知能システム開発特論:
第2回

データマイニングの基礎

ダビド (david@iwate-pu.ac.jp)
IPU
4 December 2014

データマイニングとは？

- データマイニング(データベースからの知識発見):

  興味深い(当たり前でない、潜在的、これまで知られていなかった、しかも、役に立つと思われる)情報あるいはパターンを大規模データベースから抽出すること

- データマイニングの別名

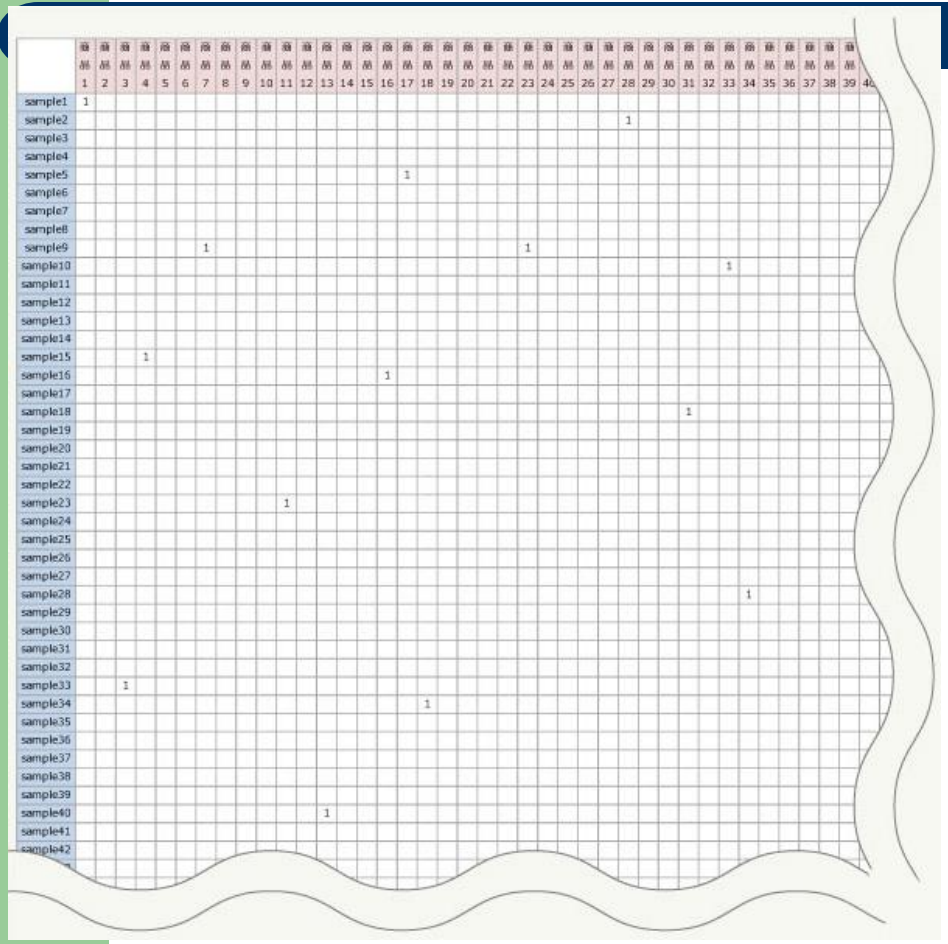
  データベースからの知識発見(Knowledge discovery in databases, KDD)、知識抽出、データ/パターン解析、データ考古学(archiving)、data dredging, 情報収穫(harvesting)、ビジネスインテリジェンス、など

2種類のデータマイニング

データマイニング	統計解析
データ量が多い 知識発見	データ量が少ない 仮説検証

- 仮説検証（目的志向）的データマイニング
  - 推定、把握（量的変数）
  - 分類、抽出（質的変数）
  - 将来の予測
- 知識発見（探索）的データマイニング
  - アソシエーションルール策定
  - クラスタリング
  - 両者で用いられるデータマイニング
  - グループの特徴を推測する（プロファイリング）

データマイニングで扱うデータ



- 数百万人以上の顧客が、数万点の商品の中からいくつかの商品を購入したデータなので、ほとんどがNULLだと言っても過言ではありません。

データ(Data)

