

1. はじめに

内容

1.1 ビッグデータの時代

近年のビッグデータ・機械学習の話題

1.2 機械学習とは何か

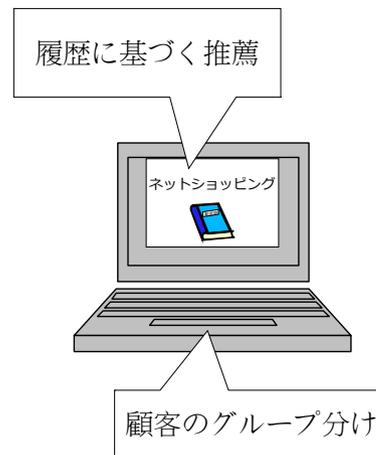
機械学習の全体像

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 ビッグデータの時代

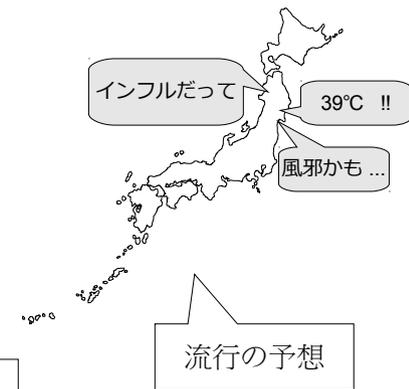
- ビッグデータとは
 - ネットワーク、センサー等の発達によって収集されたデータ
 - 大量・多様・スピードが特徴
- ビッグデータは何に使えるか
 - 有用な知見の獲得
 - 省力化
 - 将来の予測



多様な趣味・嗜好に対応



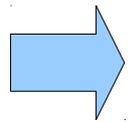
安心・安全を進化



経験や勘を超越

1.1 ビッグデータの時代

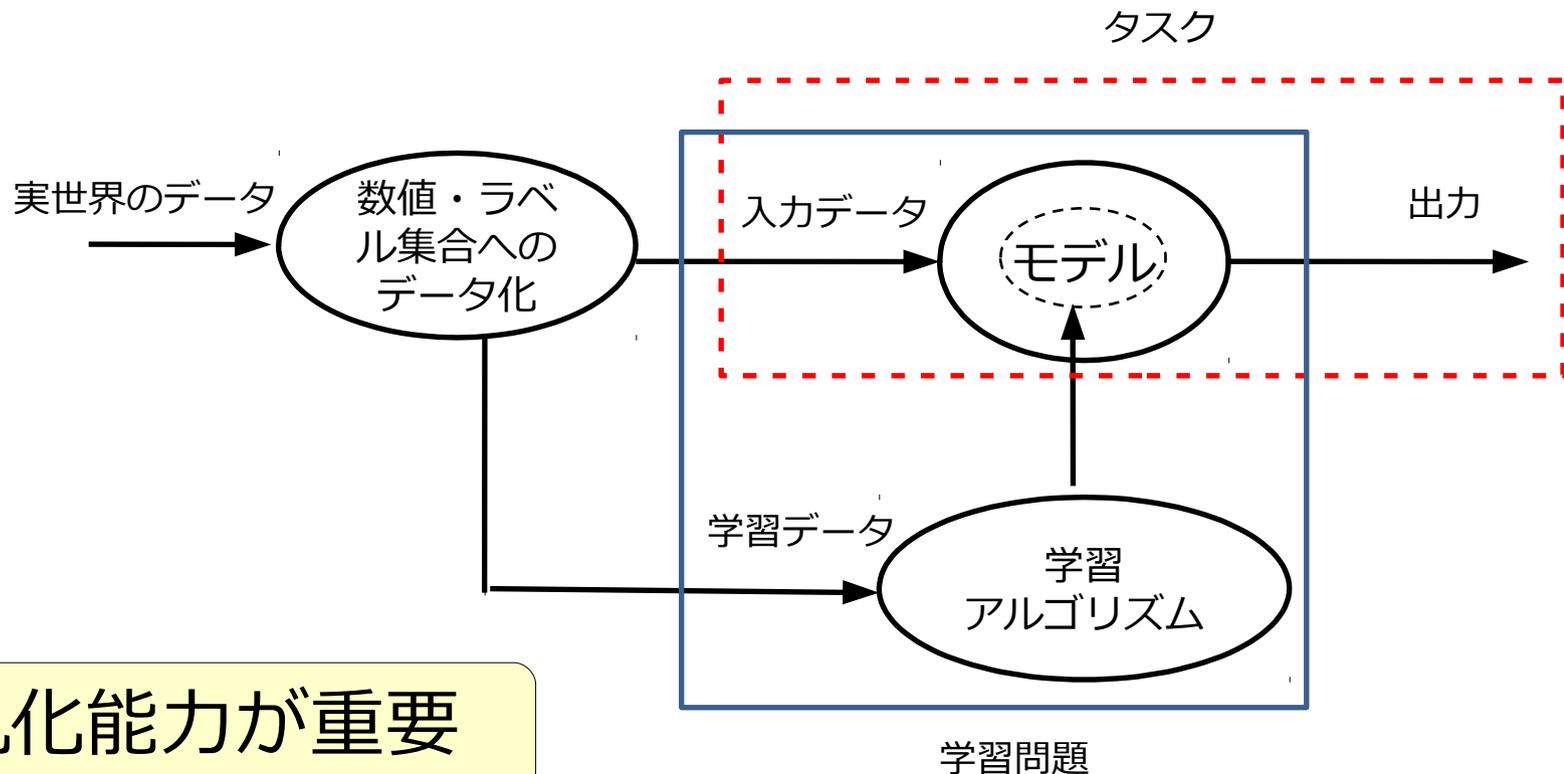
- ビッグデータ処理の問題点
 - データ量が膨大なので人手による情報抽出は不可能
 - 矛盾・曖昧性・ノイズを含むデータを処理するプログラムを記述するのは難しい



