教師なし学習

佐藤 一誠
sato@k.u-tokyo.ac.jp
http://www.ms.k.u-tokyo.ac.jp

教師なし学習とは

データに共通する部分

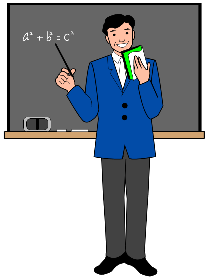
学習データ:

$$\{x_i\}_{i=1}^{n} \Rightarrow x_i = f_\theta(z_i)$$

データに固有な部分

教師付き学習は

$$\{y_i, x_i\}_{i=1}^{n} \Rightarrow y_i = f_\theta(x_i)$$



講義の流れ

1. 次元削減
2. クラスタリング
3. 非線形化
4. 生成モデル

データに共通する部分

$$x_i = f_\theta(z_i)$$

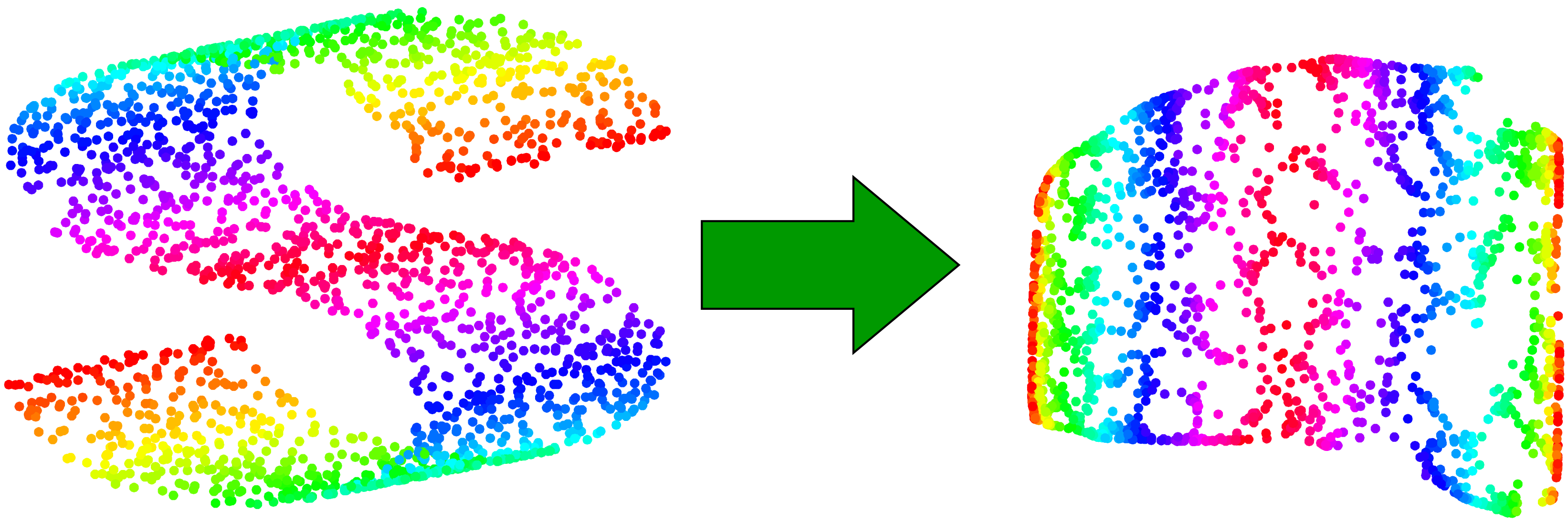
データに固有な部分

[zが連続の場合]
データの次元よりも
低い次元での表現

次元削減

- 本質的な情報を保持したまま次元を減らしたい！

もとの高次元データ　　　次元削減後のデータ(2d)



- 1～3次元に減らせば，データを可視化できる.
- 次元削減の基本的な仮定:
  - 手持ちの高次元データはある意味で冗長である

線形次元削減

- （高次元）標本：

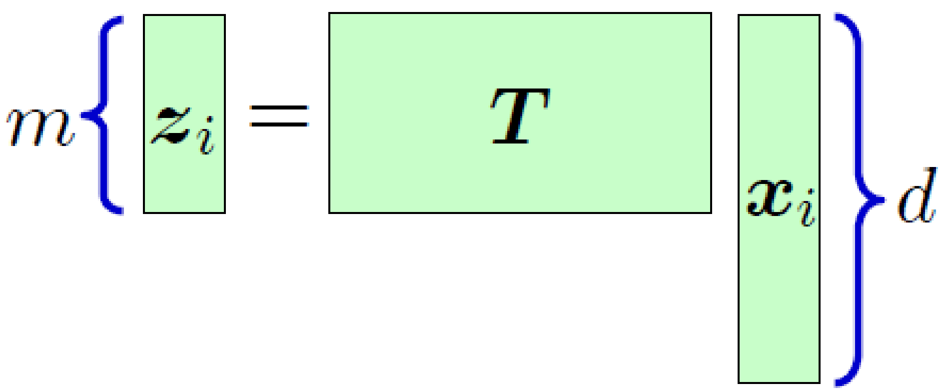
$$\{\boldsymbol{x}_i\}_{i=1}^n, \quad \boldsymbol{x}_i \in \mathbb{R}^d, \quad d \gg 1$$

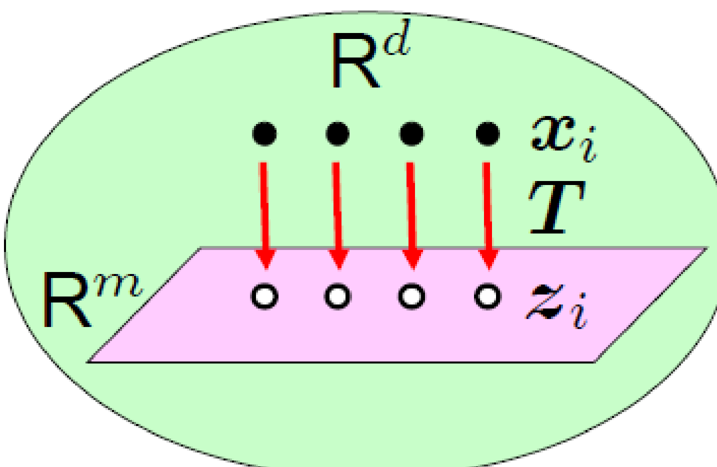
- 埋め込み行列：

$$\boldsymbol{T} \in \mathbb{R}^{m \times d}, \quad 1 \leq m \ll d$$

- 埋め込まれた標本

$$\{\boldsymbol{z}_i\}_{i=1}^n, \quad \boldsymbol{z}_i = \boldsymbol{T}\boldsymbol{x}_i \in \mathbb{R}^m$$



