機械学習の考え方

～データからの知識発見に向けて～

日野英逸

筑波大学

2015 年 3 月 11 日

立命館大学びわこ・くさつキャンパス
講演原稿を，公開用に編集したものです.

機械学習の位置づけ

基本的な立場

(本講演で採用する) 機械学習の定義

- データから自動的に規則を発見し，その規則を現状分析や将来の予測，あるいは不確定な状況での意思決定に利用する手法群の総称

今日話すこと

機械学習で扱われる種々の問題設定

- 機械学習には，「＊＊」学習と呼ばれる種々の問題設定がある．
- よく,「機械学習は魔法ではない」と注意される．
- 今日は出来ないことや限界は考えず，標語的に「こういう問題も扱える」というカタログ的な紹介を通し，機械学習の可能性を強調
- 実際の問題に適用してうまくいかないことが多いのは皆さんの方が体感しているはず．当然，機械学習にもノウハウがあり，それはデータ解析と統計学の学習の両方を通して身に付けるもの

今日話すこと

機械学習で扱われる種々の問題設定

- 機械学習には，「＊＊」学習と呼ばれる種々の問題設定がある．
- よく,「機械学習は魔法ではない」と注意される．
- 今日は出来ないことや限界は考えず，標語的に「こういう問題も扱える」というカタログ的な紹介を通し，機械学習の可能性を強調
- 実際の問題に適用してうまくいかないことが多いのは皆さんの方が体感しているはず．当然，機械学習にもノウハウがあり，それはデータ解析と統計学の学習の両方を通して身に付けるもの

今日話すこと

機械学習で扱われる種々の問題設定

- 機械学習には，「＊＊」学習と呼ばれる種々の問題設定がある．
- よく,「機械学習は魔法ではない」と注意される．
- 今日は出来ないことや限界は考えず，標語的に「こういう問題も扱える」というカタログ的な紹介を通し，機械学習の可能性を強調
- 実際の問題に適用してうまくいかないことが多いのは皆さんの方が体感しているはず．当然，機械学習にもノウハウがあり，それはデータ解析と統計学の学習の両方を通して身に付けるもの

Index

1 導入

2 種々の問題設定
- 教師付き学習

3 教師無し学習

4 その他中間的方法

5 終わりに

Index

1 導入

2 種々の問題設定
- 教師付き学習

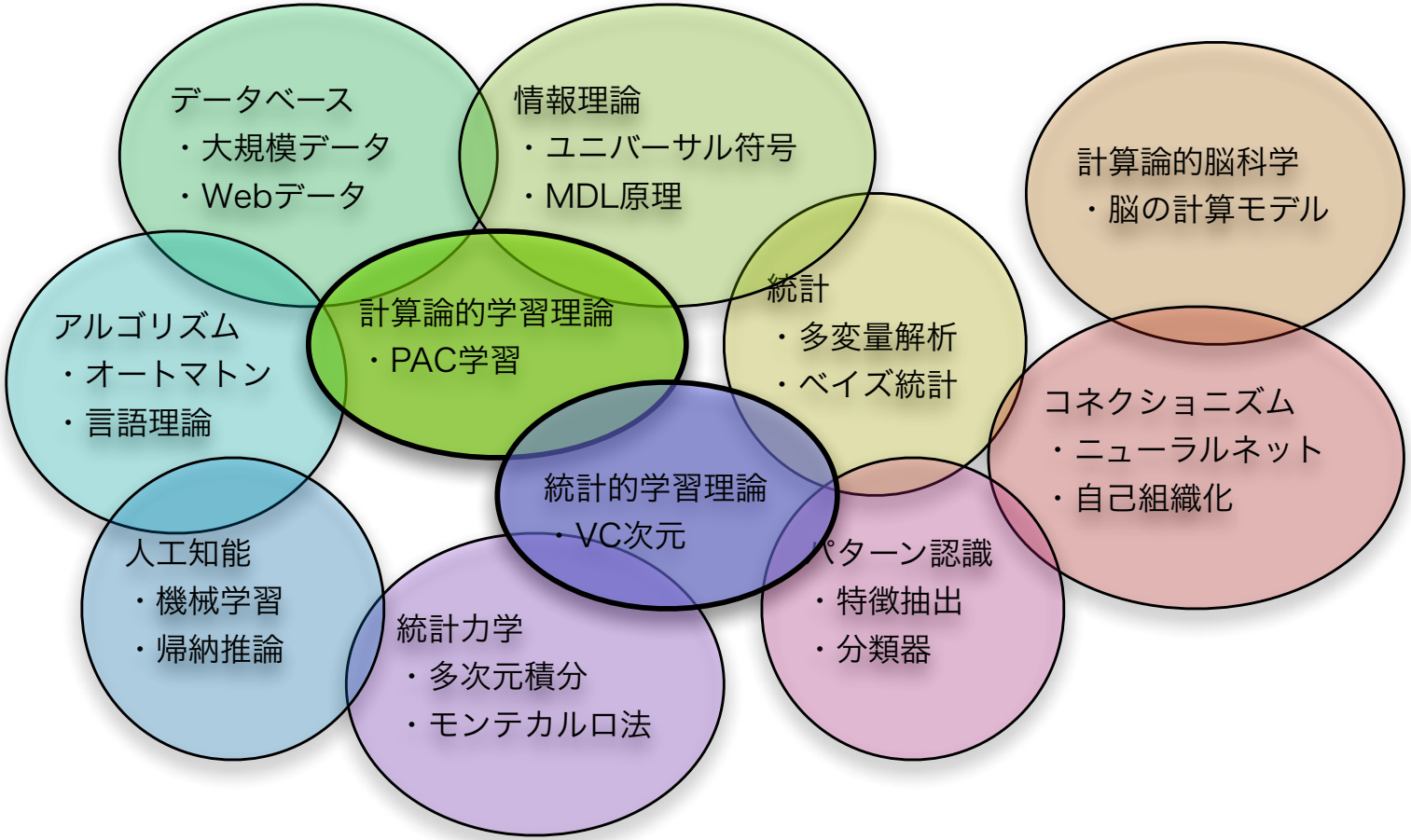
3 教師無し学習

4 その他中間的方法

5 終わりに

機械学習の位置づけ

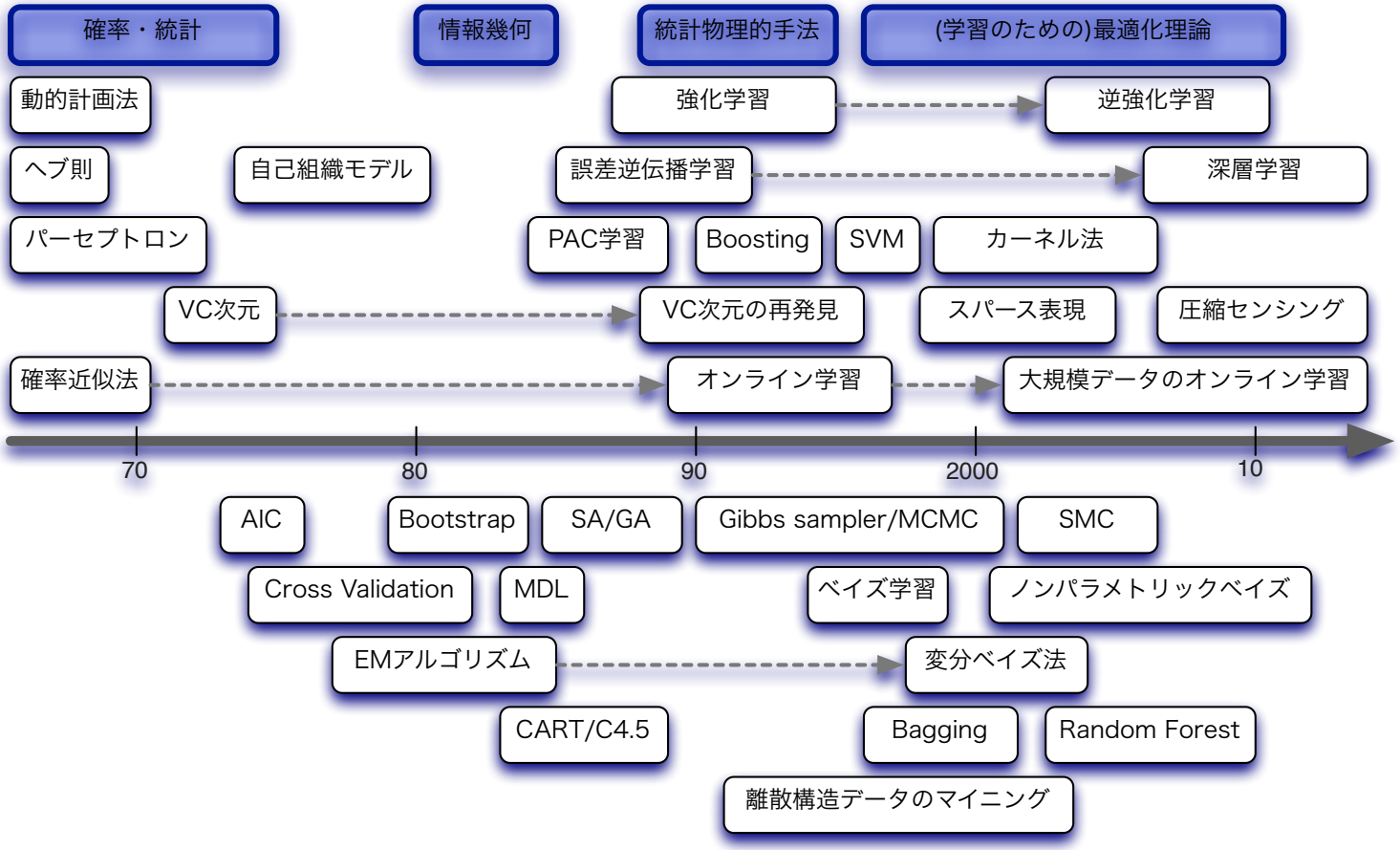
関連分野 [1]



[1] through the courtesy of Noboru Murata (Waseda Univ.)

機械学習の位置づけ

発展の歴史 [2]



[2] through the courtesy of Noboru Murata (Waseda Univ.)

機械学習の位置づけ

分類法の分類

- データ形式での分類
  - ベクトル，時系列，グラフ，画像，文字列，相対位置
- モデルでの分類
  - パラメトリックモデル
  - ノンパラメトリックモデル

  モデルの使い方
  - 単一のモデルを利用，複数のモデルを利用
- 手法・解法での分類
  - 線形計画問題，二次計画問題，半正定置計画問題，凸錘計画問題，整数計画問題，混合計画問題
  - 勾配法，劣勾配法，近接法，確率勾配法，補助関数法，粒子フィルタ，MCMC，シミュレーテッドアニーリング
- 問題設定での分類
  - 教師付き学習，教師無し学習，半教師付き学習，転移学習，強化学習，能動学習，距離学習，etc....

