T - SNE

t-distributed Stochastic Neighbor Embedding

Júlia Borges - 211039063
Ana Luiza Carneiro - 180012801

A maldição da dimensionalidade:

Dimensões = recursos ou atributos dos dados (variáveis)

Refere-se aos vários desafios e complicações que surgem ao analisar e organizar dados em espaços de alta dimensão

À medida que se adiciona mais dimensões ao conjunto de dados, o volume do espaço aumenta exponencialmente, ou seja, os **dados se tornam esparsos**

Problemas: aumento da computação, ajuste excessivo, as distâncias perdem o sentido, degradação do desempenho e desafios de visualização.

A principal solução é a "**redução da dimensionalidade**"

O que é t-SNE?

- **Técnica de redução de dimensionalidade;**
- **Não linear;**
- **Não supervisionado;**
- Particularmente útil para **exploração e visualização de dados de alta dimensão**, mas tem outros usos como identificar anomalias e padrões ou para agrupamento e classificação;
- Pode ser **usado para diversos tipos de dados** como imagens, áudio, dados biológicos e dados individuais.

Como funciona??

1

Mapeia pontos de dados de alta dimensão, **calculando a semelhança** entre cada ponto de dados e todos os outros pontos no conjunto de dados no espaço de alta dimensão e **representando essa semelhança como probabilidades.**

2

Depois, constrói uma distribuição de probabilidade semelhante no espaço de dimensão inferior e minimiza a diferença entre as duas distribuições usando a técnica **gradiente descendente.**

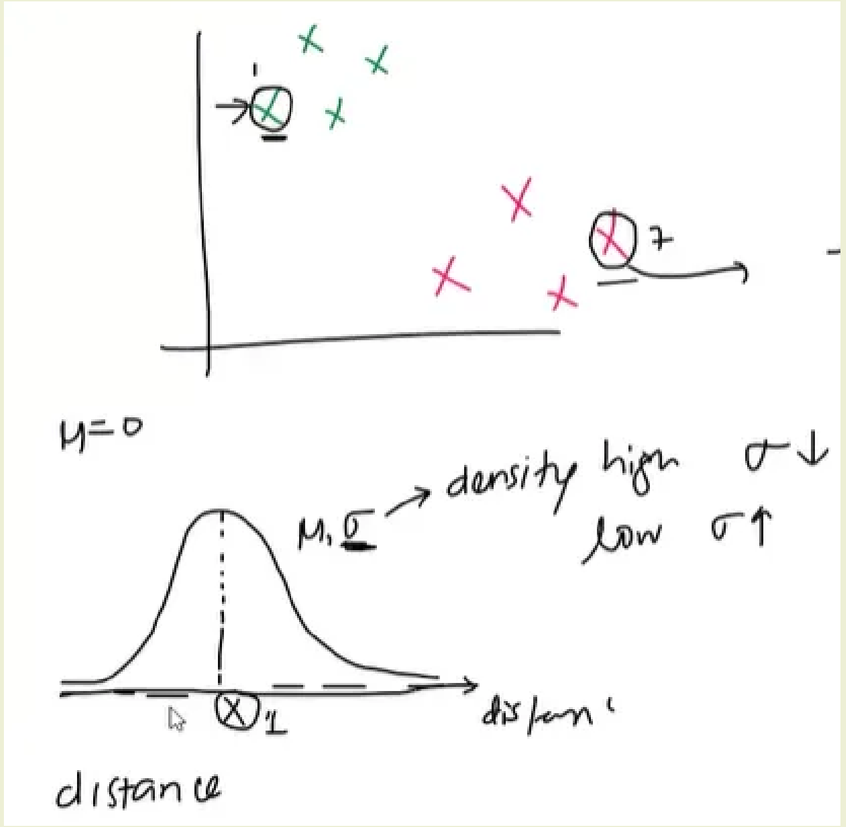
Se concentra em **preservar as relações locais** entre os pontos em um espaço de dimensão inferior dando uma sensação e uma **intuição de como os dados são organizados em dimensões mais altas.**

1 Cálculo da semelhança

Para cada ponto de dados, uma **distribuição gaussiana** é plotada em torno dele com uma média de zero e desvio padrão determinado com base na densidade de pontos próximos em torno desse ponto.