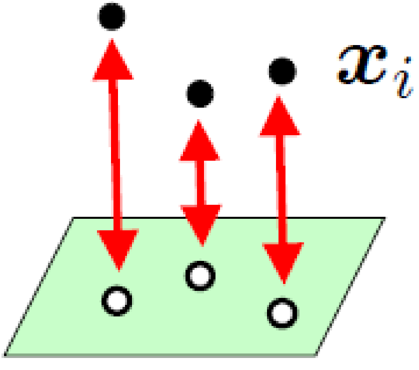


主成分分析 (PCA: Principal Component Analysis)

- 考え方:データを表現するうえで最も重要な次元(部分空間)を取り出す

$$\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}2\\0.1\end{pmatrix}, \begin{pmatrix}3\\-0.1\end{pmatrix}$$

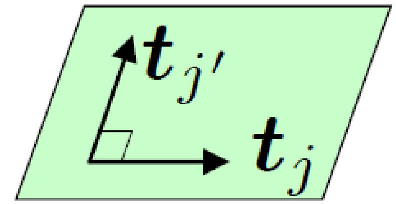
低次元の部分空間に正射影したときにできるだけデータが変化しないようにする



正射影

- $\{\boldsymbol{t}_j \mid \boldsymbol{t}_j \in \mathbb{R}^d\}_{j=1}^m$ ：m次元部分空間の正規直交基底

$$\boldsymbol{t}_j^\top \boldsymbol{t}_{j'} = \begin{cases} 1 & (j = j') \\ 0 & (j \neq j') \end{cases}$$

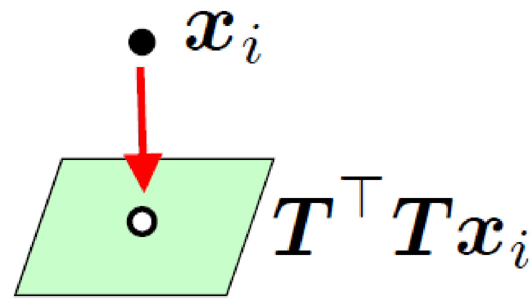


- 行列で表現すると， $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}_m$

$$\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_m)^\top \in \mathbb{R}^{m \times d}$$

- 標本 $\boldsymbol{x}_i$の正射影は

$$\sum_{j=1}^m (\boldsymbol{t}_j^\top \boldsymbol{x}_i)\boldsymbol{t}_j \quad \left(= \boldsymbol{T}^\top \boldsymbol{T} \boldsymbol{x}_i\right)$$



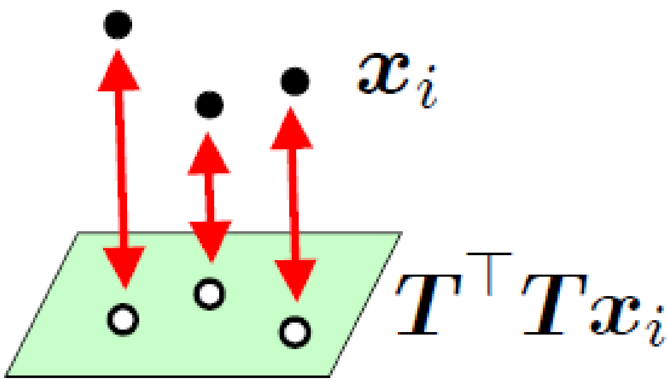
主成分分析の規準

- **射影誤差の和**を最小にする:

$$\boldsymbol{T}_{\mathrm{PCA}} = \mathop{\mathrm{argmin}}_{\boldsymbol{T}\in\mathbb{R}^{m\times d}} \left[\sum_{i=1}^{n} \|\boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{x}_i - \boldsymbol{x}_i\|^2\right]$$

$$\text{subject to } \boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}_m$$

（正規直交性の拘束条件をつける）



数学演習

■次式を証明せよ

$$\sum_{i=1}^{n} \|\boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{x}_i - \boldsymbol{x}_i\|^2 = -\mathrm{tr}\left(\boldsymbol{T}\boldsymbol{C}\boldsymbol{T}^\top\right) + \mathrm{tr}\left(\boldsymbol{C}\right)$$

$\mathrm{tr}\left(\cdot\right)$: 行列のトレース

$$\boldsymbol{C} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

標本の散布行列
（正規化されていない
共分散行列）

- 仮定: $\displaystyle \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = 0$

■ヒント:

$$\boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{T}^\top \boldsymbol{T} = \boldsymbol{T}^\top \boldsymbol{T}$$

$$\| y \|^2 = y^{\mathrm{T}} y$$

解答例

$$\sum_{i=1}^{n} \|\boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{x}_i - \boldsymbol{x}_i\|^2$$

$$= \sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{x}_i - 2\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{x}_i + \sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{x}_i$$

$\boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{T}^\top \boldsymbol{T} = \boldsymbol{T}^\top \boldsymbol{T}$: 2回射影しても変わらない

$$= -\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{T}^\top \boldsymbol{T}\boldsymbol{x}_i + \sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{x}_i$$

$\boldsymbol{a}^\top \boldsymbol{b} = \mathrm{tr}\left(\boldsymbol{b}\boldsymbol{a}^\top\right)$

$$= -\sum_{i=1}^{n} \mathrm{tr}\left(\boldsymbol{T}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{T}^\top\right) + \sum_{i=1}^{n} \mathrm{tr}\left(\boldsymbol{x}_i\boldsymbol{x}_i^\top\right)$$

$$= -\mathrm{tr}\left(\boldsymbol{T}\boldsymbol{C}\boldsymbol{T}^\top\right) + \mathrm{tr}\left(\boldsymbol{C}\right)$$

$$\boldsymbol{C} = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^\top$$

固有値問題

- 対称行列 $\boldsymbol{C}$ は，必ず

$$\boldsymbol{C} = \sum_{j=1}^{d} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^\top$$

という形に分解することができる．

- $\lambda_j, \boldsymbol{\xi}_j$ は固有値，固有ベクトルとよばれ，次の固有方程式の解として与えられる

$$\boldsymbol{C}\boldsymbol{\xi} = \lambda \boldsymbol{\xi}$$

- 固有ベクトル $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d$ は互いに直交する：

$$\boldsymbol{\xi}_j^\top \boldsymbol{\xi}_{j'} = 0 \quad \text{for } j \neq j'$$

- Octaveではeig関数で計算できる

固有値問題

$$\boldsymbol{C} = \sum_{j=1}^{d} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^\top$$

- 固有値分解を用いれば，逆行列は次式で，

$$\boldsymbol{C}^{-1} = \sum_{j=1}^{d} \lambda_j^{-1} \boldsymbol{\xi}_j \boldsymbol{\xi}_j^\top \qquad \forall \lambda_j \neq 0$$