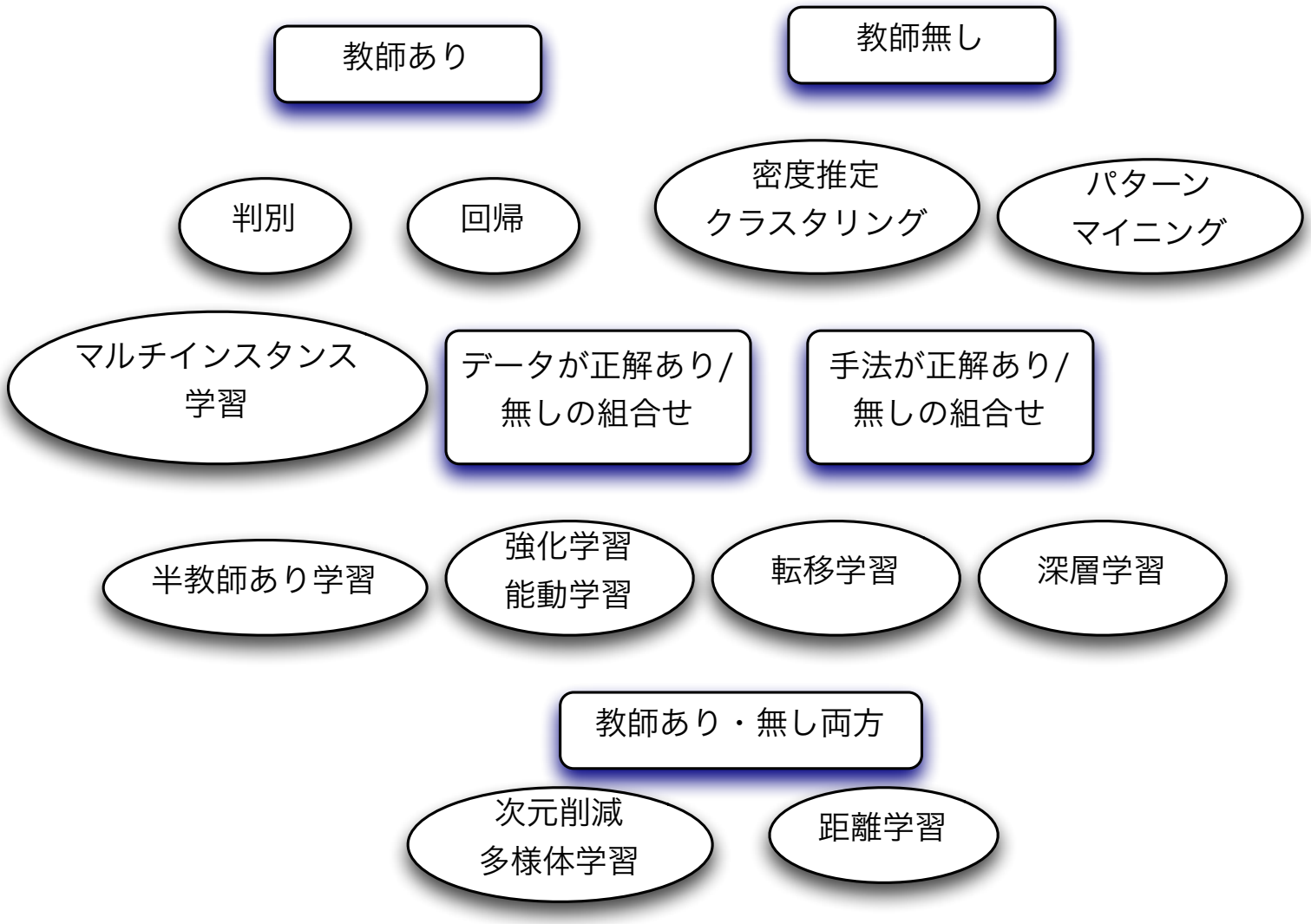
機械学習の位置づけ

分類法の分類

- データ形式での分類
  - ベクトル，時系列，グラフ，画像，文字列，相対位置
- モデルでの分類
  - パラメトリックモデル
  - ノンパラメトリックモデル

  モデルの使い方
  - 単一のモデルを利用，複数のモデルを利用
- 手法・解法での分類
  - 線形計画問題，二次計画問題，半正定置計画問題，凸錘計画問題，整数計画問題，混合計画問題
  - 勾配法，劣勾配法，近接法，確率勾配法，補助関数法，粒子フィルタ，MCMC，シミュレーテッドアニーリング
- 問題設定での分類
  - 教師付き学習，教師無し学習，半教師付き学習，転移学習，強化学習，能動学習，距離学習，etc....

機械学習:問題設定での分類



Index

1 導入

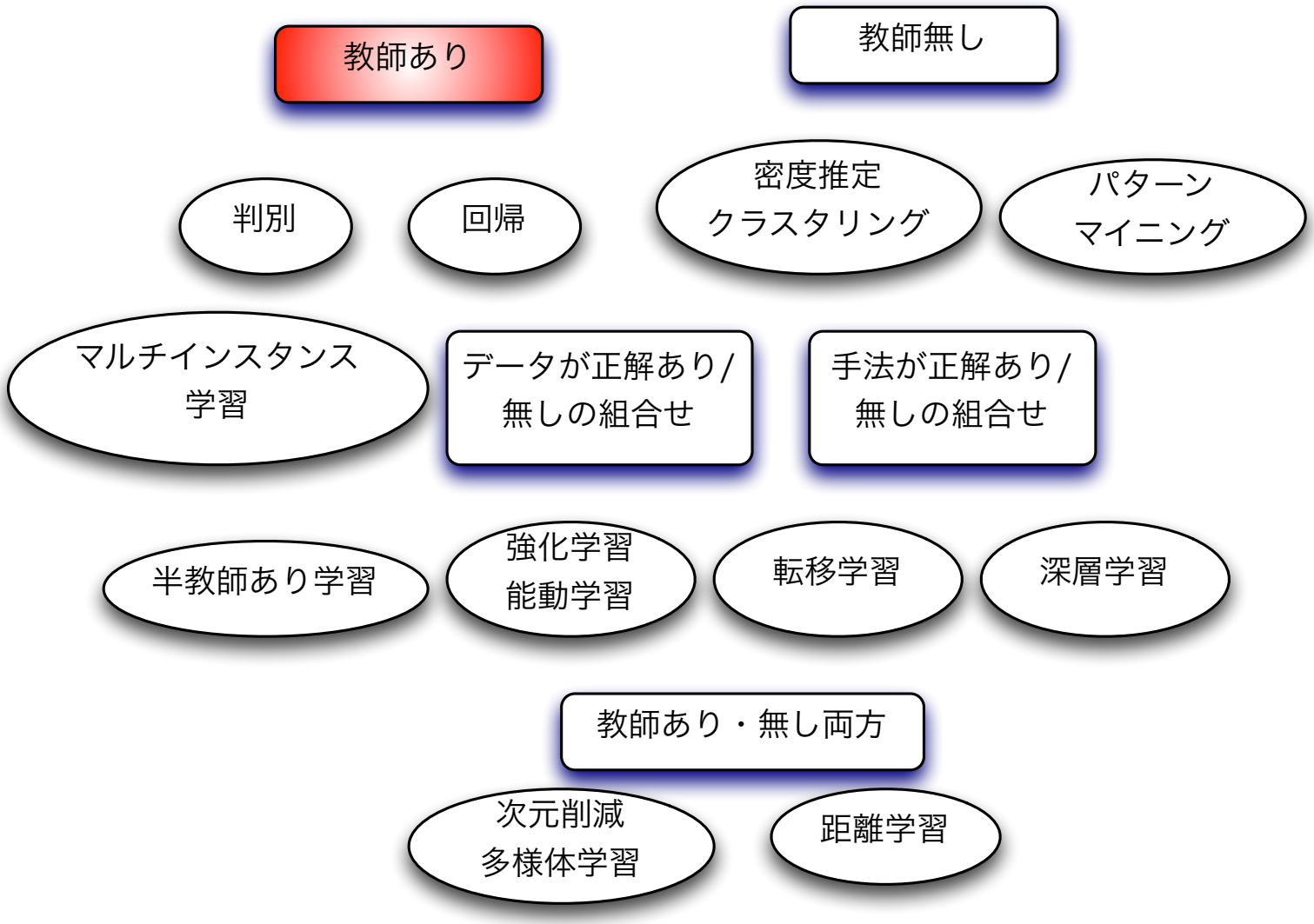
2 種々の問題設定
- 教師付き学習

3 教師無し学習

4 その他中間的方法

5 終わりに

機械学習:問題設定での分類



教師付き学習

基礎となる枠組み

- 最も典型的な問題設定
- 学習器が学ぶべき規則に従う「入力」と「出力」の例が与えられている状況
- サンプル数，特徴量の次元，例題のバランス，規則の複雑さなど，多様な状況で理論解析と実応用が進んでいる

教師付き学習

目的

入力と出力の関係 $f : \mathcal{X} \to \mathcal{Y}$ を学習（この関係 f を,「機械」と呼ぶ）

入力 学習用データ集合 $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$.

出力 $y \simeq \hat{y} = f(\boldsymbol{x})$ なる機械 f

- $\boldsymbol{x}$ は，説明変数，独立変数，入力，特徴量などと呼ばれる
- y は，応答変数，従属変数，出力，目的変数などと呼ばれる

- y が離散値：判別問題
- y が連続値：回帰問題

教師付き学習

例

- クレジットカード与信審査（判別）
  入力　年収・職業・年齢・家族構成など
  出力　カード発行の可否
- スパムメールフィルタ（判別）
  入力　メール
  出力　スパム or ハム
- 商品の推薦（回帰）
  入力　購買履歴や顧客情報
  出力　未購入商品の選好度
- 株価予測（回帰）
  入力　その日までの市場関連情報
  出力　翌日の株価

教師付き学習

定式化：リスク最小化

- 教師付き学習は，損失，リスク最小化の枠組みで議論するのがやりやすい．
- 特に，学習対象の機械 $f(\boldsymbol{x})$ が，何らかのパラメタ θ で特徴付けられており，学習がこのパラメタの最適化問題に帰着できる「パラメトリックモデル」を考える．
  - 損失関数の考え方は，教師付き学習でなくても適用可能

