Redução de Dimensionalidade

Prof. André Gustavo Hochuli

gustavo.hochuli@pucpr.br
aghochuli@ppgia.pucpr.br
github.com/andrehochuli/teaching

Plano de Aula

- Conceitos Básicos - Problemas de Alta Dimensão
- PCA
- t-SNE

About me

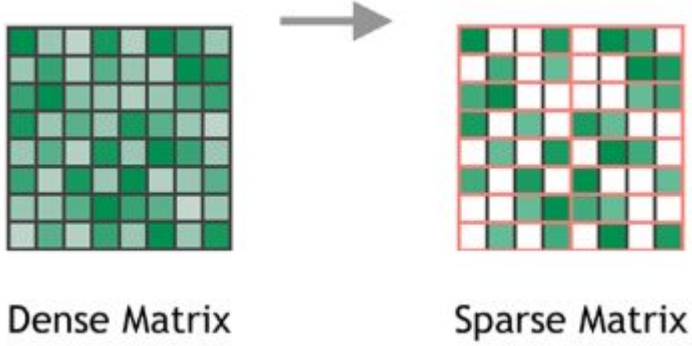
- Andre G. Hochuli
- Researcher PPGIa/PUCPR
- Research topics
  - Computer Vision
  - Deep Learning
  - Pattern Recognition

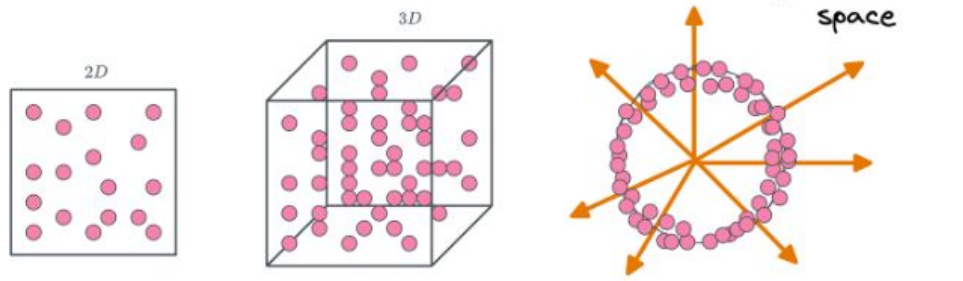
- Contact Canvas
- gustavo.hochuli@pucpr.br
- https://www.linkedin.com/in/andre-hochuli-96117b18/
- https://scholar.google.com.br/citations?user=pbekYw4AAAAJ&hl=pt-BR

Conceitos Básicos

A Maldição da Dimensionalidade

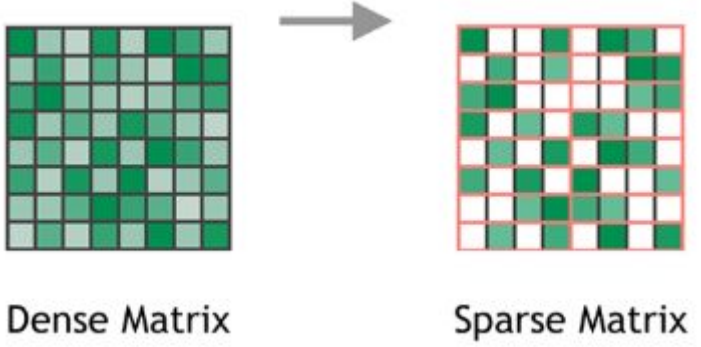
- Em espaços de alta dimensão, o volume cresce exponencialmente
- Dados tornam-se esparso
- Métricas de similaridade tornam-se menos discriminativas