Por que a distribuição gaussiana?

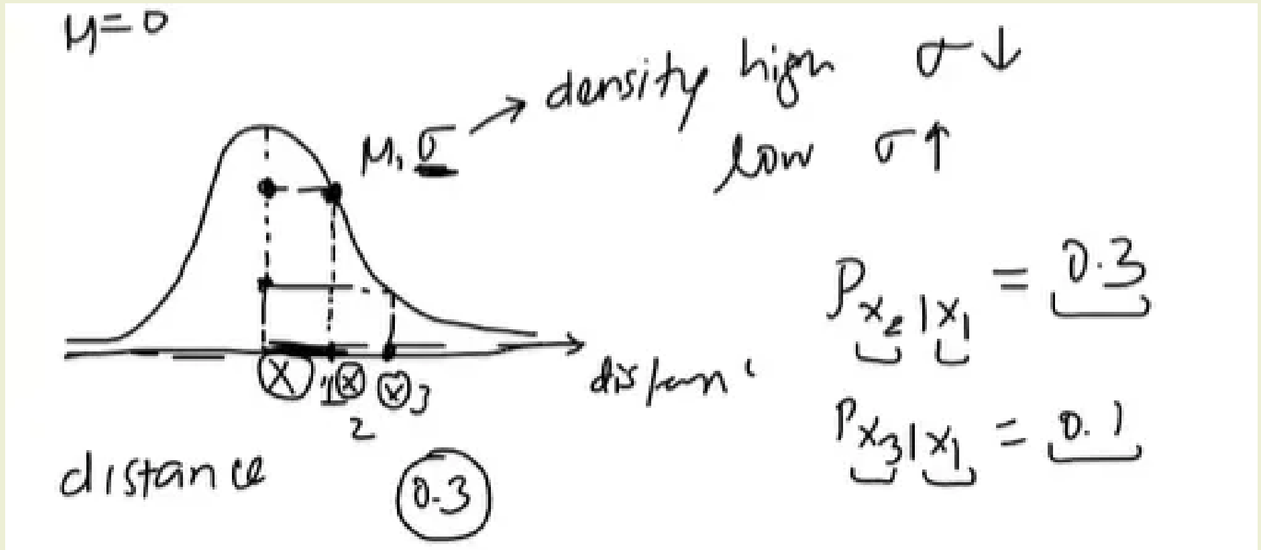
Por que abordagem probabilística?

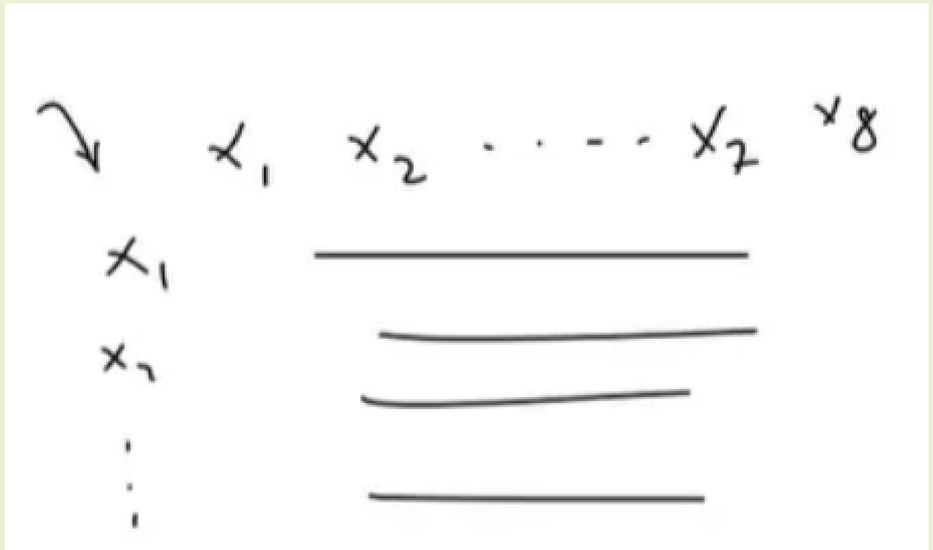


1 Cálculo da semelhança

Consideramos as distâncias de um ponto de referência, ex: x1.

