Redução de Dimensionalidade

Uma Análise Teórica e Experimental

Felipe Ramos, Nina Hirata

Instituto de Matemática, Estatística e Ciência da Computação da Universidade de São Paulo

Resumo

Em muitas aplicações de aprendizado de máquina, os dados são representados em espaços de alta dimensão, o que dificulta tanto a análise quanto a visualização. A redução de dimensionalidade oferece um meio de simplificar esses dados preservando, na medida do possível, sua estrutura geométrica. Do ponto de vista teórico, o Lema de Johnson–Lindenstrauss [2] garante que é possível projetar conjuntos de alta dimensão em espaços muito menores preservando, de forma controlada, as distâncias entre pontos. A partir desse fundamento, exploramos técnicas modernas como t-SNE [6] e UMAP [3], que vão além das projeções lineares ao buscar manter relações locais e padrões de vizinhança intrínsecos aos dados. Esses métodos têm se mostrado particularmente úteis para visualização e análise exploratória em cenários de grande complexidade.

Desafio: Redução de dimensionalidade é viável?

Se quisermos preservar todas as distâncias euclidianas exatamente, então não. Mas e se tolerarmos alguma deformidade?

Teorema (Lema de Johnson–Lindenstrauss [2]). *Para qualquer conjunto de n pontos em $\mathbb{R}^D$ existe uma aplicação linear $f : \mathbb{R}^D \to \mathbb{R}^k$ com $k = O(\log n/\varepsilon^2)$ tal que para todos x, y:*

$$(1-\varepsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\varepsilon)\|f(x)-f(y)\|^2.$$

Soluções lineares para visualização

O PCA [4] projeta cada imagem de dígito em duas dimensões buscando maximizar a variância, isto é, separar bem os exemplos mais diferentes entre si. No mapa resultante, vemos grandes clusters correspondendo a dígitos visualmente distintos, como 0 e 1, que ocupam regiões opostas.