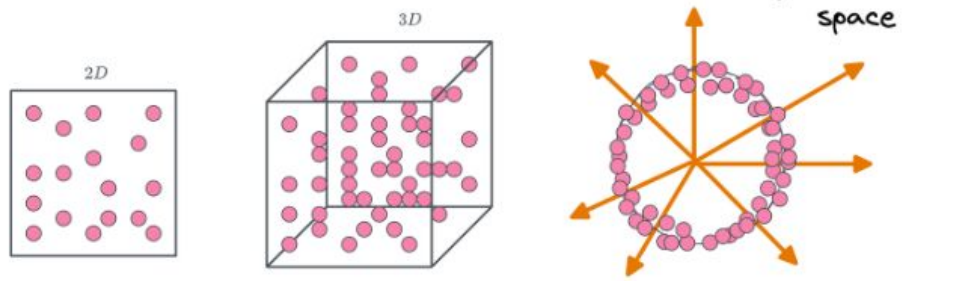




A Maldição da Dimensionalidade

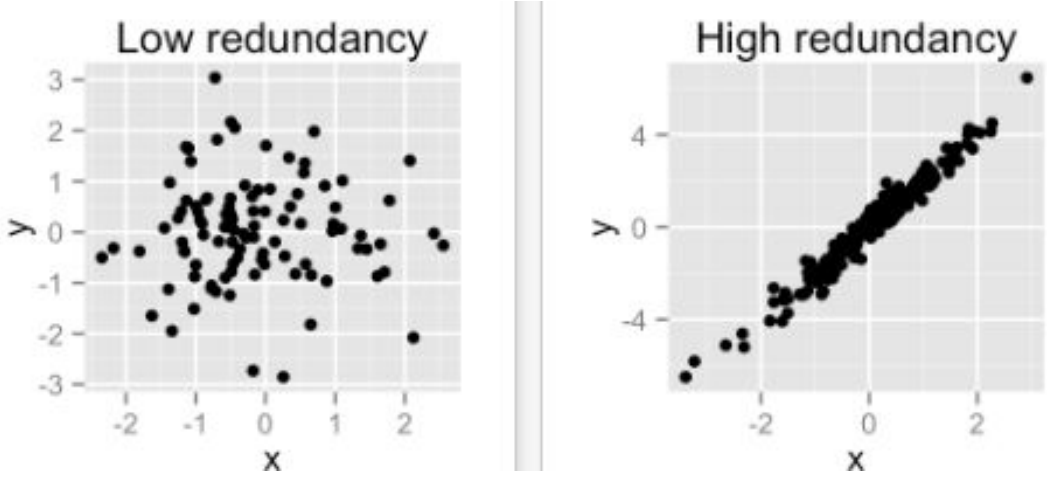
- Implicações em ML
  - Aumento da variância
  - Overfitting
  - Aprendizados baseados em distância tendem a falhar (knn, clustering)
  - Custo computacional





Redundância

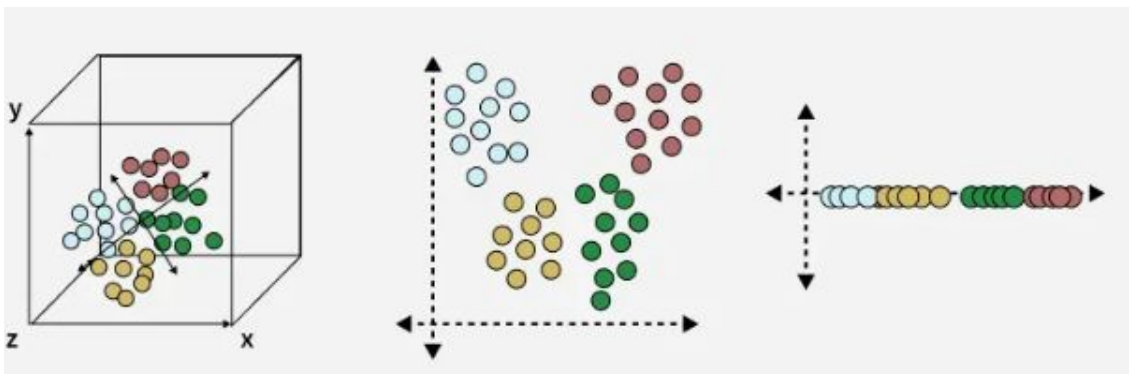
- Muitas variáveis tendem a ser altamente correlacionadas
- Dados representativos geralmente podem ser representados em subespaço menor



Redução de Dimensionalidade

- Representação Compacta
- Eliminação de Ruídos
- Redução da Variância / Regularização
- Visualização
  - 2D / 3D
  - Visualização de Clusters
  - Análise Exploratória