行列の平方根は次式で求められる

$$\boldsymbol{C}^{1/2} = \sum_{j=1}^{d} \lambda_j^{1/2} \boldsymbol{\xi}_j \boldsymbol{\xi}_j^\top \qquad \forall \lambda_j > 0$$

主成分分析の解

$$\boldsymbol{T}_{\mathrm{PCA}} = \operatorname*{argmax}_{\boldsymbol{T} \in \mathbb{R}^{m \times d}} \mathrm{tr}\left(\boldsymbol{T}\boldsymbol{C}\boldsymbol{T}^{\top}\right)$$

$$\text{subject to } \boldsymbol{T}\boldsymbol{T}^{\top} = \boldsymbol{I}_m$$

- 主成分分析の解は次式で与えられる:

$$\boldsymbol{T}_{\mathrm{PCA}} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)^{\top}$$

- $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m$:は固有値問題 $\boldsymbol{C}\boldsymbol{\xi} = \lambda\boldsymbol{\xi}$ の固有値 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ に対応する正規化 $\|\boldsymbol{\xi}_j\| = 1$ された固有ベクトル

主成分分析の解の求め方

1. 固有値問題を解く:

$$C\xi = \lambda\xi$$

標本の散布行列

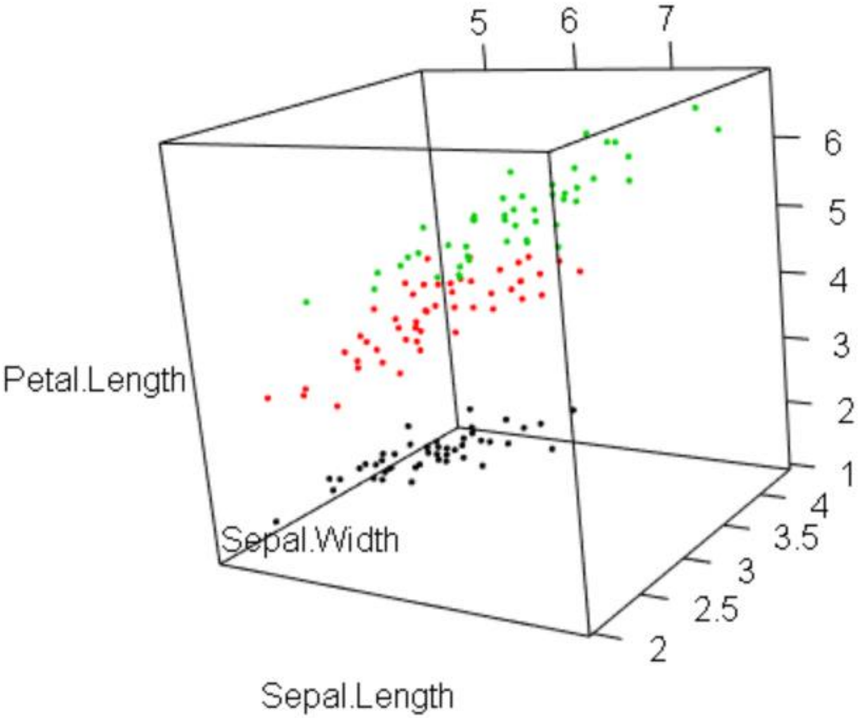
$$C = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

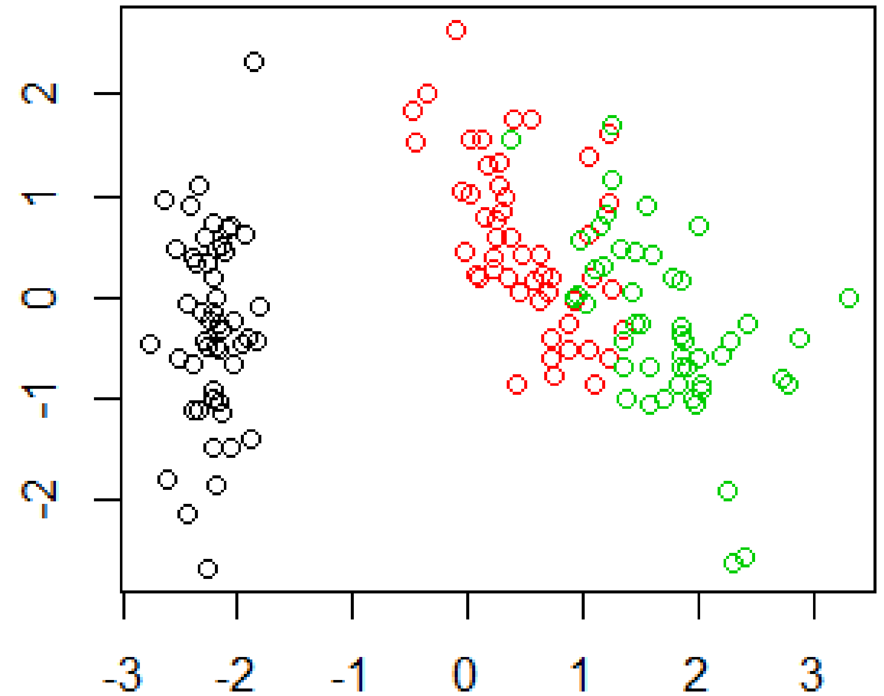
- 固有値を降順にソート: $\lambda_1 \geq \cdots \geq \lambda_d$
- 固有ベクトルを正規化: $\|\boldsymbol{\xi}_j\| = 1$

2. 上位 m 個の固有ベクトルを並べる:

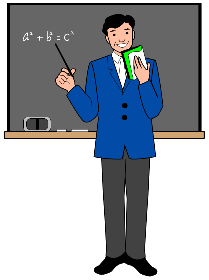
$$\boldsymbol{T}_{\mathrm{PCA}} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)^\top$$

