

Feature Learning

Finding the diamonds in the rough

High Dimensionality

Value 6, 1

Higher Dimensionality

Feature 1 Feature 2 Feature 3 Feature 4 Value 1, 1 Value 1, 2 Value 1, 3 Value 1, 4 Value 2, 1 Value 2, 2 Value 2, 3 Value 2, 4 Value 3, 1 Value 3, 2 Value 3, 3 Value 3, 4 Value 4, 1 Value 4, 2 Value 4, 3 Value 4, 4 Value 5, 1 Value 5, 2 Value 5, 3 Value 5, 4

Value 6, 3

More Data

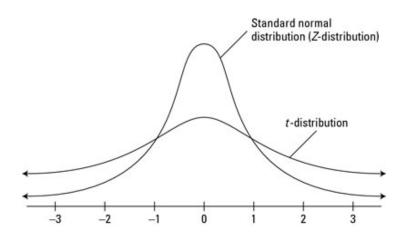
Value 6, 4

Value 6, 2

Feature Extraction

t-distributed Stochastic Neighbor Embeddings

t-SNE - Student-t vs Normal



t-SNE - Conditional Probability

$$p_{j \mid i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||_2^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-||\mathbf{x}_i - \mathbf{x}_k||_2^2 / 2\sigma_i^2)},$$

where σ_i^2 is the variance of the Gaussian distribution centered at \mathbf{x}_i .

MInDS @ Mines Feature Learning 6 / 29

t-SNE - Pairwise Probability

$$p_{ij} = \frac{p_{j\mid i} + p_{i\mid j}}{2n},$$

where n is the total number of samples in the data, $p_{ii} = 0$.

MInDS @ Mines Feature Learning 7/29

t-SNE - Embedding's Joint Probability

$$q_{ij} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||_2^2)^{-1}}{\sum_{k \neq i} (1 + ||\mathbf{y}_i - \mathbf{y}_k||_2^2)^{-1}}.$$

MInDS @ Mines Feature Learning 8 / 29

t-SNE - Objective

$$\min_{\mathbf{Y}} \sum_{i \neq j} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right).$$

MInDS @ Mines Feature Learning 9,

t-SNE - Conditional Probability

$$\rho_{j \mid i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||_2^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||\mathbf{x}_i - \mathbf{x}_k||_2^2/2\sigma_i^2)},$$

where σ_i^2 is the variance of the Gaussian distribution centered at \mathbf{x}_i .

MInDS @ Mines Feature Learning 10 / 29

t-SNE - Perplexity

Perplexity
$$(P_i) = 2^{\mathcal{H}(P_i)} = 2^{-\sum_j p_{j+i} \log p_{j+i}}$$
.

- Perplexity is used to represent the number of neighbors to consider per point.
- Higher perplexity is a higher weight to the global relationships and a lower perplexity is a higher weight to the local relationships.

MInDS @ Mines Feature Learning 11 / 29

t-SNE - Issues

- Hyperparameters in t-SNE are crucial
- Cluster sizes are meaningless
- Distance between clusters can be meaningless
- Random noise may result in some misleading patterns
- Shapes may appear in the output that are misleading
- Testing multiple perplexities may help with understanding the data

MInDS @ Mines Feature Learning 12 / 29

How to use t-SNE effectively

Demo

13 / 29

Spectral Embedding

Spectral Embedding - Objective

$$\begin{aligned} & \min_{\mathbf{Y}} & \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} ||\mathbf{y}_i - \mathbf{y}_j||_2^2, \\ & s.t. \; \mathbf{Y}^T \mathbf{Y} = \mathbf{I}. \end{aligned}$$

15 / 29

MInDS @ Mines Feature Learning

Similarity Matrix

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & \dots & s_{1n} \\ s_{21} & s_{22} & s_{23} & \dots & s_{2n} \\ s_{31} & s_{32} & s_{33} & \dots & s_{3n} \\ \dots & \dots & \dots & \dots \\ s_{n1} & s_{n2} & s_{n3} & \dots & s_{nn} \end{bmatrix}$$

