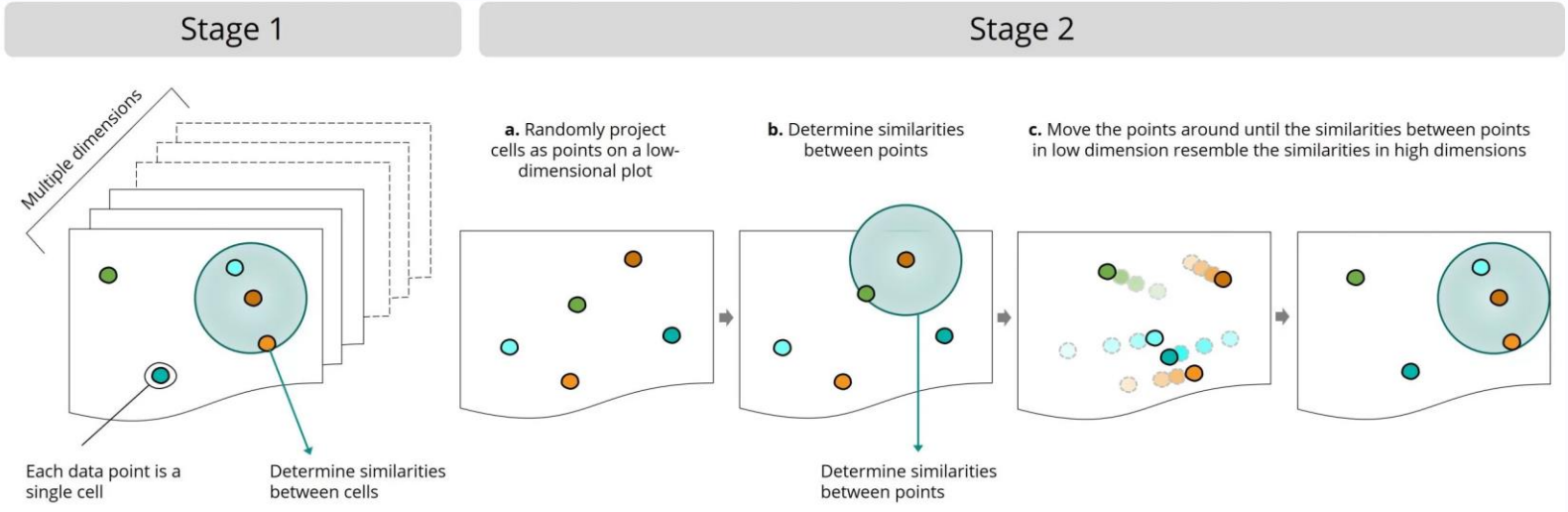
Etapas do t-SNE

5) Resultado Final

Depois de muitos ajustes, o t-SNE finaliza com uma configuração onde, em geral, pessoas que eram próximas no espaço de muitas dimensões ainda estão próximas no espaço de poucas dimensões. Isso permite visualizar e entender padrões e grupos nos dados que seriam muito difíceis de perceber em um espaço de muitas dimensões.



Etapas do t-SNE



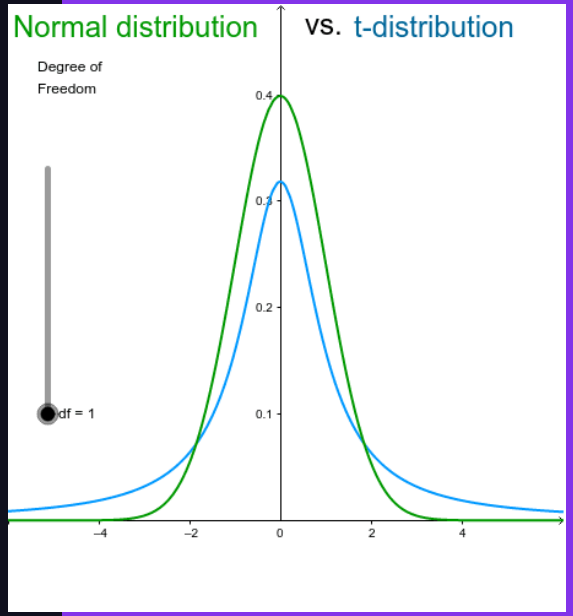
Mecânica do t-SNE

Probabilidades e Similaridades

O primeiro passo no t-SNE é converter as distâncias euclidianas entre os pontos em probabilidades que somam um para cada ponto. Isso é feito usando a distribuição Gaussiana (normal) para os dados de entrada de alta dimensão e a distribuição t de Student para o mapa de baixa dimensão.