教師付き学習

定式化：リスク最小化

- 損失関数 (loss function)：
  データ $z = (\boldsymbol{x}, y)$ に対してパラメタ θ の機械が被る罰金 $\leftrightarrow$ データに対する機械のズレ
- $l(z;\theta)$：データ z とパラメタ θ の関数
  - 教師付き学習での例: $z = (\boldsymbol{x}, y)$

$$l(z;\theta) = \|y - f(\boldsymbol{x};\theta)\|^2 \quad 2\text{ 乗誤差}$$

f は例えば線形モデルや決定木，ニューラルネットワークなど入出力関数のモデル

教師付き学習

定式化：リスク最小化

パラメトリックモデルにおける最適なパラメタの基準を考える

- 期待損失 (expected loss, risk)：データを生成している確率分布 P に対するパラメタ θ の悪さの評価

$$L_P(\theta) = E_{Z\sim P}[l(Z;\theta)] = \int p(z)l(z;\theta)\mathrm{d}z$$

- 最適モデル：期待損失に基づいた評価基準で最適な学習機械

$$\theta^* = \arg\min_{\theta\in\Theta} L_P(\theta) = \arg\min_{\theta\in\Theta} E_{Z\sim P}[l(Z;\theta)]$$

一般には，真の分布 P が未知であり，期待損失の計算は出来ないことが難しさとなる

教師付き学習

定式化：リスク最小化

観測データ $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ を用いた学習を考える

経験損失 (empirical loss)：データ D に対する学習機械 θ の悪さの評価

$$L_D(\theta) = \frac{1}{n} \sum_{z_i \in D} l(z_i; \theta)$$

- 計算可能な準最適モデル

$$\hat{\theta} = \arg\min_{\theta \in \Theta} L_D(\theta)$$

後は

- 最適化の問題
- 特定の（クラスの）損失関数に対して，最適化問題の解が真の解にどの位近いかの評価
  - 一致性（データ数無限大の時，バラツキがなくなり一定の値に収束する）
  - 不偏性（平均が真の解に一致する）

教師付き学習

識別関数，識別モデルと生成モデル

学習機械にデータの生成プロセスをどの程度組み込むかでも，判別・回帰問題は分類できる

- 生成モデル：あるクラス y が確率 $p(y)$ で選ばれ，そのクラスから特徴量 $\boldsymbol{x}$ が確率 $p(\boldsymbol{x}|y)$ で生成されたという考え方:

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})}$$

  例 naïve Bayes,
- 識別モデル：事後確率 $p(y|\boldsymbol{x})$ を，学習データ D から直接推定するアプローチ

  例 ロジスティック回帰，CRF(Conditional Random Field)
- 識別関数：事後確率のことは考えず，直接的に出力を当てる機械 f を設計するアプローチ

  例 SVM, ニューラルネットワーク

教師付き学習

識別関数，識別モデルと生成モデル

$$p(y, \boldsymbol{x}), p(y|\boldsymbol{x}), \text{ or } f : \boldsymbol{x} \mapsto y$$

- 生成モデル: クラス事後確率を求めるのに Bayes の定理を用い，結果として同時分布をモデル化しているのと等価

メリット　データの周辺分布を求めることができて，それを外れ値検出などにも利用できる．

デメリット　高次元の特徴量に対する条件付き密度を十分な精度で求めるのが困難

- 識別モデル：クラス事後確率を直接モデル化

メリット　生成モデルより計算資源を節約できる（識別に必要な事後確率を直接求めているから）

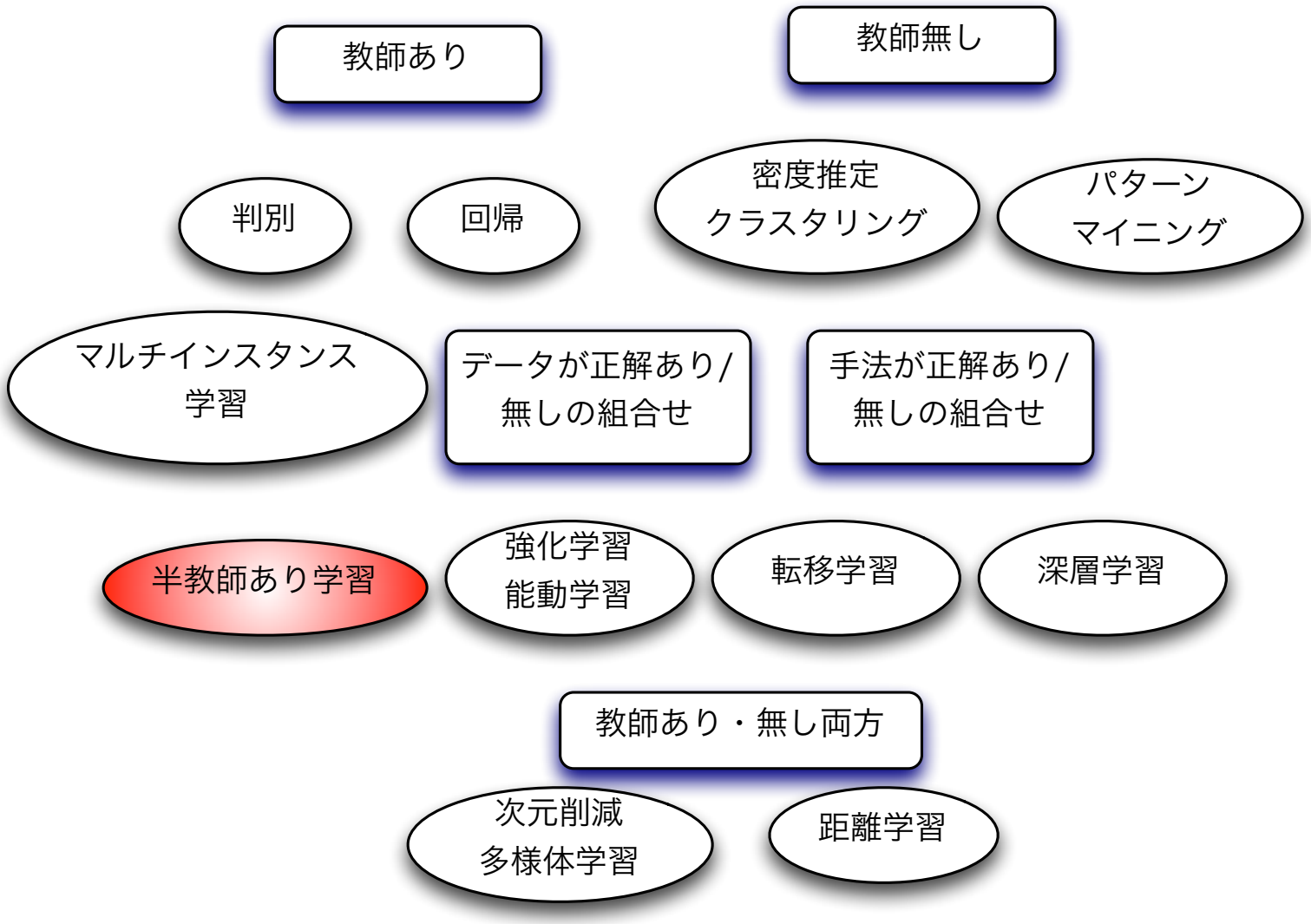
デメリット　事後確率以外の情報は得られない

- 識別関数：入力からクラスラベルに直接写像する関数を求める

メリット　分布の推定を伴わないのでデータ数が少しですむ

デメリット　出力に確率としての解釈が与えられない

機械学習:問題設定での分類



半教師付き学習

定義と例

定義 教師つきデータ集合 $D_l = \{(x_i, y_i)\}_{i=1}^{n_l}$ に加え，教師なしのデータ集合 $D_u = \{x_i\}_{i=1}^{n_u}$ もある，教師あり・なし混在データから学習することで，教師ありデータだけで学習した場合より予測精度の高い予測を実現するのが目標.

半教師付き学習

定義と例

例 1 機械の状態の監視：センサが常にデータは吐き出しているが，データに「正常」「異常」のラベルをつけるのは人間がやらなければいけないことが多い．大量の「教師なし」データと，少数の，手作業でラベルを付けた「教師付き」データ

例 2 特定の製品について書かれたブログエントリやツイートに対して,「肯定的/否定的」の分類を行ないたい．クローリングで集めたその製品に関する文書を人が読んで，クラスラベルを着けるのは大変なので，人手でラベルを着けるのは少しにしたいが，集めたデータは無駄にしたくない

半教師付き学習

難しさ

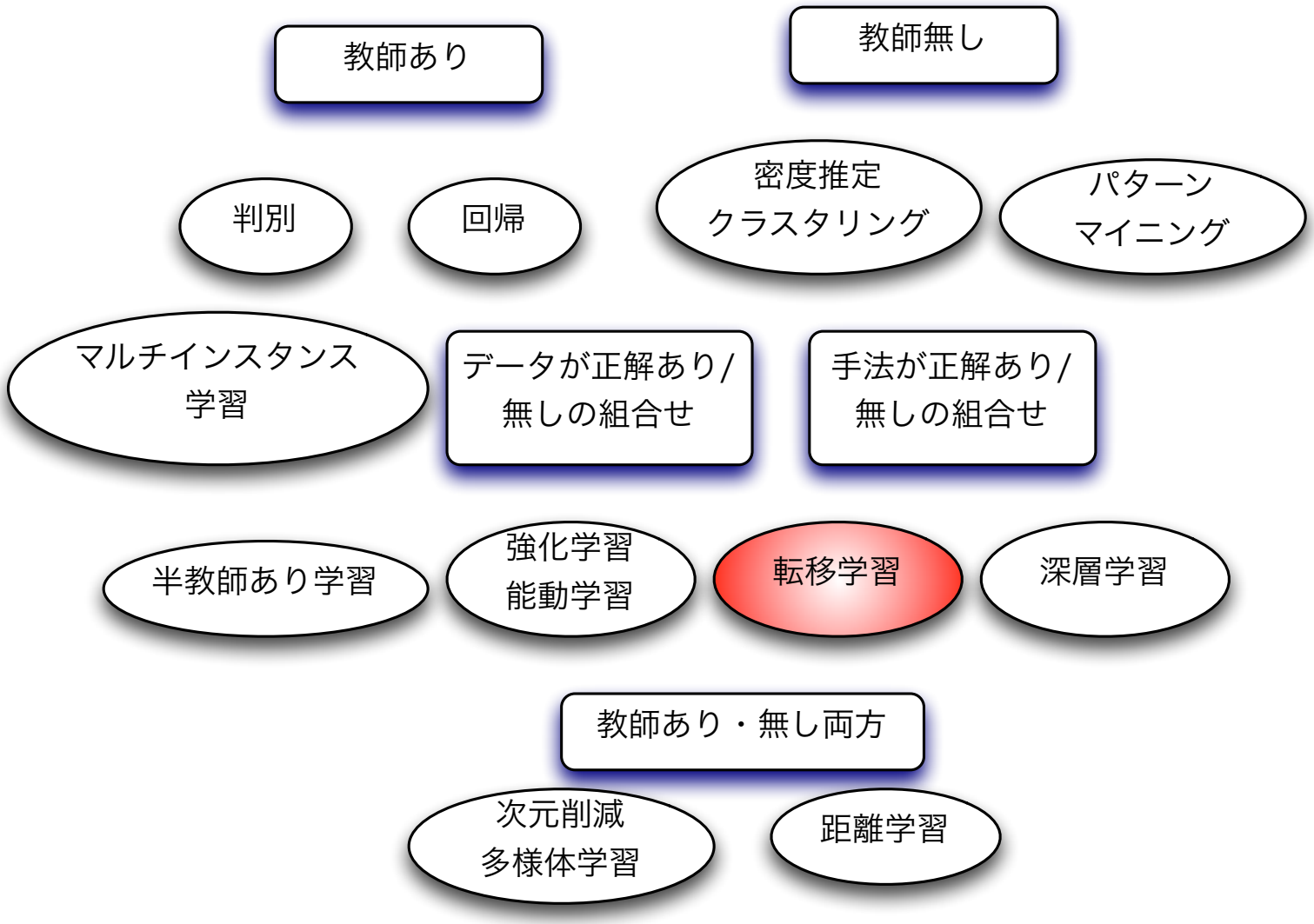
- 教師付きデータと教師無しデータをどこで使うか？どのパラメタをどちらのデータで学習すべきかの設計は個別の問題依存
- 後で教師付きデータが増えた時，教師無しデータが足を引っ張る可能性がある
- 教師無しデータは,「汚い」データの事が多く,「教師付き」「教師無し」サンプルの質の違いが機械を不安定にする可能性がある