実行例(IRISデータ):3D⇒2D





- 主成分分析によって，データの大局的な分布をよく表す部分空間が得られている
- しかし，クラスタ構造のようなデータの局所的な構造は失われることがある



講義の流れ

1. 次元削減
2. クラスタリング
3. 非線形化
4. 生成モデル

データに共通する部分

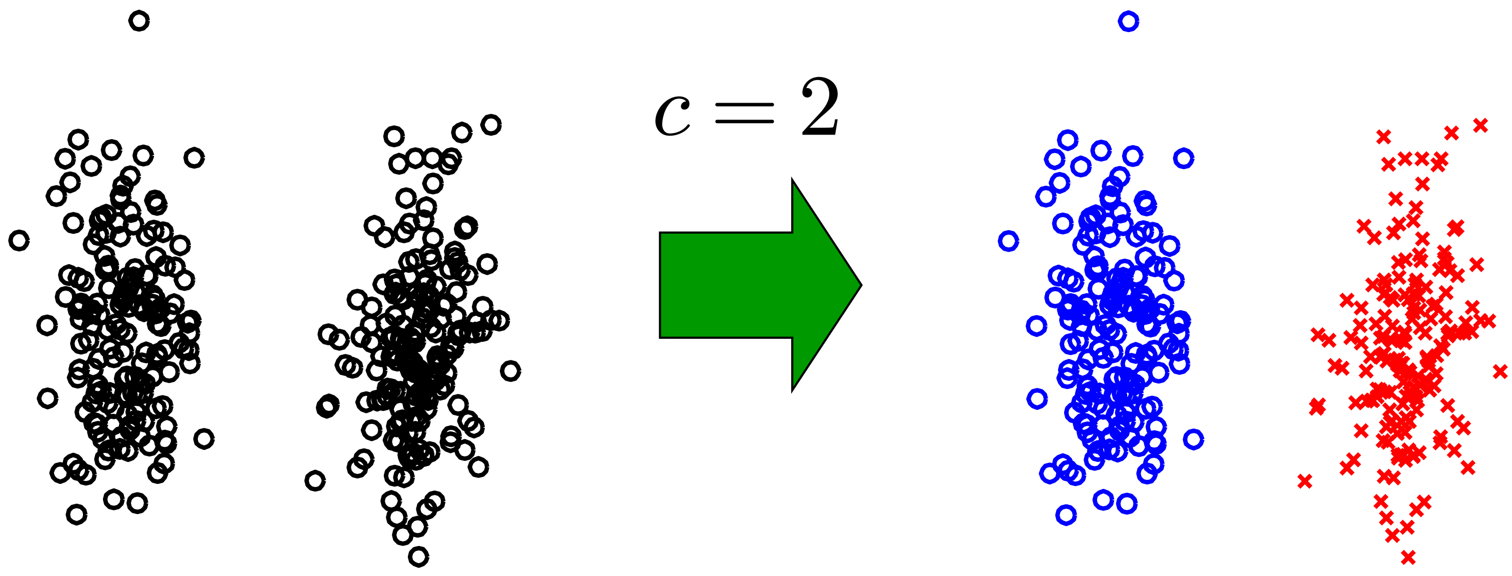
$$x_i = f_\theta(z_i)$$

データに固有な部分

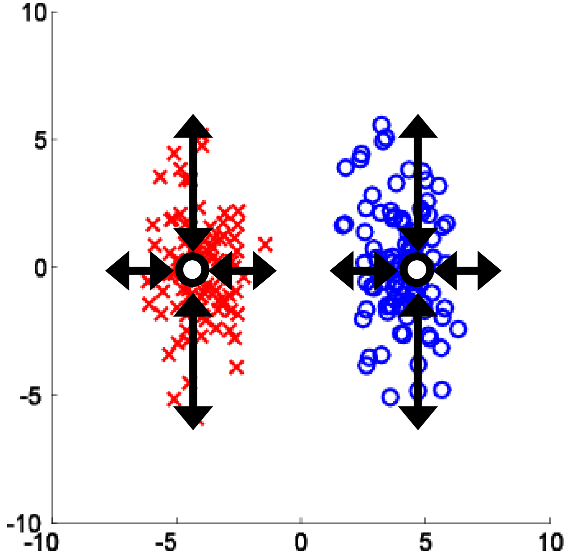
[zが離散の場合]
データのクラスタ
構造を表現

データのクラスタリング

- 目標:ラベル無しの標本 $\{\boldsymbol{x}_i\}_{i=1}^{n}$ を c 個のグループに分ける
  - 似た性質を持つデータは同じグループに
  - 異なる性質を持つデータは違うグループに
- c はあらかじめ固定しておく.



クラスタ内散布和の最小化



- クラスタ内散布:

$$\sum_{i:y_i=y} \|\boldsymbol{x}_i - \boldsymbol{\mu}_y\|^2 \qquad \boldsymbol{\mu}_y = \frac{1}{n_y}\sum_{i:y_i=y} \boldsymbol{x}_i$$

- 考え方:クラスタ内散布の和を最小にするように標本をクラスタに割り当てる

$$\min_{y_1,\ldots,y_n\in\{1,\ldots,c\}} \sum_{y=1}^{c} \sum_{i:y_i=y} \|\boldsymbol{x}_i - \boldsymbol{\mu}_y\|^2$$

- しかしこれはNP困難(計算量がO(c^n))なので実時間では解けない.

k-平均クラスタリング

1. クラスタ中心 $\{\boldsymbol{\mu}_y\}_{y=1}^{c}$ を(ランダムに)初期化する
2. 収束するまで以下を繰り返す:
   A) 各標本のクラスタの割り当てを次式で更新する:

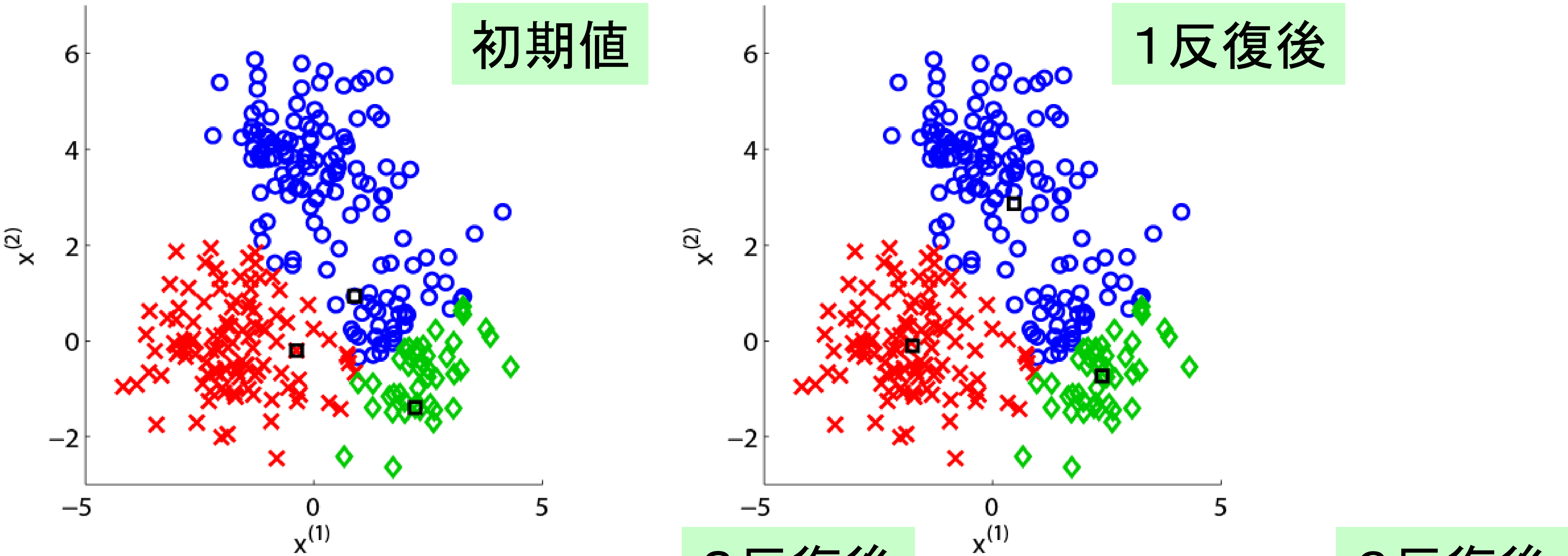
$$y_i \longleftarrow \underset{y\in\{1,\ldots,c\}}{\mathrm{argmin}} \|\boldsymbol{x}_i - \boldsymbol{\mu}_y\|^2, \quad i = 1, \ldots, n$$