- 生データ
  - 記憶装置に貯えられたファイル（画像、音声、動画、テキスト、WEBデータ、データレコード、など）
- データベース
  - ファイルの集合体（ディレクトリ、フォルダ）、多くのデータレコードの集合体
  - 規模：キロバイト、メガバイト、ギガバイト
- 関係データベース（リレーションで表現されている）
  - レコード(record)、属性(attribute)、表の行、表の列から構成されているデータベース
- LOD（リンクされた公開データ）：WWCの規格化言語から記述したデータ
- ビッグデータ：巨大、複雑なデータ集合(例：ウイキペディア、RFID)

データ(Data)

- 属性(attribute): 物の性質、特徴 (property or characteristic of an object)

例 : 人間の目の色、温度、湿度、など

別の名 : 変数、フィールド

- オブジェクト : 複数の属性から記述する。

- オブジェクトは物、レコード(record)、サンプル、エンティティ、インスタンス、などとも呼ばれる。

データ行列(DATA MATRIX)

$n \times d$ データ行列 (data matrix), n 行(rows) and d 列(columns)

行 (rows) = データセットのエンティティ(entities in the dataset)

列 (columns) = 属性、プロパーティ(attibutes or properties)

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

x_i : i-th row d-tuple : $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$

X_j : j-th column n-tuple : $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$

row=={instance, example, record, transaction, object, point, feature-vector, tuple}

column=={attribute, property, feature, dimension, variable, field}

n : データのサイズ(size of the data)

d : データの次元(dimensionality of the data)

データ(Data)

- すべてのデータセットはデータ行列ではなく、順列、テキスト、時系列、画像、音声、ビデオ、などもある。
- 特徴抽出による上記のデータはデータ行列に変換できる。
- データ分析はエンティティごとに独立ということを想定している。実世界にはそれぞれのエンティティはさまざまな関係もある。データグラフのモデルも考えられる。

実験データと観測データ

- 実験データはすべての属性を明確に定義されている。機械学習のデータセットを利用する。

- 観測データでは属性は定めてない。データマイニングのデータとして扱う。

属性の種類:量的変数 vs 質的変数

- 量的変数(quantitative or numeric):

例:重さ、温度、物の数

- 質的変数(qualitative or categorical):演算不可能

例:人間の目の色、郵便番号、IPアドレス、

数字の桁{0,1,..,9}、

Iris花のクラス{Setosa,Versicolor,Virginica}

目標変数(target variable):二つのカテゴリの表現が多い。

属性の種類:

- 量的変数(quantitative or numeric):

大小関係がある。近い値は本質的に似ている。

間隔尺度(Interval (no “true” zero)) or 比尺度(Ratio (true zero exists))

- 質的変数(qualitative or categorical):

順序あるかない属性(名義尺度(Nominal) or 順序(Ordinal))

例:

身分証明ID、郵便番号、目の色はNominal

高さ、成績、ランク付データはOrdinal

日付、カレンダー、温度数字、GPA成績はInterval

属性値の性質:

- 特殊性（distinctiveness）: = , !=
- 順序 (order): <, >
- 演算:　+, -
- 掛け算: x, /

Nominal : 特殊性

Ordinal:特殊性,順序

Interval:特殊性,順序,演算

Ratio:すべての性質

離散属性　vs 連続属性

離散属性（Numeric attribute）

属性の領域：domain(Age) = $\mathbb{N}$, domain(petal length)= $\mathbb{R}^+$

種類：離散、連続、バイナリ

カテゴリ型属性（Categorical attribute）

属性の領域：domain(血液型) ={A,B,AB,O}, domain(性別)={M,F}

演習：データの種類を定義する

a）あなたの家の電話の数
b）フライド ポテトのサイズ (小、中、大)
c）携帯電話の所有権
d）１ヶ月で行われた国内の電話
e) 最も長い電話の長さ
f)あなたの足の長さ
g)教科書の価格
h) Zip コード
i) 温度(F)
j）温度(C)
k) 温度 (K)

出力:クラスラベル、量的データ

- 予測問題の出力は二つある。
- 量的な出力を予測する問題は回帰(Regression):値を返す
- 質的な出力を予測する問題は分類(Classification):クラスラベルを返す
- 両タスクは関数近似(function approximation)の問題と呼ぶ。