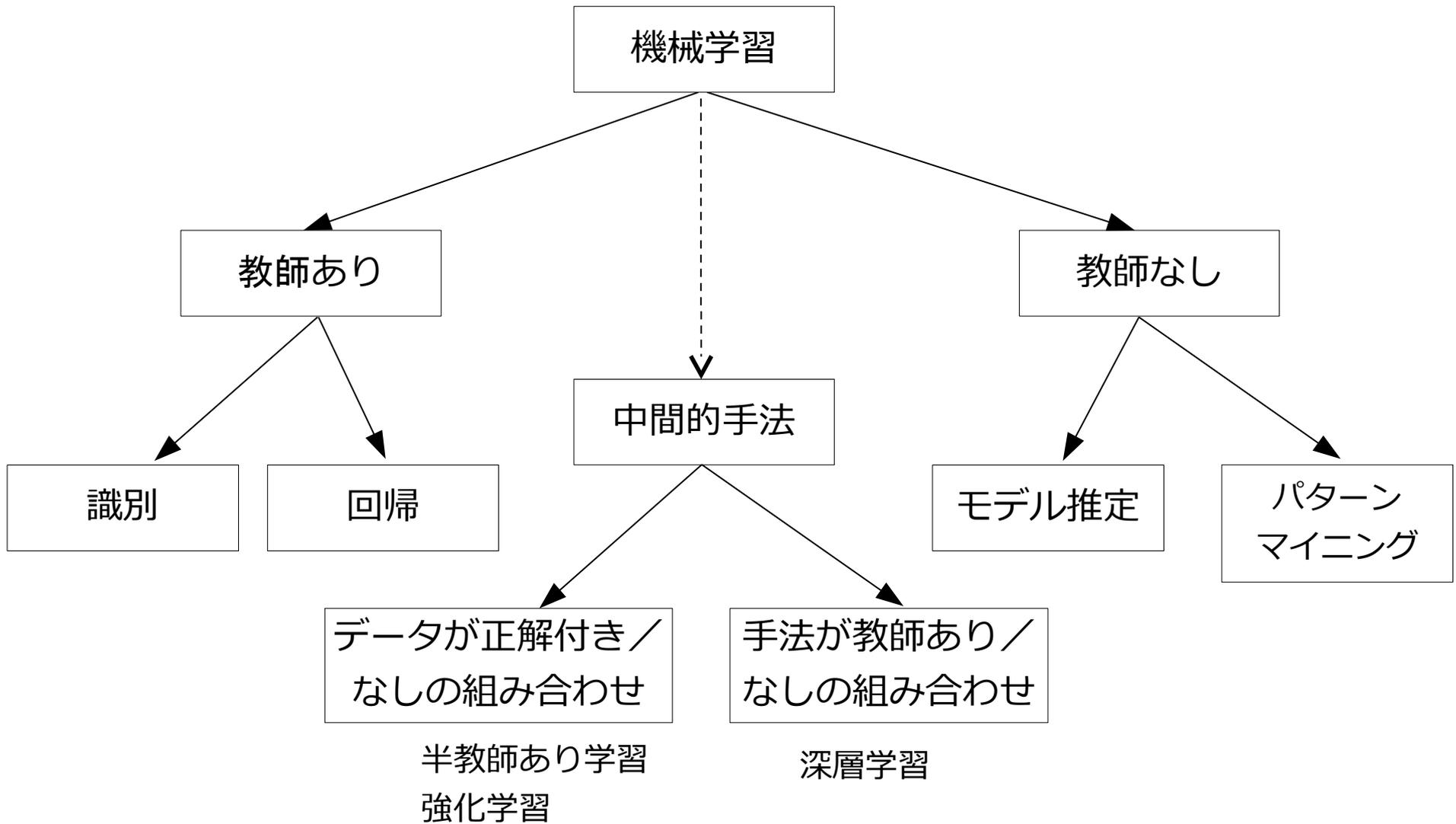
コンピュータによる**機械学習**が有望

1.2 機械学習とは何か

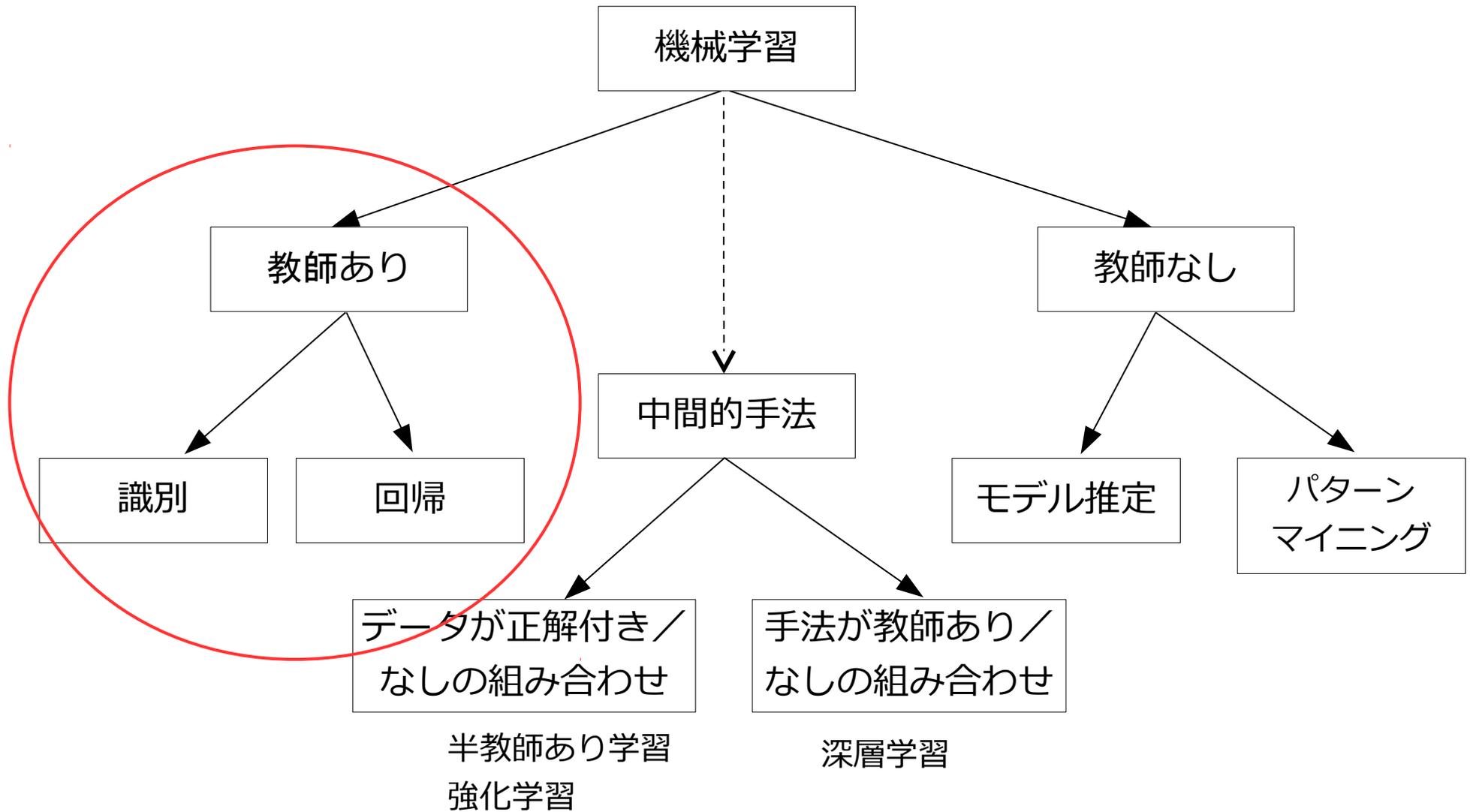
- 機械学習の定義 [Flach 2012]
 - 機械学習は、適切に**タスク**を遂行する適切な**モデル**を、適切な**特徴**から構築すること



