

機械学習講座 概要版

講師：荒木雅弘
(京都工芸繊維大学)



自己紹介

荒木雅弘

- 京都工芸繊維大学 情報工学・人間科学系 准教授
- 専門：音声対話処理
- 著書



本講座の目的

- 機械学習技術の全体像を広く・浅く知ること
で、自社製品・サービスが機械学習技術によっ
てどのように変革できるか、またそのためにど
のような技術者・開発者を育成するべきかが見
通せるようになることを目的とします。

本日の予定

- 9:30-10:30 機械学習の概要、識別 1
- 10:45-11:45 識別 2、機械学習ライブラリの紹介
(昼休憩)
- 13:00-14:00 回帰、教師なし学習
- 14:15-15:15 高度な機械学習、深層学習
- 15:30-16:30 機械学習エンジニア育成のために

Section 1

- 機械学習の概要（1,2章）
- 識別1（3章）

1. はじめに

内容

1.1 人工知能・機械学習・深層学習

何が違うか、何ができるか

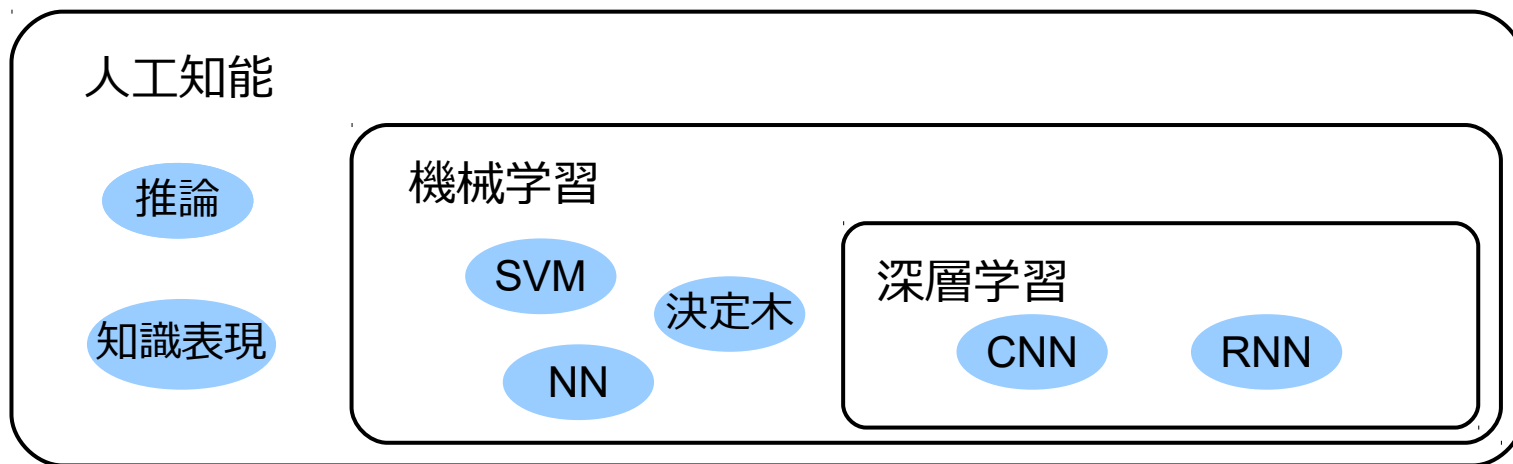
1.2 機械学習とは何か

機械学習の全体像

1.3 機械学習の分類

教師あり学習、教師なし学習、中間的手法

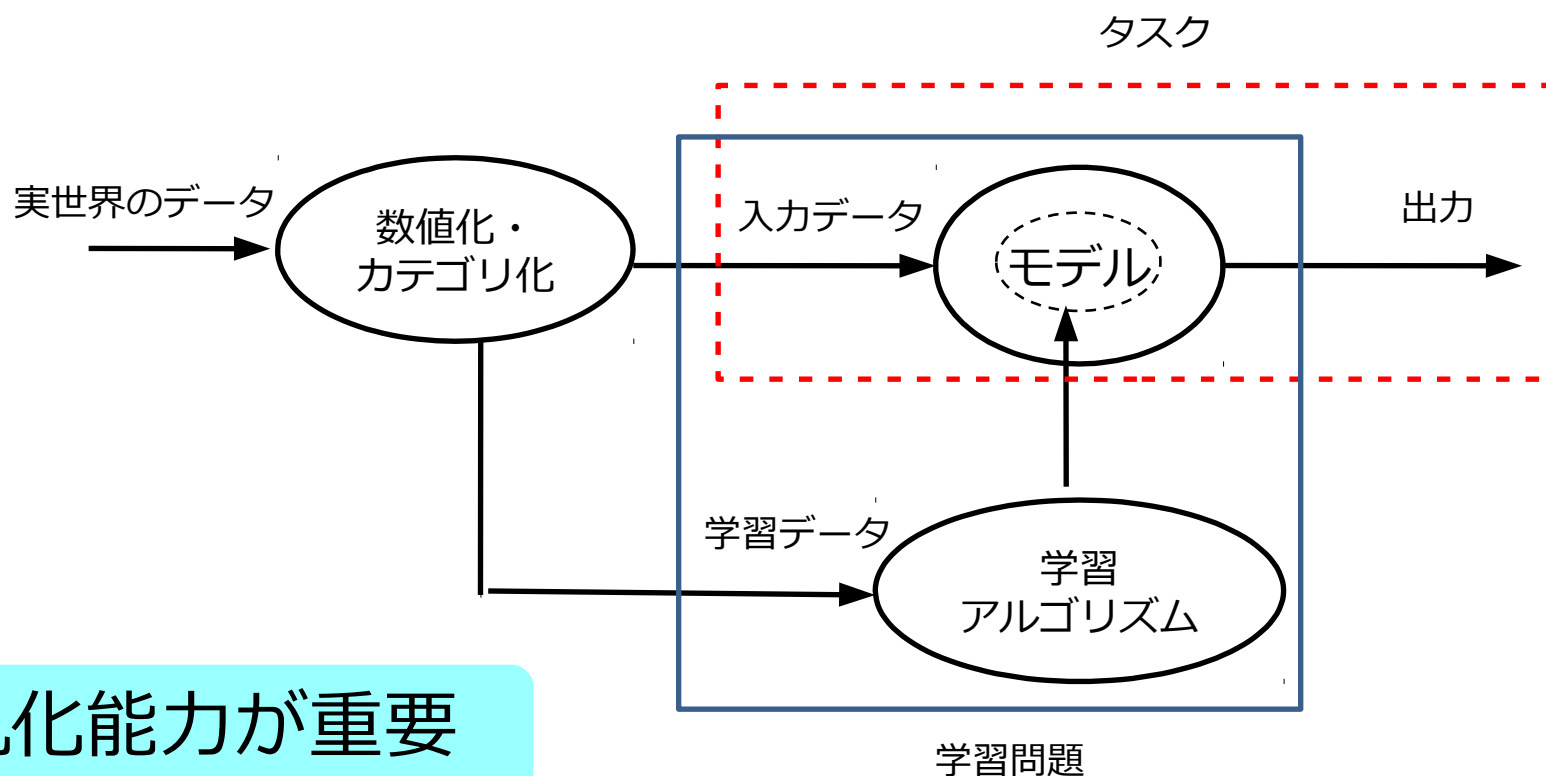
1.1 人工知能・機械学習・深層学習



- 人工知能とは
 - 現在、人が行っている知的な判断を代わりに行う技術
 - 技術が普及すると人工知能とはみなされなくなる
 - 例) 文字認識、乗換案内
- 探索・知識表現・推論・機械学習などを含む