プロセスの各ステップ

- 応用領域についての習熟:
  - –適切な前提知識と応用の目的の明確化
- 目標データセットの作成: データ選択
- データ洗浄と前処理: (場合によっては全体の60%の労力を要する!)
- データの縮小と変換:
  - –有効な特徴の抽出、次元/変数の縮小、平滑化、集約、一般化、正規化、新属性の構築
- データマイニングの機能の選択:
  - –要約、分類、回帰分析、関連ルール、クラスタリング
  - –マイニングアルゴリズムの選択
- データマイニング: 興味あるパターンの探索
- パターンの評価と知識の表現:
  - –視覚化、変換、冗長パターンの除去など。
- 発見された知識の利用

データマイニングの機能

- 分類と予測
  - –予測を目的として、クラスや概念を区別するための記述を求める。
  - –例国を気候により分類する。自動車を燃費で分類する。
  - –表現: 決定木、分類規則、ニューラルネットワーク
  - –予測:未知の、あるいは欠落した数値(missing values)を予測する。
- クラスター分析
  - –分類カテゴリが未知: 新しいクラスを作るためにグループ化する。たとえば、分布パターンを見つけるために顧客の住居をクラスター化する。
  - –クラスタリングの原理: クラス内の類似性を最大化し、同時に、クラス間の類似性を最小にする。

サンプリング

- 母集団から標本（サンプル）を無作為抽出する（ランダムサンプリング）
- すべてのデータを使用すると、処理が遅すぎるとすべてのデータも不要かもしれない。
- 理想：母集団とサンプルが同じ性質を持っている。
- 例：文章コーパスから構造化データのサンプリング

サンプリング

- 量的属性の扱い、一つの変数(univariate)

サンプリング Univariate

f, F, μ, σ, r: 母集団の確率分布,累積分布関数,平均,標準偏差,レンジ

$\hat{f}, \hat{F}, \hat{\mu}, \hat{\sigma}, \hat{r}$:サンプルの内容

サンプリング

- 量的属性の扱い、二つの変数(bivariate)

サンプリング Bivariate

$\hat{f}, \hat{F}, \hat{\mu}, \hat{\sigma}, \hat{\Sigma}, \hat{\rho}$：ベクトル平均、共分散、相関係数

サンプリング

- 量的属性の扱い、二つ以上の変数 (multivariate)

サンプリング Multivariate

$\hat{f}, \hat{F}, \hat{\mu}, \hat{\sigma}, \hat{\Sigma}$:ベクトル平均、共分散、行列共分散

データの正規化

- レンジ正規化 (range normalization)
- 標準スコアー正規化(standard score normalization)

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \cdots & x_d \end{pmatrix}^T$$

Range Normalization range in [0,1]

$$x_i' = \frac{x_i - min_i x_i}{\hat{r}} = \frac{x_i - min_i x_i}{max_i x_i - min_i x_i}$$

