t-distributed Stochastic Neighbor Embedding (t-SNE)

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

t-SNE とは？

✓非線形の可視化手法の一つ

✓PCA や GTM のように、元の空間から低次元空間 (基本的には二次元平面) に写像させる関数が得られるわけではないので注意

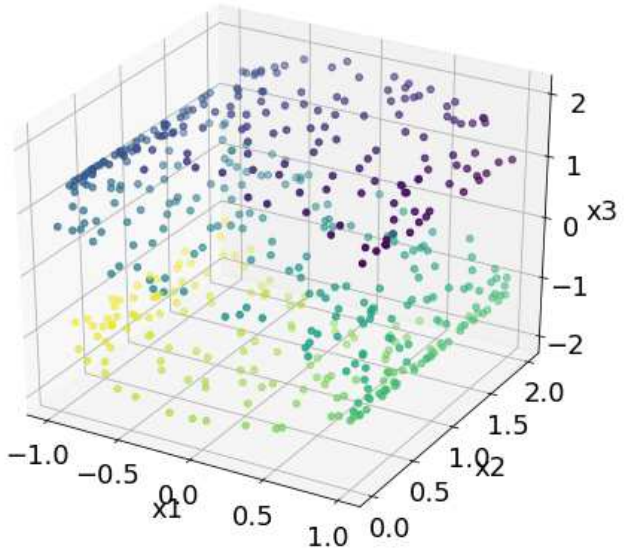
✓写像というよりは、サンプル全体が見やすいように二次元平面に配置するイメージ

✓可視化に特化した手法

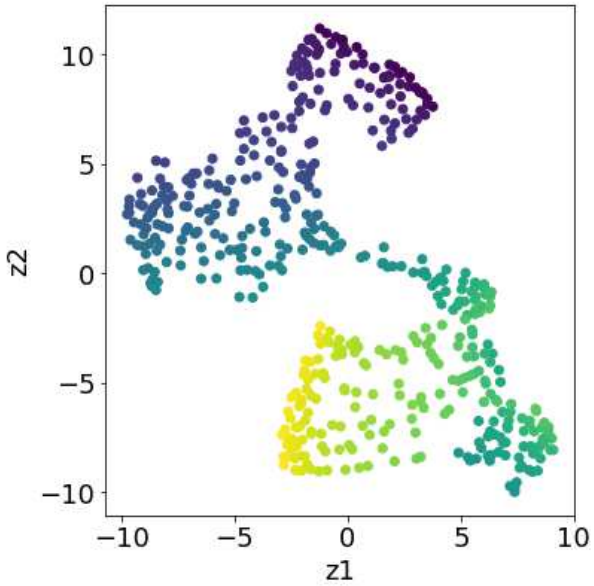
✓元の空間におけるサンプル間の距離関係が二次元平面におけるサンプル間の距離関係として保持されるほど値が小さくなる目的関数を準備して、それが小さくなるように二次元平面にサンプルを配置させる