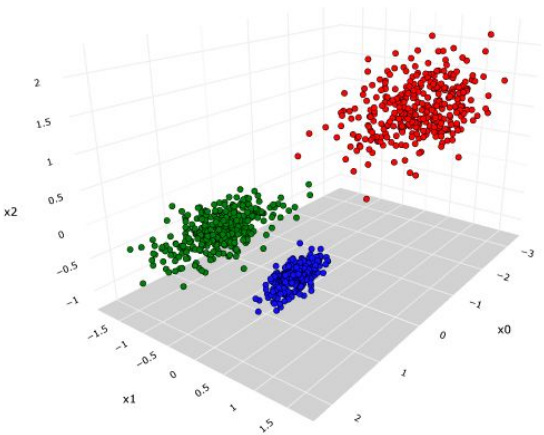
- "A questão não é apenas reduzir dimensão, mas preservar a informação relevante para a tarefa."

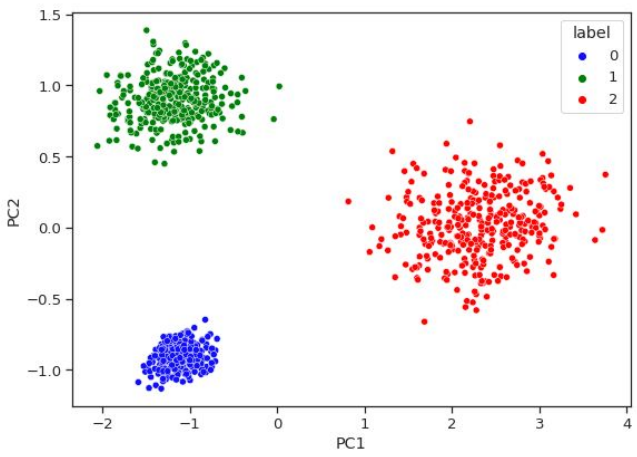


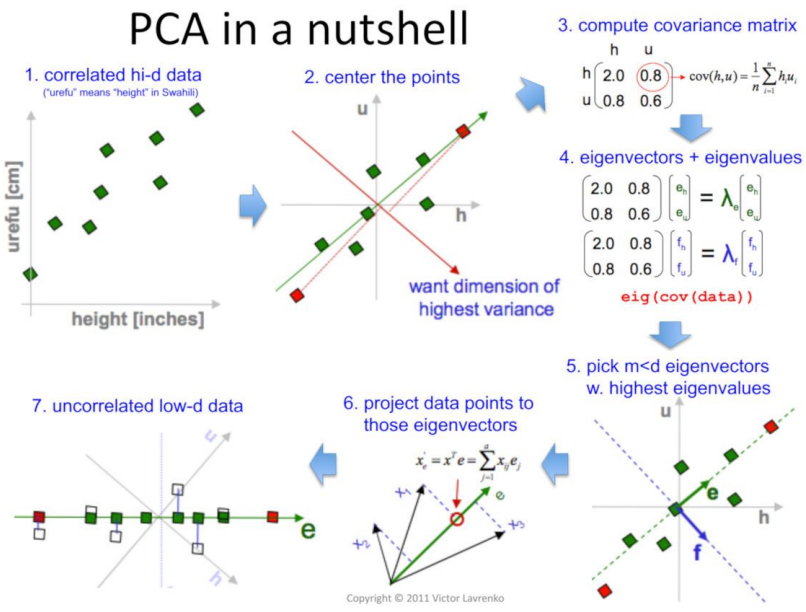
Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- Reorganiza os dados para capturar o máximo de variação possível
- Elimina variáveis de alta correlação
- Descarta informações (dimensões) pouco relevantes
- Baseado na transformação linear do espaço (algebra linear)







PCA

- Matriz de covariância
  - Mede a **variância individual** de cada variável (elementos da diagonal)
  - Mede a **covariância entre pares de variáveis**, isto é, como duas variáveis **variam conjuntamente** ou o grau de sua **dependência linear** (elementos fora da diagonal).

$$\begin{array}{c} \\ x \\ y \end{array} \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ x \\ y \\ z \end{array} \begin{array}{c} \begin{array}{ccc} x & y & z \end{array} \\ \begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix} \end{array}$$

PCA

- Componente Principal:
  - Combinação Linear das variáveis maximizando a variância sob restrição de ortogonalidade