1.1 人工知能・機械学習・深層学習

- 機械学習とは
 - 機械学習は、適切に**タスク**を遂行する適切な**モデル**を、適切な**特徴**から構築すること [Flach 2012]

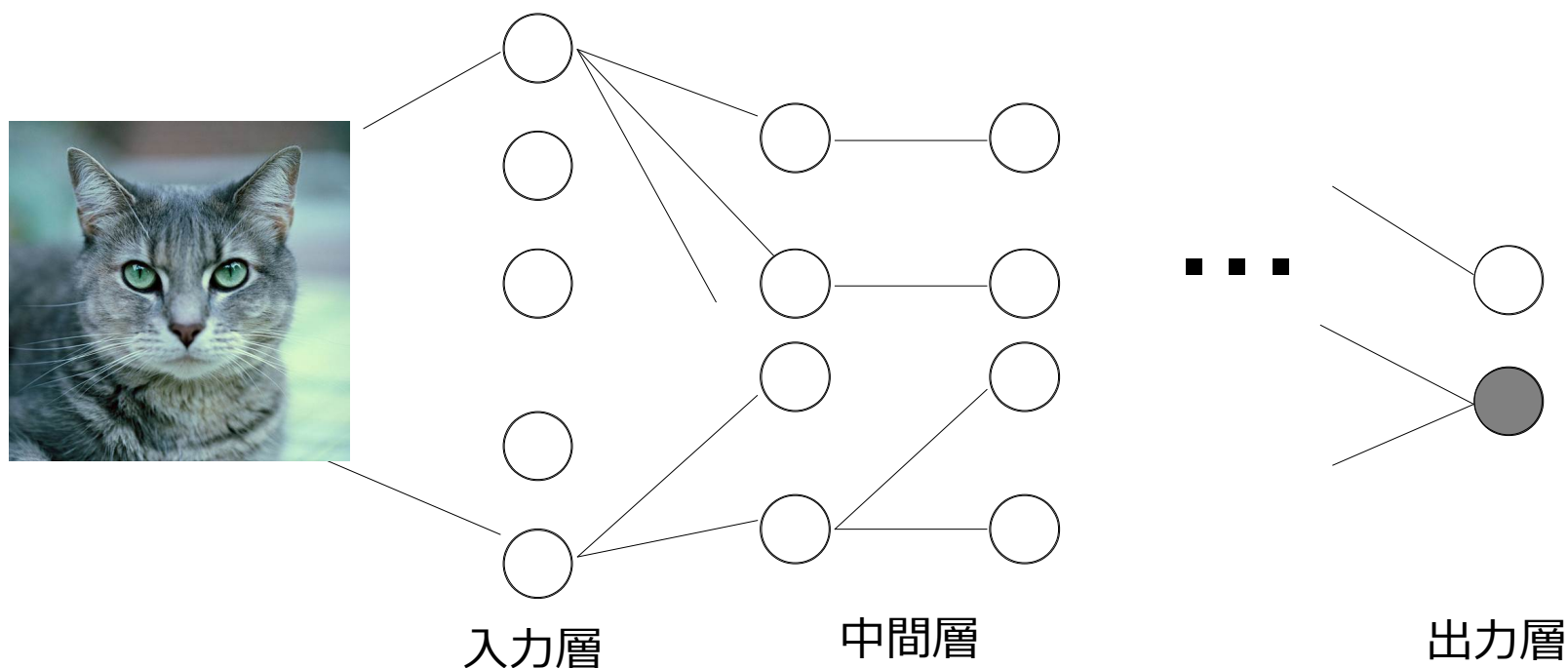


汎化能力が重要

1.1 人工知能・機械学習・深層学習

- 深層学習とは

- 多層に非線形変換を重ねる手法による機械学習
 - 一般的には中間層を多くもつニューラルネットワーク
 - 特徴抽出処理も学習対象とすることができる点が特長
 - 学習には大量のデータが必要



1.2 機械学習とは何か

- 機械学習の位置づけ



観測データ

(134.1, 34.6, 12.9)

(135.5, 30.1, 43.0)

...

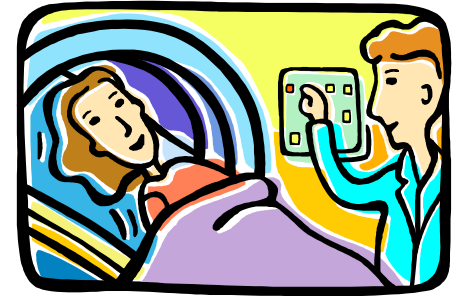


売り上げデータ

(パン、ハム)

(パン、牛乳、バター)

...



診療データ

(男, 28, 178, 75, yes)

(女, 68, 165, 44, no)

...