t-SNE のイメージ

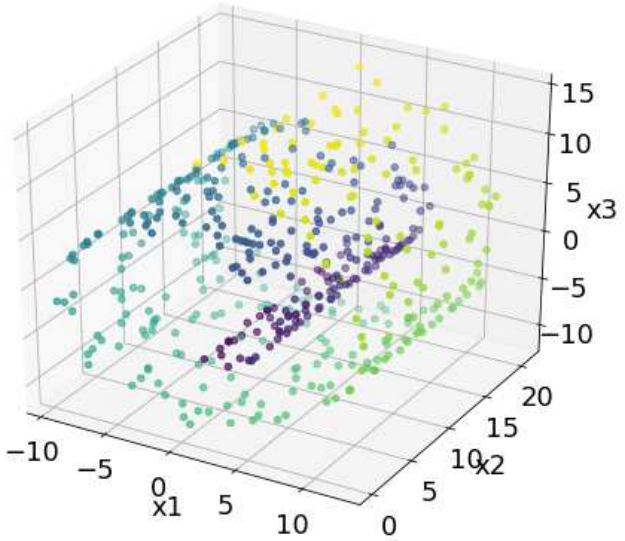
S-curve



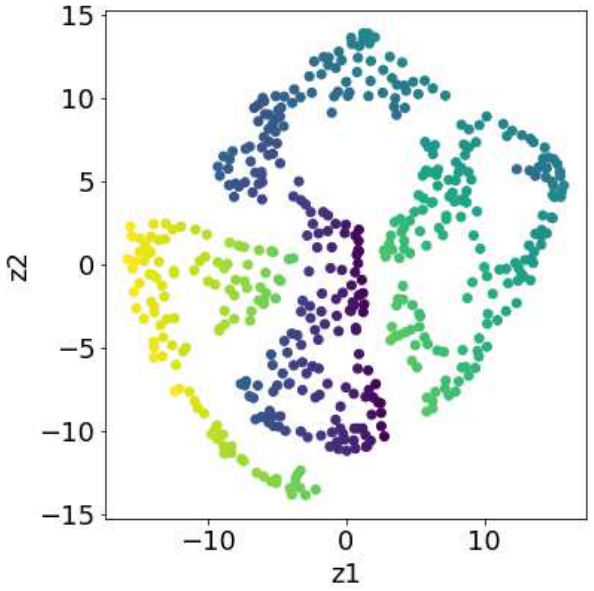
t-SNE



Swiss roll



t-SNE



文字の定義

✓元の空間におけるデータセットを $\mathbf{X}$ とする

- 変数の数：m
- サンプルの数：n
- つまり、$\mathbf{X}$ は $n \times m$ の行列
- i 番目のサンプルを $\mathbf{x}^{(i)}$ を表す

✓低次元空間におけるデータセットを $\mathbf{Z}$ とする

- 今回は二次元に低次元化するため、変数の数：2
- サンプルの数：n
- $\mathbf{Z}$ は $n \times 2$ の行列
- i 番目のサンプルを $\mathbf{z}^{(i)}$ を表す

前処理

✓基本的に、X を標準化 (オートスケーリング) する
- 標準化についてはこちら
  https://datachemeng.com/basicdatapreprocessing/

✓X (を標準化したあと) に前処理として主成分分析 (Principal Component Analysis, PCA) を行うこともある
- PCA についてはこちら
  https://datachemeng.com/principalcomponentanalysis/

t-SNEでは何をしているか？

目的関数 C を最小化している

$$C = \sum_{i=1}^{n} \sum_{j=1}^{n} p\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) \log \frac{p\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)}{p\left(\mathbf{z}^{(i)}, \mathbf{z}^{(j)}\right)}$$

ざっくりというと、C は $\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ の距離関係と、$\mathbf{z}^{(i)}$ と $\mathbf{z}^{(j)}$ との距離関係が似ているほど小さくなる

⇨ 元の空間におけるサンプル間の距離関係と
二次元空間におけるサンプル間の距離関係とが
同じになるように、$\mathbf{Z}$ を作成できる！

詳しくは、次のページ以降

$p(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ って何？

$p(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ は、$\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ の同時確率分布

$\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ の “近さ” をあらわす

つまり、$\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ が似ている (距離が小さい) ほど、大きくなる

t-SNE では下のように定義している

$$p\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \frac{p\left(\mathbf{x}^{(j)} \mid \mathbf{x}^{(i)}\right) + p\left(\mathbf{x}^{(i)} \mid \mathbf{x}^{(j)}\right)}{2}$$

$p(\mathbf{x}^{(j)} \mid \mathbf{x}^{(i)})$ って何？

$p(\mathbf{x}^{(j)} \mid \mathbf{x}^{(i)})$ は、$\mathbf{x}^{(i)}$ が与えられたときの $\mathbf{x}^{(j)}$ の事後確率分布

t-SNE では正規分布を仮定して下のように定義している

$$p\left(\mathbf{x}^{(j)} \mid \mathbf{x}^{(i)}\right) = \frac{\exp\left(-\frac{\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right\|^2}{2\sigma_i^2}\right)}{\sum_{k=1}^{n} \exp\left(-\frac{\left\|\mathbf{x}^{(i)} - \mathbf{x}^{(k)}\right\|^2}{2\sigma_i^2}\right) - 1}$$

σ_i : i 番目のサンプルに対応する標準偏差

分母の -1 は、$k = 1$ から n まで和を取るときに、$k = i$ が含まれてしまうため
その分を引くためのもの

$\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ の距離と $p(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ の関係

$\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ の距離が近い (似ている)

$p(\mathbf{x}^{(j)} \mid \mathbf{x}^{(i)})$ や $p(\mathbf{x}^{(i)} \mid \mathbf{x}^{(j)})$ が大きくなる

$p(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ が大きくなる

また、 $p\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \dfrac{p\left(\mathbf{x}^{(j)} \mid \mathbf{x}^{(i)}\right) + p\left(\mathbf{x}^{(i)} \mid \mathbf{x}^{(j)}\right)}{2}$

により、$\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ に対して対象になっている
($\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ を入れ替えても同じ値)