1.3 機械学習の分類



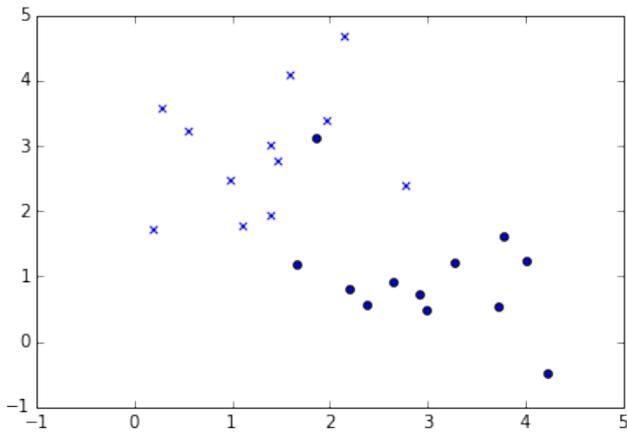
1.3 機械学習の分類



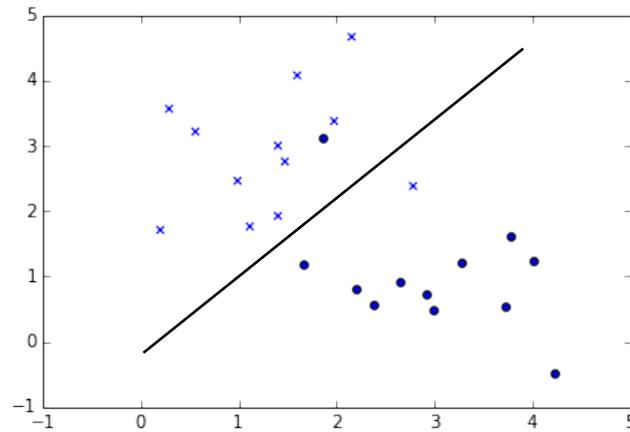
1.3.1 教師あり学習

- 識別

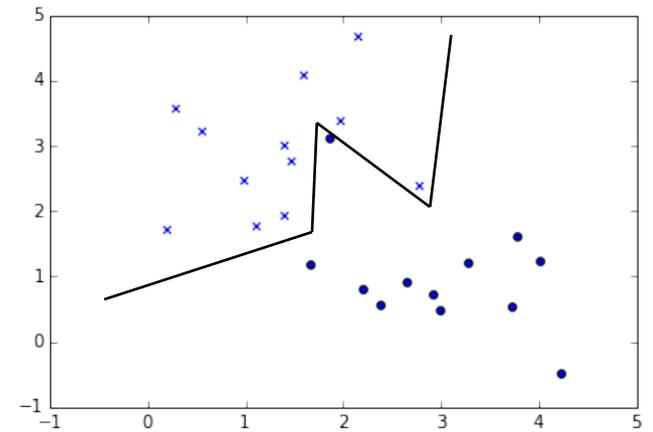
- 学習データに対するエラーが最小となるような特徴空間上の分離面を求める



(a) 入力が2次元数値ベクトルの識別問題



(b) 学習結果 1



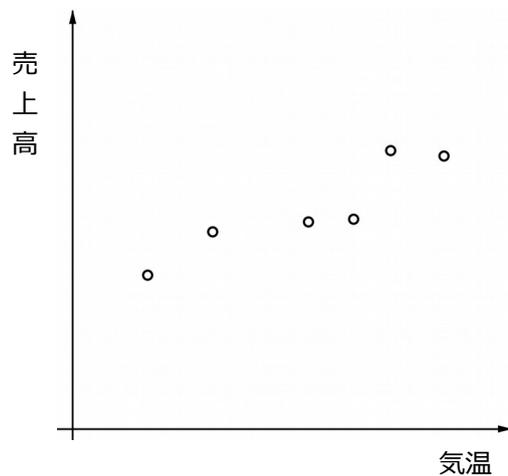
(c) 学習結果 2

一般化という視点でどちらが適しているか

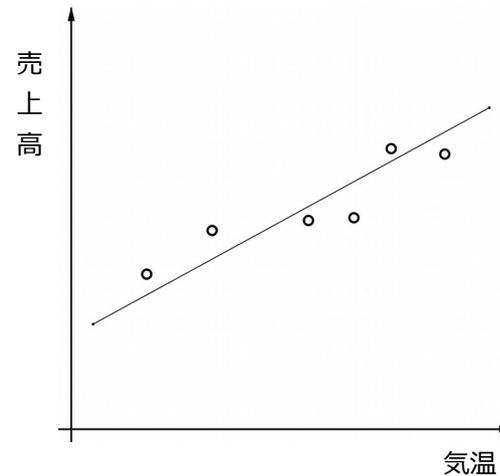
1.3.1 教師あり学習

- 回帰

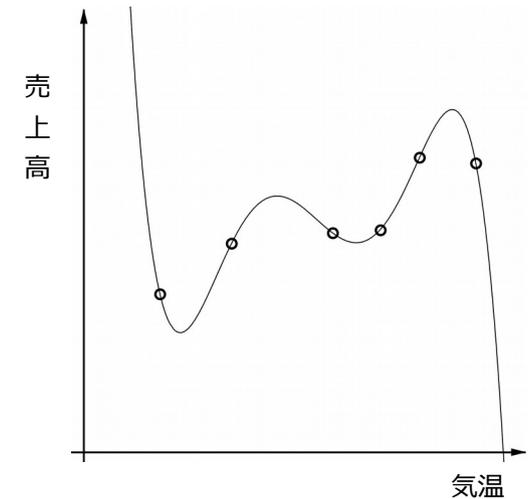
- 学習データに対するエラーが最小となるような近似関数を求める



(a) 過去の気温と売上高の関係



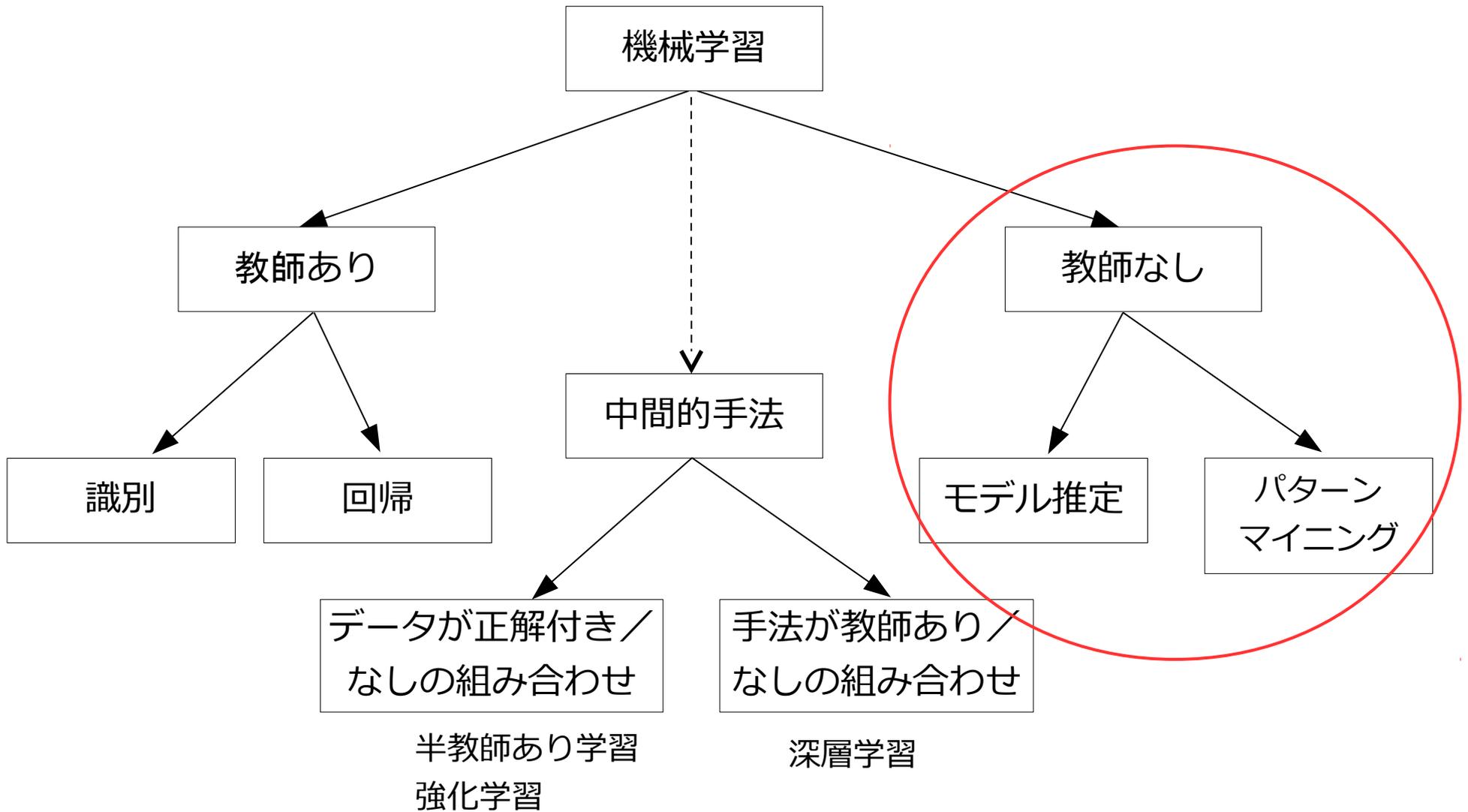
(b) 1次式による回帰



(c) 高次の式による回帰

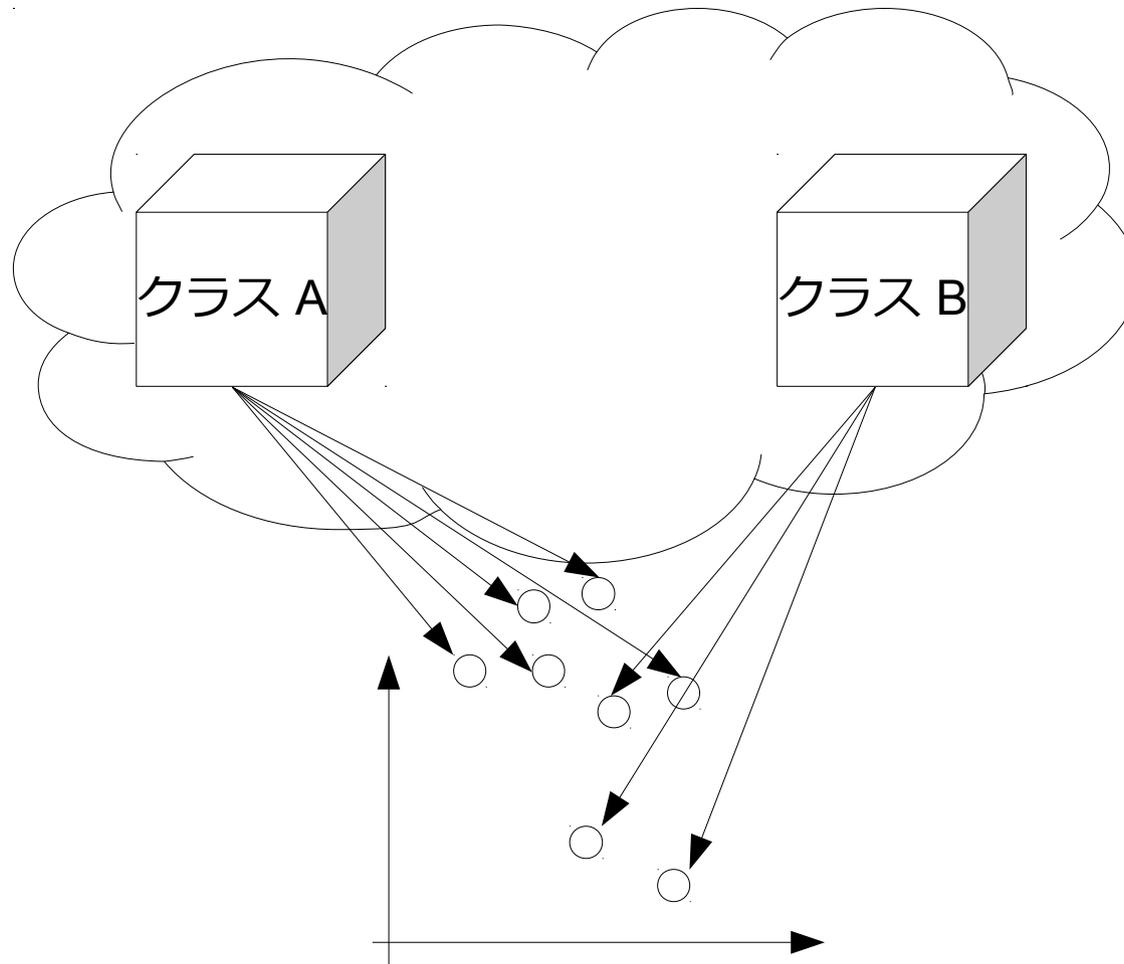
一般化という視点でどちらが適しているか

1.3 機械学習の分類



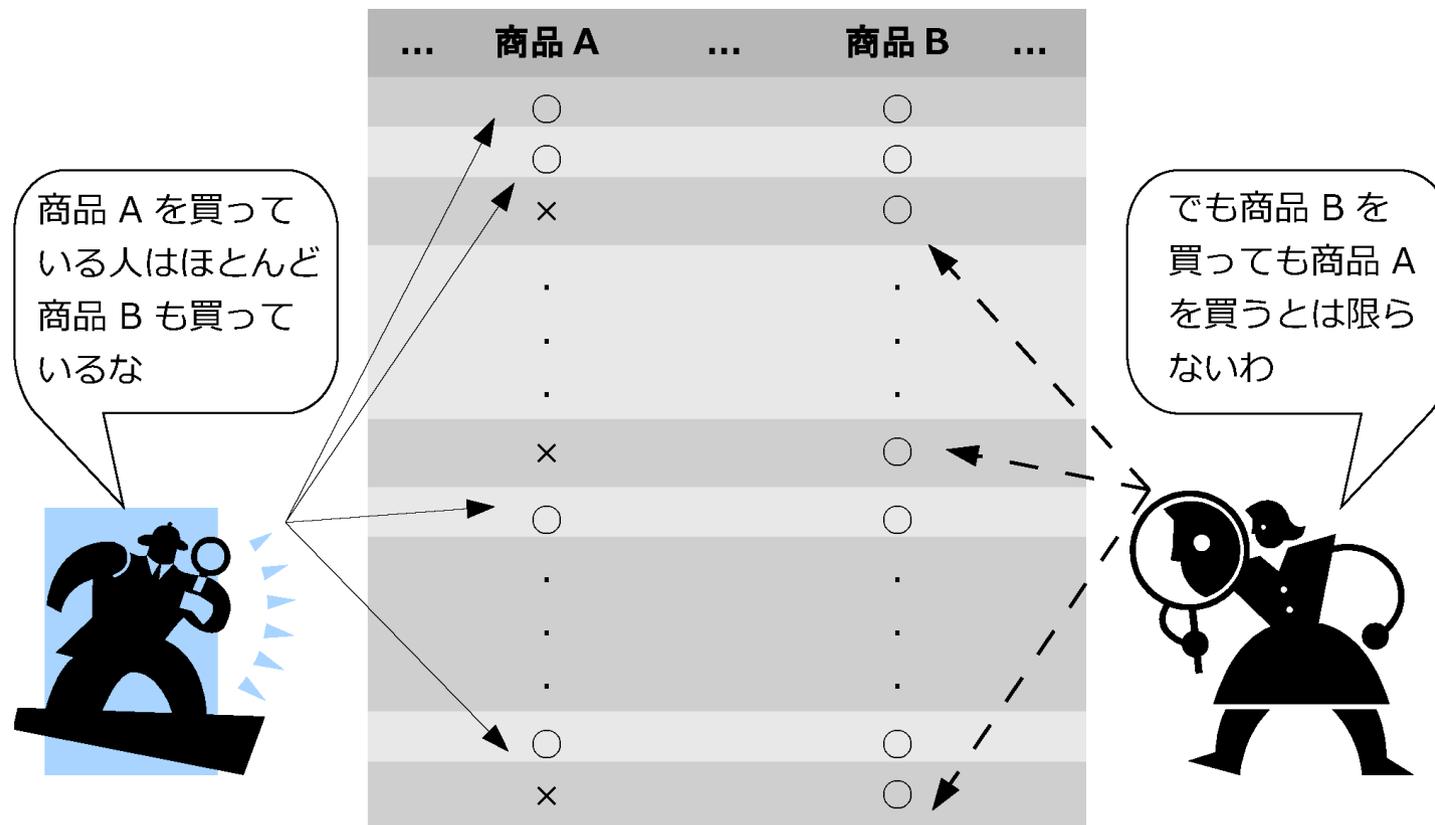
1.3.2 教師なし学習

- モデル推定
 - データを生じさせたクラスを推定



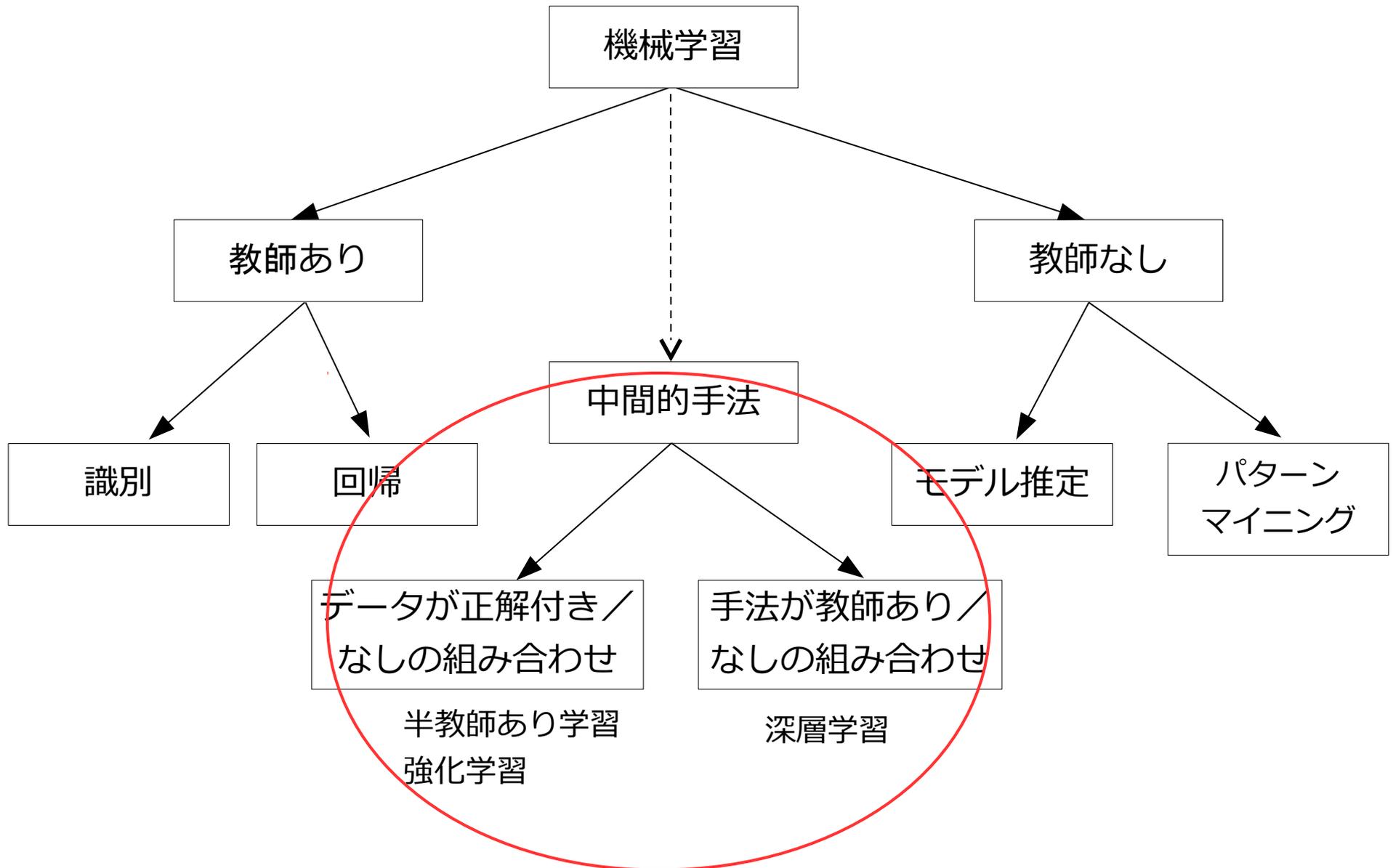
1.3.2 教師なし学習

- パターンマイニング
 - 頻出項目や隠れた規則性を発掘