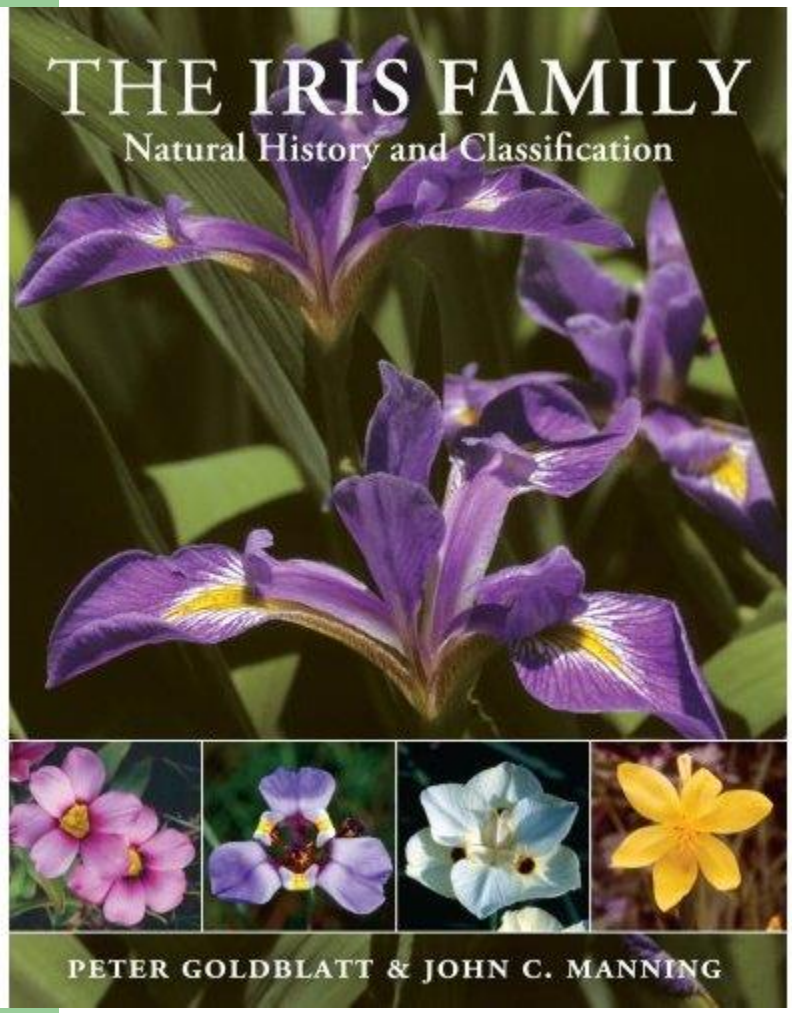
Standard Score Normalization

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

学習の手法

- 情報から従来使えそうな知識をどんな方法で見つけ出す
- 例題（事例）からアルゴリズムに基づいて概念記述（structural descriptions)を取得
- 概念記述には明示的にパターンを定義される

問題定式化



- 入力:
- C: クラス（classes）
- F: 特徴(Features)
- 出力:
- 特徴データとクラスの分類

データセット例(Iris dataset)

Setosa

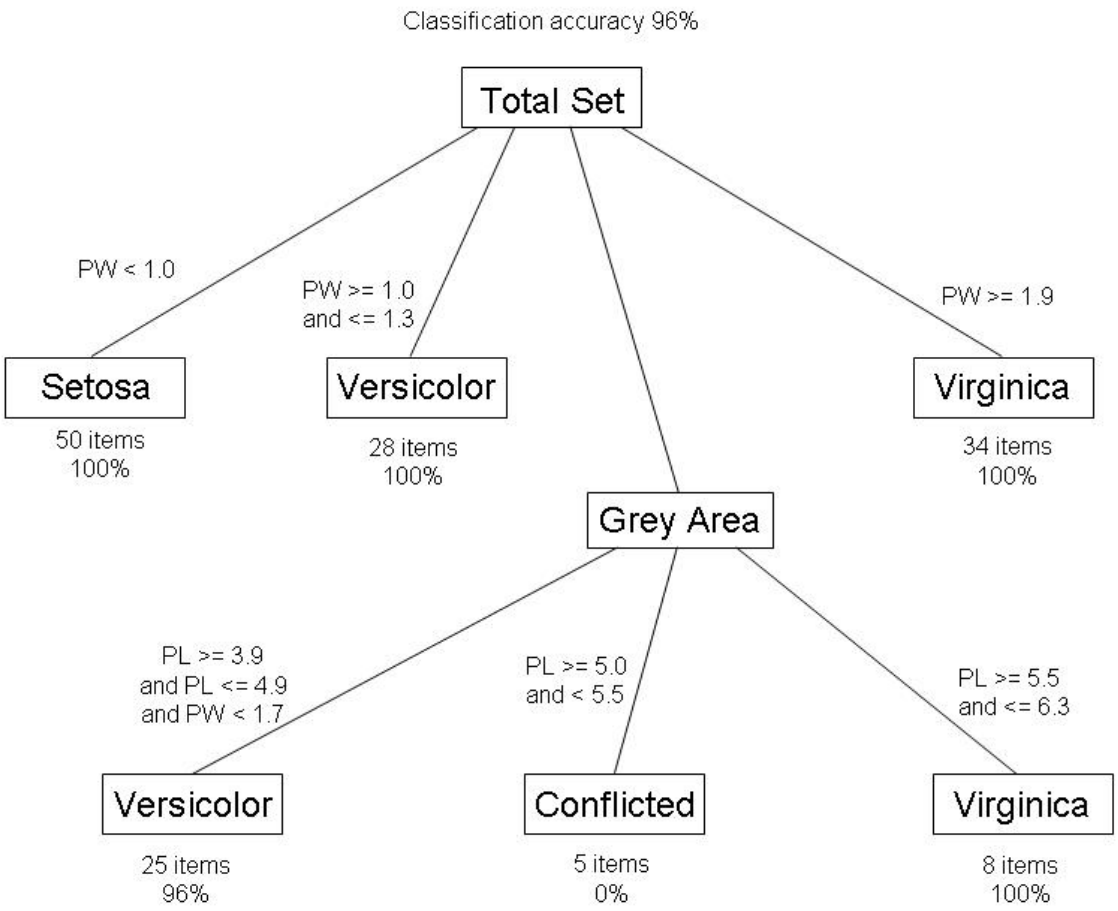
D1:{PL=1.4,PW=0.2,SL=5.1,SW=3.5}
D2:{PL=1.4,PW=0.2,SL=4.9,SW=3.0}
D3:{PL=1.3,PW=0.2,SL=4.7,SW=3.2}
…