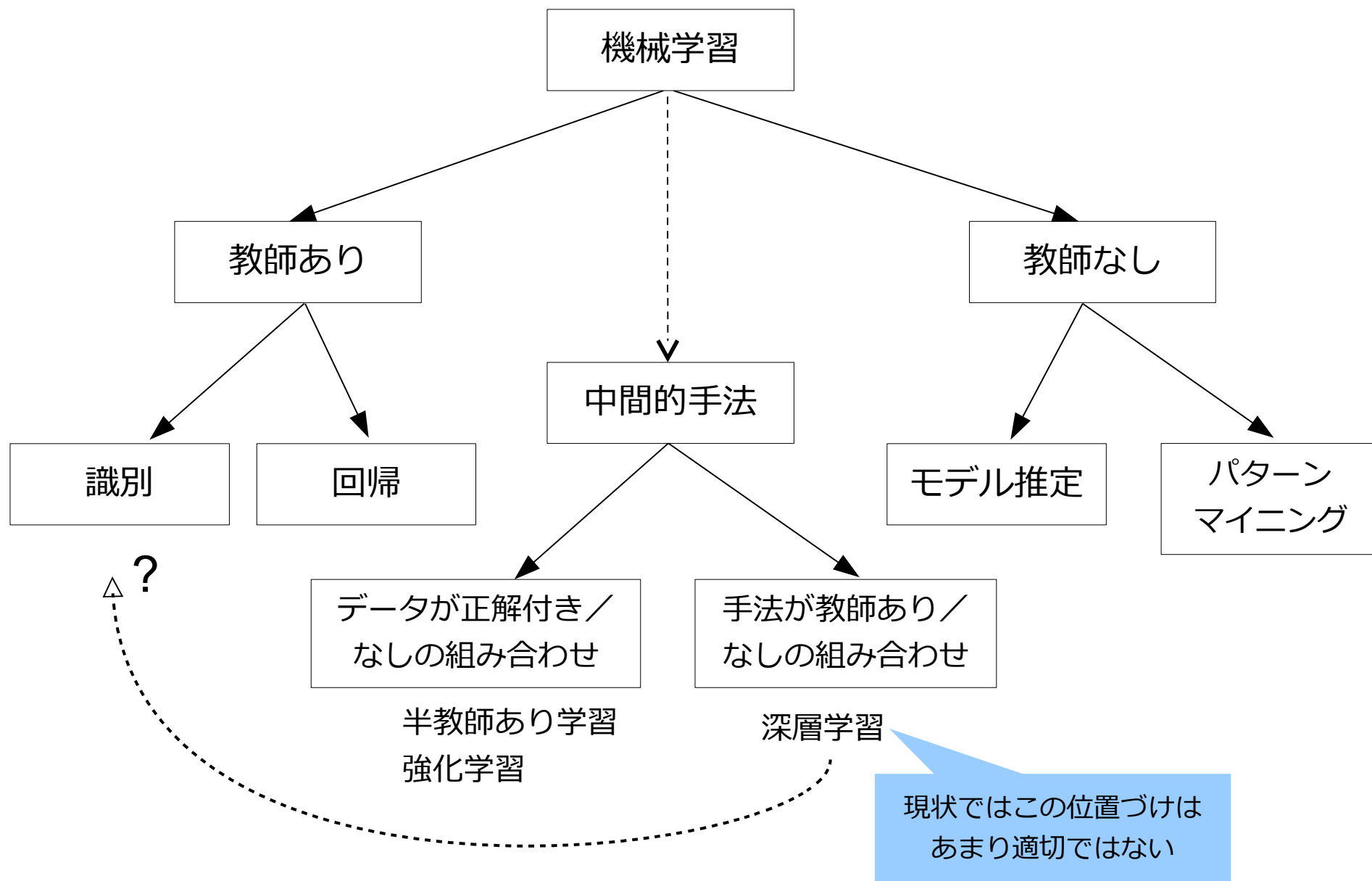
機械学習

規則

関数

分類

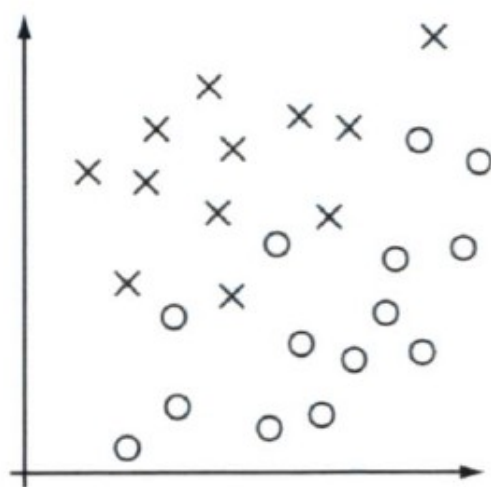
1.3 機械学習の分類



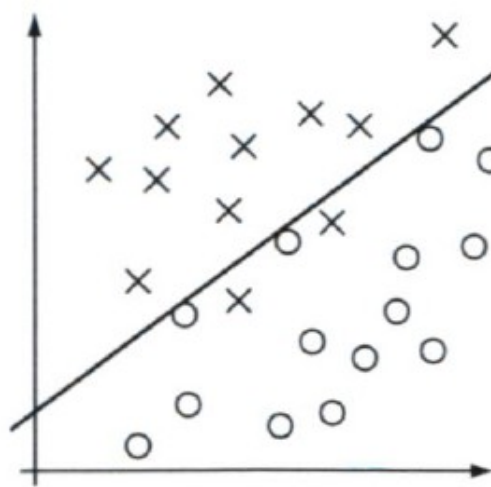
1.3.1 教師あり学習

- 識別

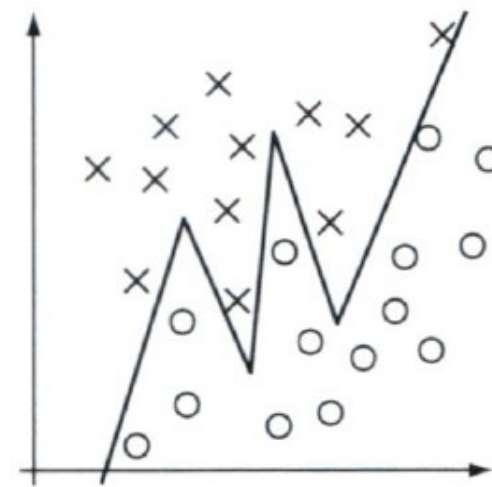
- 正解情報がカテゴリデータ
- 汎化誤差が最小となるような特徴空間上の識別面を求める



(a) 入力が 2 次元数値ベクトルの識別問題



(b) 学習結果 1

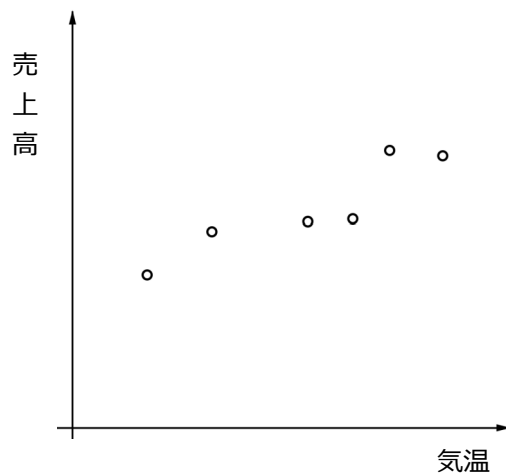


(c) 学習結果 2

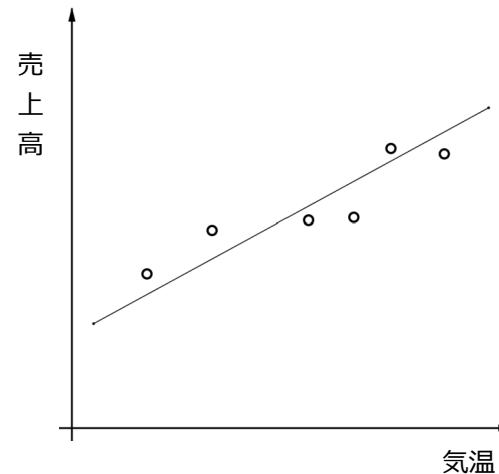
一般化という視点でどちらが適しているか

1.3.1 教師あり学習

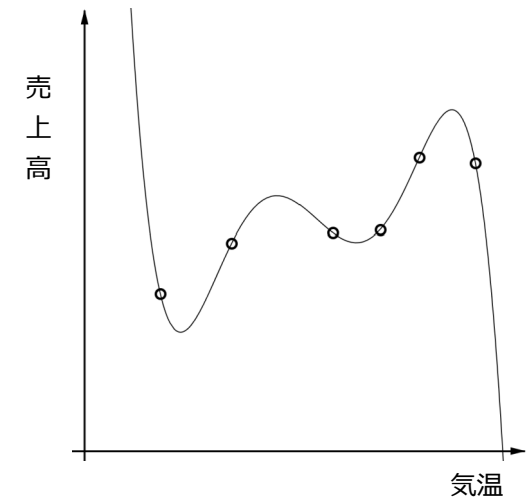
- 回帰
 - 正解情報が数値データ
 - 汎化誤差が最小となるような近似関数を求める



