UMAP: Lý thuyết và ứng dụng

Nguyễn Minh Ánh Nguyệt

Trình bày từ tiểu luận cuối kỳ môn *Xử lí đa chiều*

Đại học Khoa học Tự nhiên

Ngày 13 tháng 6 năm 2023

Outline

Giới thiệu

- Mục tiêu của giảm chiều: tìm đặc trưng ẩn của dataset ở chiều thấp.
- Trước đây, t-SNE là SOTA của các thuật toán giảm chiều. Hạn chế:
  - Chỉ có thể hoạt động ngẫu nhiên trên dữ liệu mà nó nhìn thấy.
  - Chỉ sử dụng cho trực quan hóa.
  - Chậm, chi phí tính toán tốn kém.
  - Chỉ bảo toàn cấu trúc cục bộ.

Tại sao UMAP được dùng nhiều?

- Chất lượng tương đương hoặc tốt hơn.
- Ứng dụng rộng hơn.
- Nhanh hơn.
- Bảo toàn được cấu trúc cục bộ và toàn cục.

Lấy thông tin địa phương

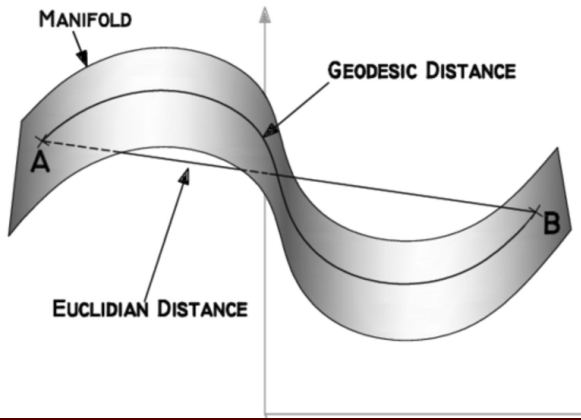
- Giả sử dữ liệu nằm trên một đa tạp M.
- Lấy thông tin "địa phương" bằng cách tìm khoảng cách của mỗi điểm tới lân cận của nó .

Chọn lân cận như thế nào?

Dùng KNN!!!

Tính khoảng cách trắc địa như thế nào?

Thêm giả thiết (phân bố đều trên $M, \ldots$) thì tính xấp xỉ được.



Từ địa phương tới toàn cục

- Với mỗi điểm, thông tin địa phương $\leftrightarrow$ không gian "gần"metric.
- Chuyển và ghép các thông tin thành thông tin toàn cục.

Chuyển không gian về dưới dạng nào?

Đồ thị có trọng số!!

Làm sao để ghép chúng lại với nhau?

Thay đổi trọng số dựa vào ý tưởng xác suất

$$a \perp b = a + b - ab.$$

Thuật toán UMAP

Algorithm 1 UMAP(X, k, d, min-dist, n-epochs)

Input Tập dữ liệu X, k là số điểm lân cận sẽ chọn ứng với một điểm dữ liệu, số chiều d sau khi giảm, min-dist là tham số để kiểm soát độ phân tán của dữ liệu ở số chiều thấp, n-epochs là số lần duyệt qua toàn bộ tập dữ liệu.

Output Y là đại diện cho cấu trúc liên kết của dữ liệu sau khi giảm chiều.

procedure UMAP
    # Xây dựng đại diện cấu trúc liên kết của dữ liệu ở số chiều cao.
    for all $x \in X$ **do**
        fs-set[x] $\leftarrow$ LocalFuzzySimplicialSet(X, x, k) # Lấy thông tin cục bộ
    top-rep $\leftarrow \cup_{x \in X}$ fs-set[x] # Hợp các kết quả để có được đại diện ở số chiều cao
    # Tối ưu hóa và trả kết quả là đại diện cấu trúc liên kết của dữ liệu ở số chiều thấp.
    $Y \leftarrow$ SpectralEmbedding(top-rep, d) # Khởi tạo
    $Y \leftarrow$ OptimizeEmbedding(top-rep, Y , min-dist, n-epochs) # Tối ưu hóa
    return Y

Tìm không gian gần metric

Dùng KNN, ta tìm được lân cận $N(x)$ của x và tính được $d_{x,y}$ là khoảng cách từ x tới y trong không gian "gần"metric tương ứng, cho bởi

$$d_{x,y} = \begin{cases} \frac{d_{\mathbb{R}^N}(x,y)-\rho}{\sigma}, & \text{nếu } y \in N(x) \\ +\infty, & \text{khác} \end{cases},$$

trong đó $d_{\mathbb{R}^N}$ là metric trên $\mathbb{R}^N$ và ρ là khoảng cách từ x đến lân cận gần nhất theo metric $d_{\mathbb{R}^n}$.