D51:{PL=4.7,PW=1.4,SL=7.0,SW=3.2}
D52:{PL=4.9,PW=1.5,SL=6.4,SW=3.2}
D53:{PL=4.0,PW=1.5,SL=6.9,SW=3.1}
…

D101:{PL=6.0,PW=2.5,SL=6.3,SW=3.3}
D102:{PL=5.1,PW=1.9,SL=5.8,SW

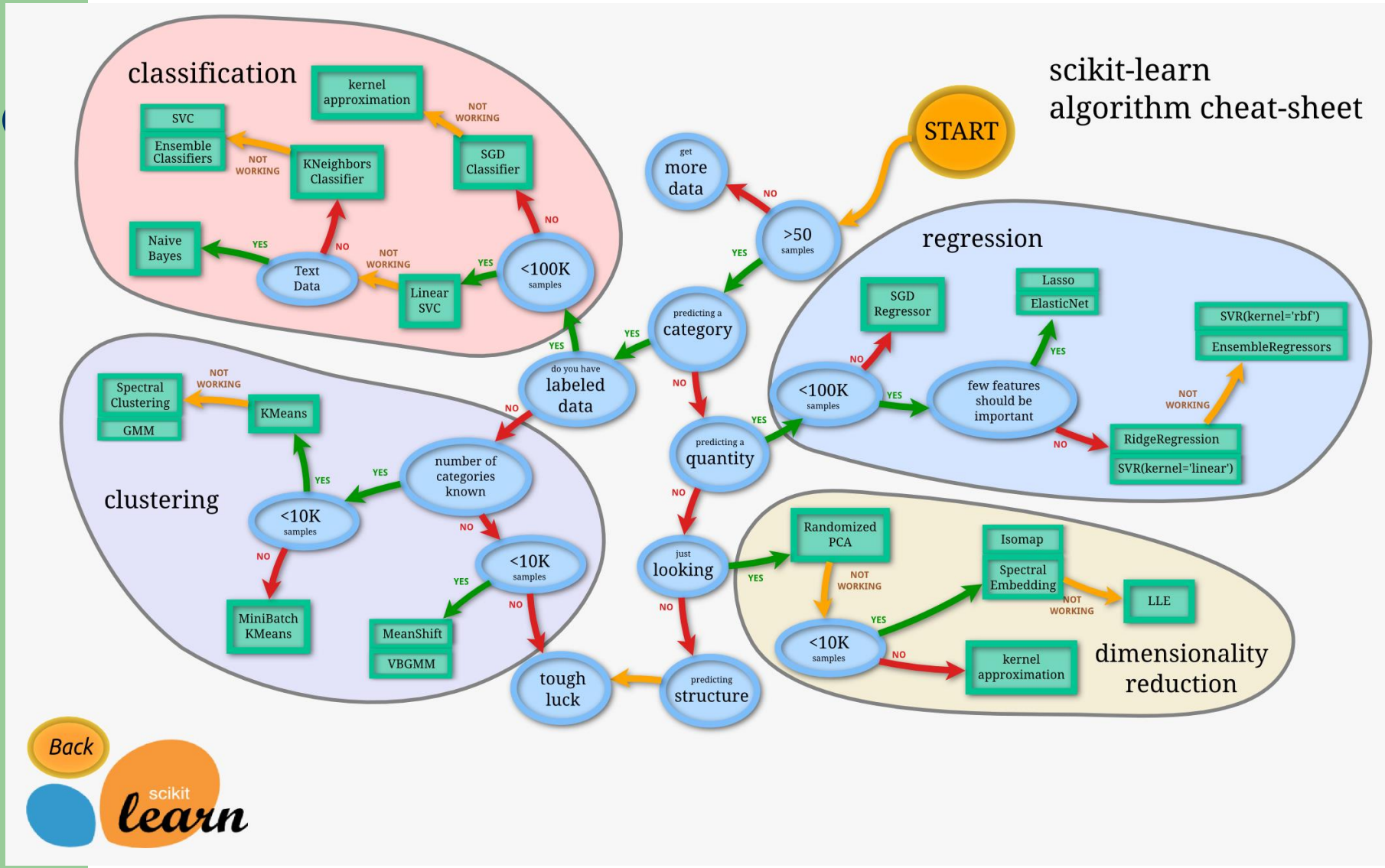
- 3つのクラスに50サンプル毎に収集
- 特徴
  - **PW: Petal Width**
  - **PL: Petal Length**
  - **SW: SepalWidth**
  - **SL: Sepal Length**