(a) 夏の平均気温とビールの売上高の関係



(b) 1 次式による回帰

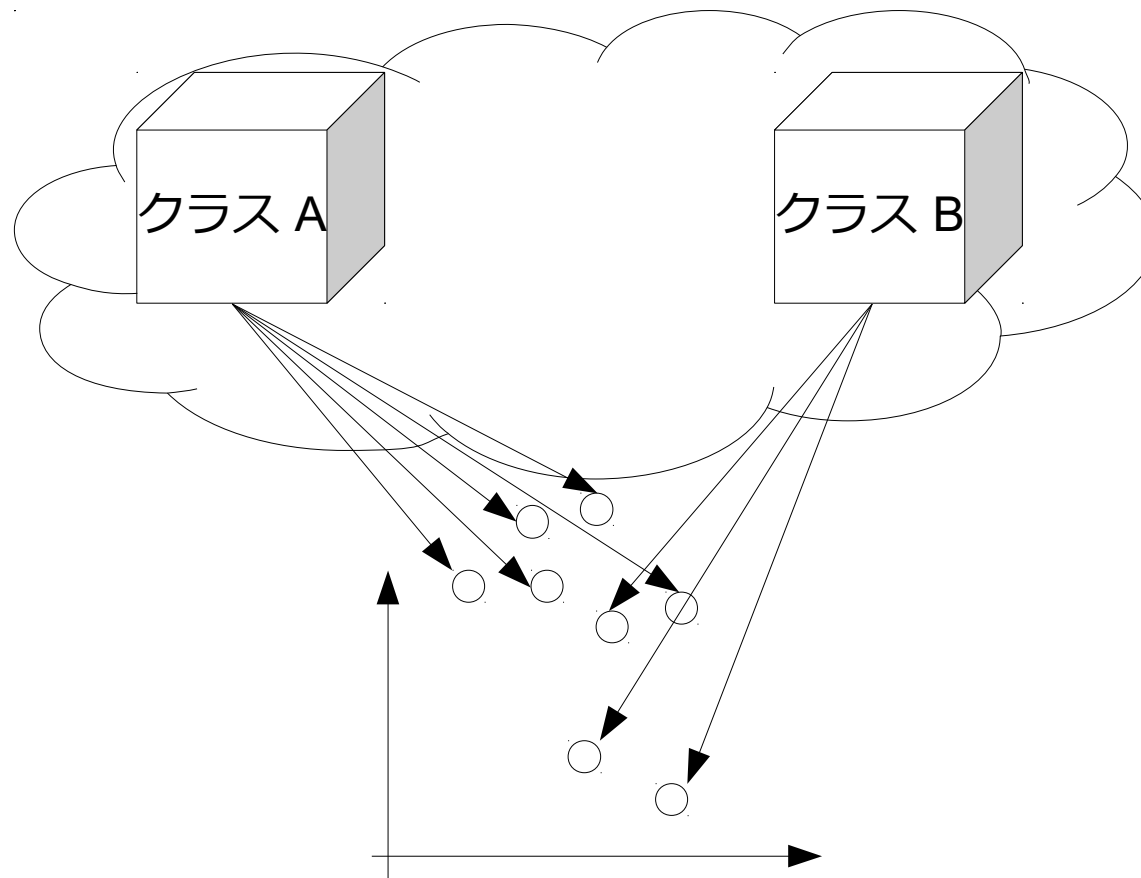


(c) 高次の式による回帰

一般化という視点でどちらが適しているか

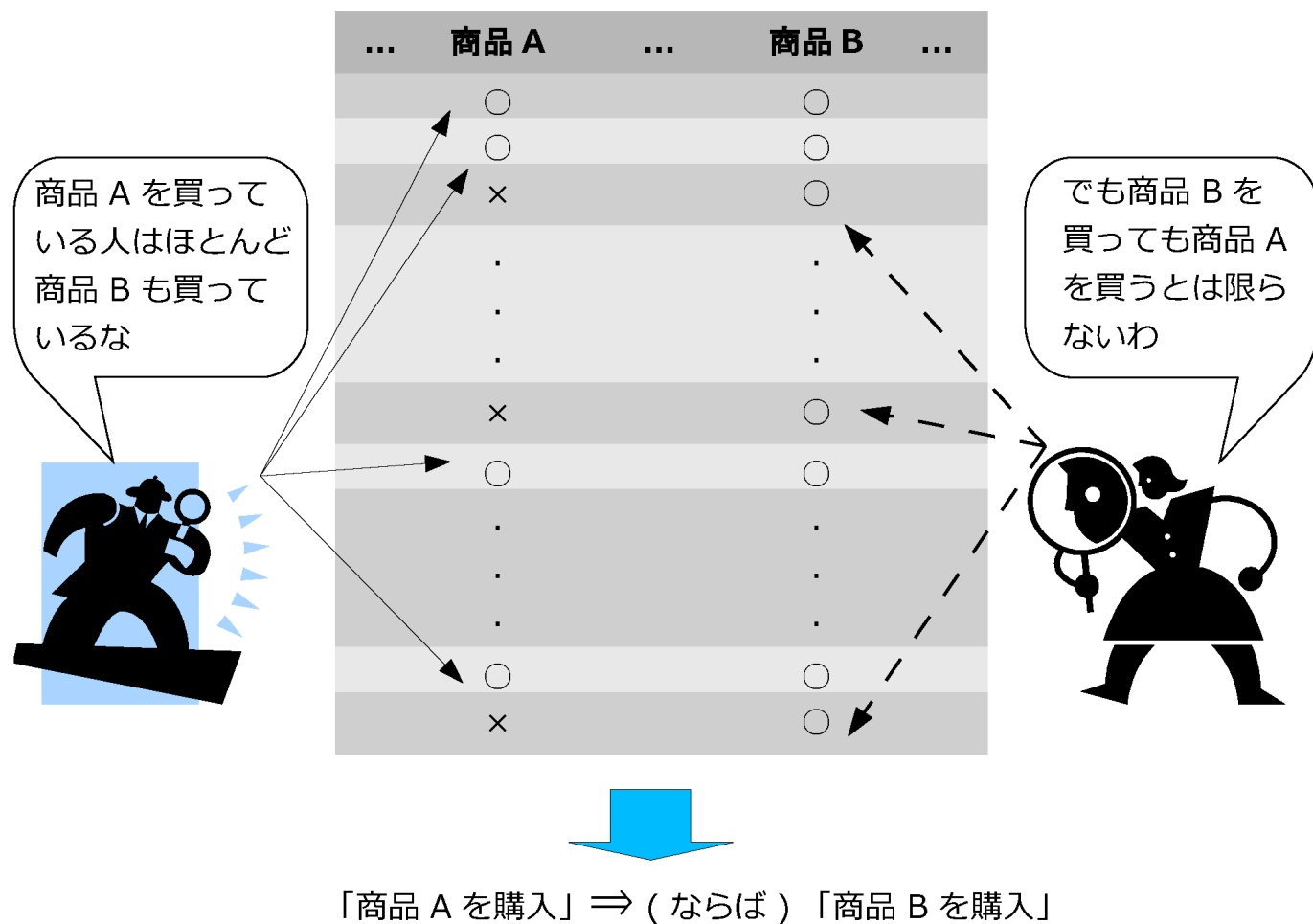
1.3.2 教師なし学習

- モデル推定
 - データを生じさせたクラスを推定
 - 特徴ベクトルは主として数値データ



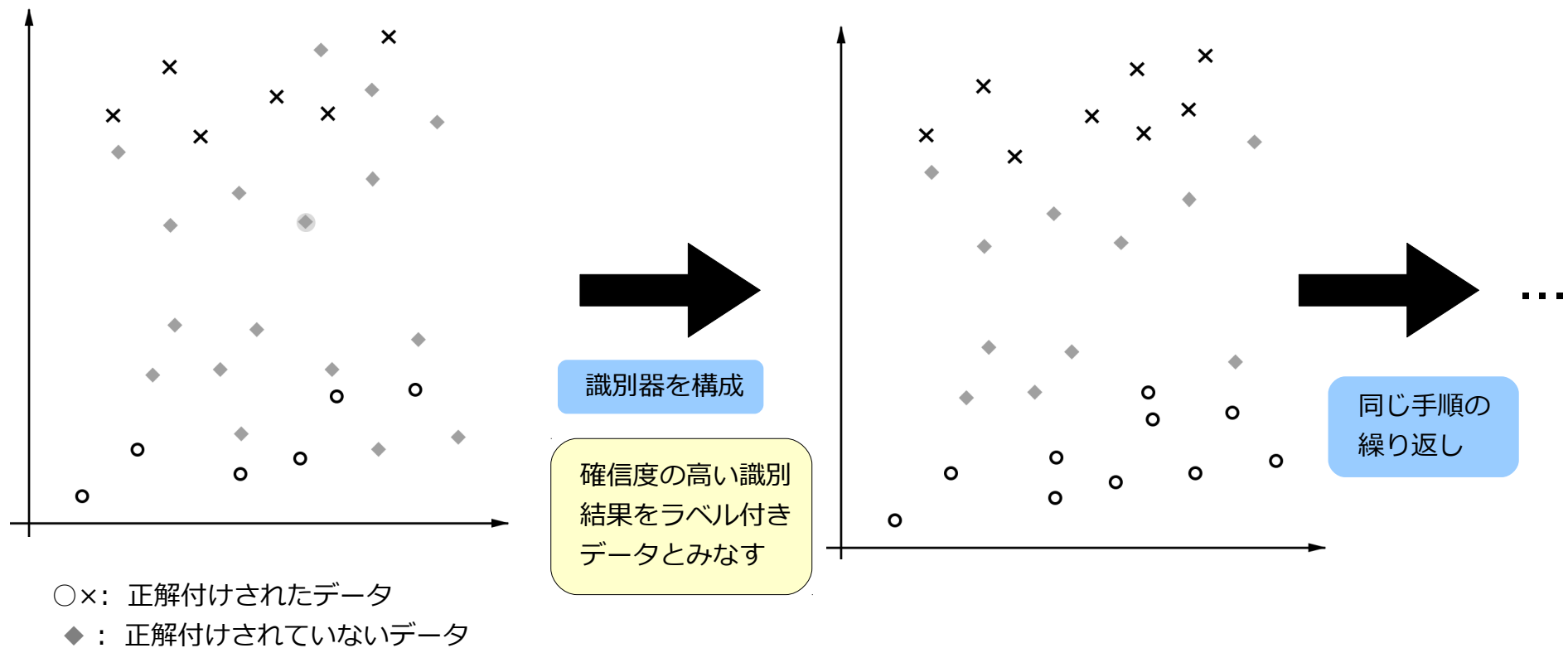
1.3.2 教師なし学習

- パターンマイニング
 - 頻出項目や隠れた規則性を発掘
 - 特徴ベクトルは主としてカテゴリーカルデータ



1.3.3 中間的手法

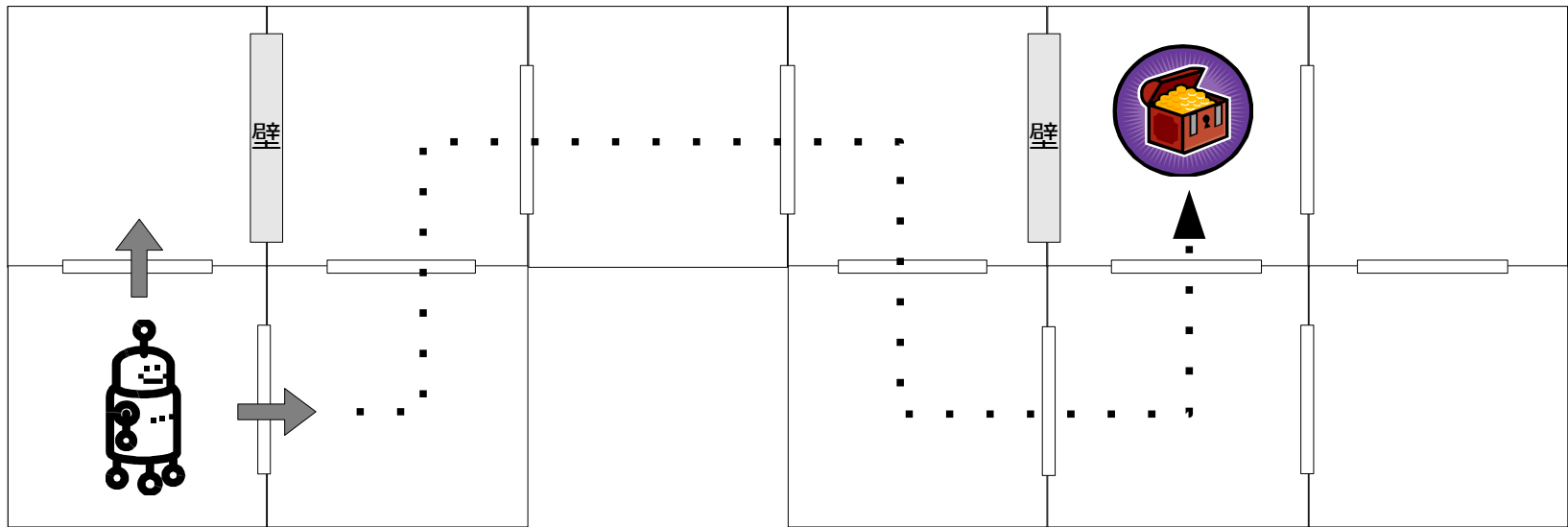
- 半教師あり学習
 - 繰り返しによる学習データの増加