半教師付き学習

定石

- 自己学習（Self-Training）：教師付きデータで作成した機械を用いて教師なしデータを識別し，確信度が高い結果を信じてそのデータを教師付きデータに取り込んで再学習を繰り返す．初期の誤りの影響をずっと受ける危険性が有る．
- YATSI(Yet Another Two-Stage Idea): 繰り返しによる誤差の増幅を避けるため，教師付きデータのみで一度だけ識別器を学習し，その識別器で全教師なしデータを識別してラベルと付与する．初めから教師があったデータを重視した重み付き k-NN で，新しい入力を識別する．
- 共学習（Co-Training）：ラベル付き入力ベクトル $\boldsymbol{x}$ の一部分で機械 A を，残りのベクトルで機械 B を学習．機械 A でラベルなしデータを識別し，高い確信度で識別されたデータを教師付きデータとみなして機械 B を学習．逆も行う．これを，機械 A, B それぞれの精度が安定するまで繰り返す．

機械学習:問題設定での分類



転移学習

もったいないの精神

- 過去のシステムで収集したデータ，あるいは過去に学習した学習器は，システムリプレースによって「ゴミ」になるのか？
  - 夏のみの環境・土壌等のデータを用いて構築した「虫害発生予測システム」は，他の季節における予測には使えないのか？(一年分のデータをまたなければいけないか？)
  - ある品種のキュウリ農園のために構築した「虫害発生予測システム」は，新品種のキュウリ農園には適用できないか？
  - キュウリ農園のために構築した「虫害発生予測システム」は，トマト農園のための「虫害発生予測システム」には使えないか？
  - キュウリ農園のための虫害発生システムは，キュウリ（or トマト）農園のための「病気発生予測システム」に使えないか？

転移学習

定義と例

定義 ある問題を効果的かつ効率的に解くために，別の関連した問題のデータや学習結果を再利用する方法の総称．２つの問題に共通したメタ知識を抽出・利用する．

例 1 手話の単語認識を，画像から行ないたい．実際の手話の画像に，正解がついているデータはあまり多くない．一方，手話の手形状を CG で生成するプログラムにより，CG 画像としてのラベル付きデータはいくらでも生成できる．CG 画像で学習した結果を，実写画像を入力した時の手話単語の判別に利用できるか？

例 2 英語の文章を用いて学習した高性能なスパムフィルタの性能を，新しく学習する日本語メール向けのフィルタに反映させることが出来るか？

転移学習

特徴

- 半教師付き学習では「少量の教師ありデータと大量の教師なしデータ」という，入力 $\boldsymbol{x}$ としては同じ種類のものを考えていた．
- 転移学習では，本来学習したい入力データが属する「ドメイン」とは異なる「ドメイン」に属するサンプル（学習例，あるいは学習済みモデル）がある状況を考える

色々な呼び方：帰納転移 (inductive transfer)，ドメイン適応 (domain adaptation)，マルチタスク学習 (multi-task learning)，knowledge transfer，learning to learn，lifetime learning，共変量シフト (covariate shift)，標本選択バイアス (sample selection bias)

転移学習

難しさ

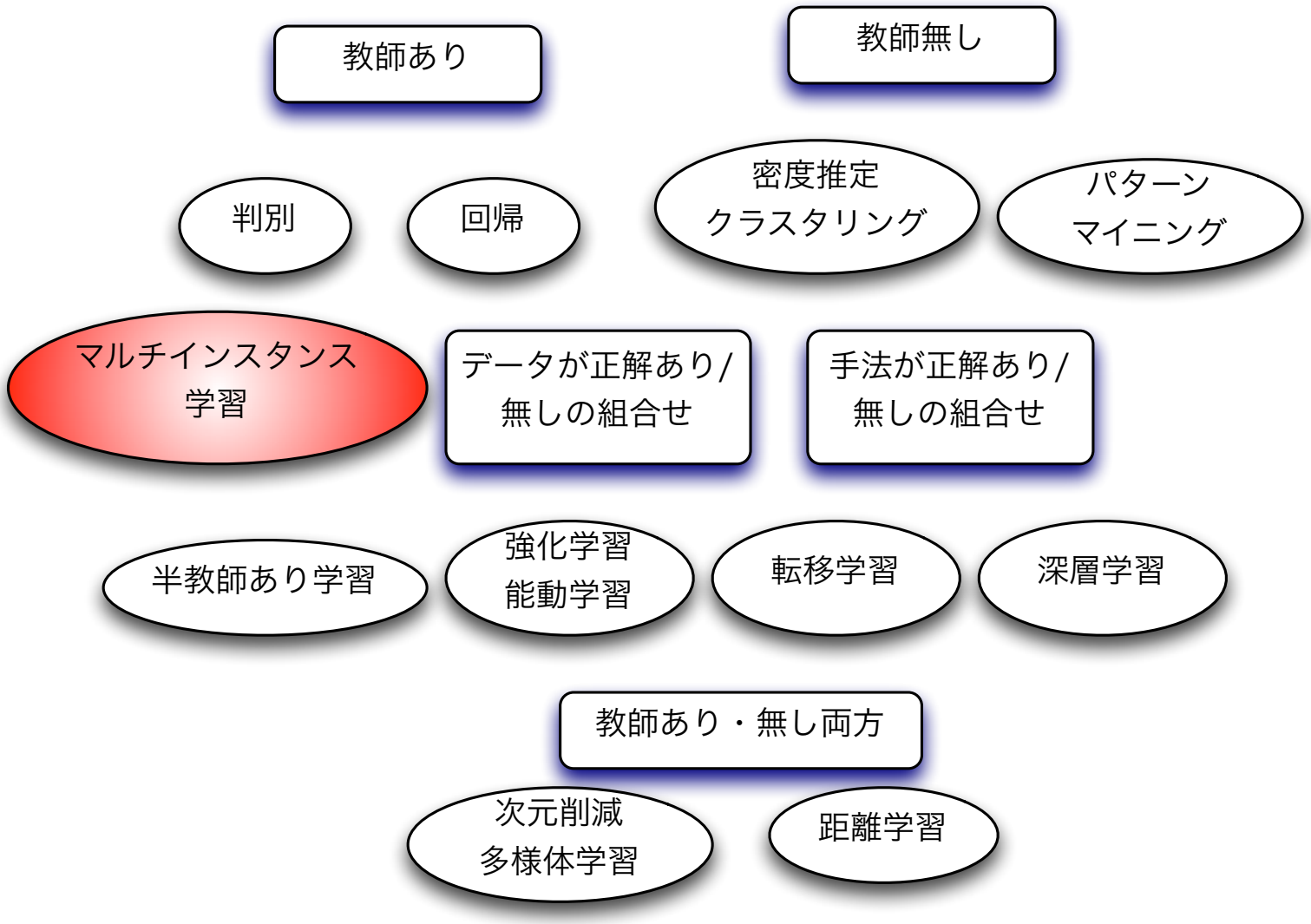
- メタ知識の抽出・利用方法の統一した方法論の欠如
- 負の転移の存在

転移学習

分類

1. 帰納的転移学習：両方のドメインで入出力の組が与えられている．同時確率が一致するように $(p(X^S, Y^S) = p(X^T, Y^T))$，サンプルに重みをつけるなど
2. トランスダクティブ転移学習：目標ドメインにはラベルが無い．$p(Y^S|X^S) = p(Y^T|X^T)$ を仮定して，$p(X^S)$ と $p(X^T)$ が一致するようにサンプルに重みをつけるなど
3. 自己教示学習：目標ドメインにはラベルが有る．$p(X^S) = p(X^T)$ を仮定して，データの事前分布や，データを効率よく表現する表現方法を元ドメインのデータを用いて獲得する
4. 教師無し転移学習：便宜上分類はしたものの，この設定がうまく行った例は？

機械学習:問題設定での分類



マルチインスタンス学習

この部屋の中に犯人がいる．誰かはわからない．

- ひと塊のデータが与えられ，その塊の中に少なくとも一つは「当たり」がある，あるいは一つも「当たり」は含まれないという粒度でしか教示されないことがある
- 実応用として重要な問題が多い，現実に則した問題設定
  - 薬剤活性分析
    化合物の中で，どの要素が薬剤の活性（あるいは副作用）に寄与しているかわからない
  - 画像のカテゴリ分類
    一つの画像に複数のオブジェクトが写っている場合の，画像の分類
  - 車のセンサデータからの不具合検出・識別
    車が不具合を起こしたことはわかるが，どの時点で不具合が起きたかはわからない状況での，不具合発生時点の同定

マルチインスタンス学習

定式化

定義 複数のインスタンスのまとまりをバッグ（bag）$B_i = \{\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_{n_i}^{(i)}\}$ と呼び，各バッグに対してひとつのラベルが割り当てられている設定

$$D = \{B_i, y_i\}_{i=1}^n.$$

バッグの中のインスタンスの数は一般に異なってもよい

マルチインスタンス学習

定式化

- バッグ: $\{X_1, X_2, \ldots, X_n\}$. $X_i = \{x_{i1}, x_{i2}, \ldots, x_{im_i}\}, x_{ij} \in \mathcal{X}$
- バッグラベル：$\{y_1, y_2, \ldots, y_n\}, y \in \{0, 1\}$
- クラスラベル $y_{ij} \in \{0, 1\}$

$$y_i = \max_j \{y_{ij}\} = \begin{cases} 1 & \exists j \quad \text{s.t.} \quad y_{ij} = 1, \\ 0 & \quad otherwise \end{cases}$$

MIL の目的は，インスタンスの判別機 $h(X) : \mathcal{X} \to \mathcal{Y}$ か，バッグの判別機 $H(X) : \mathcal{X}^m \to \mathcal{Y}$ を構成すること.