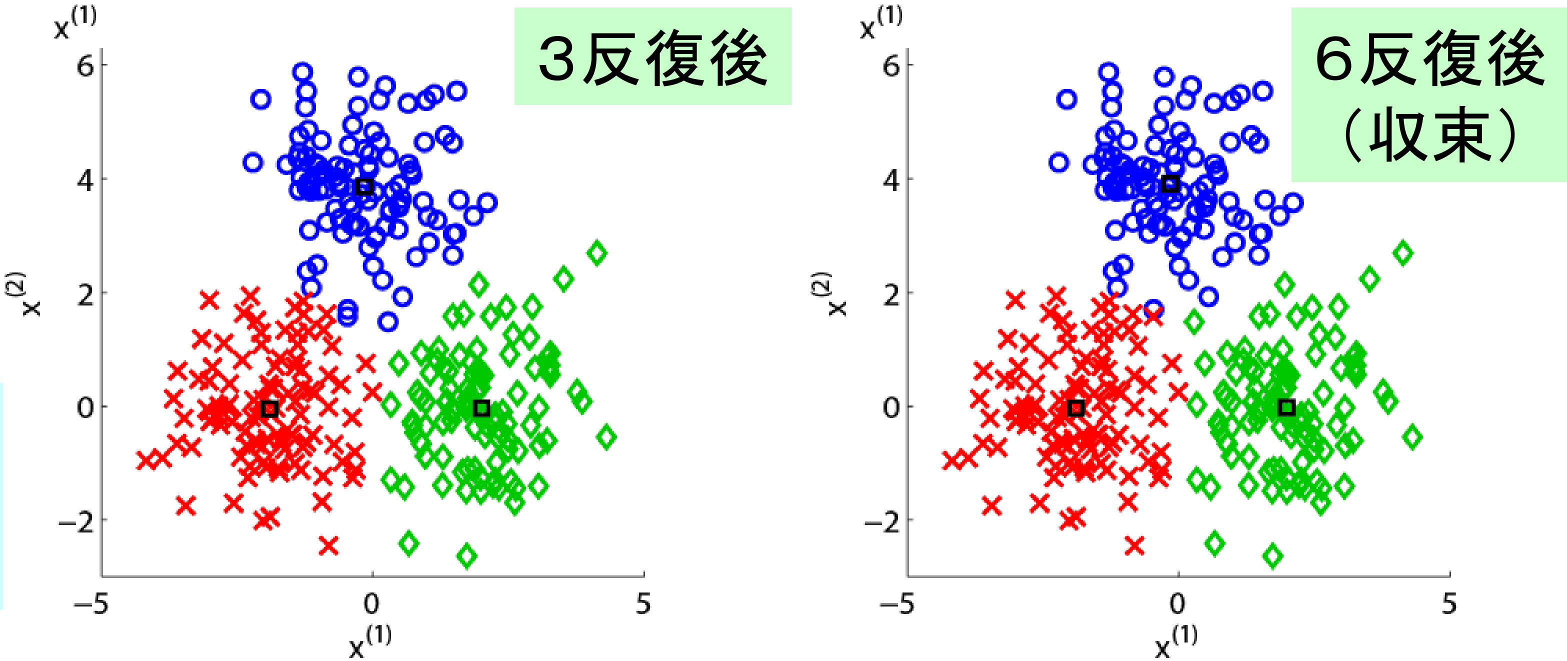
   B) 各クラスタの中心を更新する:

$$\boldsymbol{\mu}_y \longleftarrow \frac{1}{n_y} \sum_{i:y_i=y} \boldsymbol{x}_i, \quad y = 1, \ldots, c$$

- これは, クラスタ内散布和の最小化問題の局所最適解を求めるアルゴリズムになっている.

実行例

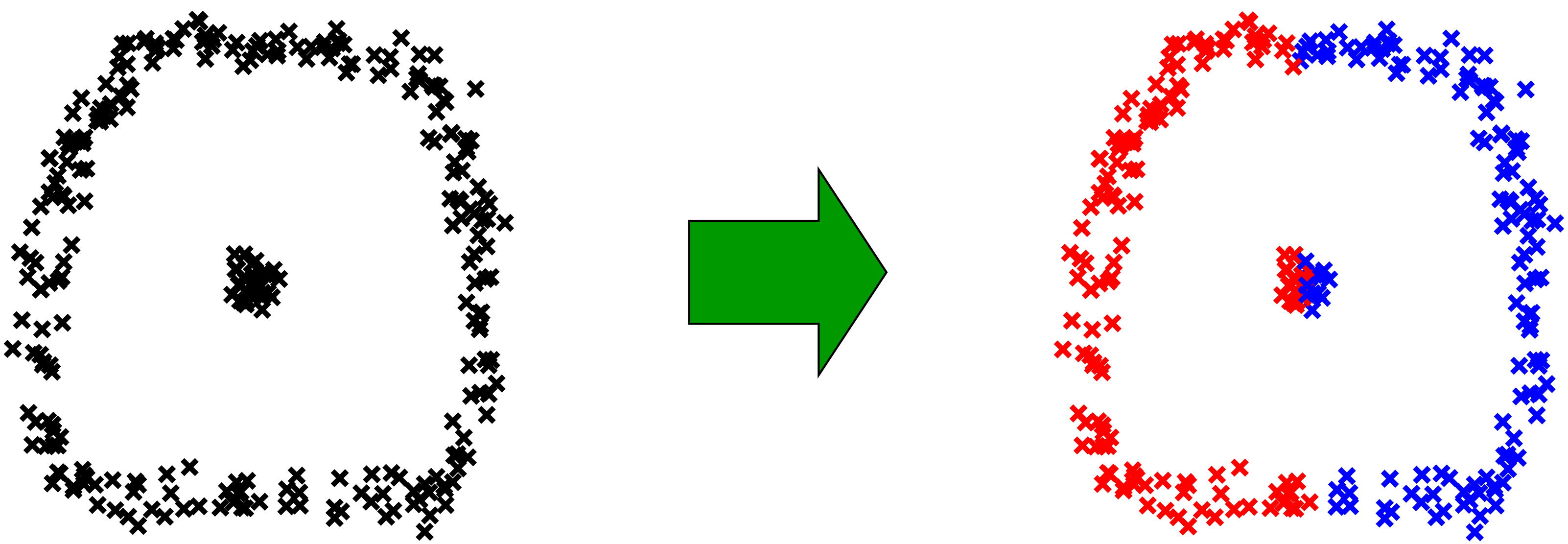


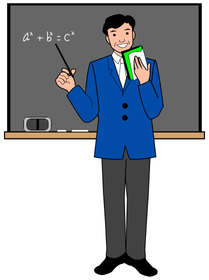


実装は
宿題

クラスタリング：まとめ

- クラスタ内散布の和を最小にする
- そのままではNP困難なので近似解を求める
  - クラスタリング結果が初期値に依存する.
- クラスタの形が凸でないとき, うまくいかない.
- クラスタ数をあらかじめ決める必要がある.