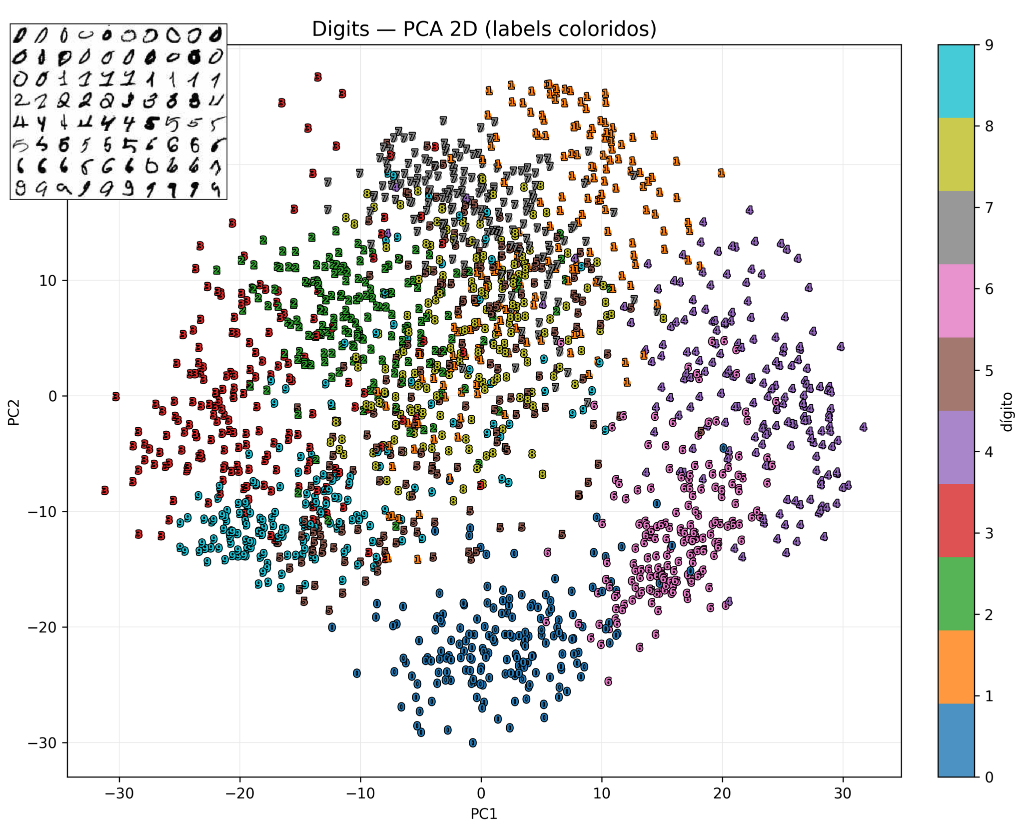
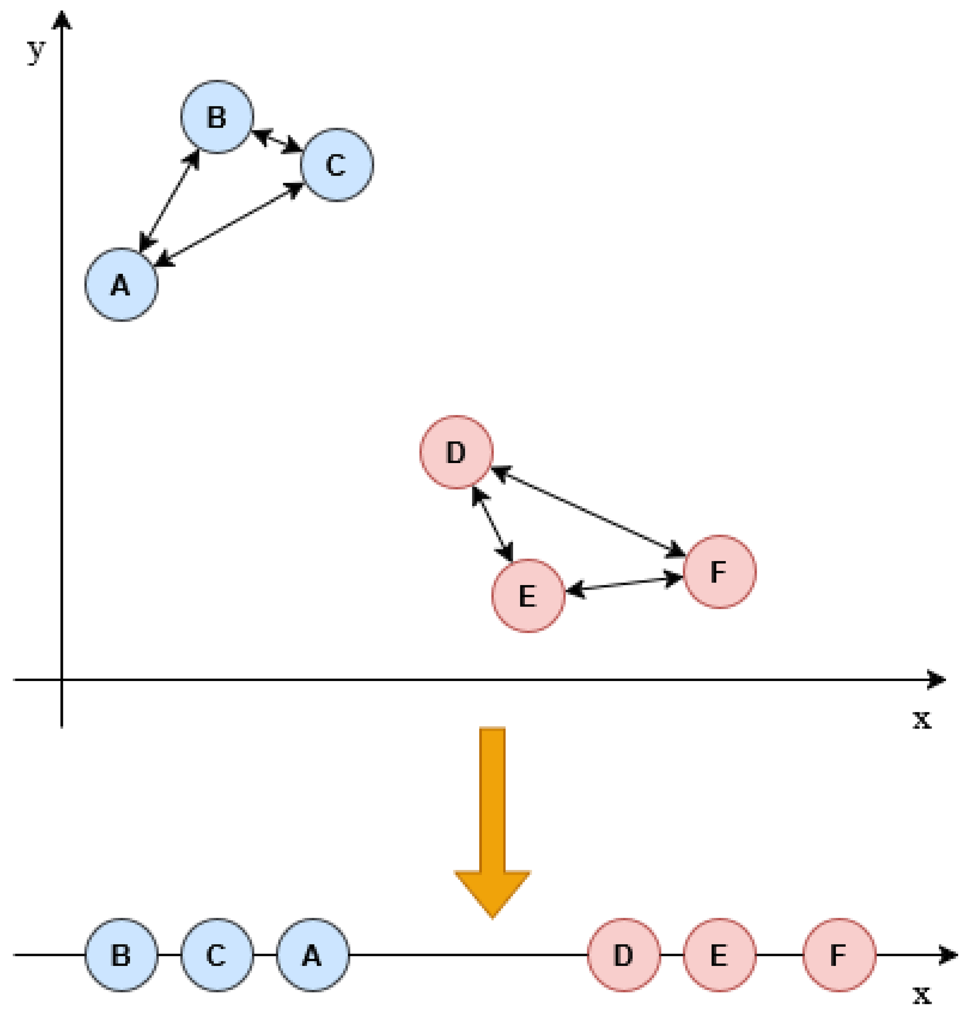


Figura 1: Representação em duas dimenções do banco de dígitos MINST, utilizando o PCA.

O método capta tendências globais dos dados, mas não organiza bem a estrutura local: dígitos semelhantes podem se misturar e, sem cores, o mapa se torna uma única nuvem. Isso revela a limitação central do PCA para visualização: ele privilegia distâncias grandes, enquanto padrões finos exigem métodos que preservem vizinhanças locais.

Olhar para os vizinhos: t-SNE e UMAP

Métodos não lineares baseados em vizinhança, como o t-SNE, buscam preservar relações locais entre pontos ao projetar dados em duas dimensões. O UMAP segue a mesma filosofia, mas constrói uma noção mais explícita de similaridade no espaço original: para cada ponto (por exemplo, A–F), calculam-se as distâncias aos vizinhos e aplica-se uma curva exponencial decrescente para convertê-las em *escores de similaridade*. Esses escores representam o quão próximos os pontos realmente são nos dados originais, formando vizinhanças fortes (como os grupos {A,B,C} e {D,E,F}).



Em seguida, o UMAP inicializa uma representação em baixa dimensão e move os pontos gradualmente: aproxima aqueles com alta similaridade e afasta aqueles com baixa similaridade, até que o mapa 2D preserve a mesma estrutura de agrupamentos observada no espaço original. O resultado é uma visualização rápida, interpretável e capaz de revelar tanto clusters quanto outliers.