データセットの分類(Iris dataset)



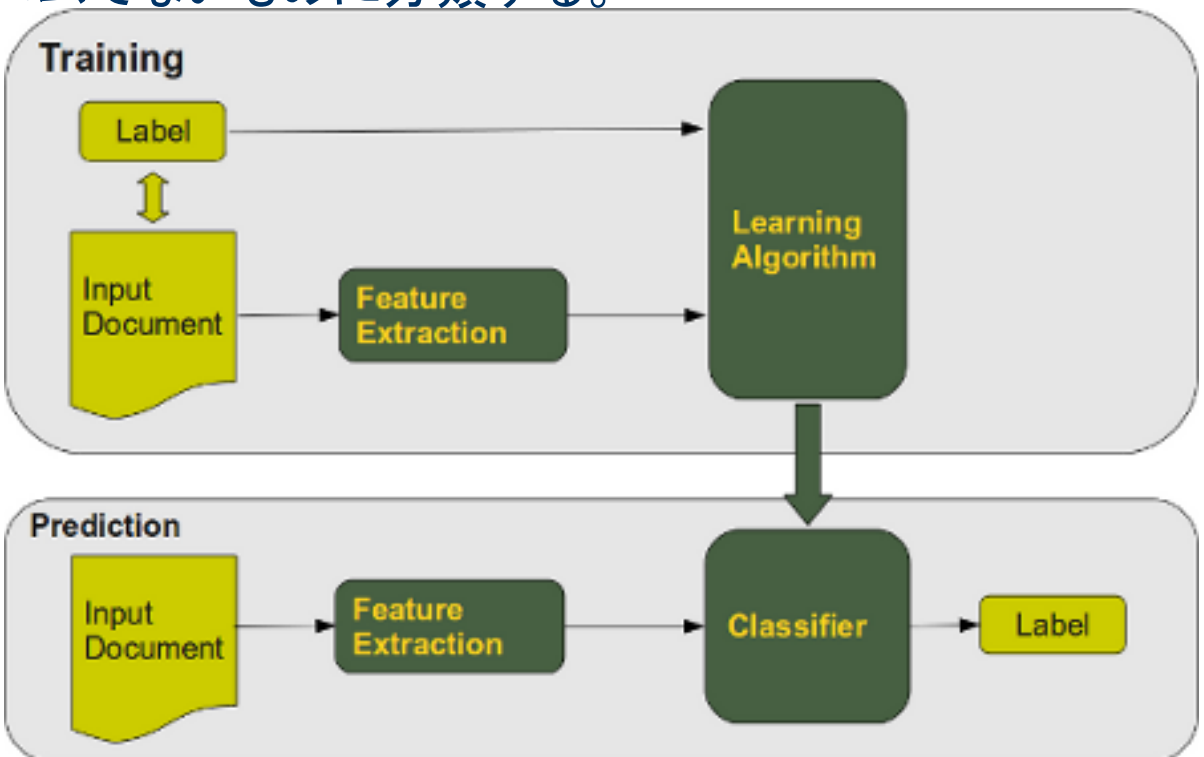
- 特徴
  - PW: Petal Width
  - PL: Petal Length