- Para cada ponto no conjunto de dados, calculamos sua distância para x1 e traçamos essas distâncias ao longo do eixo x.
- O eixo y representa a densidade de probabilidade correspondente, permitindo determinar a probabilidade de cada ponto de dados em relação a x1.
- Resulta em uma matriz n * n em que a **pontuação de similaridade** para cada ponto de dados é registrada em relação a todos os outros pontos de dados.
- **Maior o valor de P entre dois pontos significa que eles são vizinhos um do outro, enquanto valor baixo indica que eles são pontos diferentes.**





2 Redução

1. **Reduz a dimensão** para uma dimensão inferior de modo que os pontos são distribuídos aleatoriamente no eixo x. **Recalcula-se a pontuação de similaridade para cada ponto em relação aos outros**, resultando em **outra matriz n * n**;
2. Para garantir que os pontos vizinhos permaneçam próximos na dimensão inferior, busca-se **alinhar a matriz de dimensão inferior com a de dimensão superior**, **ajustando a posição dos pontos iterativamente** até que a matriz de similaridade na dimensão inferior se assemelhe à da dimensão superior o mais próximo possível;
3. Isso é obtido **minimizando a divergência** entre a distribuição de probabilidade da dimensão alta original e da dimensão baixa. O algoritmo usa a **descida do gradiente para minimizar a divergência**. A incorporação de dimensão inferior é otimizada para um estado estável.

2 Redução

Queremos minimizar a diferença entre as semelhanças de pontos no espaço de alta dimensão e suas contrapartes no espaço de baixa dimensão. **Como?**

Divergência de Kullback-Leibler (KL) entre a distribuição de probabilidade de alta dimensão e a distribuição de probabilidade de baixa dimensão.

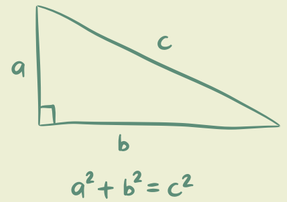
Quantifica o quão diferentes são as semelhanças entre pares entre os espaços de alta e baixa dimensão. **A divergência KL baixa é um sinal de melhores resultados.**

Para minimizar a divergência KL, usamos **gradiente descendente**. Essa **técnica de otimização iterativa** ajusta as posições dos pontos no espaço de baixa dimensão.

Cada iteração calcula o gradiente da função de custo em relação às posições dos pontos no espaço de baixa dimensão. Ao atualizar as posições dos pontos, gradualmente converge para a melhor configuração de semelhança.

O gradiente indica a direção em que se deve mover cada ponto para reduzir a diferença entre as semelhanças de alta e baixa dimensão.

Formulação matemática



- Distribuição de **Similaridade no Espaço Original**

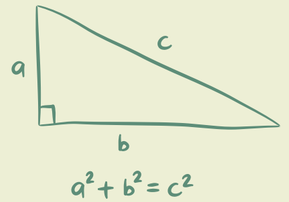
Dado um conjunto de dados $X = \{x_1, x_2, ..., x_N\}$ com similaridades par a par p_{ij} , tal que p_{ij} representa a probablidade de escolher x_j como vizinho de x_i sob uma distribuição Gussiana centrada em x_i , em que $i \neq j$, e é definida por:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Para obter uma distribuição simétrica, usamos:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Formulação matemática

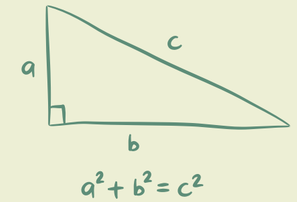


- Distribuição de **Similaridade no Espaço Reduzido:**

Para o espaço de dimensão reduzida $Y = \{y_1, y_2, \ldots, y_N\}$, as similaridades q_{ij} são definidas de forma análoga, mas nesse caso utiliza-se a distribuição t-Student com 1 grau de liberdade. Assim é dado:

$$q_{ij} = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i}(1+\|y_i - y_k\|^2)^{-1}}$$

Formulação matemática



- Função de Custo (Divergência KL)

A função de custo que o t-SNE minimiza é a divergência de Kullback-Leibler (KL) entre as distribuições **P** (original) e **Q** (reduzida), ponderada pela Perplexidade:

$$C = \sum_i KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Para encontrar como os pontos y_i devem ser atualizados para minimizar **C,** tomamos a **derivada parcial** de **C (gradiente da função de custo):**

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}$$

Esse gradiente é usado para atualizar as posições dos pontos no espaço reduzido.