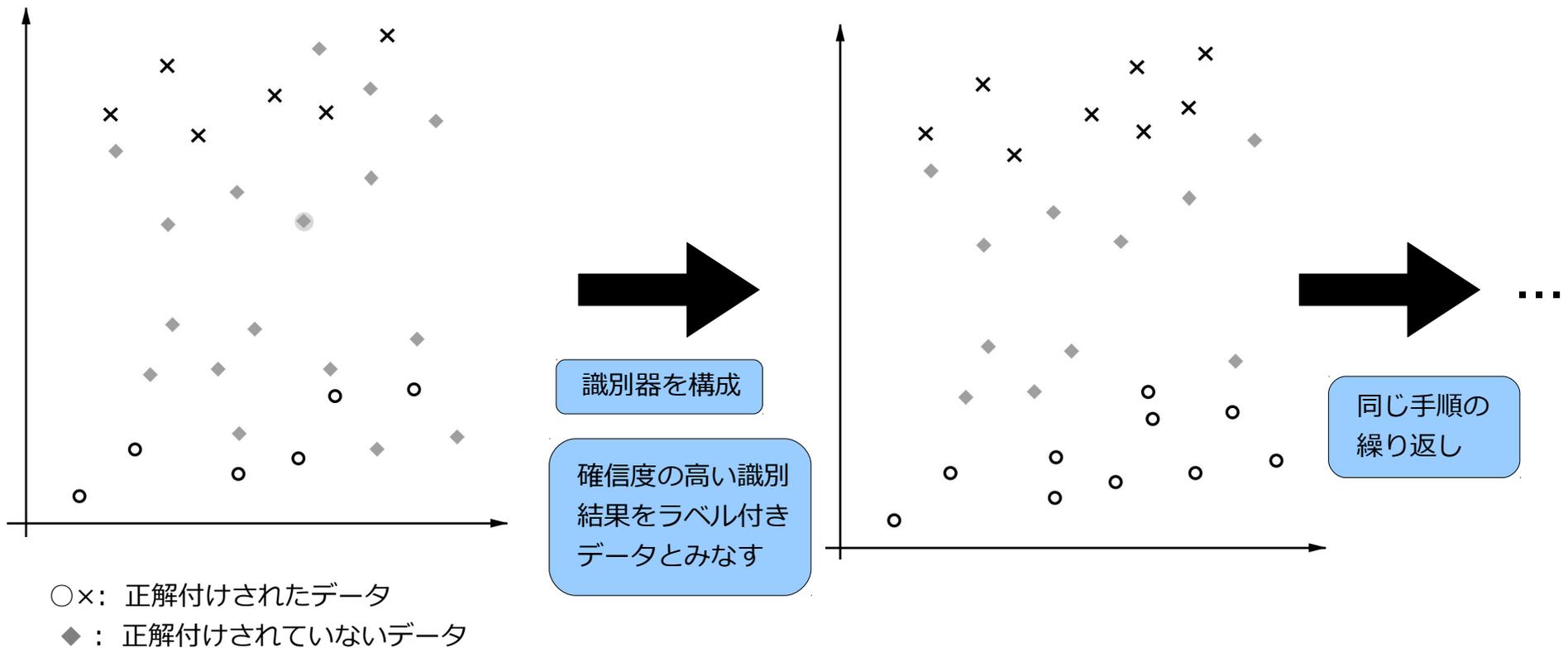
「商品 A を購入」 ⇒ (ならば) 「商品 B を購入」

1.3 機械学習の分類



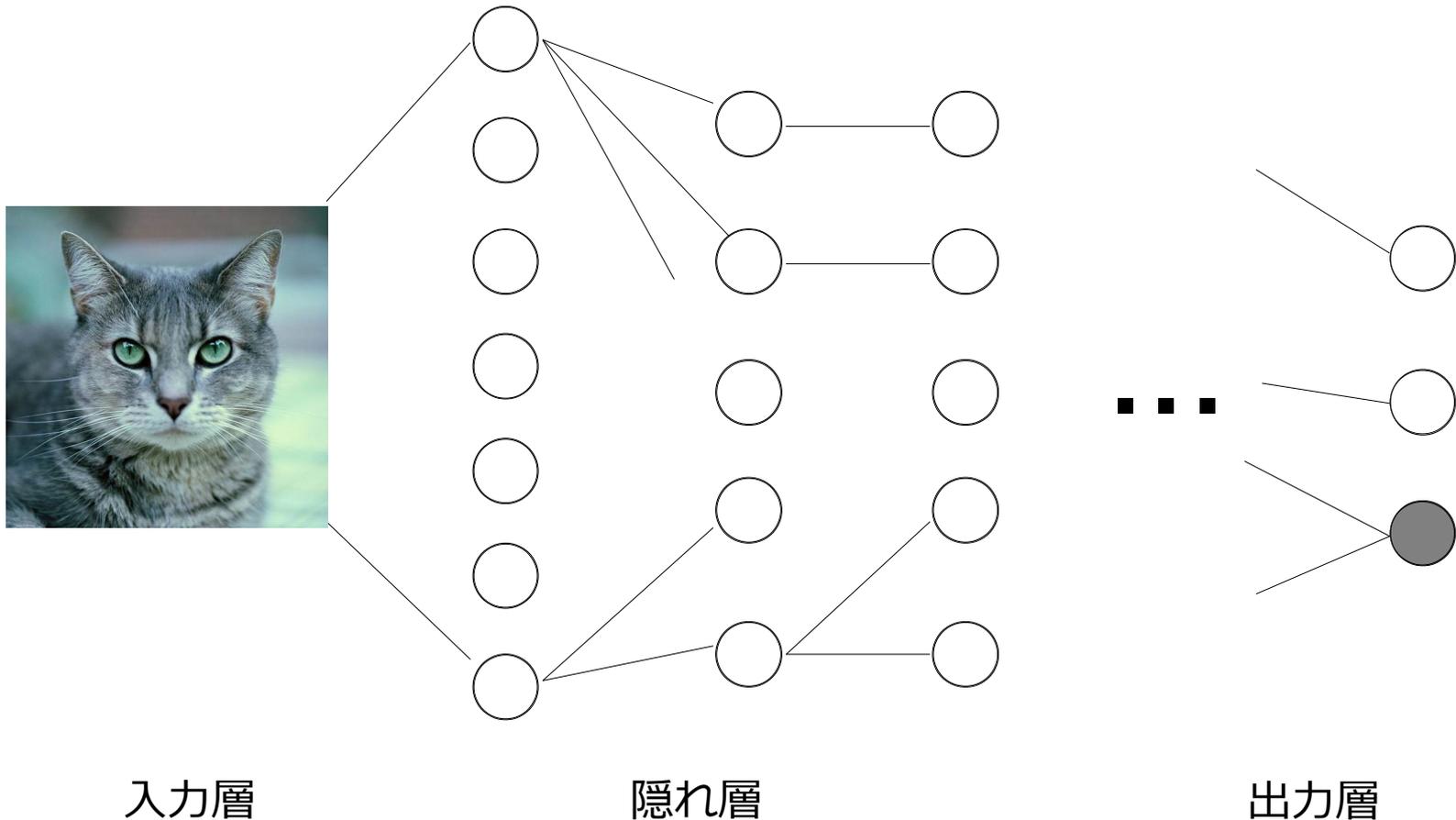
1.3.3 中間的手法

- 半教師あり学習
 - 繰り返しによる学習データの増加

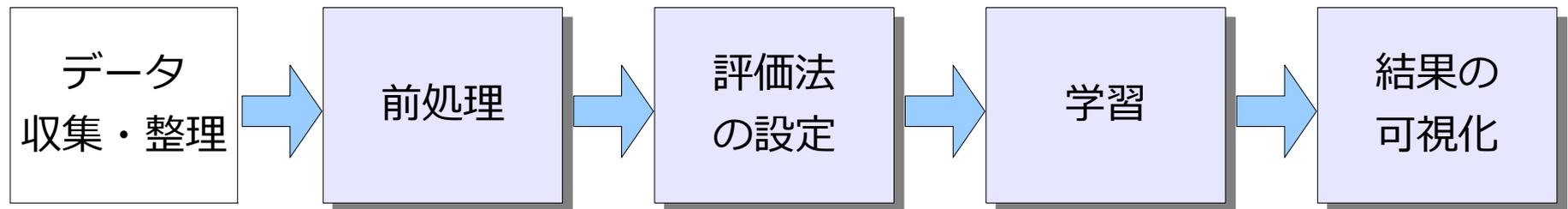


1.3.3 中間的手法

- 深層学習
 - 教師なし学習で初期値を設定→表現学習
 - 教師あり学習で識別能力を学習

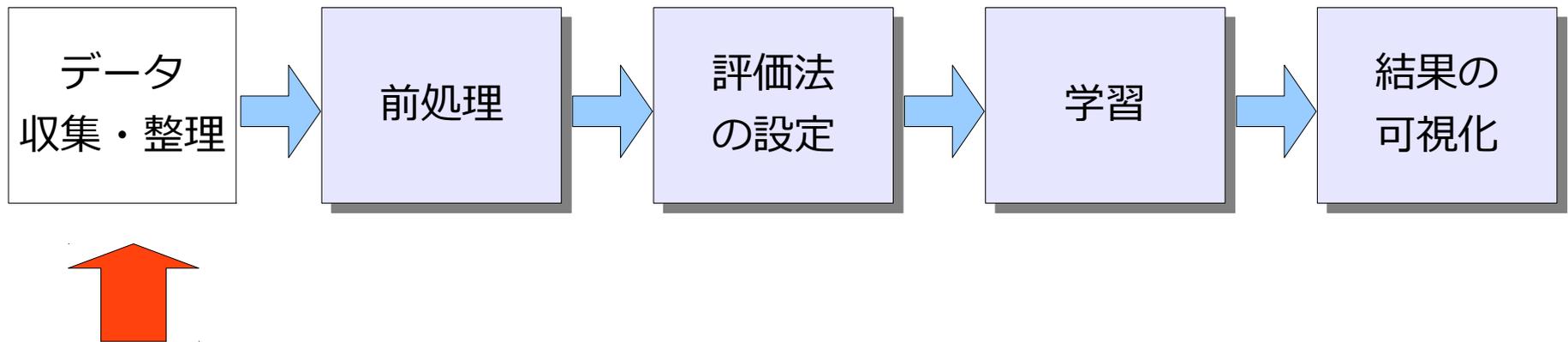


2. 機械学習の基本的な手順



 : ツールによる支援が可能

2. 機械学習の基本的な手順



2.1 データ収集・整理

- (勉強用の) 機械学習のデータ
- データマイニングツール Weka に付属

表 2.2 Weka 付属のデータ

データ名	内容	特徴	正解情報
breast-cancer	乳癌の再発	ラベル	クラス (2 値)
contact-lenses	コンタクトレンズの推薦	ラベル	クラス (3 値)
cpu	CPU の性能評価	数値	数値
credit-g	融資の審査	混合	クラス (2 値)
diabetes	糖尿病の検査	数値	クラス (2 値)
iris	アヤメの分類	数値	クラス (3 値)
Reuters-Corn	記事分類	テキスト	クラス (2 値)
supermarket	スーパーの購買記録	ラベル	なし
weather.nominal	ゴルフをする条件	ラベル	クラス (2 値)
weather.numeric	ゴルフをする条件	混合	クラス (2 値)

2.1 データ収集・整理