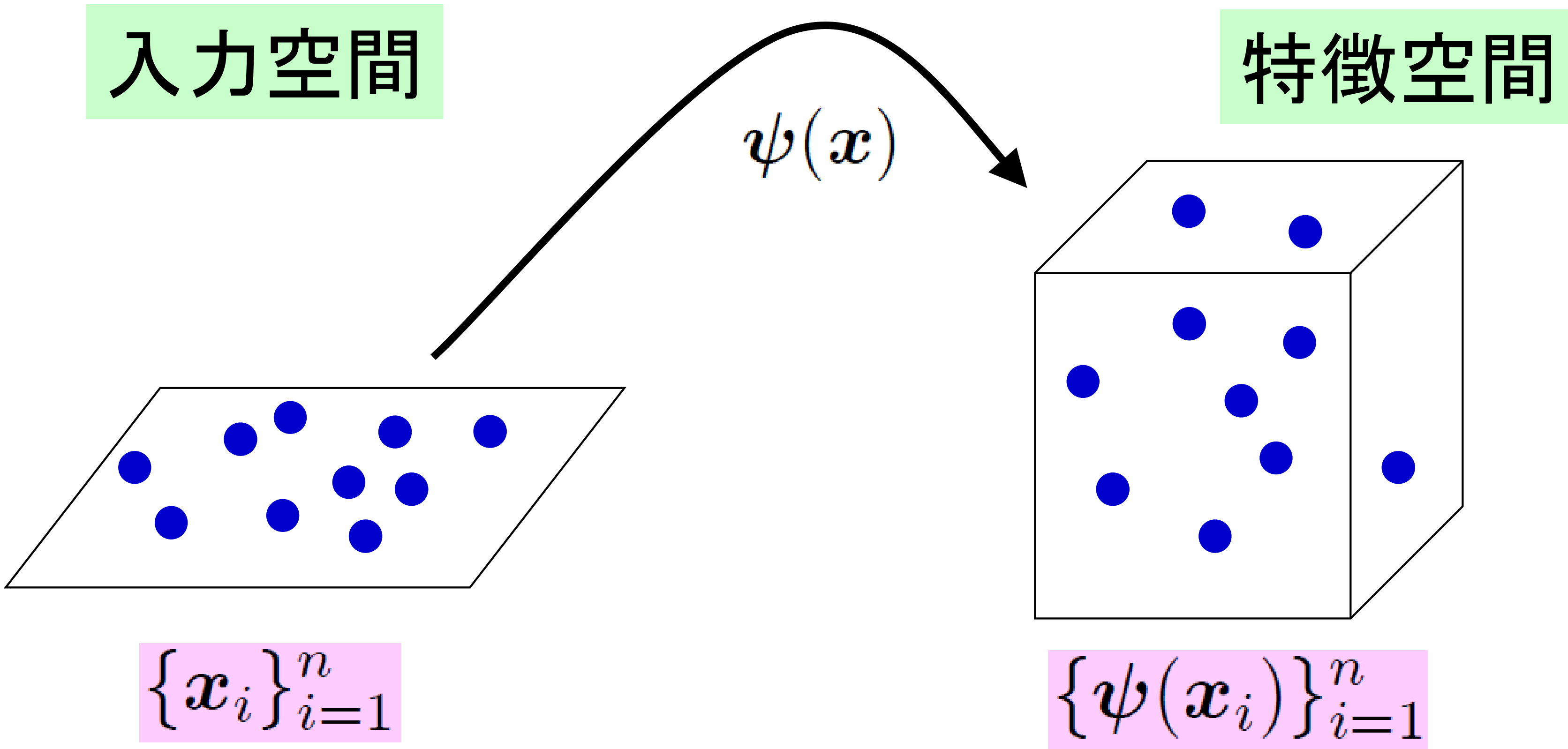


講義の流れ

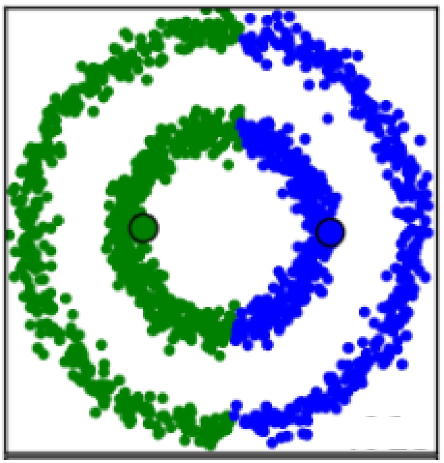
1. 次元削減
2. クラスタリング
3. 非線形化
4. 生成モデル

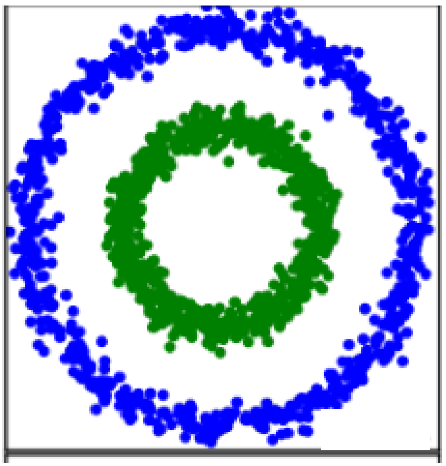
非線形への拡張

- 非線形関数 $\psi(x)$ で標本を特徴空間へ写像し，特徴空間内でアルゴリズムを実行

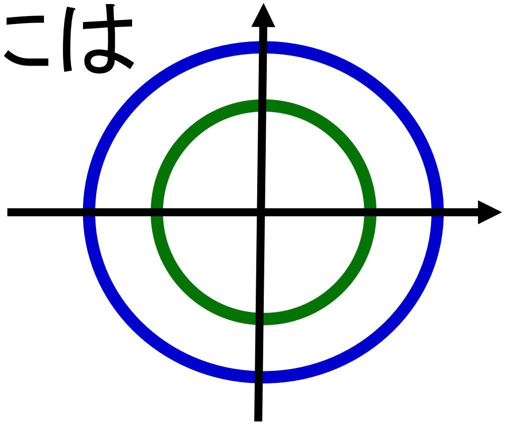


K-meansの非線形拡張





直感的には



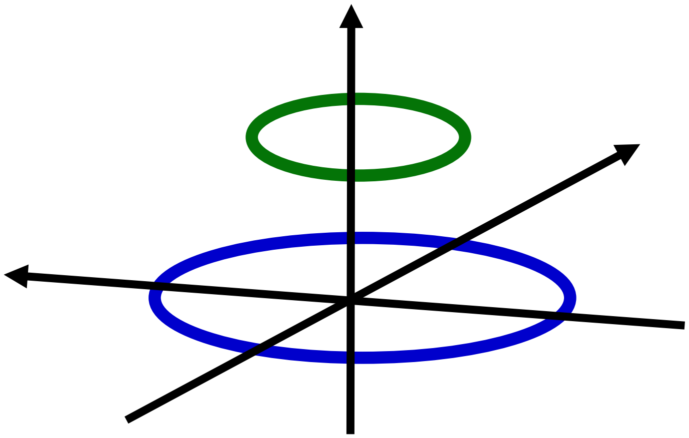


Image from http://scikit-learn.org/stable/auto_examples
/cluster/plot_cluster_comparison.html

カーネルトリック

- 特徴空間内での内積を半正定値カーネル関数で計算：

$$\psi(\boldsymbol{x}_i)^\top \psi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\forall \boldsymbol{x}, \boldsymbol{x}', \quad K(\boldsymbol{x}, \boldsymbol{x}') \geq 0$$

例えばガウシアンカーネル

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2 / (2h^2)\right)$$

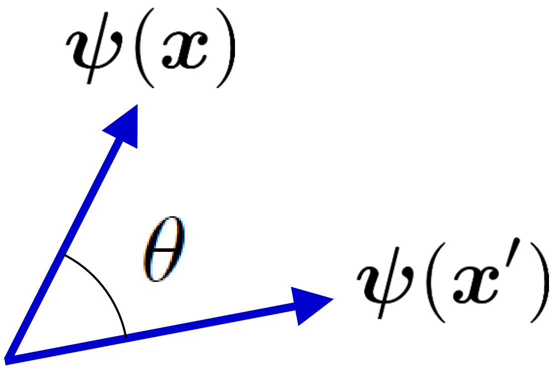
- 主成分分析，k平均クラスタリングはカーネル非線形化できる
  - しかし，カーネルの選択が自明でない

数学演習

$$\boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\psi}(\boldsymbol{x}') = K(\boldsymbol{x}, \boldsymbol{x}')$$

■ 以下の量をカーネル関数を用いて計算せよ

- ノルム: $\|\boldsymbol{\psi}(\boldsymbol{x})\|$
- 距離: $\|\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}(\boldsymbol{x}')\|$
- 角度: $\cos\theta$



$$\boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\psi}(\boldsymbol{x}') = \|\boldsymbol{\psi}(\boldsymbol{x})\| \|\boldsymbol{\psi}(\boldsymbol{x}')\| \cos\theta$$

解答例

- $$\|\boldsymbol{\psi}(\boldsymbol{x})\| = \sqrt{\|\boldsymbol{\psi}(\boldsymbol{x})\|^2} = \sqrt{\boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\psi}(\boldsymbol{x})} = \sqrt{K(\boldsymbol{x},\boldsymbol{x})}$$