マルチインスタンス学習

定石:最尤法に帰着する

サンプル x_i と判別器 h に対して，このサンプルのクラス事後確率を $p_i = \Pr(y_i = 1|x_i, h)$ とする．データセット $\mathcal{D} = \{(x_i, y_i)\}$ の対数尤度を

$$L(h) = \sum_{i|y_i=1} \ln(p_i) + \sum_{i|y_i=0} \ln(1 - p_i) \tag{1}$$

で定義する．MIL においてはサンプルがバッグに対応するが，サイズの異なるバッグの判別器を直接構成するのは難しいので，バッグの中身の各インスタンスの判別器をベースに尤度関数を計算する．

マルチインスタンス学習

定石:最尤法に帰着する

改めて，$p_i = \Pr(y_i = 1|X_i, h)$ をバッグのクラス事後確率，$p_{ij} = \Pr(y_{ij} = 1|x_{ij}, h)$ をインスタンスのクラス事後確率とする．インスタンスのクラス事後確率から，バッグのクラス事後確率は

$$p_i = \max_j \{p_{ij}\} \tag{2}$$

で計算できる．インスタンスに対する適当な判別器から p_{ij} を計算し，上式からバッグの事後確率を計算することで，判別器を固定した時の尤度が計算できる．原理的には尤度が最大になるように判別器を最適化すればよいが，max 関数の非凸性から，最尤推定が（普通なら凸問題を与えるような判別器を採用しても）非凸な問題となるのが難しいところ．

Index

1 導入

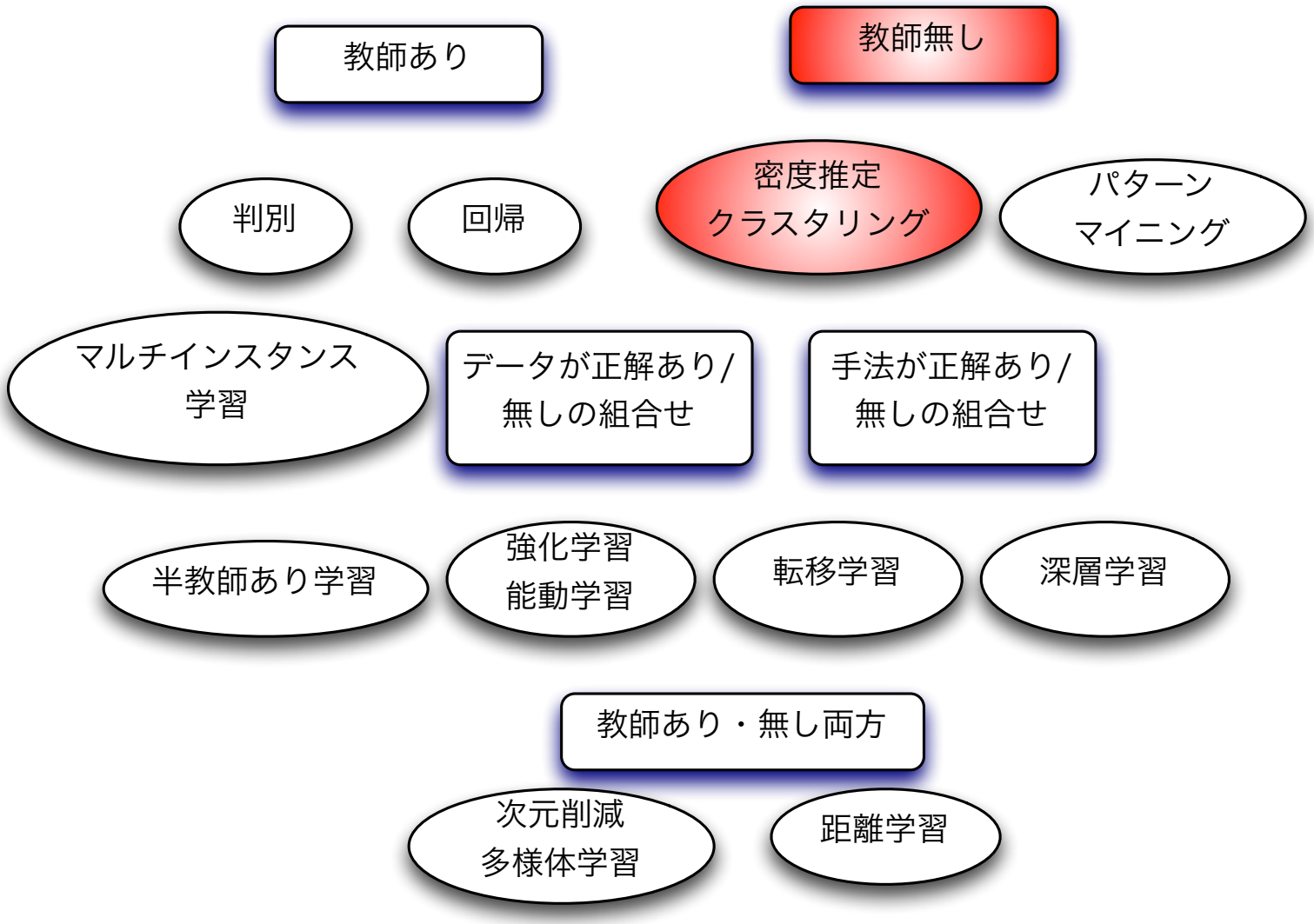
2 種々の問題設定
- 教師付き学習

3 教師無し学習

4 その他中間的方法

5 終わりに

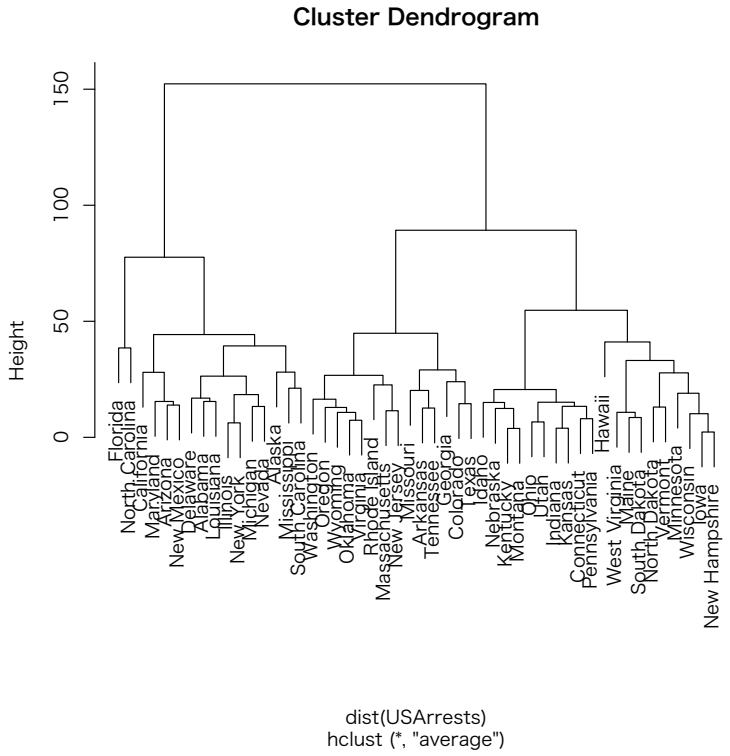
機械学習:問題設定での分類

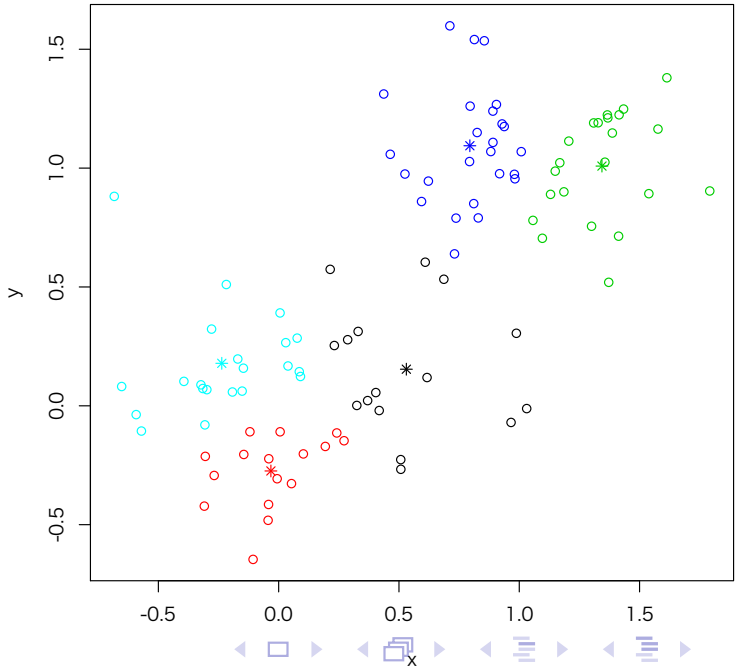


クラスタリング

定義

定義 分類対象の集合を，内的結合と外的分離が達成されるような部分集合に分割すること





クラスタリング

特徴・難しさ

- 厳密に最適化しようとすると組合せ最適化問題を扱う必要がある
- データ間の類似度・距離尺度で結果が大きく異なる
  - distance metric learning for clustering
- クラスタ数は幾つが良いか？
  - AIC, BIC, X-means, Gap-Statistics, Dip-means, etc.