Nhận xét

- Mỗi điểm x nối với ít nhất một điểm có trọng số bằng 1.
- σ là tham số chuẩn hóa.

Xây dựng đồ thị có trọng số

Ta cần lấy chuyển thông tin từ lân cận của $x \in X$ thành một đồ thị.

- Tập đỉnh: X.
- Tập cạnh: các cạnh nối từ x với những đỉnh khác.
- Trọng số: $w(x, y) = e^{-d_{x,y}}$ nếu y là một trong k lân cận gần nhất của x và bằng 0 trong trường hợp khác.

Xây dựng đồ thị có hướng có trọng số

Algorithm 2 LocalFuzzySimplicialSet(X, x, k)

Input Tập dữ liệu X, $x \in X$, k là số lân cận ứng với một điểm dữ liệu.
Output fs-set là 0-simplicies và 1-simplices ứng với x.

procedure LocalFuzzySimplicialSet
    # Xây dựng không gian giả metric ứng với điểm dữ liệu x.
    knn, knn-dists $\leftarrow$ ApproxNearestNeighbors(X, x, k) # Thu được kết quả knn là k điểm lân cận của x và knn-dists tương ứng là khoảng cách của x đến k điểm đó.
    $\rho \leftarrow$ knn-dists[1] # Khoảng cách đến lân cận gần nhất.
    $\sigma \leftarrow$ SmoothKNNDist(knn-dists, k, ρ) # Tham số chuẩn hóa.
    fs-set$_0 \leftarrow X$. # Ứng với tập dữ liệu ban đầu
    fs-set$_1 \leftarrow \{([x, y], 0) | y \in X\}$ # Lấy tất cả cạnh có một đỉnh là x và trọng số khởi tạo là 0.
    for all $y \in$ knn **do**
        $d_{x,y} \leftarrow \max(0, \text{dist}(x, y) - \rho)/\sigma$ # Tính khoảng cách trong không gian giả mêtríc mở rộng đã được chuẩn hóa
        fs-set$_1 \leftarrow$ fs-set$_1 \cup ([x, y], \exp(-d_{x,y}))$ # Thay thế trọng số của các cạnh xây dựng từ x và một trong k điểm lân cận của nó.
    return fs-set

Giảm chiều dữ liệu

- Khởi tạo n điểm dữ liệu trong $\mathbb{R}^D \Rightarrow$ Tìm được đồ thị tương ứng.
- Sử dụng cross entropy để đo sự sai khác giữa hai đồ thị.
- Di chuyển các điểm để được biểu diễn số chiều thấp.

Khởi tạo

Từ X, ta đã tìm được G là đồ thị chứa thông tin toàn cục.

Khởi tạo

Dùng Spectral Embedding để tìm được biểu diễn Y của G trong $\mathbb{R}^D$.

Xây dựng đồ thị ở số chiều thấp

Xây dựng đồ thị

- Đồ thị xây dựng đẳng cấu với G như là đồ thị không có trọng số.
- Trọng số: $\Phi(x, y) = (1 + a(\|x - y\|_2^2)^b)^{-1}$ với a, b là các tham số để fit Φ gần với Ψ với

$$\Psi(x, y) = \begin{cases} 1, & \text{nếu } \|x - y\|_2 \leq \text{min\_dist} \\ e^{-\|x-y\|_2 + \text{min\_dist}}, & \text{khác} \end{cases},$$

trong đó min_dist đảm bảo được việc không có điểm nào bị cô lập giống như ở số chiều ban đầu.

Tối ưu hoá

Di chuyển các điểm

- Ở mỗi vòng lặp, ta di chuyển từng điểm sao cho hai đồ thị tương ứng có cross entropy nhỏ nhất có thể.
- Trọng số lớn từ y tới x càng lớn (nhỏ) thì di chuyển y lại gần (xa) x.

Tối ưu hóa

Algorithm 4 OptimizeEmbedding(top-rep, Y , min-dist, n-epochs))

Input top-rep là đại diện cho cấu trúc liên kết ở số chiều cao, Y là khởi tạo cho đại diện ở số chiều thấp sử dụng Spectral embedding, min-dist là tham số để kiểm soát layout hay để khởi tạo hàm Ψ, n-epochs là số lượng lần duyệt qua tập dữ liệu.