- Weka のデータ形式 ARFF フォーマット

```
% 1. Title: Iris Plants Database
@RELATION iris
@ATTRIBUTE sepallength REAL
@ATTRIBUTE sepalwidth REAL
@ATTRIBUTE petallength REAL
@ATTRIBUTE petalwidth REAL
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1, 3.5, 1.4, 0.2, Iris-setosa
4.9, 3.0, 1.4, 0.2, Iris-setosa
...
7.0, 3.2, 4.7, 1.4, Iris-versicolor
6.4, 3.2, 4.5, 1.5, Iris-versicolor
...
6.3, 3.3, 6.0, 2.5, Iris-virginica
5.8, 2.7, 5.1, 1.9, Iris-virginica
...
```

データセット名

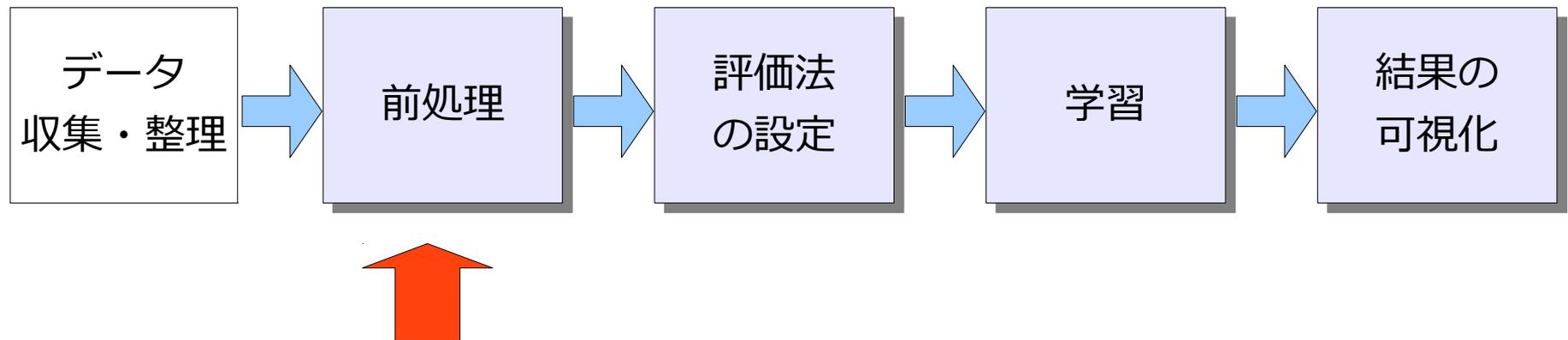
特徴名と型

萼・花びらの
長さ・幅

アヤメの
種類

これ以降、1行に1事例
(ExcelのCSV形式と同じ)

2. 機械学習の基本的な手順



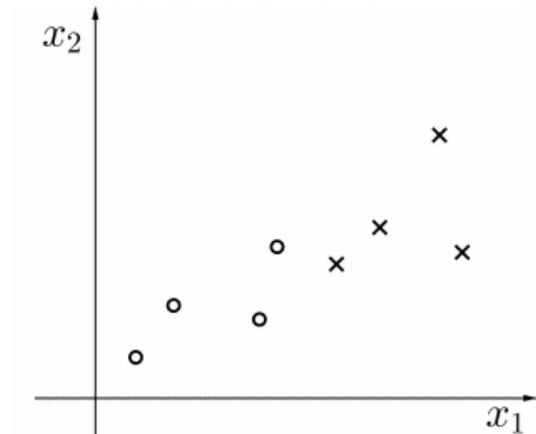
2.2 前処理

なぜ必要？

- データの標準化
 - 各次元に対して平均値を引き、標準偏差で割る
 - その結果、平均 0、分散 1 の標準正規分布に従う
- 分析
 - 主成分分析（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
 - データの可視化に有効