異常・変化検出

定義と例

異常・外れ値検出 所与のデータの中で，大多数のデータと異なる性質のインスタンスを見つける

変化検出 時系列データから，時系列の性質が変化した時点を見つける

問題設定別に多くのアプローチがある

バッチ的 one-class SVM, 局所異常因子 (Local Outlier Factor)

時系列が対象 CUMSUM, Change Finder, 相関異常検知

異常・変化検出

定式化

- 正常データのみが与えられた学習問題と捉える
- 正常時に対応する統計モデルの構築と，検定の枠組みに落としこむのが「正当」な方法？
  - （逐次）尤度比検定 (CUMSMU)

$$\log \frac{p(x_t; \theta^f)}{p(x_t; \theta^p)}$$

- 機械学習でもわりとアドホックなアプローチが多く，基本的には
  1. 外れ値・異常値スコアの算出
  2. 閾値との比較

  という手順

異常・変化検出

難しさ

- 非定常性，共変量シフト
- 性能評価の難しさ
- 異常・正常の閾値の設定

Index

1 導入

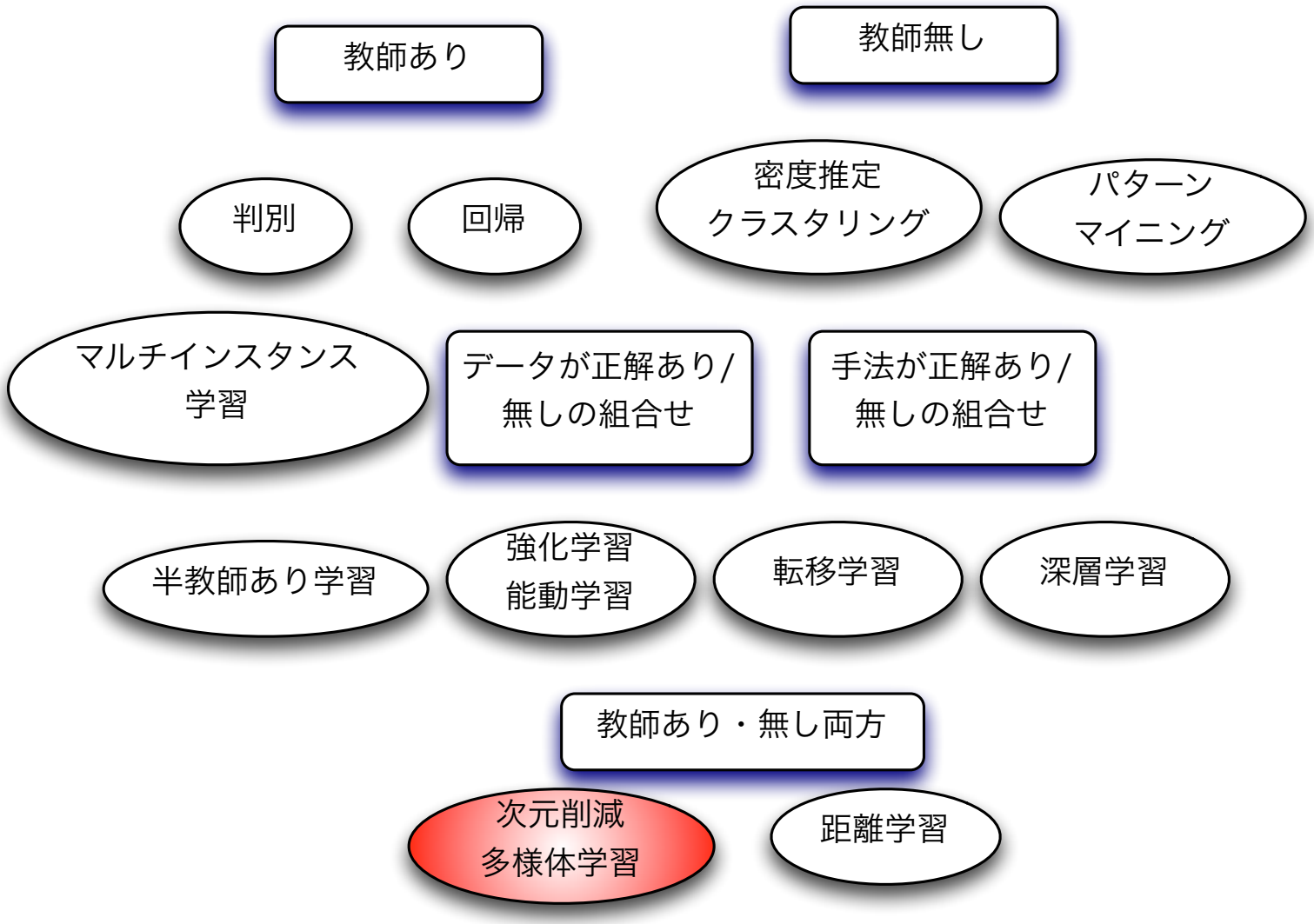
2 種々の問題設定
- 教師付き学習

3 教師無し学習

4 その他中間的方法

5 終わりに

機械学習:問題設定での分類



次元削減・多様体学習

次元の呪いの回避・可視化

- 高次元データも，実質的には数次元のパラメタで記述できることが多い
- 高次元データは次元の呪いが生じる．種々の非直観的な現象
  - Marcenko-Pastur 則
  - 球面集中現象
- 可視化のためには，低次元に落とす必要が生じる

次元削減・多様体学習

- 教師無し: (kernel) 主成分分析，多次元尺度構成法，SOM，ラプラシアン固有マップ，diffusion map, ISOMAP, LLE(Locally Linear Embedding)，LOE(Local Ordinal Embedding)
- 教師あり：フィッシャー線形判別分析，局所フィッシャー判別分析，層別逆回帰
- 組合せ：PLS(Partial Least Squares)，一段階主成分回帰

次元削減・多様体学習

典型的な定式化

もとの距離 $d_o(x_i, x_j)$ を保ちつつ，低次元に．その際に補助情報やグラフ構造をどう入れるか

$$\min_{f:x\to y} \sum_{i,j} \|d_o(x_i, x_j) - d_t(y_i, y_j)\| \tag{3}$$

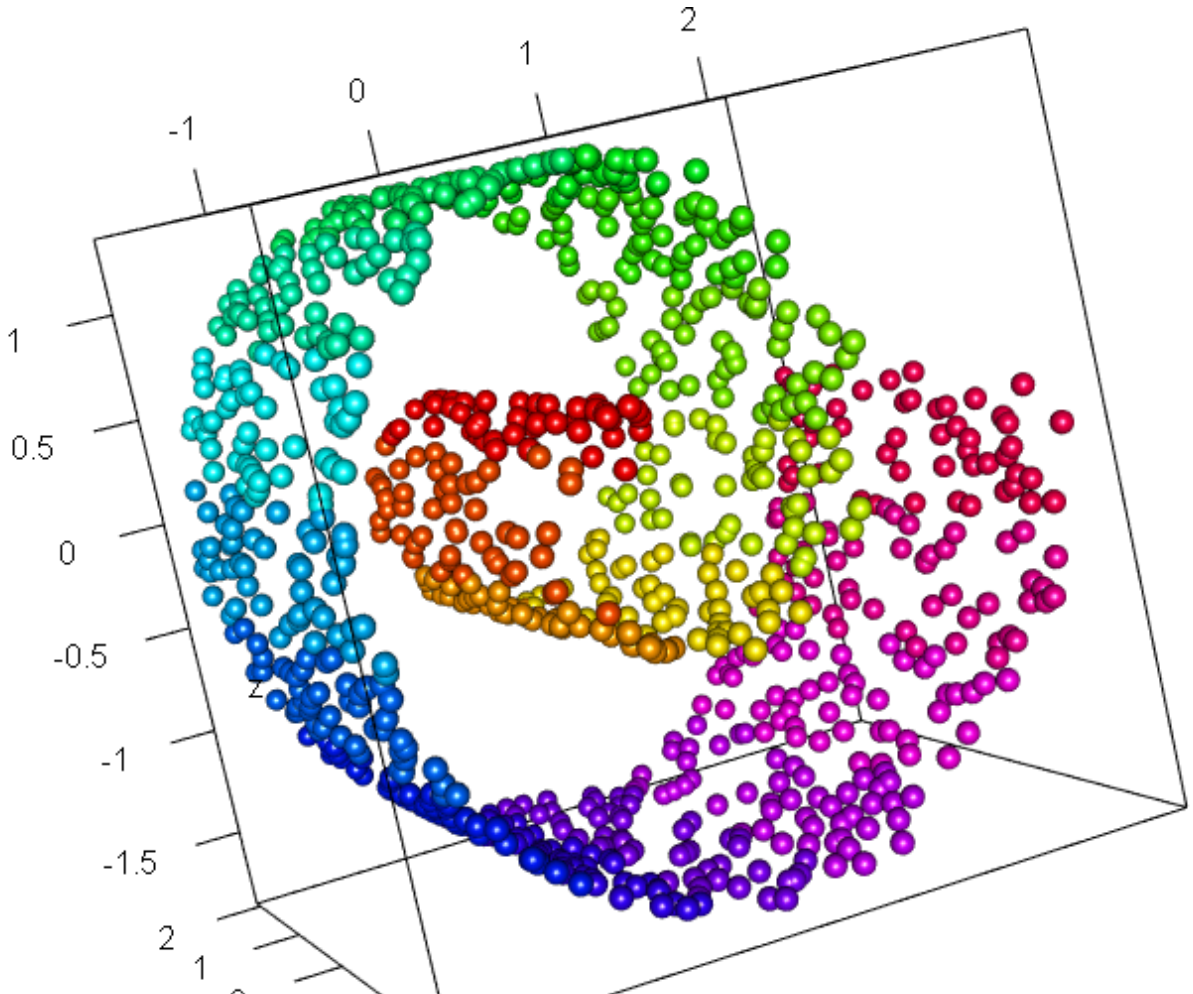
$$+\lambda \sum_{(i,j)\in\mathcal{S}} d_t(y_i, y_j) \tag{4}$$