Output Y là đại diện ở số chiều thấp.

procedure OPTIMIZEEMBEDDING
  $\alpha \leftarrow 1.0$
  Từ min-dist ta xây dựng hàm Ψ, sau đó cố gắng cho hàm Φ gần với Ψ nhất có thể để tìm ra hai tham số của hàm Φ
  for $e \leftarrow 1, \ldots,$ n-epochs **do**
    for all $([a, b], p) \in$ top-rep$_1$ **do**
      if RANDOM()$\leqslant p$ **then**
        $y_a \leftarrow y_a + \alpha \nabla(\log(\Phi))(y_a, y_b)$
        for $i \leftarrow 1, \ldots,$ n-neg-samples **do**
          $c \leftarrow$ random sample from Y
          $y_a \leftarrow y_a + \alpha \nabla(\log(1 - \Phi))(y_a, y_c)$
    $\alpha \leftarrow 1.0 - e/$n-epochs
  return Y

Giới thiệu

- Trong NLP, chúng ta có rất nhiều phương pháp để chuyển đổi từ dạng văn bản về dạng véctơ số.
- Vậy khi bắt đầu với một dữ liệu mới (cụ thể tập dữ liệu fetch_20newsgroups), phương pháp nào là phù hợp?

Mục tiêu

Sử dụng UMAP để hỗ trợ việc đánh giá phương pháp nào là tốt nhất.

Mô tả dữ liệu

Bộ fetch_20newsgroups tải từ thư viện sklearn. Dữ liệu dưới dạng văn bản bao gồm hơn 18000 bản tin của 20 chủ đề tương ứng là nhãn của chúng.

	text	source
0	\n\nI am sure some bashers of Pens fans are pr...	10
1	My brother is in the market for a high-perform...	3
2	\n\n\n\n\tFinally you said what you dream abou...	17
3	\nThink!\n\nIt's the SCSI card doing the DMA t...	3
4	1) I have an old Jasmine drive which I cann...	4

Hình 1: 5 dòng đầu trong tập dữ liệu

Xử lý dữ liệu

- Xóa đi các kí tự đặc biệt.
- Tách đoạn văn bản thành các từ.
- Xóa những từ quá ngắn và stopword.
- Đưa từ về dạng gốc của nó.

Ý tưởng chính

- Chia dữ liệu thành hai tập lần lượt là train - test.
- Sử dụng 4 phương pháp nhúng lần lượt là Bag of word, Tf-idf, Pre-train model, RNN-LSTM trên tập train.
- Lấy mô hình đã được huấn luyện để chuyển đổi tập test.
- Dùng UMAP giảm chiều dữ liệu trên tập train và chuyển đổi tương tự với tập test.
- Đánh giá phương pháp nhúng nào là phù hợp dựa vào Trustworthiness score, trực quan hóa và Silhouette score.

Sử dụng UMAP

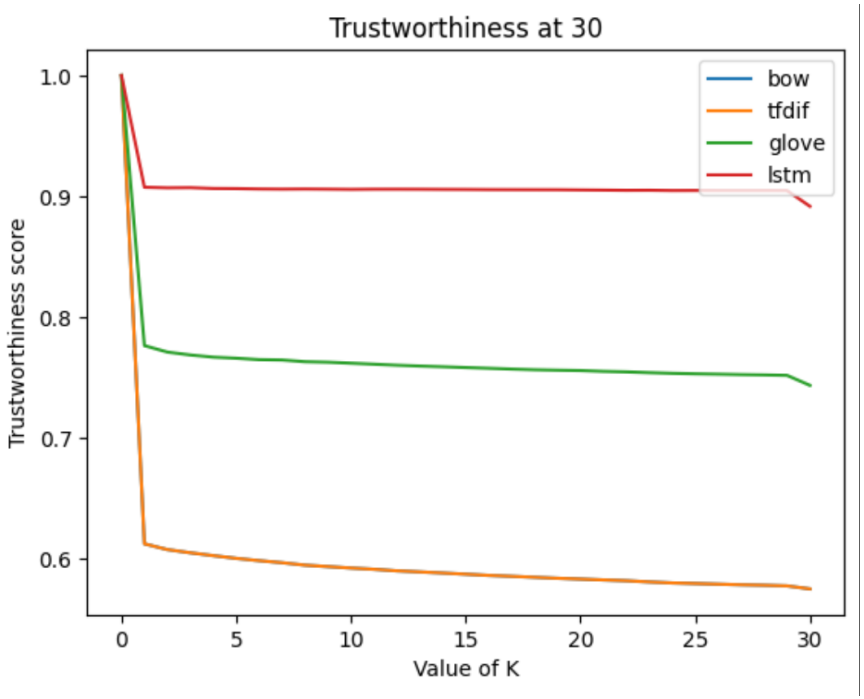
- Thông tin cụ thể về các phương pháp
  - BoW, Tf-idf: số chiều tối đa 1000.
  - Pre-train model: Sử dụng bộ glove 100d nên số chiều là 100.
  - LSTM: Sử dụng bi-LSTM với số chiều hidden state là 100 nên kết quả có số chiều 200.
- Sử dụng UMAP để giảm chiều kết quả sau khi áp dụng các phương pháp trên với cái siêu tham số được lựa chọn là
  - n_neighbors=30.
  - min_dist=0.1.
  - n_components=2.
  - metric='cosine'.
  - random_state=42