Hiperparâmetros

Perplexidade:

- Controla o **número efetivo de vizinhos que cada ponto considera durante o processo de redução** de dimensionalidade;
- Não existe um valor único para todos, valores diferentes podem revelar estruturas diferentes, portanto, deve-se experimentar um intervalo de valores (usualmente entre 5 e 50);
- Executa-se um loop para obter a métrica de divergência KL em várias perplexidades com determinado intervalo de pontos. De forma que **usaremos o valor de perplexidade no algoritmo t-SNE no qual a divergência KL tornou-se constante**.

Outros parâmetros como a **taxa de aprendizado** e o **número de iterações** também afetam os resultados.

t-SNE X PCA

Ambas são técnicas de redução de dimensionalidade

- **PCA (Análise de Componentes Principais):**

1. **Técnica Linear**
2. Busca identificar os componentes principais subjacentes nos dados, projetando-os em dimensões inferiores, minimizando a variação e preservando grandes distâncias entre pares, ou seja, se preocupa manter grandes distâncias entre pares para maximizar a variação.

- **t-SNE:**

1. **Técnica não linear**
2. Se concentra em preservar as semelhanças entre os pares de pontos de dados em um espaço de dimensão inferior, se preocupando em preservar pequenas distâncias entre pares.

- **PCA preserva a variação nos dados**, enquanto o **t-SNE preserva as relações entre os pontos de dados!**

t-SNE X PCA

PCA	t-SNE
Redução linear	Redução não-linear (e linear também)
Preserva a variância global	Preserva similaridades locais entre pares
Extração de características, redução de ruído	Visualização de dados complexos e de alta dimensão, detecção de padrões e anomalias
Componentes principais	Representação de baixa dimensão
Baixo custo computacional	Alto custo computacional
Fácil de interpretar	Difícil de interpretar

Hora de praticar !

O banco de dados : Wine

Baseado em um estudo químico sobre diferentes variedades de vinho cultivadas na região da Itália.

Características do Dataset

- Número de Linhas: 178 (amostras de vinho).
- Número de Colunas: 14 (atributos químicos + classe).

- Class: Tipo de vinho (3 classes: 1, 2, 3).
- Alcohol: Porcentagem de álcool.
- Malic Acid: Quantidade de ácido málico.
- Ash: Quantidade de cinzas.
- Alcalinity of Ash: Alcalinidade das cinzas.
- Magnesium: Quantidade de magnésio.
- Total Phenols: Fenóis totais.
- Flavanoids: Flavonoides.
- Non-Flavanoid Phenols: Fenóis não flavonoides.
- Proanthocyanins: Proantocianinas.
- Color Intensity: Intensidade da cor.
- Hue: Matiz.
- OD280/OD315: Proporção de absorção.
- Proline: Quantidade de prolina.

O banco de dados : Expressão de genes

Contém expressões gênicas de pacientes diagnosticados com diferentes tipos de tumores:

- BRCA: Câncer de mama
- KIRC: Carcinoma de células renais claras
- COAD: Câncer colorretal
- LUAD: Adenocarcinoma de pulmão
- PRAD: Adenocarcinoma de próstata

Características do Dataset

- Número de Linhas: 801 (pacientes).
- Número de Colunas: 20.531 (expressões gênicas).

Antes de começar

- Reprodutibilidade: o t-SNE começa com uma inicialização aleatória, levando a resultados diferentes cada vez que você o executa.
- Limpeza de Dados
  - Tratamento de valores ausentes: substituição (e.g., média, mediana) ou exclusão de linhas/colunas.
  - Eliminação de valores duplicados
- Redução de Dimensionalidade: dados de alta dimensionalidade podem conter redundâncias ou ruído que prejudicam a visualização
- Normalização dos Dados: t-SNE é sensível às escalas das variáveis.

Dúvidas ?