O Desafio da "Maldição da Dimensionalidade"

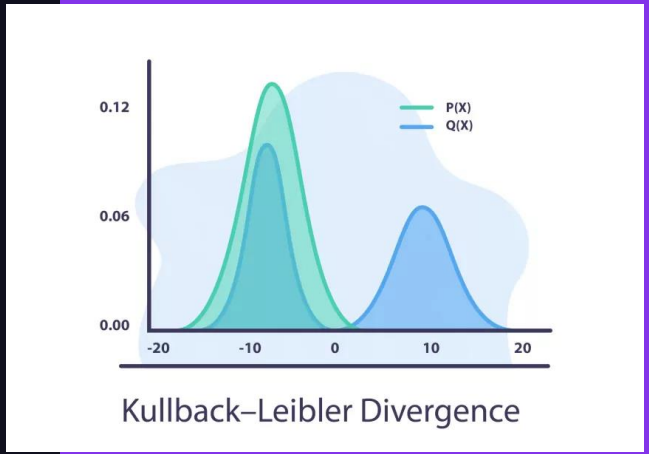
A "maldição da dimensionalidade" refere-se ao fenômeno onde o aumento das dimensões dos dados leva a problemas como o aumento do espaço vazio e a distorção das relações de distância. O t-SNE aborda este desafio focando nas probabilidades de similaridade em vez de apenas replicar as distâncias métricas, o que ajuda a preservar a estrutura local dos dados.



Mecânica do t-SNE

Otimização por Gradiente Descendente

O t-SNE utiliza uma técnica chamada gradiente descendente para minimizar a diferença entre as probabilidades no espaço de alta dimensão e no espaço projetado de baixa dimensão. Este processo é conhecido como minimização da divergência de Kullback-Leibler entre as duas distribuições de probabilidades, que mede quão bem o espaço de baixa dimensão representa o de alta dimensão.



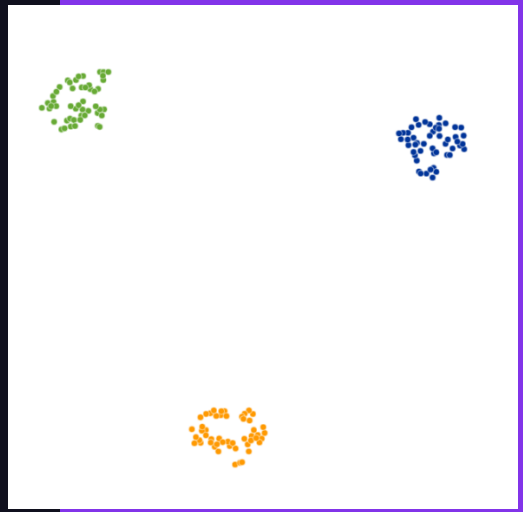
Desafios e Limitações do t-SNE

Sensibilidade aos Hiperparâmetros

Um dos desafios ao usar o t-SNE é sua sensibilidade a hiperparâmetros como o "perplexity", que indica quantos vizinhos próximos cada ponto considera. A escolha deste parâmetro pode afetar significativamente a aparência dos gráficos finais, e não existe um valor único ideal para todos os conjuntos de dados.

Limitações na Interpretação de Clusters

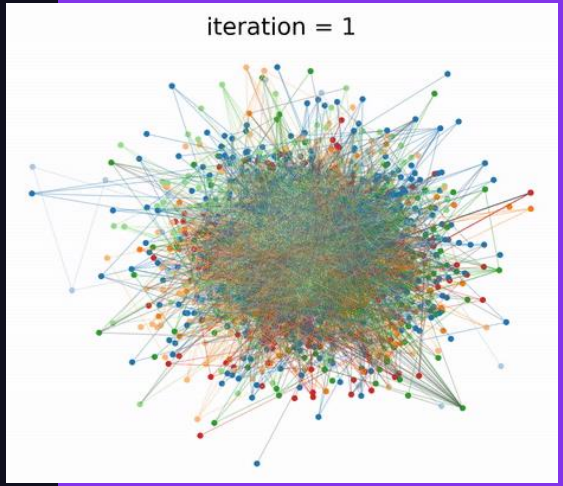
Apesar de o t-SNE ser excelente para visualizar agrupamentos, é importante não tirar conclusões precipitadas sobre os "clusters" visualizados. O t-SNE pode às vezes criar ou exagerar clusters devido à sua forma de otimização e não necessariamente representar divisões naturais nos dados.