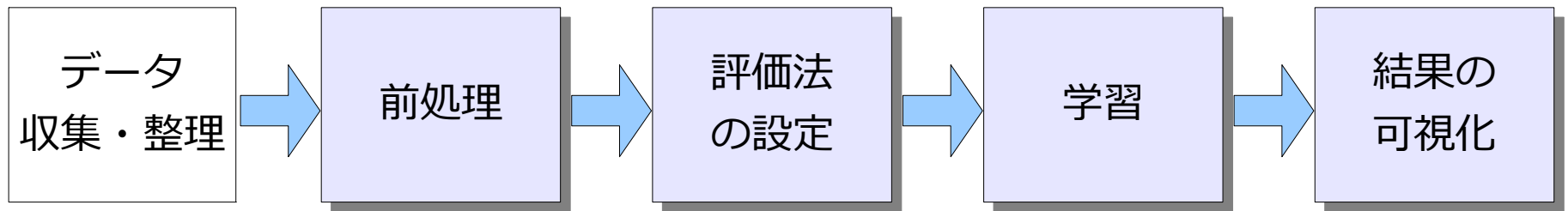
1.3.3 中間的手法


- 強化学習

- 教師信号が、間接的に、ときどき、確率的に与えられる



2. 機械学習の基本的な手順



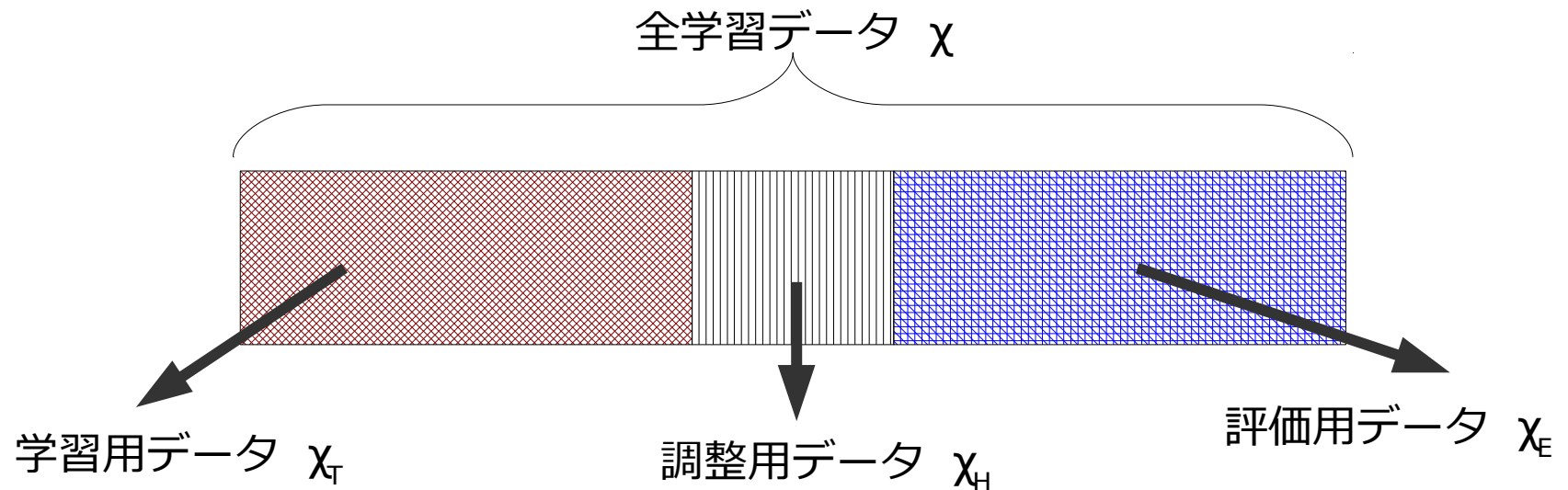
 : ツールによる支援が可能

2.2 前処理

- 欠損値、異常値の処理
 - 欠損値は、その割合などを考慮し適切に補う
 - 異常値の定義は難しく、教師なし学習が有効
- 分析
 - 次元削減
 - データの散らばりをできるだけ保存する低次元空間へ写像 → 主成分分析
 - データの可視化に有効

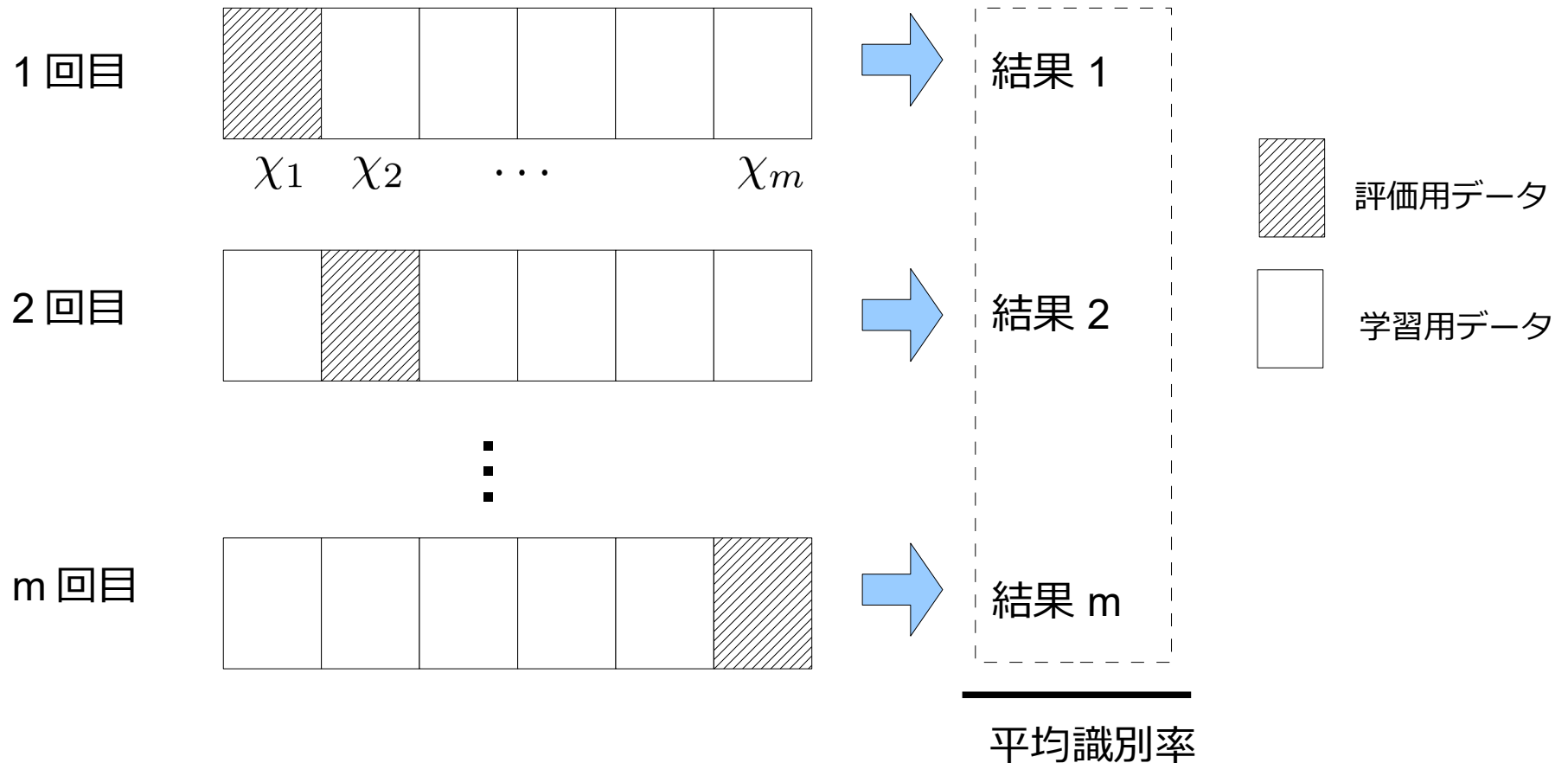
2.3 評価法の設定

- 学習データで評価
 - 動作確認にのみ用いる
- 分割法
 - データが大量にある場合
 - 手元のデータを学習用・調整用・評価用に分ける



2.3 評価法の設定

- 交差確認法
 - アルゴリズムやパラメータの選択を行う



2.4 学習

- 識別問題
 - 決定木、ナイーブベイズ、ロジスティック識別、サポートベクトルマシン、ニューラルネットワーク
- 回帰問題
 - 線形回帰、回帰木、モデル木
- 教師なし学習
 - クラスタリング、確率密度推定、バスケット分析

2.5 結果の可視化

- 学習したモデル
 - 式、木構造、ネットワークの重み、 etc.
- 性能
 - 正解率、精度、再現率、 F 値
 - グラフ
 - パラメータを変えたときの性能の変化
 - 異なるモデルの性能比較

2.5 結果の可視化

- 混同行列

	予測+	予測-
正解+	true positive(TP)	false negative(FN)
正解-	falsepositive(FP)	true negative(TN)