2.2 前処理

主成分分析の考え方



共分散行列 Σ の計算

\bar{x}_1, \bar{x}_2 : 平均値、 N : データ数

$$\Sigma = \frac{1}{N} \begin{pmatrix} \sum (x_1 - \bar{x}_1)^2 & \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \\ \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) & \sum (x_2 - \bar{x}_2)^2 \end{pmatrix}$$

対角成分は分散、
非対角成分は相関を表す

Σ は

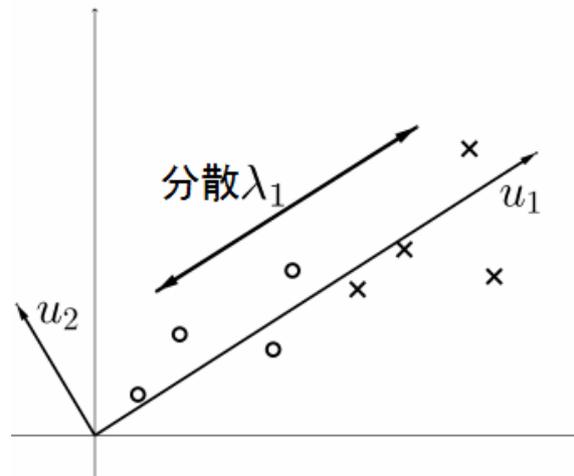
半正値 (→ 固有値が全て 0 以上の実数)

対称行列 (→ 固有ベクトルが実数かつ直交)

であるので

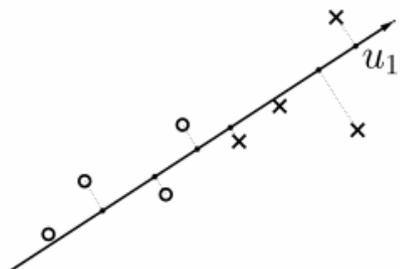
$$\Sigma' = U^T \Sigma U = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

λ は固有値の大きい順、
 U は対応する固有ベクトルを並べたもの

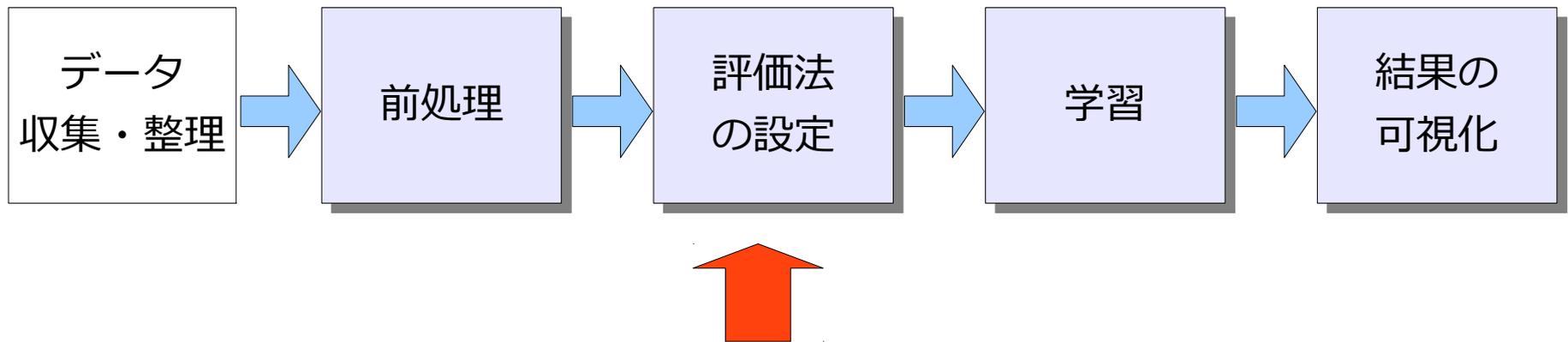


λ_1 に対応する固有ベクトルからなる行列 U_1 で
2次元データを1次元に射影

$$u_1 = U_1^T \mathbf{x}$$



2. 機械学習の基本的な手順

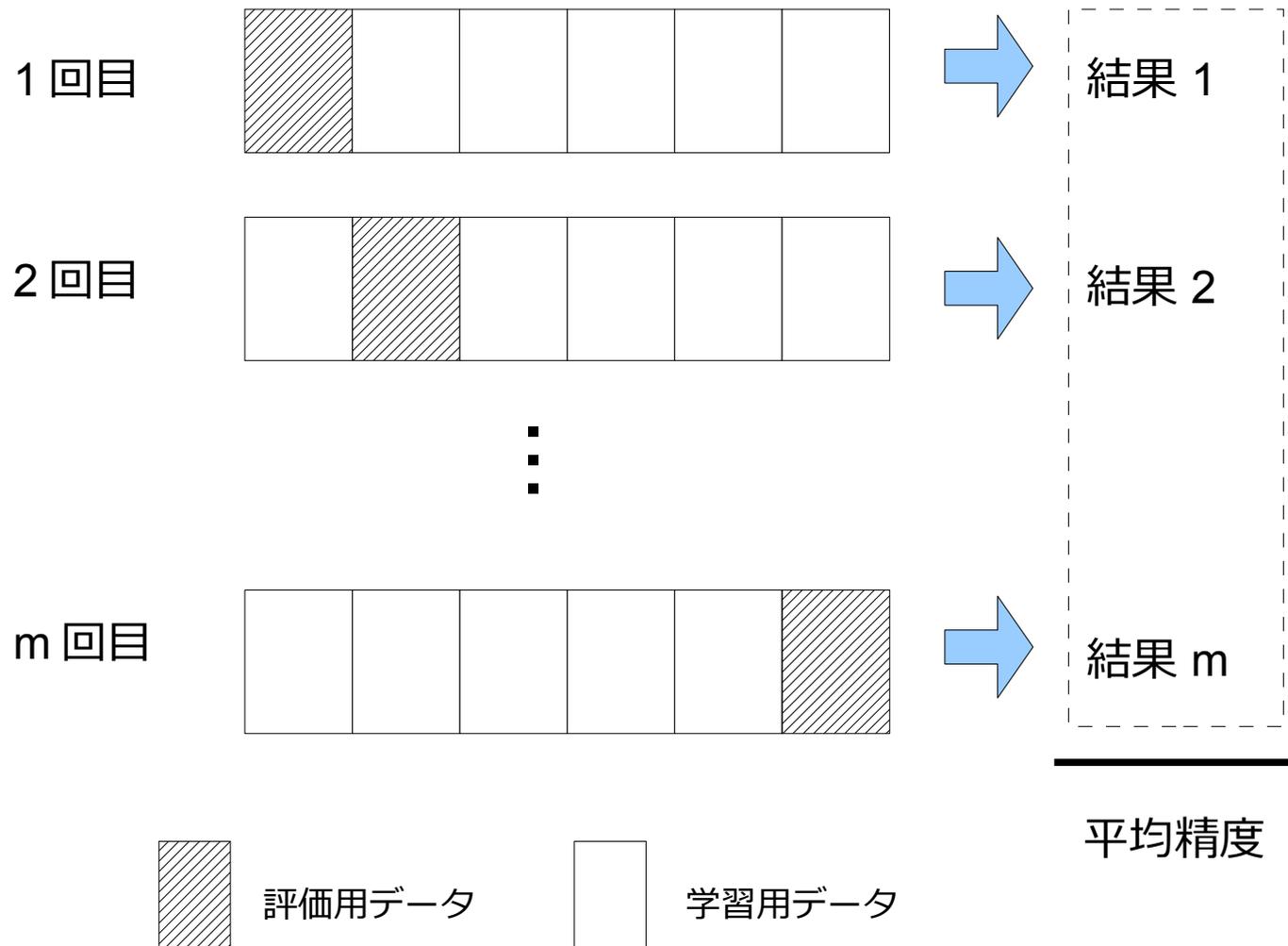


2.3 評価基準の設定

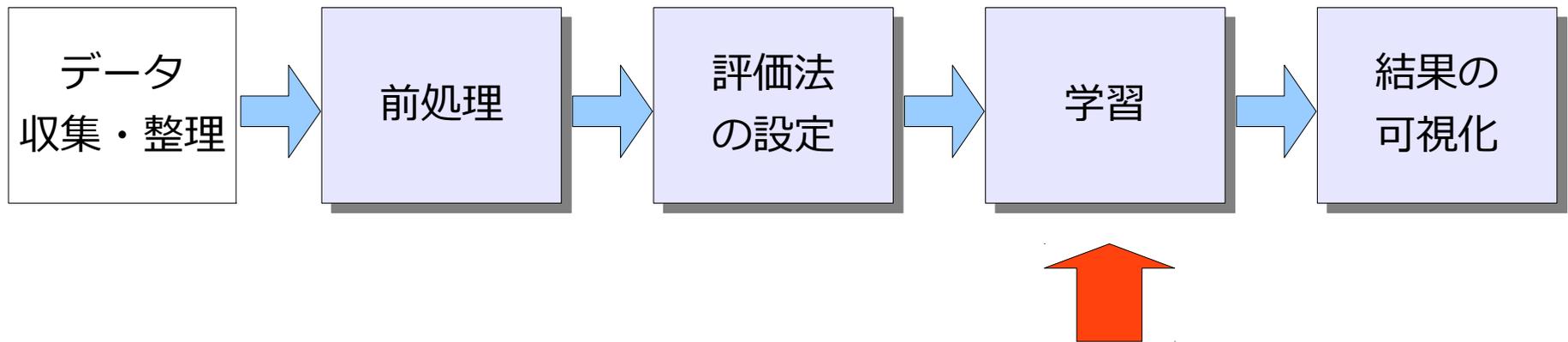
- 学習したモデルの評価
 - 学習データに適合しすぎては意味がない
⇒ 過学習の問題
 - 汎用性の評価が必要
⇒ 未知のデータに対する識別能力
- 分割法
 - データを学習用と評価用で半々に分ける
もったいない ...