Desafios e Limitações do t-SNE

Custo Computacional e Escalabilidade

O t-SNE pode ser computacionalmente intensivo, especialmente para conjuntos de dados muito grandes. Isso ocorre porque o algoritmo compara todos os pares de pontos para calcular as probabilidades, o que pode se tornar impraticável com o aumento do número de pontos.



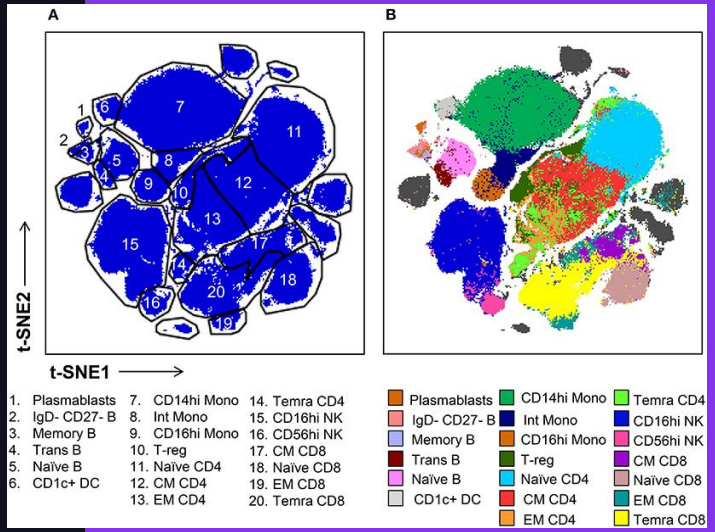
Aplicações do t-SNE

Biologia Computacional

No campo da biologia computacional, o t-SNE é frequentemente utilizado para analisar dados de sequenciamento de células únicas, ajudando os pesquisadores a identificar tipos de células com base em seus perfis de expressão gênica. Este tipo de análise pode revelar novos subtipos celulares e potenciais alvos terapêuticos.

Reconhecimento de Imagens

O t-SNE também é aplicado no reconhecimento de imagens, onde pode ser usado para reduzir a dimensão de representações de imagem antes de classificações ou outras análises. Visualizar essas representações pode ajudar a entender como diferentes categorias de imagens são percebidas pelo modelo de IA.



t-SNE vs PCA

PCA (Análise de Componentes Principais)

Método: Linear, o que significa que ele projeta os dados originais em direções que maximizam a variância, sem tentar preservar relações não-lineares.
Vantagens: Rápido e eficiente para grandes conjuntos de dados; bom para reduzir ruídos e destacar as características mais importantes dos dados.
Desvantagens: Não é eficaz para capturar complexidades e padrões não-lineares nos dados; pode perder informações importantes em dados que têm estrutura intrincada.
Melhores usos: Análise inicial para obter uma visão geral das principais variações nos dados; útil em campos como finanças e outras áreas onde as relações lineares são predominantes.

t-SNE

Método: Não-linear, focado em preservar a estrutura local dos dados ao mapear pontos próximos em alta dimensão para pontos próximos em baixa dimensão.
Vantagens: Excelente para visualizar agrupamentos e padrões em dados de alta complexidade; muito eficaz em preservar a vizinhança local, permitindo visualizações intuitivas.
Desvantagens: Computacionalmente intensivo, especialmente com grandes datasets; sensível a parâmetros como perplexidade, o que pode exigir ajustes finos para resultados ótimos.
Melhores usos: Análise exploratória de dados para descobrir agrupamentos e padrões ocultos, especialmente útil em biologia, marketing e qualquer campo onde as relações não-lineares são importantes.