16 / 29

Degree Matrix

$$\mathbf{D} = \begin{bmatrix} \sum_{i=1}^{n} s_{1i} & 0 & 0 & \dots & 0 \\ 0 & \sum_{i=1}^{n} s_{2i} & 0 & \dots & o \\ 0 & 0 & \sum_{i=1}^{n} s_{3i} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sum_{i=1}^{n} s_{ni} \end{bmatrix}$$

17 / 29

Laplacian Matrix

$$\mathbf{L} = \mathbf{D} - \mathbf{S}$$

$$= \begin{bmatrix} (\sum_{i=2}^{n} s_{1i}) & -s_{12} & -s_{13} & \dots & -s_{1n} \\ -s_{21} & (\sum_{i=1, i\neq 2}^{n} s_{2i}) & -s_{23} & \dots & -s_{2n} \\ -s_{31} & -s_{32} & (\sum_{i=1, i\neq 3}^{n} s_{3i}) & \dots & -s_{3n} \end{bmatrix}$$

$$= \begin{bmatrix} -s_{n1} & -s_{n2} & -s_{n3} & \dots & (\sum_{i=1, i\neq n}^{n} s_{ni}) \end{bmatrix}$$

MInDS @ Mines Feature Learning 18 / 29

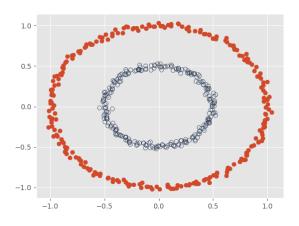
Spectral Embedding - Solution

$$\mathbf{Y} = eig(\mathbf{L}, r)$$
.

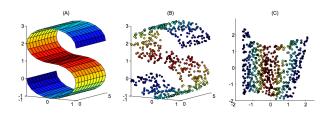
The embeddings are the r eigenvectors of the Laplacian matrix corresponding to the smallest r non-zero eigenvalues.

MInDS @ Mines Feature Learning 19 / 29

Spectral Clustering



Spectral Clustering



Linear Discriminant Analysis

LDA - Projection

$$\mathbf{y}_i = \mathbf{W}^T \mathbf{x}_i,$$

where **W** is the $r \times d$ projection matrix by which we transform the original data from d dimensions into the new r-dimensional space.

MInDS @ Mines Feature Learning 23 / 29

LDA - Class Means

$$\mathbf{m}_k = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \mathbf{x}_k,$$

where C_k is the set of all points belonging to class k.

MInDS @ Mines Feature Learning 24 / 29

LDA - Variances

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T,$$

$$\mathbf{S}_W = \sum_{i \in \mathcal{C}_1} \left((\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T \right) + \sum_{j \in \mathcal{C}_2} \left((\mathbf{x}_j - \mathbf{m}_2)(\mathbf{x}_j - \mathbf{m}_2)^T \right).$$

MInDS @ Mines Feature Learning 25 / 29

LDA - Objective

$$\max_{\mathbf{W}} \mathbf{tr} \left(\frac{\mathbf{W} \mathbf{S}_{B} \mathbf{W}^{T}}{\mathbf{W} \mathbf{S}_{W} \mathbf{W}^{T}} \right).$$

26 / 29

Dimensionality Reduction Methods

- PCA maximize variance per embedding dimension
- t-SNE maximize relative entropy between embedding and original
- Spectral Embedding minimize the distance between embeddings weighted by original point similarities
- LDA maximize class separation (supervised)

MInDS @ Mines Feature Learning 27 / 29

Questions

These slides are designed for educational purposes, specifically the CSCI-470 Introduction to Machine Learning course at the Colorado School of Mines as part of the Department of Computer Science.

Some content in these slides are obtained from external sources and may be copyright sensitive. Copyright and all rights therein are retained by the respective authors or by other copyright holders. Distributing or reposting the whole or part of these slides not for academic use is HICHLY prohibited, unless explicit permission from all copyright holders is granted.