- 正解率 $Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$

- 精度 $Precision = \frac{TP}{TP + FP}$

- 再現率 $Recall = \frac{TP}{TP + FN}$

- F 値 $F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

正解の割合
クラスの出現率に
偏りがある場合は不適

正例の判定が
正しい割合

正しく判定された
正例の割合

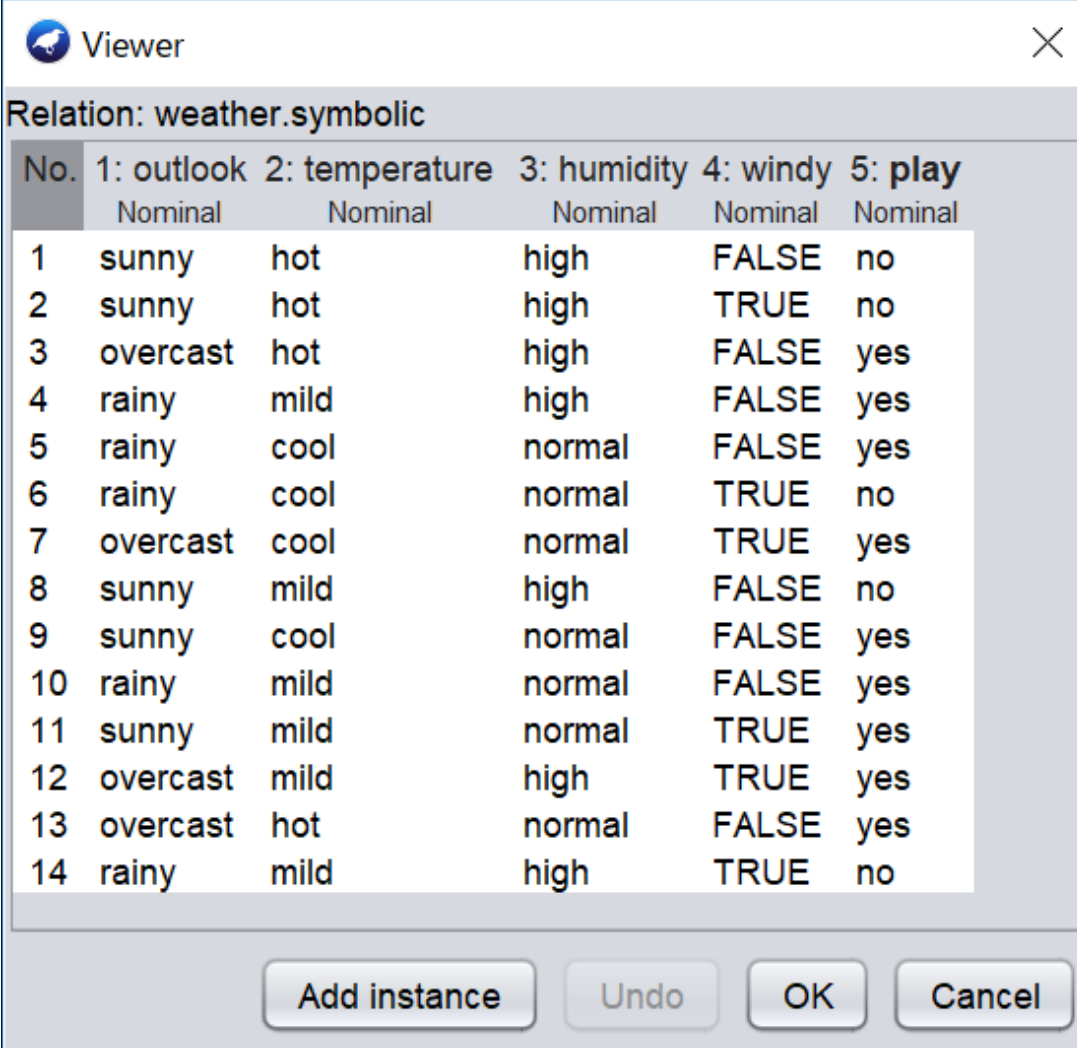
トレードオフ

精度と再現率の
調和平均

3. 識別 —概念学習—

- 学習データ

(テニスをする日 ; weather.nominal.arff)



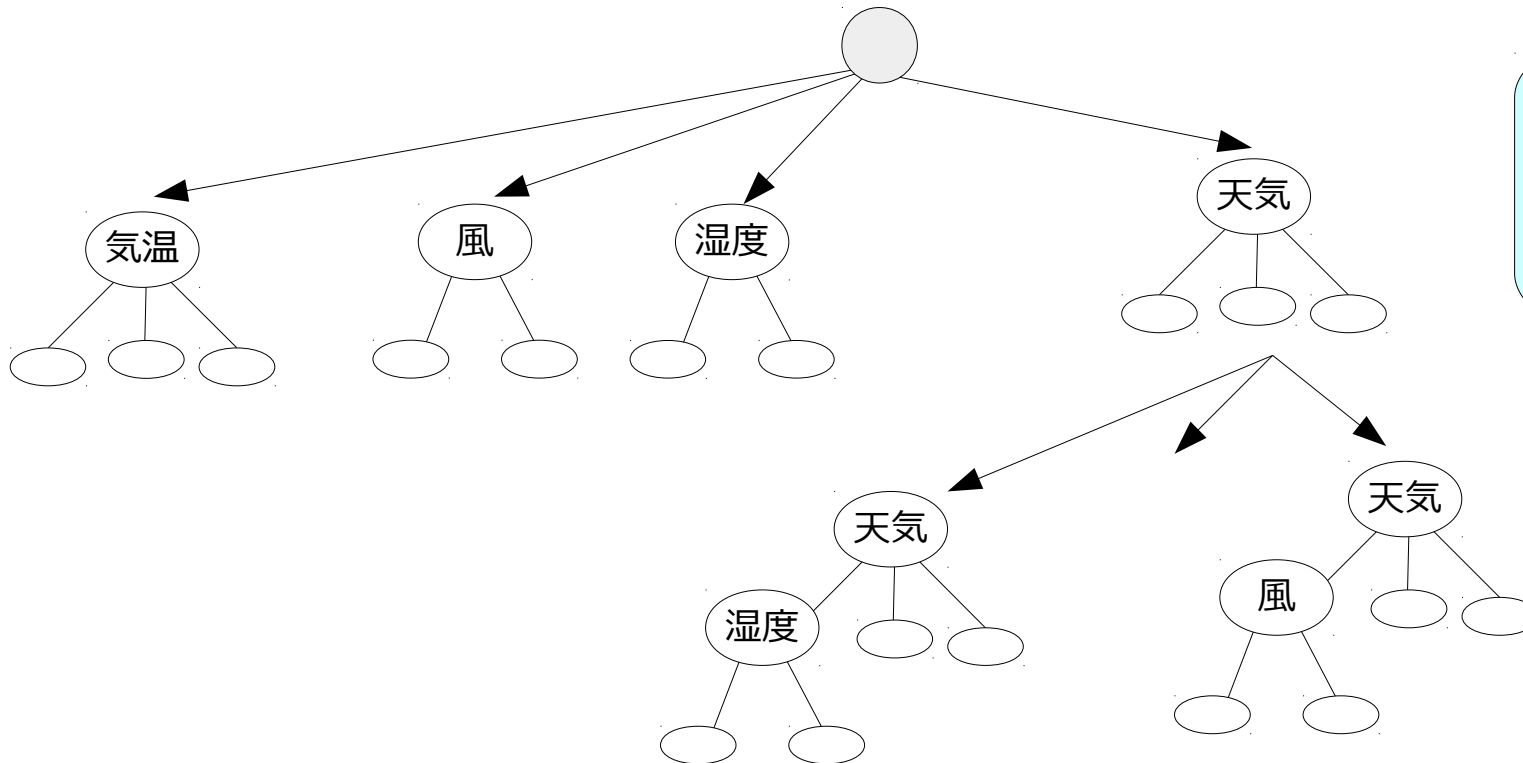
Relation: weather.symbolic

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Buttons: Add instance, Undo, OK, Cancel

3.4 決定木の学習

- 決定木学習の考え方
 - 節はデータを分割する条件を持つ
 - できるだけ同一クラスのデータが偏るように
 - 分割後のデータ集合に対して、同様の操作を行う
 - 全ての葉が単一クラスの集合になれば終了