  Các siêu tham số này được lựa chọn dựa trên thử nghiệm.

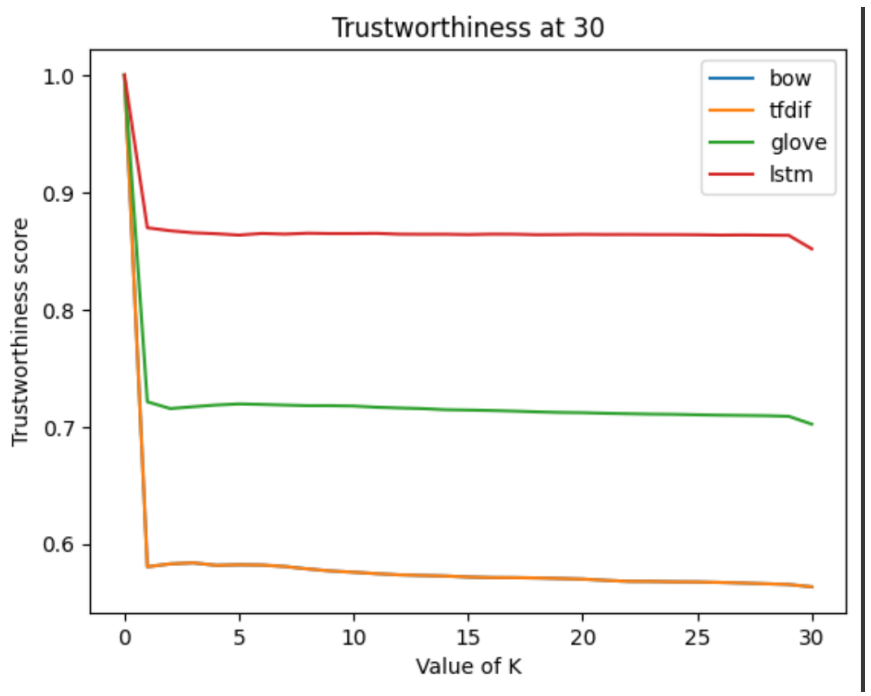
Phương pháp đánh giá

1. Trustworthiness score:
   - Đánh giá chất lượng biểu diễn của dữ liệu ở số chiều thấp. Một ánh xạ được coi là đáng tin cậy nếu k điểm này ở không gian có số chiều cao cũng gần điểm x trong không gian có số chiều thấp.
   - Giá trị Trustworthiness score thuộc $[0, 1]$ trong đó giá trị càng lớn thì cấu trúc cục bộ của dữ liệu được bảo tồn tốt.
2. Trực quan hóa: Đánh giá xem các đại diện của cùng một nhóm có được phân bố gần nhau không.
3. Silhouette score:
   - Phương pháp được sử dụng để đánh giá chất lượng của các cụm với giá trị thuộc khoảng $[-1, 1]$.
   - Giá trị càng lớn thì cụm của điểm dữ liệu i càng tốt.