- $$\begin{aligned}\|\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}(\boldsymbol{x}')\| &= \sqrt{\|\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}(\boldsymbol{x}')\|^2} \\ &= \sqrt{\|\boldsymbol{\psi}(\boldsymbol{x})\|^2 - 2\boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\psi}(\boldsymbol{x}') + \|\boldsymbol{\psi}(\boldsymbol{x}')\|^2} \\ &= \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - 2K(\boldsymbol{x},\boldsymbol{x}') + K(\boldsymbol{x}',\boldsymbol{x}')}\end{aligned}$$

- $$\cos\theta = \frac{\boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\psi}(\boldsymbol{x}')}{\sqrt{\|\boldsymbol{\psi}(\boldsymbol{x})\|^2 \|\boldsymbol{\psi}(\boldsymbol{x}')\|^2}} = \frac{K(\boldsymbol{x},\boldsymbol{x}')}{\sqrt{K(\boldsymbol{x},\boldsymbol{x})K(\boldsymbol{x}',\boldsymbol{x}')}}$$

数学演習

$$\boldsymbol{\psi}(\boldsymbol{x})^\top \boldsymbol{\psi}(\boldsymbol{x}') = K(\boldsymbol{x}, \boldsymbol{x}')$$

- カーネル関数を用いてk-平均法の以下の式を表現せよ

$$y_i \longleftarrow \underset{y \in \{1,\ldots,c\}}{\operatorname{argmin}} \|\boldsymbol{x}_i - \boldsymbol{\mu}_y\|^2, \quad i = 1, \ldots, n$$

ヒント:

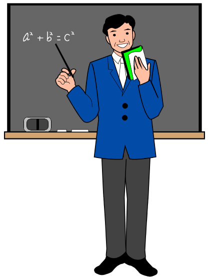
$$k = \arg\min_k \left\| \psi(x_i) - \frac{1}{n} \sum_{j:z_j=k} \psi(x_j) \right\|^2$$

解答例

$$\left\|\psi(x_i) - \frac{1}{n}\sum_{j:z_j=k}\psi(x_j)\right\|^2$$

$$= \|\psi(x_i)\|^2 - \frac{2}{n}\psi(x_i)^T \sum_{j:z_j=k}\psi(x_j) + \frac{1}{n^2}\sum_{j:z_j=k}\sum_{j':z_{j'}=k}\psi(x_j)\psi(x_{j'})$$

$$= K(x_i, x_i) - \frac{2}{n}\sum_{j:z_j=k}K(x_i, x_j) + \frac{1}{n^2}\sum_{j:z_j=k}\sum_{j':z_{j'}=k}K(x_j, x_{j'})$$