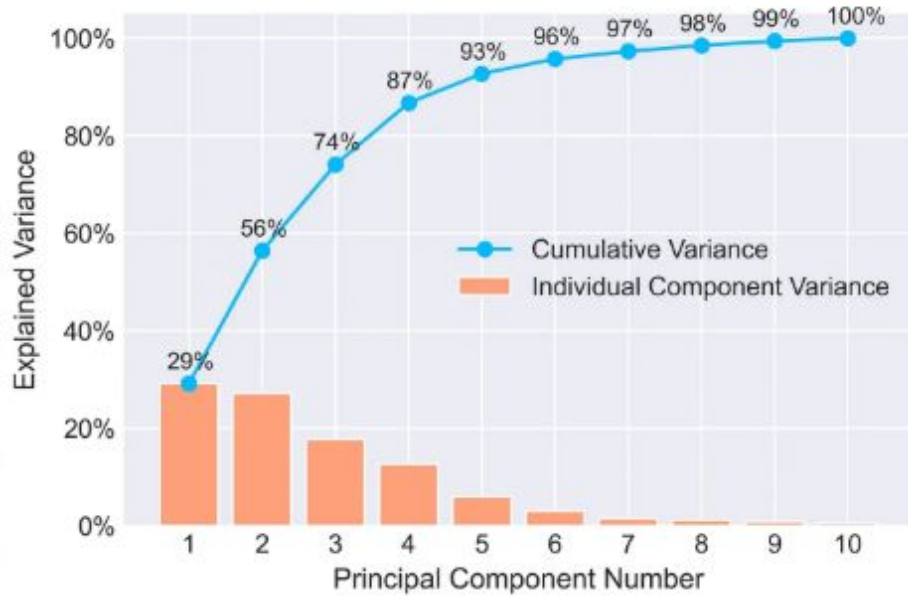
$$\begin{bmatrix} 3 & 4 & -2 \\ 1 & 4 & -1 \\ 2 & 6 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 6 \end{bmatrix}$$

- Autovetor (eigenvector)
  - Representa a direção de um componente principal
  - Define um eixo ortogonal no espaço transformado
  - Os dados são projetos nessa direção para obter a componente principal

- Autovalores (eigenvalues)
  - Mede a importância relativa do componente principal
  - Determina a variância explicada daquele componente

PCA

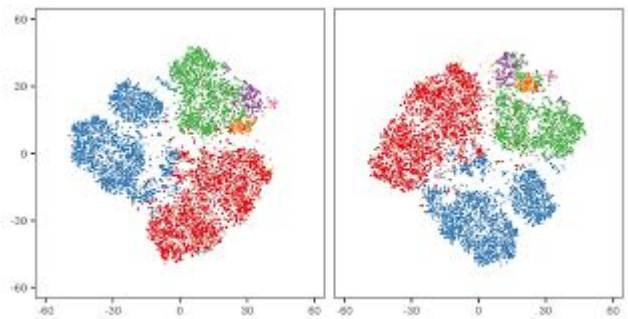
- Análise de Distribuição da Variância e Variância Acumulada
  - Identifica **quantos componentes** são necessários para representar os dados.
  - Quantifica a **importância relativa** de cada componente.
  - Permite detectar **direções pouco informativas (ruído)**.
  - Fundamenta a **redução de dimensionalidade com preservação de informação**.



t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE

- Visualização de representações de alta dimensionalidade (embeddings latentes)
- Preservação de estrutura local via minimização (KL Divergence)
- Preserva a geometria intra-cluster (estrutura de vizinhança local)
- Mitiga o “crowding problem”
  - Permite que pontos moderadamente distantes no espaço original sejam posicionados adequadamente no espaço reduzido, reduzindo compressão excessiva.
- Aplicações Técnicas em ML
  - Análise qualitativa de separabilidade de classes
  - Avaliação qualitativa de embeddings (CNN, CAEs, STL)