この手順に従うと、
一般には小さな木
ができる

バイアス

複雑な説明よりも
単純な説明の方が
汎用性が高い

計算例

$$E(D) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$Gain(D, a) \equiv E(D) - \sum_{v \in Values(a)} \frac{|D_v|}{|D|} E(D_v)$$

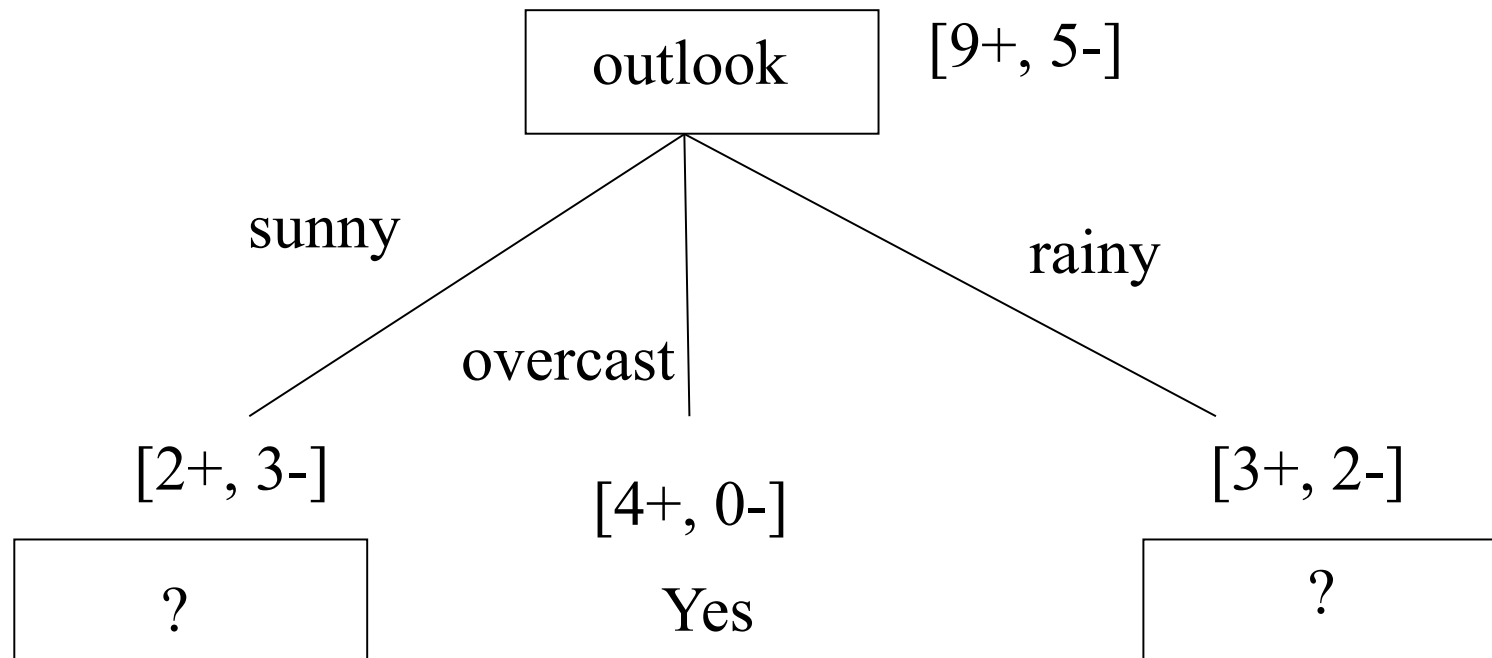
- 情報獲得量

Gain(S, outlook)=0.246

Gain(S, humidity)=0.151

Gain(S, windy)=0.048

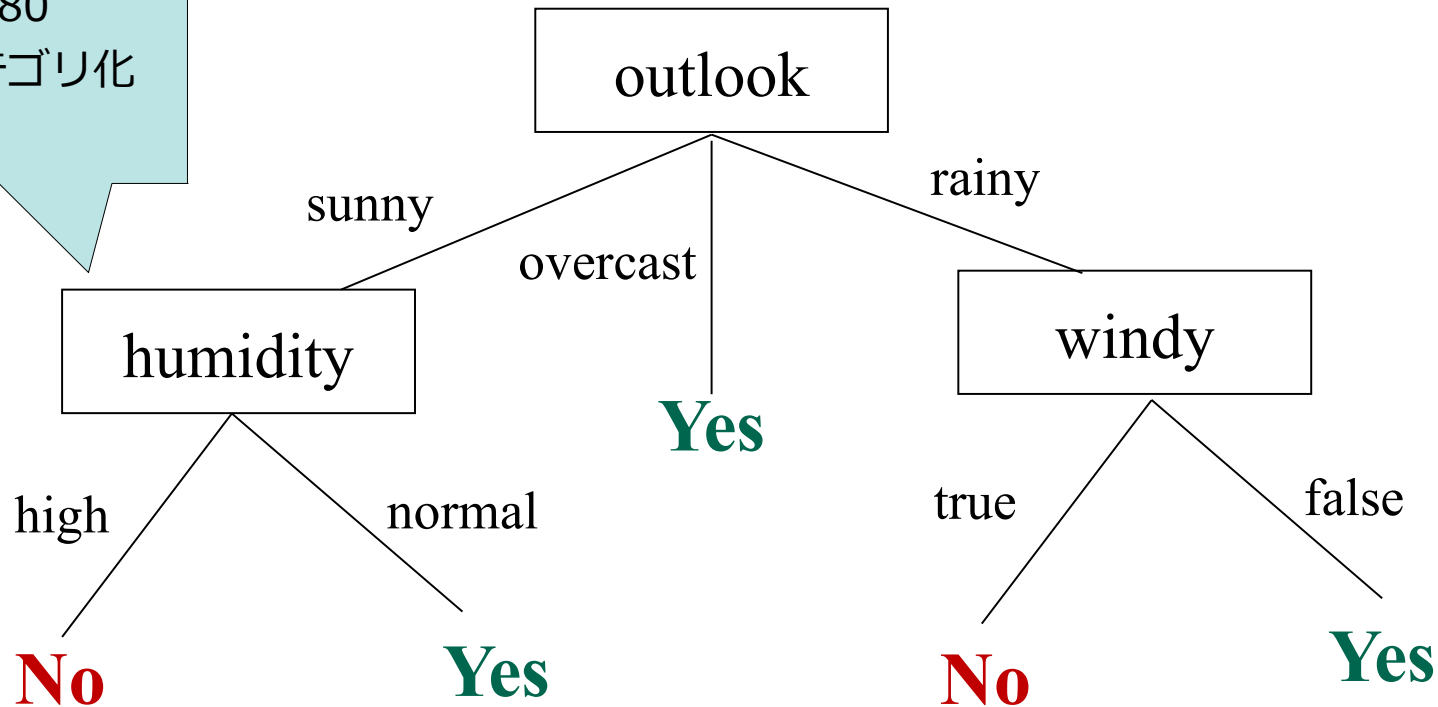
Gain(S, temperature)=0.029



3.4 決定木の学習

- 結果として得られる決定木

数値データの場合は、
humidity > 80
のようにカテゴリ化
する。



Section1 のまとめ

- 人工知能 ⊃ 機械学習 ⊃ 深層学習
- 機械学習とは
 - 適切にタスクを遂行する適切なモデルを、適切な特徴から構築すること
- 機械学習の分類
 - 教師あり・教師なし・中間的手法
- 評価基準の理解が重要
 - 正解率・精度・再現率・F 値
- 汎用性の観点からの評価が必要
 - 決定木は、多くのデータを説明できる小さい木を求めることで、汎用性を確保している