$p(\mathbf{x}^{(i)} \mid \mathbf{x}^{(j)})$ の σ_i はどうする？

✓周辺にサンプルが密に存在しているサンプル $\mathbf{x}^{(i)}$ にとっては、
σ_i は小さくあるべき

✓周辺にサンプルがあまりないサンプル $\mathbf{x}^{(i)}$ にとっては、σ_i は大きくあるべき

$\mathbf{x}^{(i)}$ に対応する確率密度分布 ($p(\mathbf{x}^{(1)} \mid \mathbf{x}^{(i)})$, $p(\mathbf{x}^{(2)} \mid \mathbf{x}^{(i)})$,…,
$p(\mathbf{x}^{(n)} \mid \mathbf{x}^{(i)})$) の情報エントロピー(シャノンエントロピー) H を
固定して、σ_i を決めよう！

$$H = -\sum_{j=1}^{n} p\left(\mathbf{x}^{(j)} \mid \mathbf{x}^{(i)}\right) \log_2 p\left(\mathbf{x}^{(j)} \mid \mathbf{x}^{(i)}\right)$$

$$perplexity = 2^{H}$$

t-SNE では二分探索 [1] で σ_i を計算
(*perplexity* は σ_i に対して単調増加なのでOK)

perplexity は事前に設定する必要がある

[1] https://ja.wikipedia.org/wiki/二分探索

$p(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})$ って何？

評価関数　$$C = \sum_{i=1}^{n} \sum_{j=1}^{n} p\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) \log \frac{p\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)}{p\left(\mathbf{z}^{(i)}, \mathbf{z}^{(j)}\right)}$$

について、$p(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ は OK、次は $p(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})$

$p(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})$ は、$\mathbf{z}^{(i)}$ と $\mathbf{z}^{(j)}$ の同時確率分布

$\mathbf{z}^{(i)}$ と $\mathbf{z}^{(j)}$ の “近さ” をあらわす

つまり、$\mathbf{z}^{(i)}$ と $\mathbf{z}^{(j)}$ が似ている (距離が小さい) ほど、大きくなる

$p(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})$ の式

t-SNE では自由度 1 の (スチューデントの) t 分布 [1] を仮定して
下のように定義している

$$p\left(\mathbf{z}^{(i)}, \mathbf{z}^{(j)}\right) = \frac{\left(\frac{1}{1+\left\|\mathbf{z}^{(i)} - \mathbf{z}^{(j)}\right\|^2}\right)}{\sum_{k=1}^{n}\sum_{l=1}^{n}\left(\frac{1}{1+\left\|\mathbf{z}^{(k)} - \mathbf{z}^{(l)}\right\|^2}\right) - n}$$

分母の $-n$ は、和を取るときに、$k = l$ が含まれてしまうため
その分を引くためのもの

[1] https://ja.wikipedia.org/wiki/T分布

目的関数 C の最小化

目的関数 C を $\mathbf{z}^{(i)}$ で偏微分すると、

$$\frac{\partial C}{\partial \mathbf{z}^{(i)}} = 4\sum_{j=1}^{n} \frac{\left\{ p\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) - p\left(\mathbf{z}^{(i)}, \mathbf{z}^{(j)}\right) \right\}\left(\mathbf{z}^{(i)} - \mathbf{z}^{(j)}\right)}{1 + \left\| \mathbf{z}^{(i)} - \mathbf{z}^{(j)} \right\|^2}$$

確率的勾配降下法 [1] で $\mathbf{z}^{(i)}$ を更新していく

t-SNE ではモメンタム法を利用

[1] https://ja.wikipedia.org/wiki/確率的勾配降下法

$\mathbf{z}^{(i)}$ の初期値

$\mathbf{z}^{(i)}$ の初期値は、平均 0, 標準偏差 10^{-4} の正規分布に従うように、乱数で生成

t-SNE をやってみる

こちら https://github.com/hkaneko1985/tsne

をご利用ください

perplexity をどう決めるか？

① 試行錯誤によって決める

目安は、5 ～ 50 の間の値

② k3n-error (k-Nearest Neighbor Normalized Error for visualization and reconstruction) [1] によって最適化する

こちら https://github.com/hkaneko1985/tsne

に k3n-error を用いた *perplexity* の最適化のデモがあります

[1] H. Kaneko, Chemom. Intell. Lab. Syst., 176, 22-33, 2018

参考文献

- ✓L. van der Maaten, G. Hinton, Visualizing High-Dimensional Data Using t-SNE, . Journal of Machine Learning Research, 9, 2579–2605, 2008
  http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

- ✓sklearn.manifold.TSNE — scikit-learn 0.19.1 documentation
  http://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

- ✓H. Kaneko, K-Nearest Neighbor Normalized Error for Visualization and Reconstruction – A New Measure for Data Visualization Performance, Chemometrics and Intelligent Laboratory Systems, 176, 22-33, 2018
  https://www.sciencedirect.com/science/article/abs/pii/S0169743918300698