$$+\eta \sum_{(i,j)\in\mathcal{D}} 1/d_t(y_i, y_j) \tag{5}$$

次元削減・多様体学習

例

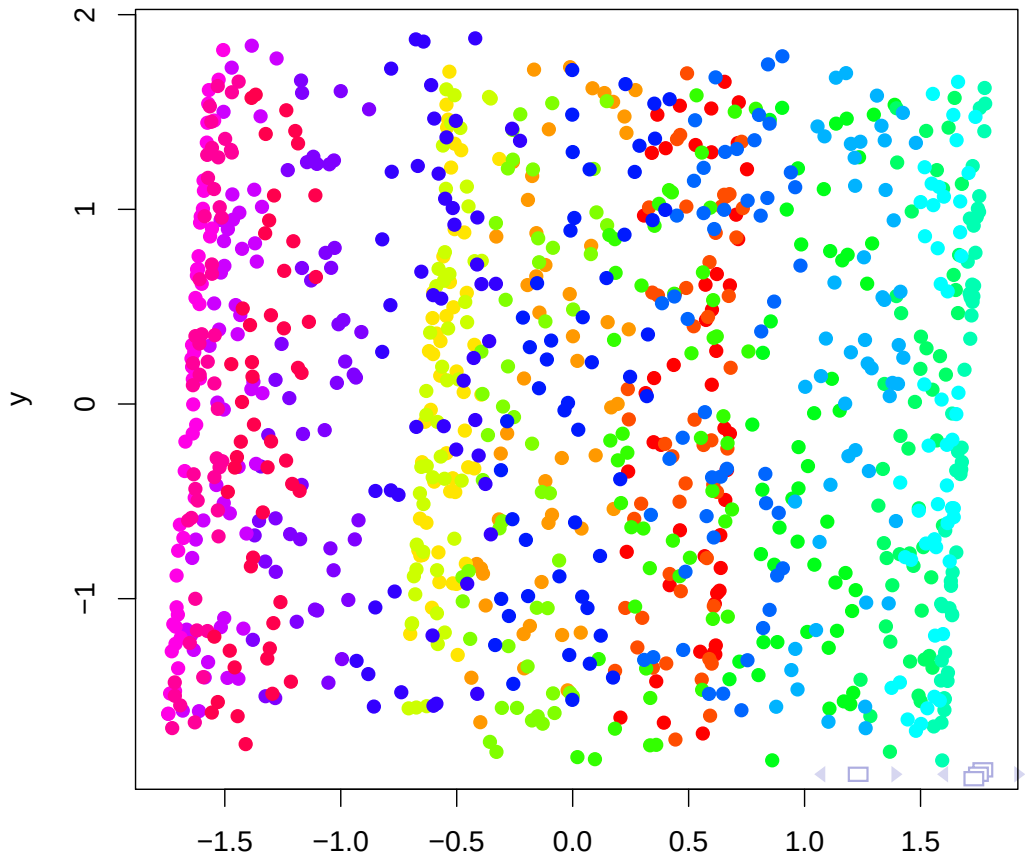
Swiss Roll データ



次元削減・多様体学習

例

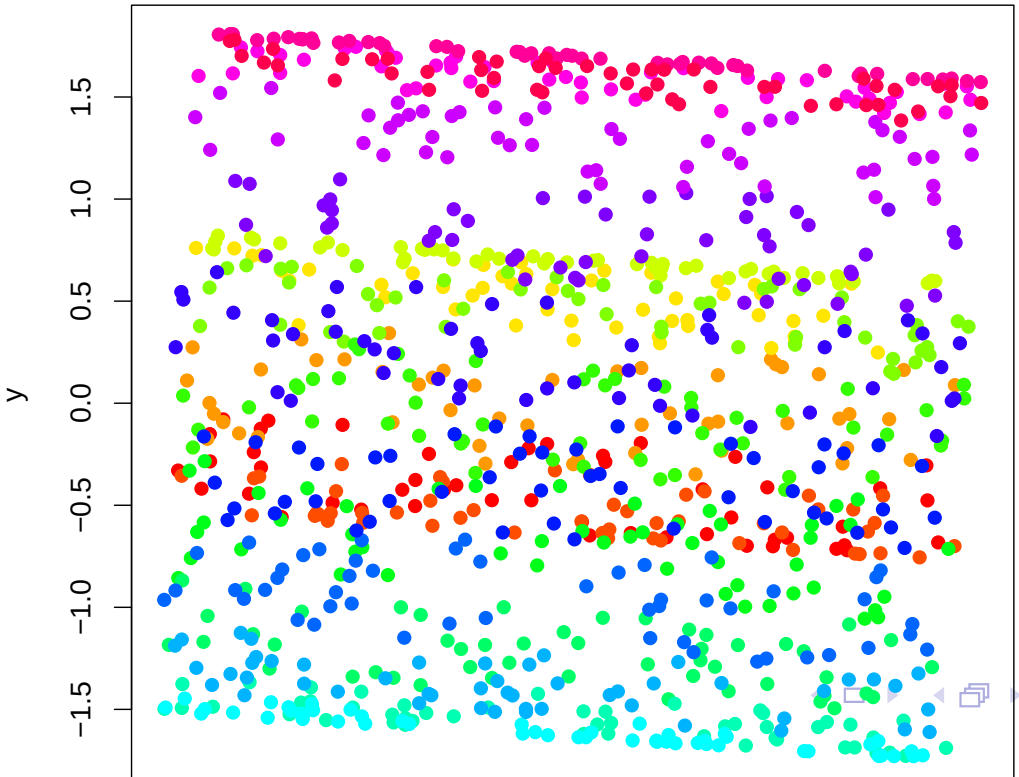
主成分分析:まず試す方法として



次元削減・多様体学習

例

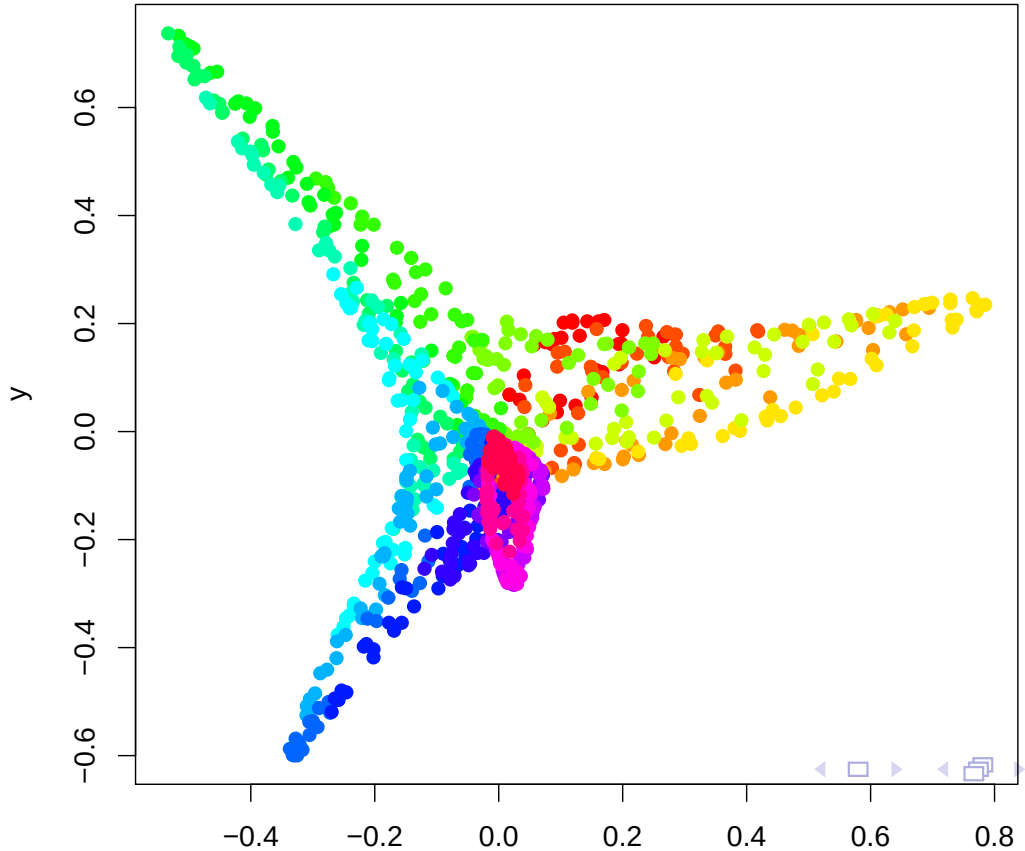
古典的多次元尺度構成法:距離を保存する埋込み．距離がユークリッドなら PCA と同じ