2.3 評価基準の設定

- 交差確認法 (cross validation)

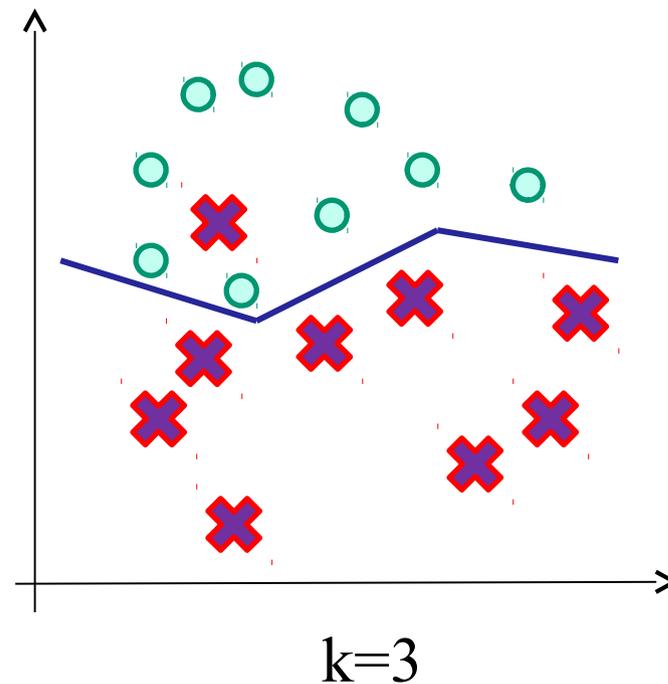
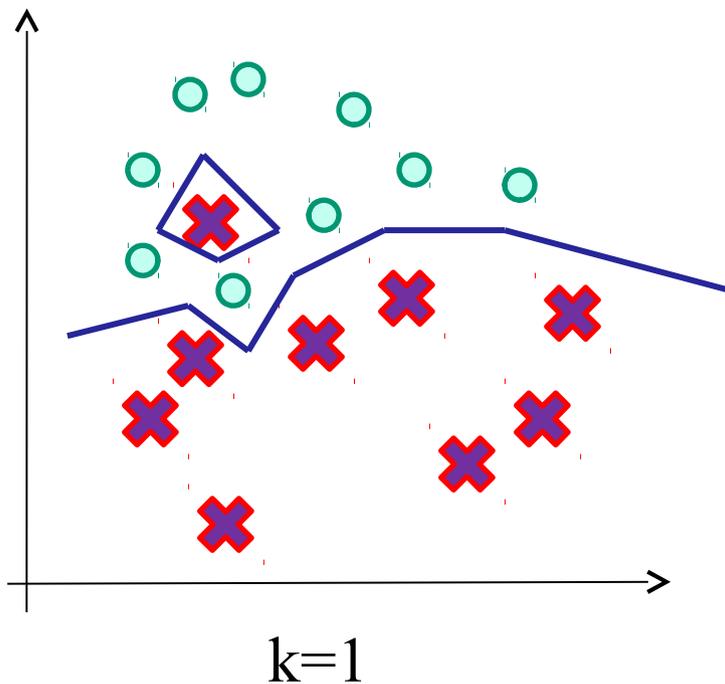


2. 機械学習の基本的な手順

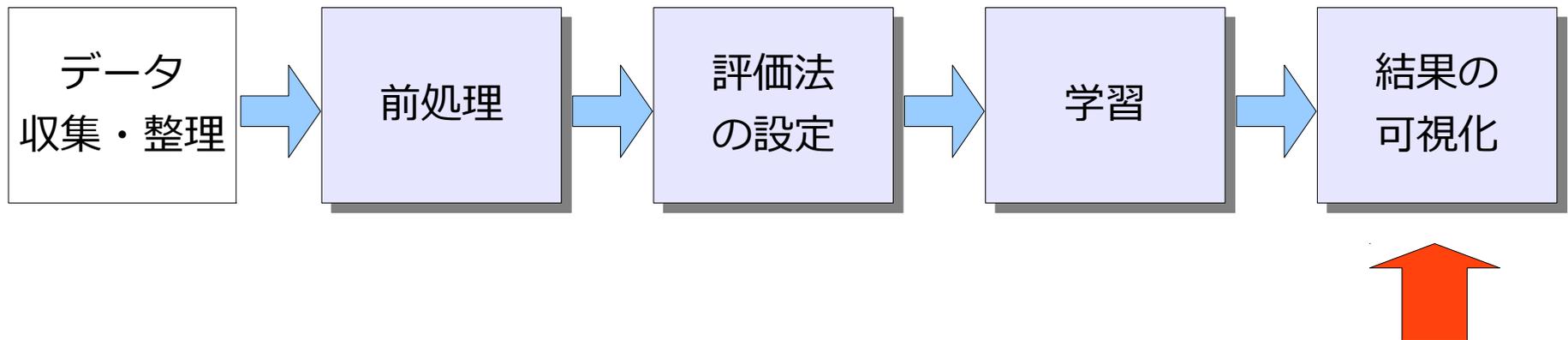


2.4 学習

- k-NN 法
 - 入力データと最も近い学習データのクラスに分類する
 - ノイズに強くするためには k- 近傍の多数決を取る



2. 機械学習の基本的な手順



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

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

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

精度と再現率の
調和平均

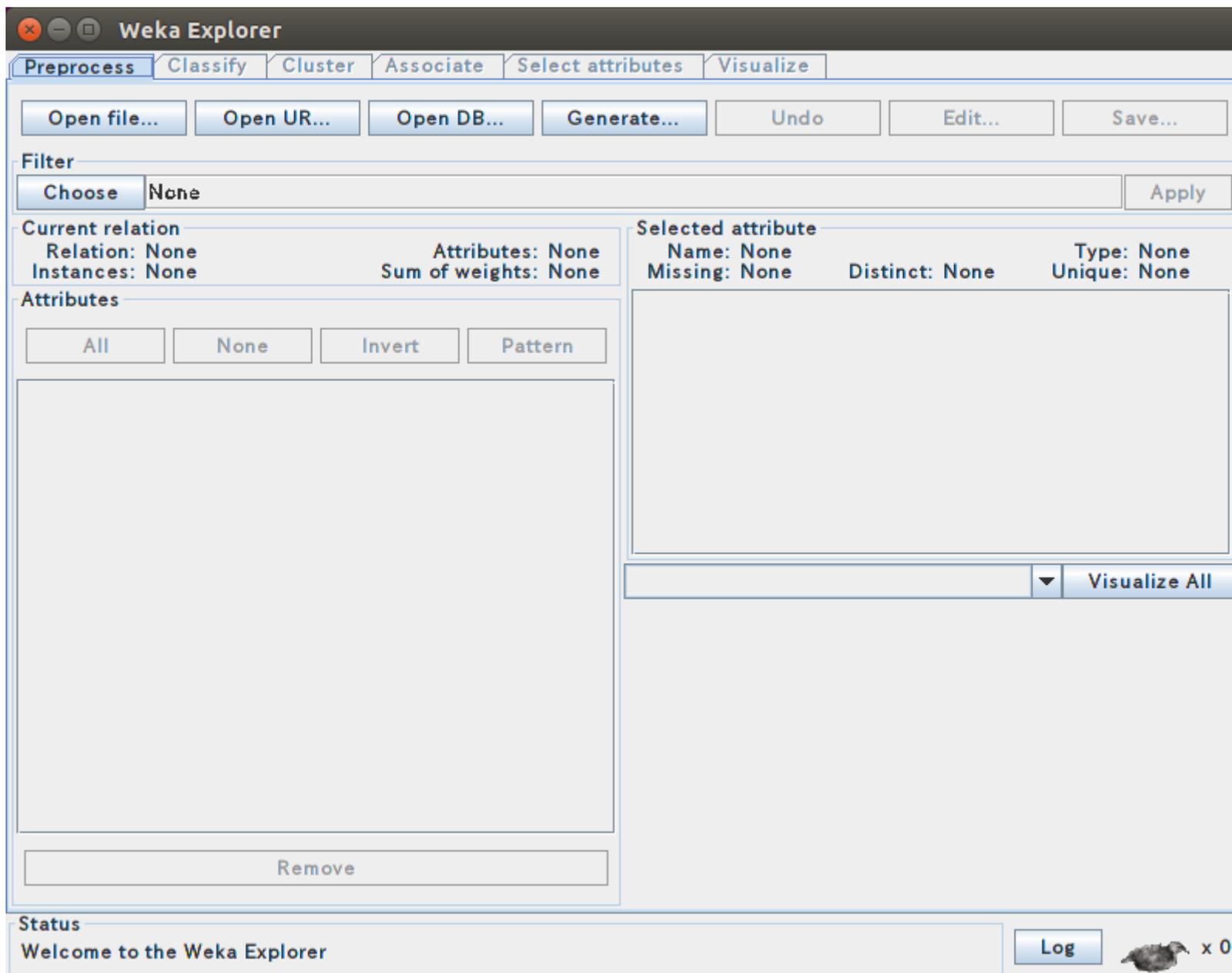
トレードオフ

Weka の起動画面



3.7.X: 開発版
機能が豊富
3.6.X: 安定版
日本語の UI

エクスプローラー



学習データを開く

読み込んだ
データの表示

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is active. The 'Open file...' button is circled in red. Below the buttons, the 'Filter' section shows 'Choose' and 'None' options. The 'Current relation' section displays 'Relation: iris' and 'Instances: 150'. The 'Selected attribute' section shows 'Name: sepallength' with statistics: Minimum (4.3), Maximum (7.9), Mean (5.843), and StdDev (0.828). The 'Class: class (Nom)' is selected, and a histogram is displayed below it. The histogram shows the distribution of sepal length values across three classes, with counts of 16, 30, 34, 28, 25, 10, and 7 for different bins.