Kết quả Trustworthiness score

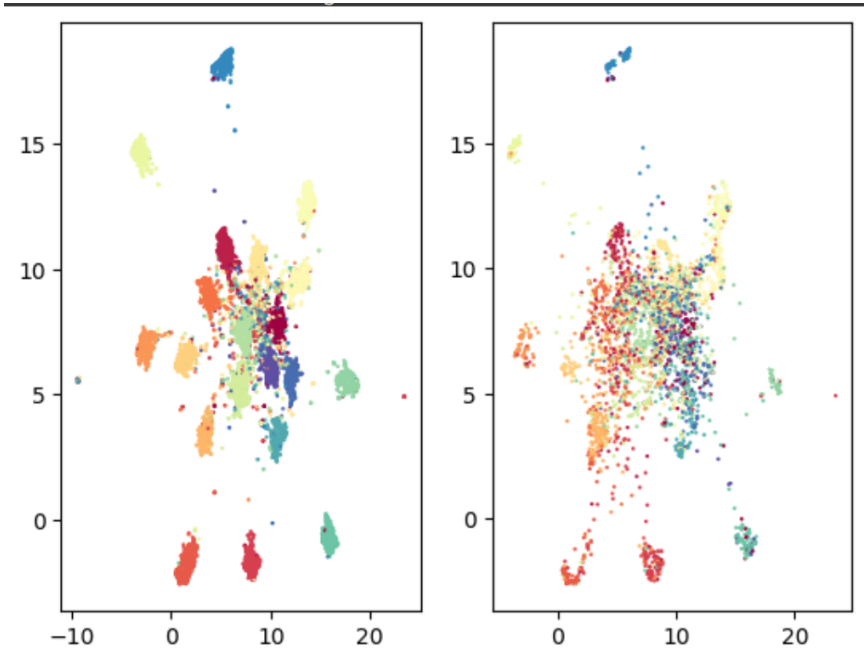


Hình 2: Tập train

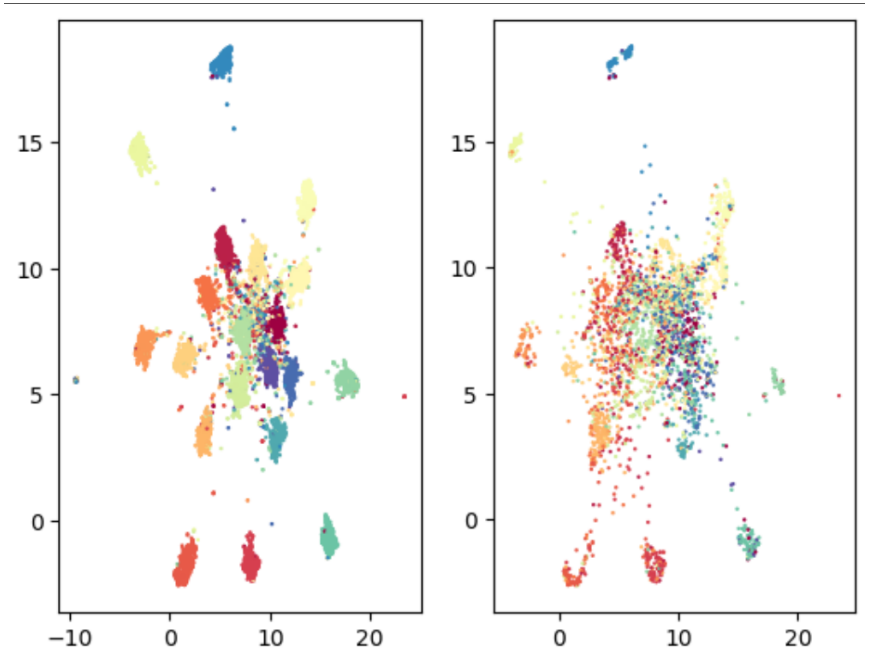


Hình 3: Tập test

Kết quả trực quan hóa

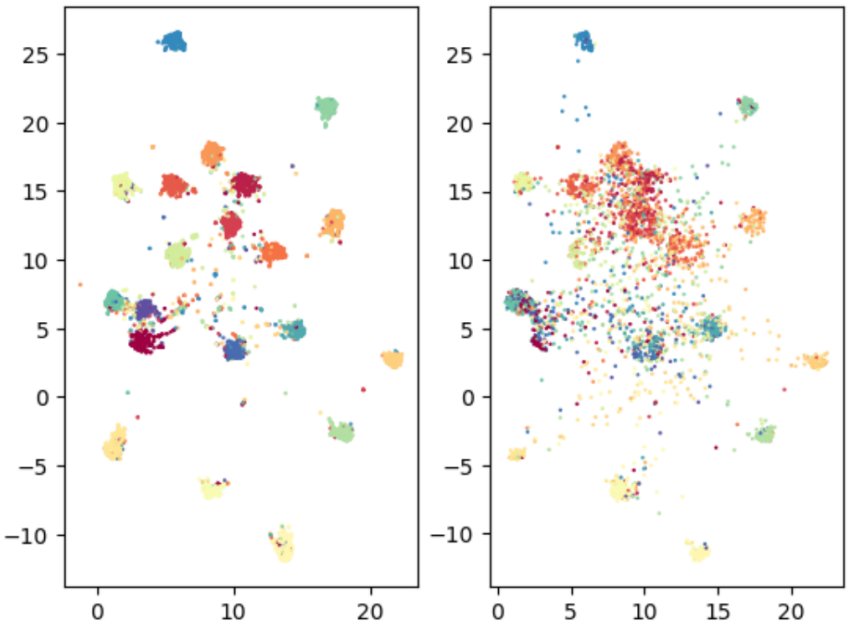


Hình 4: Bag of word

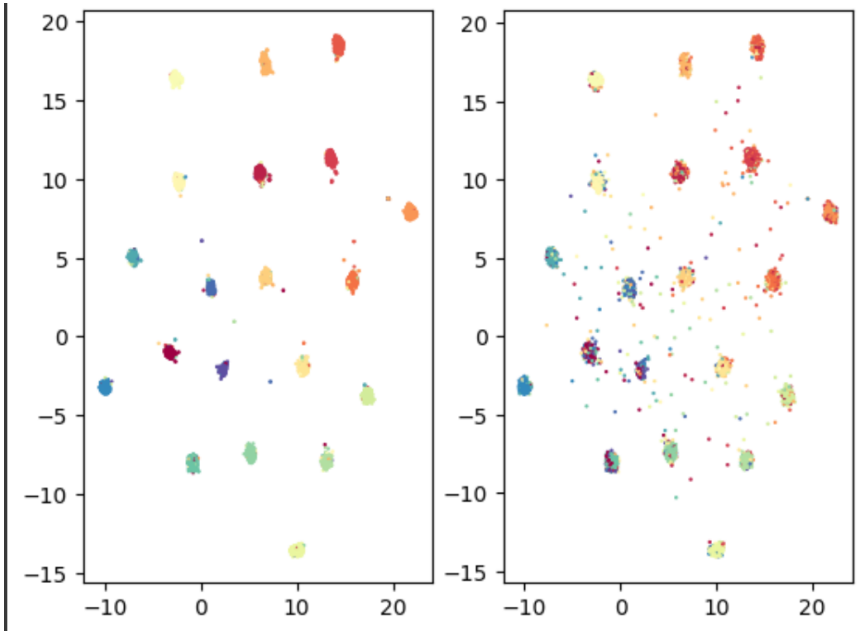


Hình 5: Tf-idf

Kết quả trực quan hoá



Hình 6: Pre-train model



Hình 7: LSTM

Kết quả Silhouette score

Phương pháp	Tập huấn luyện	Tập thử nghiệm
Bag of word	0.47847196	-0.1658458
Tf-idf	0.47847196	-0.1658458
Pre-train model	0.5970492	-0.109202586
LSTM	0.8254086	0.09736216

Hướng mở rộng trong tương lai

- Tạo pipeline để tiền xử lý dữ liệu phù hợp cho mỗi phương pháp.
- Mở rộng giới hạn chiều cho phương pháp bag of word và tf-idf. Kết hợp với SVD hoặc PCA giảm thiểu số chiều trước khi đưa vào UMAP.
- Thực hiện các bước để lựa chọn được bộ siêu tham số của UMAP cho kết quả tốt nhất.
- Sử dụng UMAP với ứng dụng khác như kết hợp với K-means và Decision Tree để khai thác insight của tập dữ liệu.

References

[1] L. McInnes, J. Healy, J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.

[2] M. Belkin, P. Niyogi, *Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in neural information processing systems*.

[3] M. Belkin, P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*.

[4] W. Dong, C. Moses, K. Li, *Efficient k-nearest neighbor graph construction for generic similarity measures*.

[5] A. Jackson, *The mathematics of UMAP*.

[6] Curve fitting, Spicy Doccument of Curve fitting, Spicy

List of References

[7] Source code UMAP, Souce code

[8] Huỳnh Quang Vũ, Bài giảng Tôpô.

[9] T. Leinster, Basic Category Theory, Cambridge University Press, 2014.

[10] Wiki, LSTM LSTM

[11] G. Friedman, An elementary illustrated introduction to simplicial sets.

[12] B. Andrews, Lectures on Differential Geometry.

[13] D. Spivak, Metric realization of fuzzy simplicial sets.

Cảm ơn thầy, cô và mọi người đã lắng nghe!