次元削減・多様体学習

例

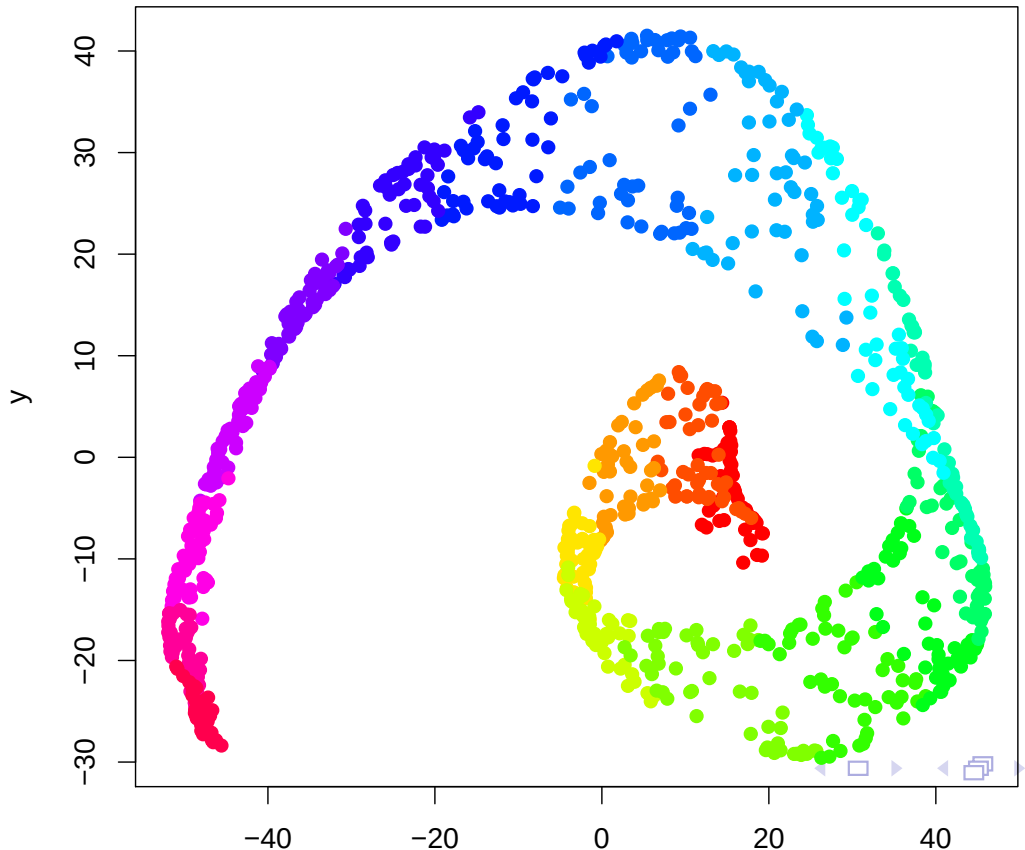
カーネル主成分分析:PCA の非線形化



次元削減・多様体学習

例

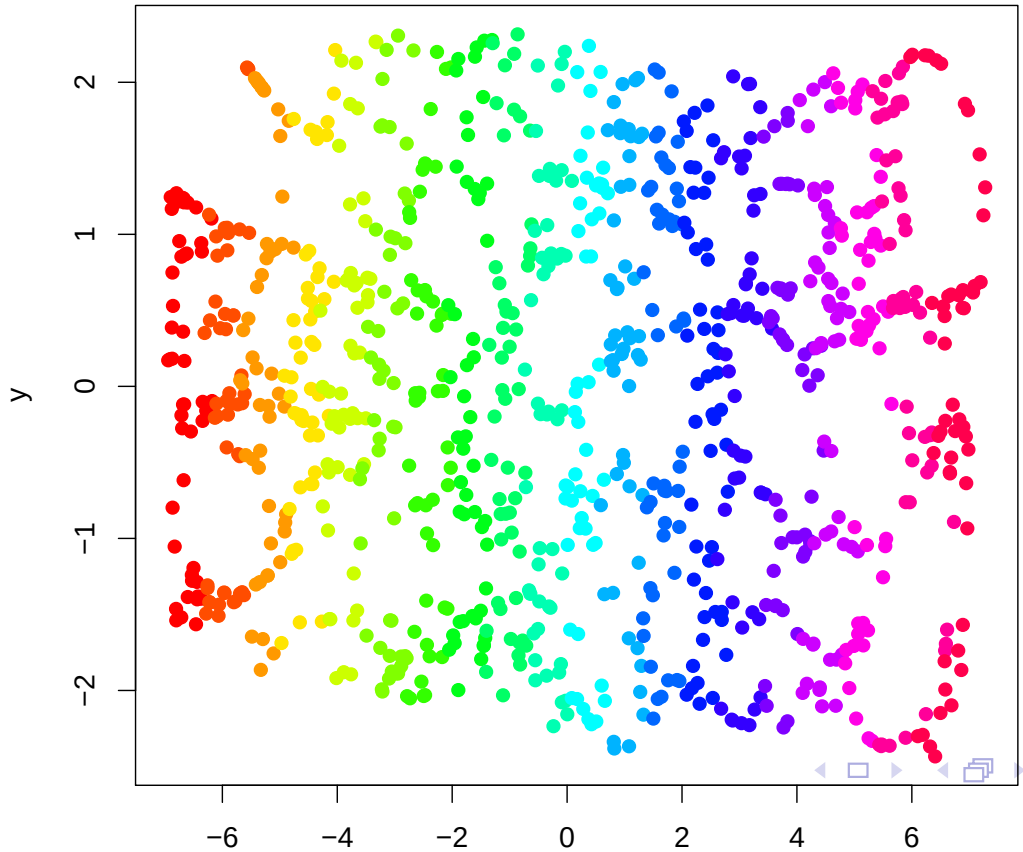
diffusion map:グラフ上の物理モデルに基づく次元削減



次元削減・多様体学習

例

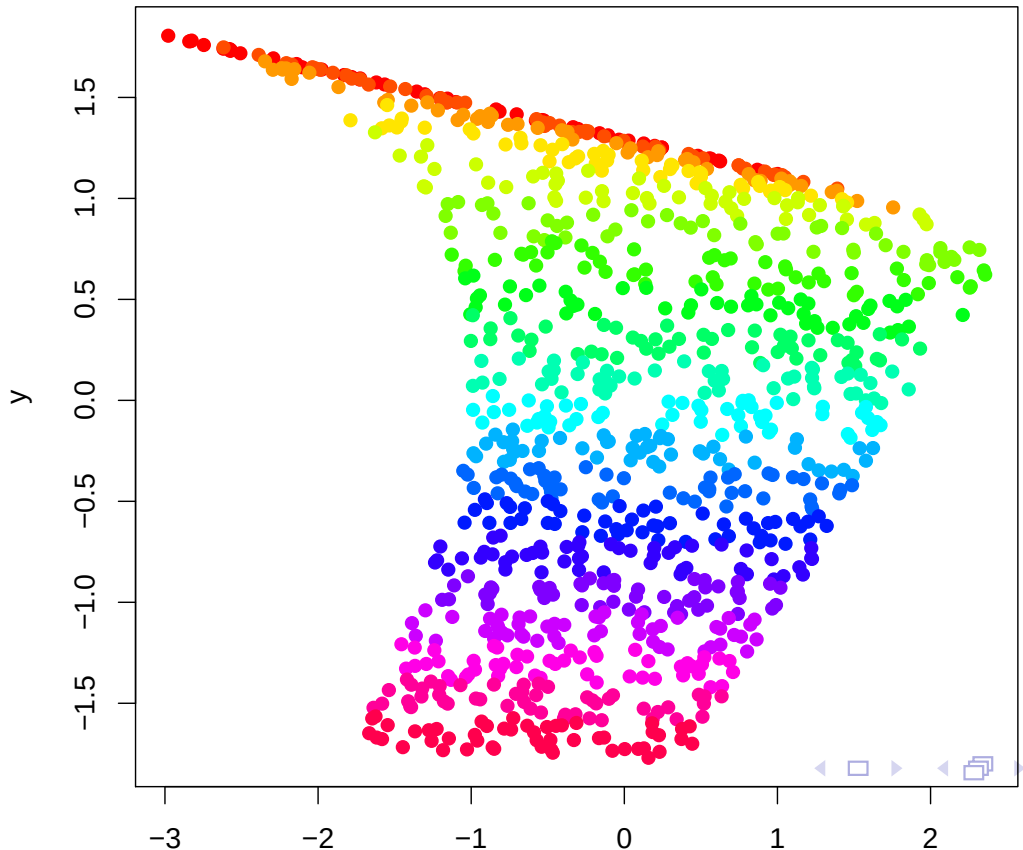
ISOMAP:多様体上の距離に基づく次元削減



次元削減・多様体学習

例

locally linear embedding:ISOMAP の線形近似

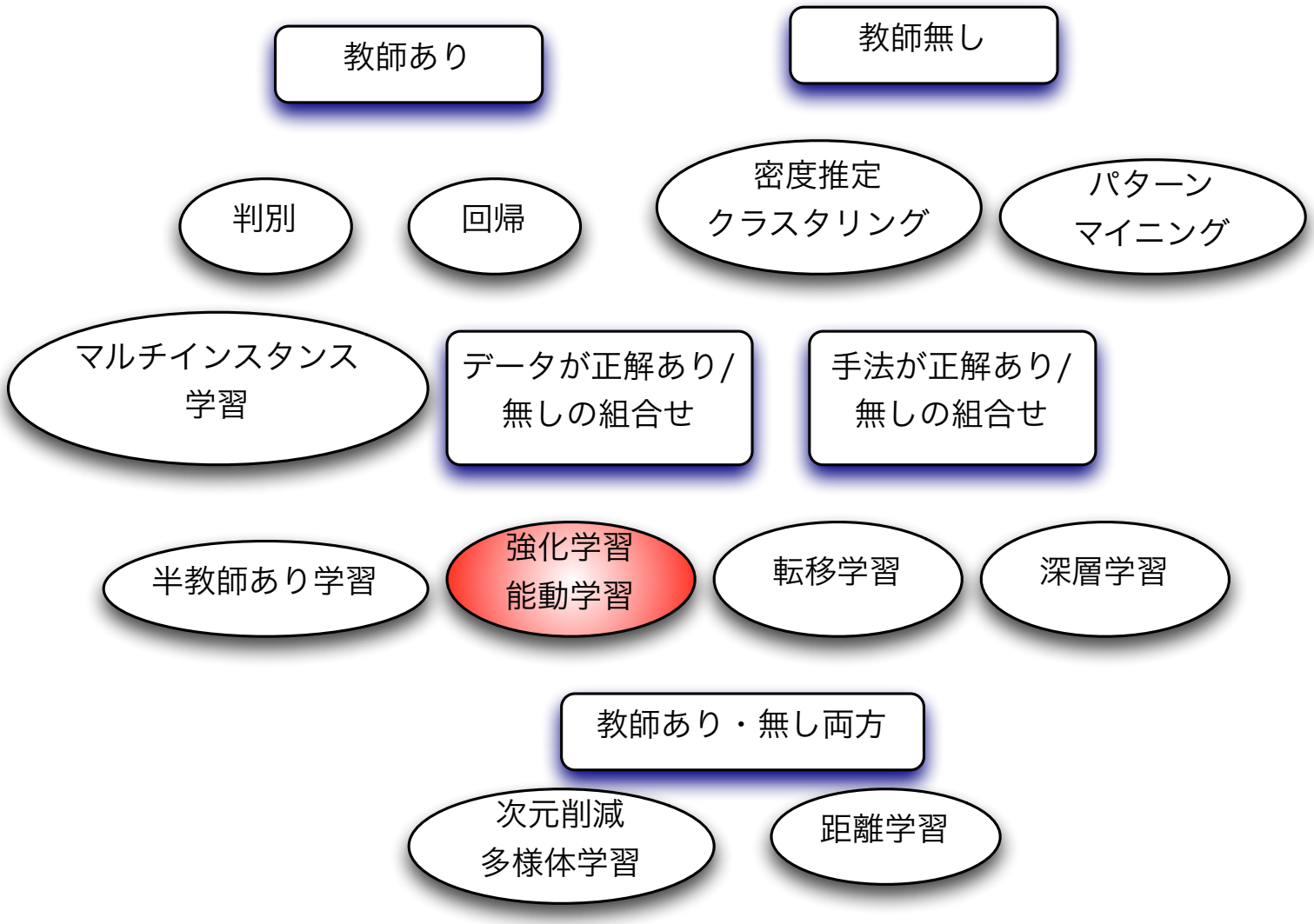


次元削減・多様体学習

難しさ

- 何次元にするか？可視化が目的なら 2,3 次元だが，一般に intrinsic dimension の推定は難しく，一つの研究分野
- データ点だけから次元削減する場合，初めは恣意的な距離に頼ることになる（ユークリッド空間における距離など）
  - 近傍の定義

機械学習:問題設定での分類



強化学習

定義と例

「報酬を得るために，環境に対して何らかの行為を行う意思決定エージェントの学習」

- 自律システム構築の基盤（ロボットの歩行動作獲得等）
- 知能を獲得の基盤（コンピュータチェス将棋・囲碁）
- データ収集・行動計画のコスト最適化（インターネット広告の最適化，A/B テスト）

強化学習

特徴

- 状態遷移を繰り返した後，遅れて報酬がエージェントに与えられる
  入力に対する出力を明示的に示す教師が存在しない代わりに，一連の行動終了後にその結果の良し悪しの評価値が与えられる
- エージェントは予め環境に関する知識を持たない（持っていても限られている）
- 環境の状態遷移は確率的
- 報酬の与えられ方も確率的
- 試行錯誤を通して，エージェントが置かれた環境における最適な行動を学習
- まだデータがない時に，どのようにデータを集めるかという方針も含めて決定する (能動学習とも関連)

強化学習

難しさ

HMM で表現出来そうだが …

- 遷移行列・報酬に関する情報がない
- 情報を更新しながら行動を繰り返す（試行錯誤）
  - 観測できるのは行動とその結果のみ
  - 問題は，この方法で十分な情報が収集できるか（世界の探索と，方策の学習を両方行わないといけない）

強化学習

成功例

- 強化学習による滞納税金採りたて最適化 [Abe+,KDD2010]
- 強化学習+深層学習で，ビデオゲームのプレー戦略の学習 [Riedmiller+,NIPS2013, Nature2015]
- 逆強化学習によるラジコンヘリ自動操縦 [Abbeel+,NIPS2007] （エキスパートの演示からエキスパートのもつ未知の報酬関数を推定する設定．徒弟学習とも）

終わりに

No-Free lunch theorem

- 機械学習で研究されている問題設定をいくつか紹介した
- 他の工学分野と同様，定式化が最重要.
- 「適切な問題設定」と「適切なアルゴリズム」を選ぶことが必要.
  - 定式化の幅を広げる一助となることを期待
  - アルゴリズムについては問題設定のキーワードで調べて下さい