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Bin Range	Count
4.3 - 4.7	16
4.7 - 5.1	30
5.1 - 5.5	34
5.5 - 5.9	28
5.9 - 6.3	25
6.3 - 6.7	10
6.7 - 7.1	7

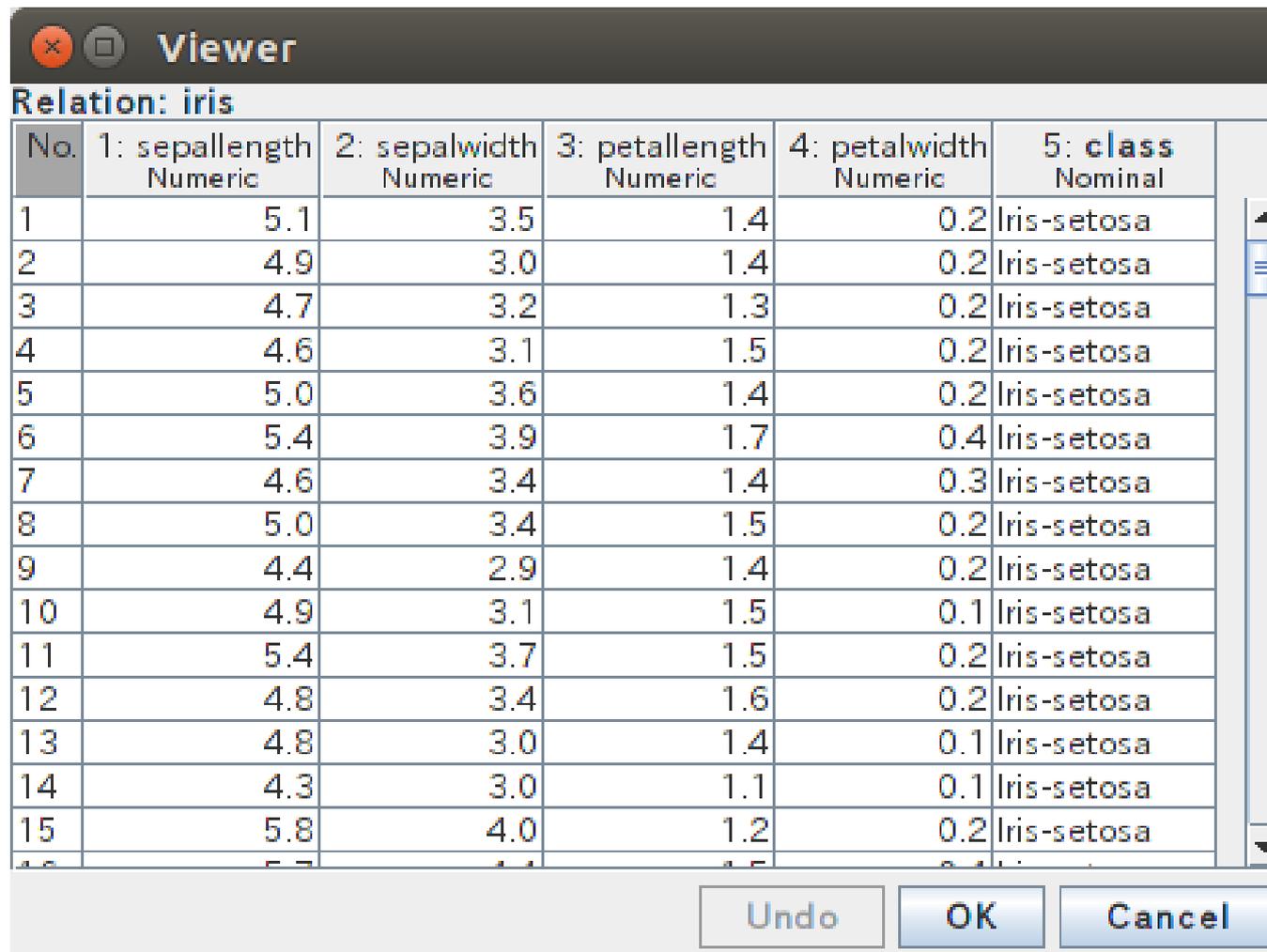
前処理はここで

Status
OK

Log

x 0

iris データ

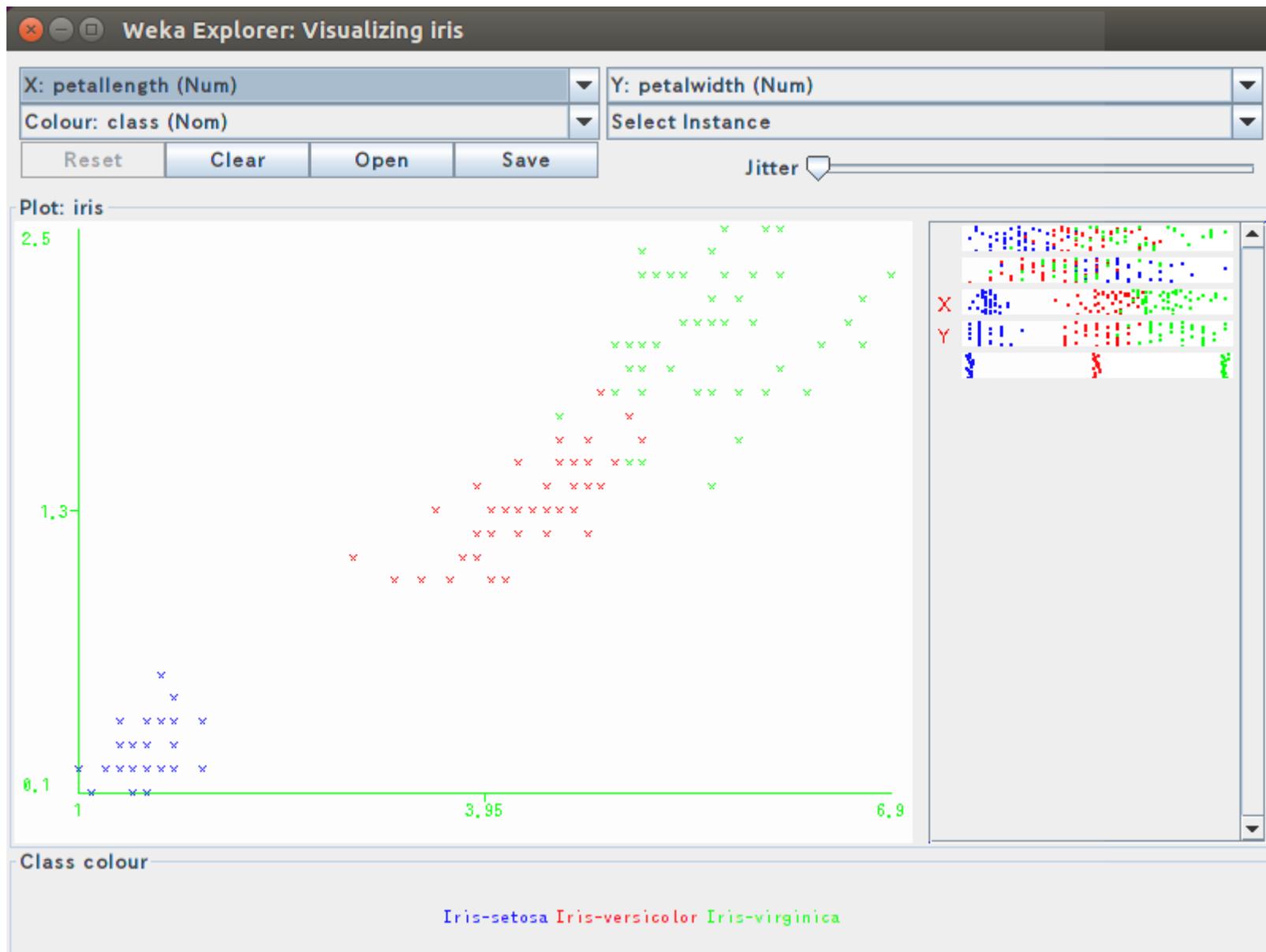


Relation: iris

No.	1: sepallength Numeric	2: sepalwidth Numeric	3: petallength Numeric	4: petalwidth Numeric	5: class Nominal
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3.0	1.4	0.1	Iris-setosa
14	4.3	3.0	1.1	0.1	Iris-setosa
15	5.8	4.0	1.2	0.2	Iris-setosa

Undo OK Cancel

データの可視化



学習実験

1. 学習アルゴリズムを選択

1. パラメータを調整

2. 評価方法を選択

3. 学習開始

4. 結果

The screenshot shows the Weka Explorer interface. The 'Classify' tab is active. The classifier is set to 'IBk' with parameters: `-K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A %*weka.core.EuclideanDistance -K first-last%"`. Under 'Test options', 'Cross-validation' is selected with 'Folds' set to 10. The 'Start' button is circled in red. The 'Classifier output' pane shows the following results:

Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	143	95.3333 %
Incorrectly Classified Instances	7	4.6667 %
Kappa statistic	0.93	
Mean absolute error	0.0399	
Root mean squared error	0.1747	
Relative absolute error	8.9763 %	
Root relative squared error	37.0695 %	
Coverage of cases (0.95 level)	95.3333 %	
Mean rel. region size (0.95 level)	33.3333 %	
Total Number of Instances	150	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area
1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
0.940	0.040	0.922	0.940	0.931	0.896	0.952	
0.920	0.030	0.939	0.920	0.929	0.895	0.947	
Weighted Avg.	0.953	0.023	0.953	0.953	0.953	0.930	0.966

=== Confusion Matrix ===

```
a b c <-- classified as
50 0 0 | a = Iris-setosa
0 47 3 | b = Iris-versicolor
0 4 46 | c = Iris-virginica
```

Status: OK

パラメータの調整