Resultados

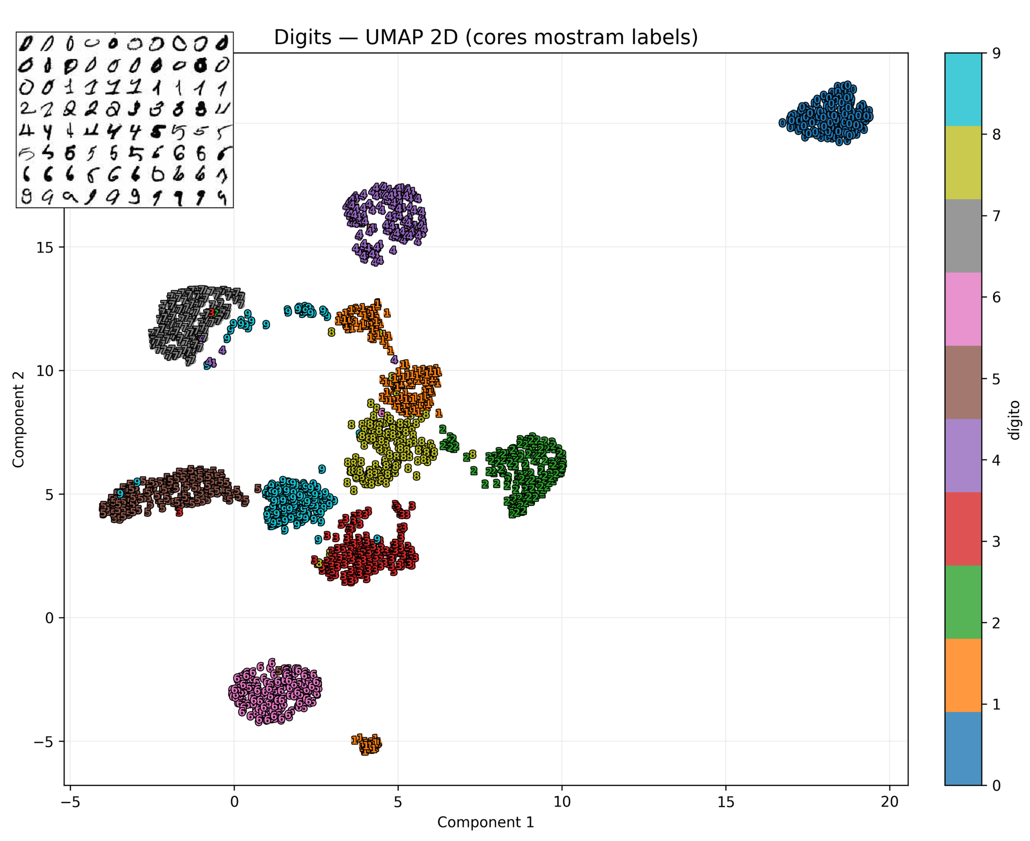


Figura 2: Representação em duas dimenções do banco de dígitos MINST, utilizando o UMAP com 15 vizinhos

Figura 3: Mapa 2D de 400 imagens do conjunto CIFAR-10 obtido pela redução de dimensionalidade via UMAP. Cada miniatura representa uma imagem original do conjunto. Inicialmente, suas representações de alta dimensão foram extraídas por uma ResNet-50 pré-treinada (vetores de 2048 dimensões). Em seguida, o UMAP foi aplicado para preservar as relações de vizinhança no espaço bidimensional. O mosaico organiza as imagens em uma grade 20×20 minimizando a distância entre as coordenadas UMAP e os centros da grade, revelando padrões visuais e agrupamentos implícitos aprendidos pelo modelo.

Conclusões

Métodos lineares, como o PCA, oferecem visões globais simples e rápidas, mas tendem a perder estrutura local: pontos semelhantes podem se misturar, e nuances importantes desaparecem. Já métodos não lineares, como t-SNE e UMAP, preservam vizinhanças e revelam clusters mais nítidos, destacando padrões que o PCA não captura.

Os experimentos mostram isso com clareza: nos dígitos e no CIFAR, as projeções não lineares organizam melhor grupos naturais e transições visuais. Assim, para análise exploratória, UMAP e t-SNE produzem mapas mais informativos, enquanto o PCA continua útil para captar tendências gerais.

Referências

[1] Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021.

[2] William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 26, 1984.

[3] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[4] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11), 1901.

[5] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Comput. Graph. Forum*, 34(3):201–210, June 2015.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.