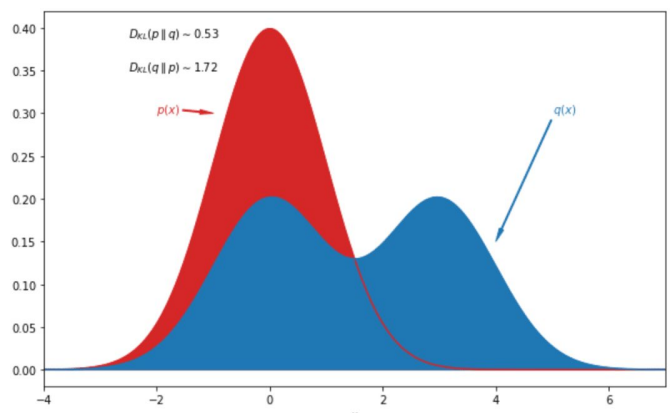


t-SNE

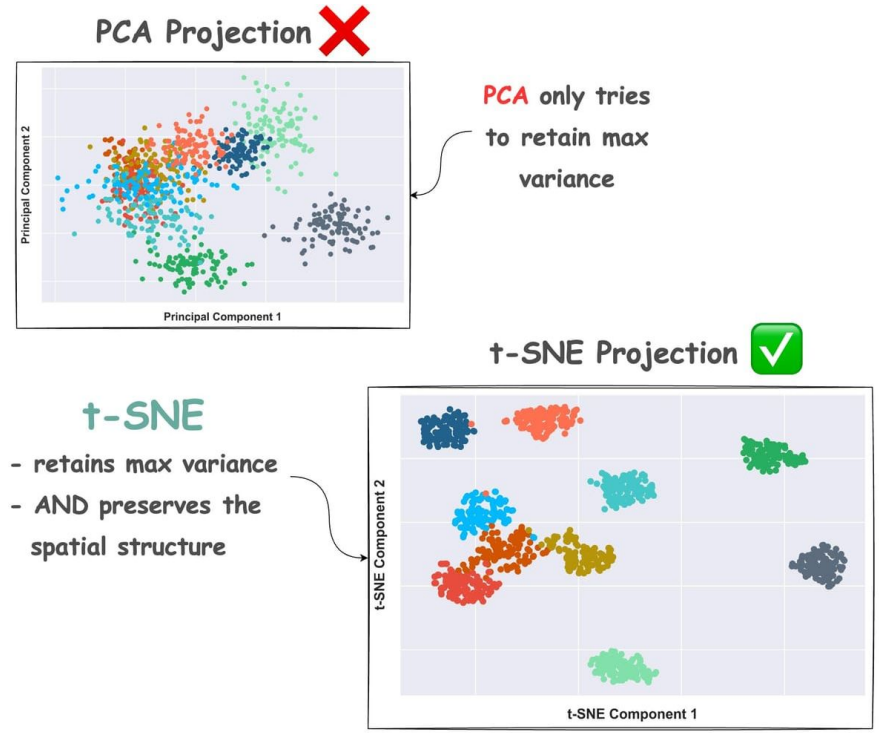
- Estrutura Local
  - Preservação das relações (i.e distâncias) entre vizinhos próximos no espaço de alta dimensão
  - Coesão Intra-Cluster

- Para tal, t-SNE minimiza a divergência KL (Kullback-Leibler)
  - P_ij = probabilidade de similaridade no espaço original
  - Q_ij = probabilidade no espaço projetado

$$\mathcal{L} = D_{KL}(P\|Q) = \sum_{i \neq j} P_{ij} \log \frac{P_{ij}}{Q_{ij}}$$



Considerações Finais



- PCA: Qual é a melhor forma de resumir esses dados mantendo o máximo possível de informação?

- t-SNE: “Quais pontos são parecidos entre si e como posso visualizar esses grupos?

PCA vs t-SNE

Critério	PCA (Principal Component Analysis)	t-SNE (t-Distributed Stochastic Neighbor Embedding)
Natureza do método	Redução de dimensionalidade **linear**	Visualização **não linear**
Base matemática	Álgebra linear (autovalores/autovetores, variância)	Probabilidade + otimização (minimiza divergência KL)
Objetivo principal	Preservar a **variância global** dos dados	Preservar **vizinhanças locais** (clusters)
Interpretabilidade	Alta (componentes têm significado linear)	Baixa (eixos não possuem interpretação direta)
Estabilidade	Determinístico e estável	Pode variar entre execuções
Custo computacional	Rápido e escalável	Mais custoso
Uso típico	Pré-processamento, compressão, entrada para modelos	Visualização exploratória de embeddings e clusters