講義の流れ

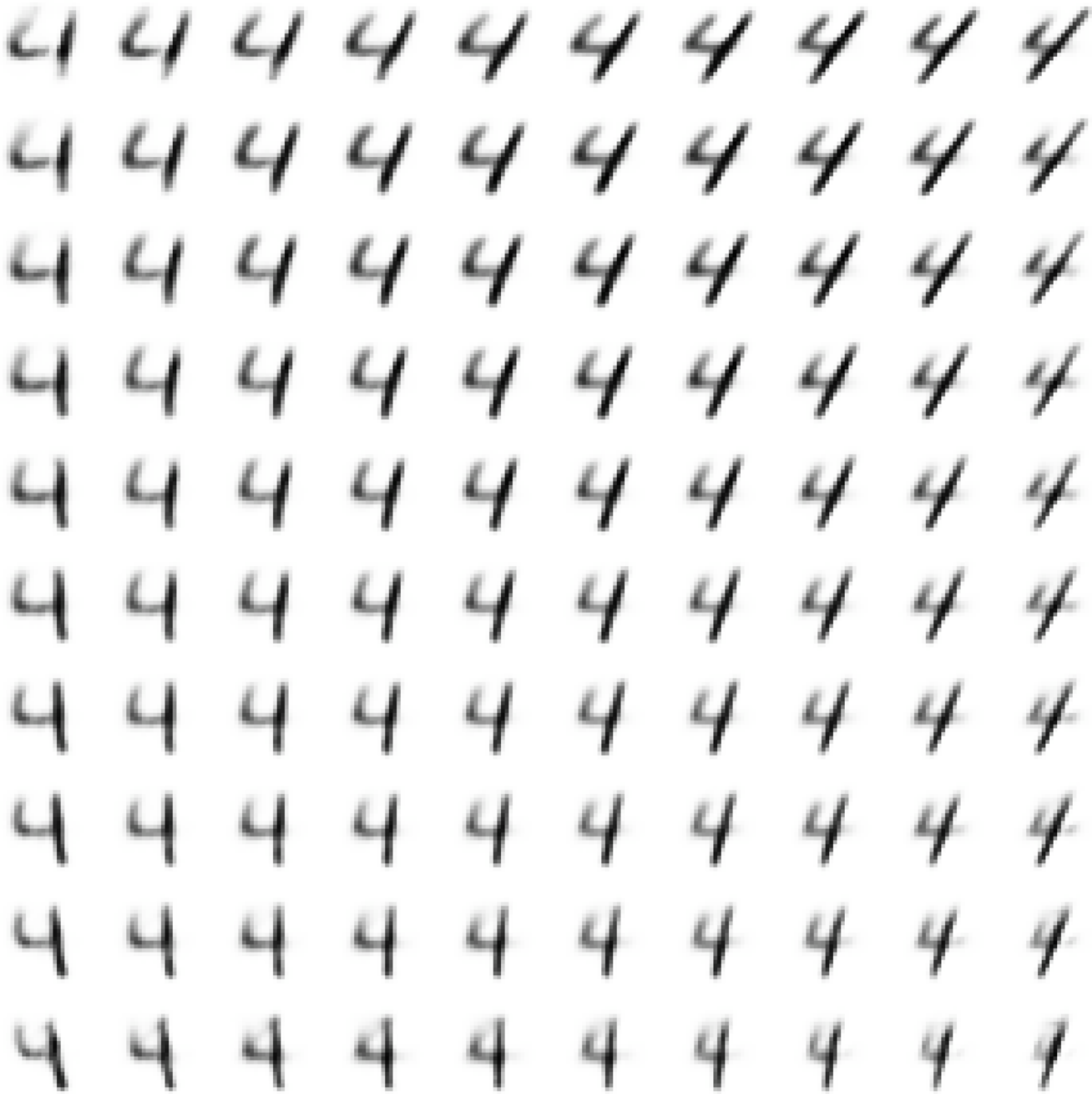
1. 次元削減
2. クラスタリング
3. 非線形化
4. 生成モデル

データに共通する部分

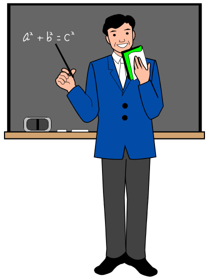
$$x_i = f_\theta(z_i)$$

データに固有な部分

zを入力すれば
xを生成できる



Kingma+, Semi-supervised Learning with Deep Generative Models , NIPS2014



講義の流れ

1. 次元削減
2. クラスタリング
3. 非線形化
4. 生成モデル

データに共通する部分

$$x_i = f_\theta(z_i, y_i)$$

連続 離散

データに固有な部分連続と離散の両方を用いる

離散部分はカテゴリ情報があるデータ
に関しては予め入力しておく

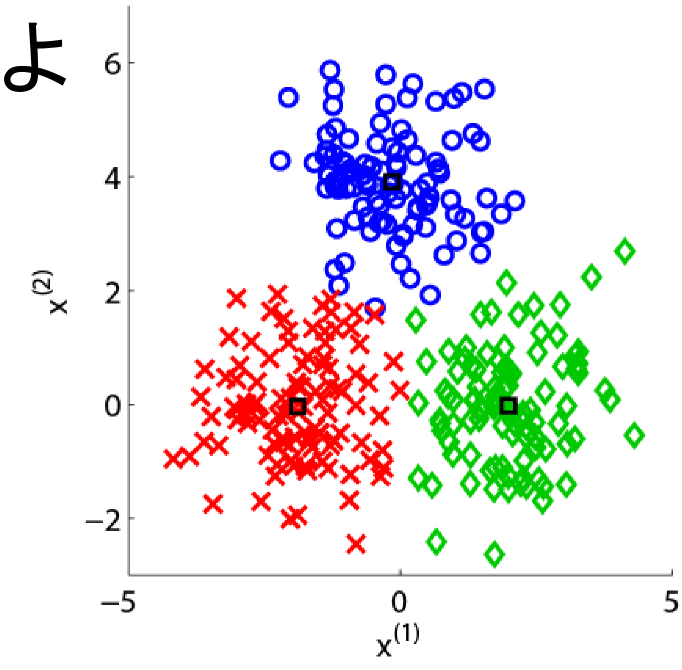
Kingma+, Semi-supervised Learning with Deep Generative Models , NIPS2014

文章の生成

「レンタ・カーは空のグラスを手にとり、蛇腹はすっかり暗くなっていた。それはまるで獲物を咀嚼しているようだった。彼は僕と同じようなものですね」と私は言った。「でもあなたはよく女の子に爪切りを買った。そしてその何かを振り払おうとしたが、今では誰にもできやしないのよ。私は長靴を棚の上を乗り越えるようにした。…

宿題

- 3次元以上のデータ(任意)に対してPCAを適用し2次元へ次元圧縮せよ(実装例を示す)
- 2次元に次元圧縮したデータを用いてk平均クラスタリングを適用せよ(実装例を示す)
- クラスタリング結果を可視化せよ



可視化例

まとめ

- 教師なし学習：入力データだけから学習
  - 次元削減：高次元データに含まれる本質的な情報を保持したまま次元数を削減
  - クラスタリング：データをグループに分割
  - 生成モデル：訓練データから新たなデータを生成
- カーネルトリックにより非線形化できるが，結果がカーネルの選び方に依存する