O hiperparâmetro Perplexity

Imagine que você está numa festa e quer decidir com quantas pessoas você deve conversar. Se você conversar com muitas pessoas, vai ter uma boa ideia geral de quem está na festa. Se você conversar apenas com algumas pessoas, vai entender muito bem essas poucas pessoas, mas não tanto sobre a festa como um todo. A **perplexidade** no t-SNE é semelhante a isso: é como se fosse um número que você escolhe para decidir quantas "pessoas" (ou pontos de dados) cada ponto deve considerar como seus vizinhos próximos.

A **perplexidade** ajuda a determinar:

Quantos vizinhos próximos cada ponto deve considerar: Um número pequeno significa que cada ponto só olha para seus vizinhos muito próximos (conversa apenas com seus amigos mais próximos). Um número grande significa que cada ponto considera muitos outros pontos como seus vizinhos (conversa com muitas pessoas na festa).

Como Escolher a **Perplexidade**?
Não muito alta, não muito baixa: Se a perplexidade é muito alta, o t-SNE pode perder os pequenos detalhes, porque cada ponto está tentando ser amigo de quase todos. Se é muito baixa, os pontos podem acabar formando muitos pequenos grupinhos isolados que podem não representar bem a festa (os dados).
Depende dos dados: Não existe um número mágico perfeito para todos os conjuntos de dados. Você normalmente tem que experimentar diferentes valores para ver qual dá a melhor visão geral dos seus dados.

Projeto – t-SNE

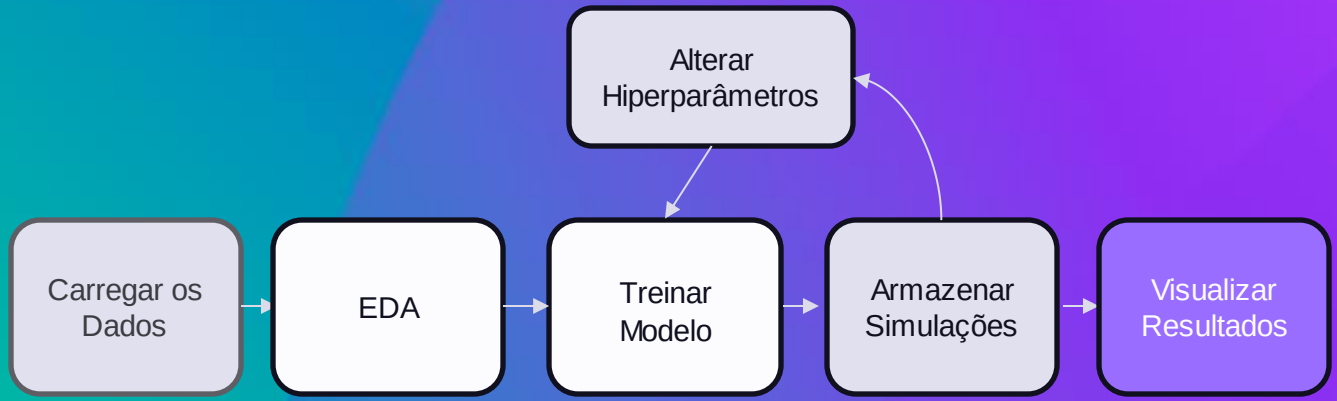
Uma empresa de cosméticos deseja trabalhar num algoritmo de recomendação para seus clientes de e-commerce, de forma a indicar produtos com base em características similares, tais como tipo de pele, marca e preço.

Porém, como a combinação entre marcas, tipo de produtos e demais características é muito grande, analisar estes dados antes de propor um algoritmo é extremamente desafiador e a empresa não gostaria de criar um algoritmo que recomende produtos que não façam sentido.

Desta forma, iremos desenvolver uma ferramenta visual para que esta empresa possa visualizar seus produtos num chart 2D, de forma a identificar se produtos próximos em termos de características estão próximos e, pra isso, usaremos o algoritmo t-SNE.

Como comentado anteriormente, este algoritmo possui uma sensibilidade a hiperparâmetros, especialmente o perplexity e apresentamos este chart de forma animada, com diversos parâmetros deste hiperparâmetro.

Estrutura do Projeto



Code Time ...

rocketseat.com.br