BÀI 8 - t-SNE & UMAP

Hạn nộp bài: hết ngày 27/06/2024.

1 Bài tập thực hành

Bài 1. Dữ liệu trong Penguins chứa thông tin của 344 cá thể chim cánh cụt thuộc ba loài khác nhau.

a) Đọc dữ liệu và tiến hành phân tích, tiền xử lý dữ liệu. Cho biết số lượng cá thể chim cánh cụt ở mỗi loài.

b) Trực quan hoá từng cặp biến định lượng trong dữ liệu với màu là loài chim. Có sự tách biệt giữa các loài chim khác nhau hay không?

c) Trực quan hoá dữ liệu trong 2 chiều bằng t-SNE và bằng UMAP. Nhận xét.

Bài 2. Tập dữ liệu Digits (trong `sklearn.datasets.load_digits`) chứa hình ảnh của các chữ số viết tay, gồm 10 lớp trong đó mỗi lớp là một chữ số.

a) Trực quan hoá dữ liệu trong 2 chiều bằng t-SNE (dùng `sklearn.manifold.TSNE`) với màu là các lớp.

b) Trực quan hoá dữ liệu trong 2 chiều bằng UMAP (dùng `umap.umap_.UMAP`) với màu là các lớp.

c) So sánh và nhận xét kết quả trực quan hoá và thời gian thực thi của hai phương pháp.

d) Điều chỉnh siêu tham số `perplexity` trong t-SNE.

e) Điều chỉnh các siêu tham số `n_neighbors` và `min_dist` trong UMAP.

Bài 3. Trong bài báo UMAP :Uniform Manifold Approximation and Projection for Dimension Reduction, các tác giả trình bày Thuật toán 3 để tính hệ số chuẩn hóa cho khoảng cách σ như sau

Algorithm 3 Tính hệ số chuẩn hóa cho khoảng cách σ

function SmoothKNNDist(knn-dists, k, ρ)
Tìm σ sao cho $\sum_i^n \exp(-(\text{knn-dists}_i - \rho)/\sigma) = \log_2(k)$.
return σ

Hãy viết hàm `SmoothKNNDist(knn_dists, k, rho, tol=0.001)` để tìm σ bằng thuật toán tìm kiếm nhị phân:

- **Đầu vào:**

  `knn-dists` là các khoảng cách từ điểm đang xét đến các điểm hàng xóm;

`k` là số lượng điểm lân cận của điểm đang xét;

`rho` là khoảng cách từ điểm đang xét đến điểm gần nhất

`tol` là sai số cho phép.

- **Đầu ra:** Giá trị σ.

Chạy hàm vừa viết cho mẫu dữ liệu đầu tiên trong tập Digits ở bài 2 với $k = 5, 10, 20$. Kiểm tra kết quả bằng cách tính giá trị biểu thức

$$\sum_{i}^{n} \exp(-(\text{knn-dists}_i - \rho)/\hat{\sigma}) - \log_2(k).$$