分類器/クラスタ分析器/回帰など

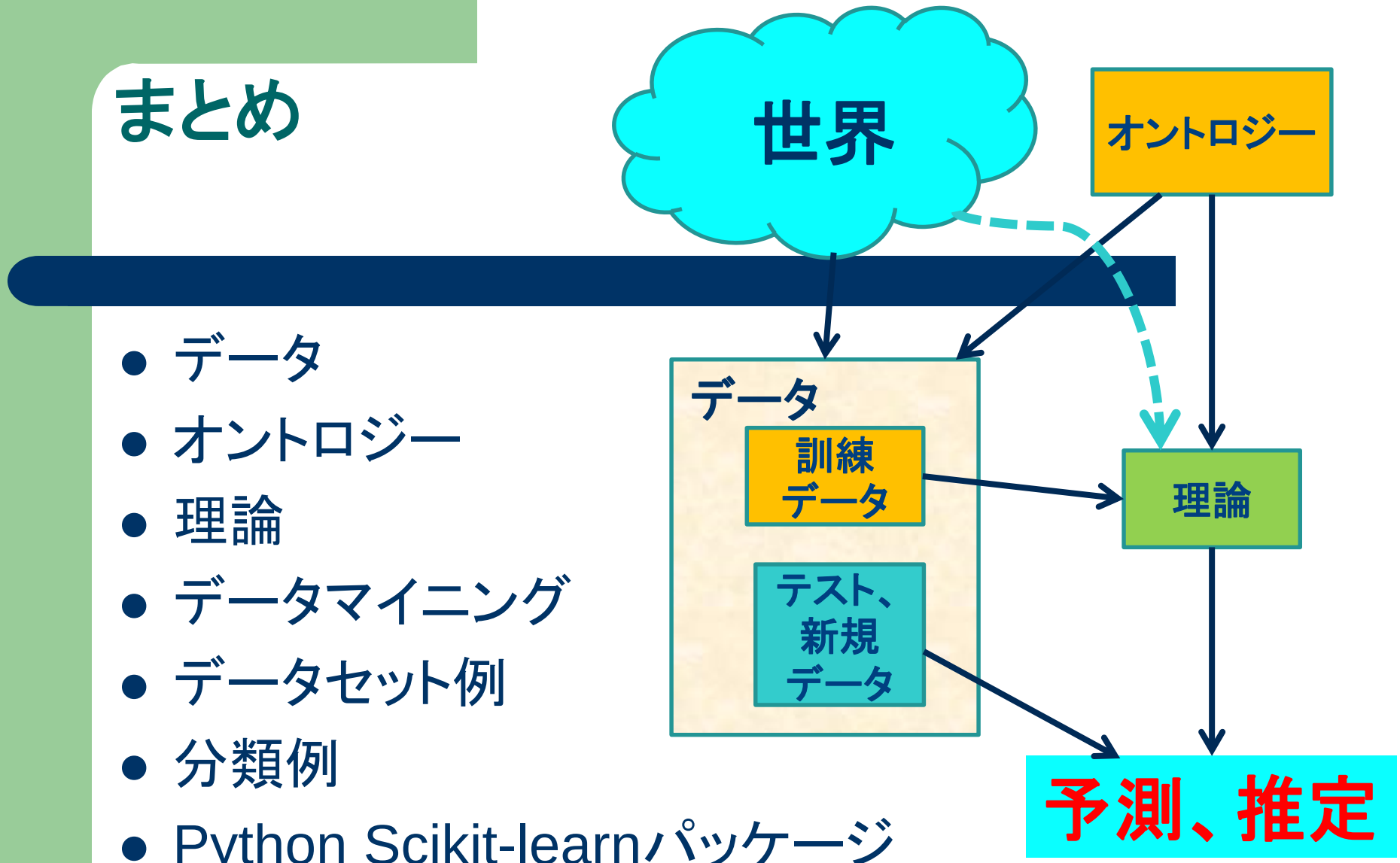


文書の自動分類

- 単純ベイズ分類器を文書分類問題に適用した例を示す。文書群をその内容によって分類する問題であり、例えば、電子メールをスパムとスパムでないものに分類する。



まとめ



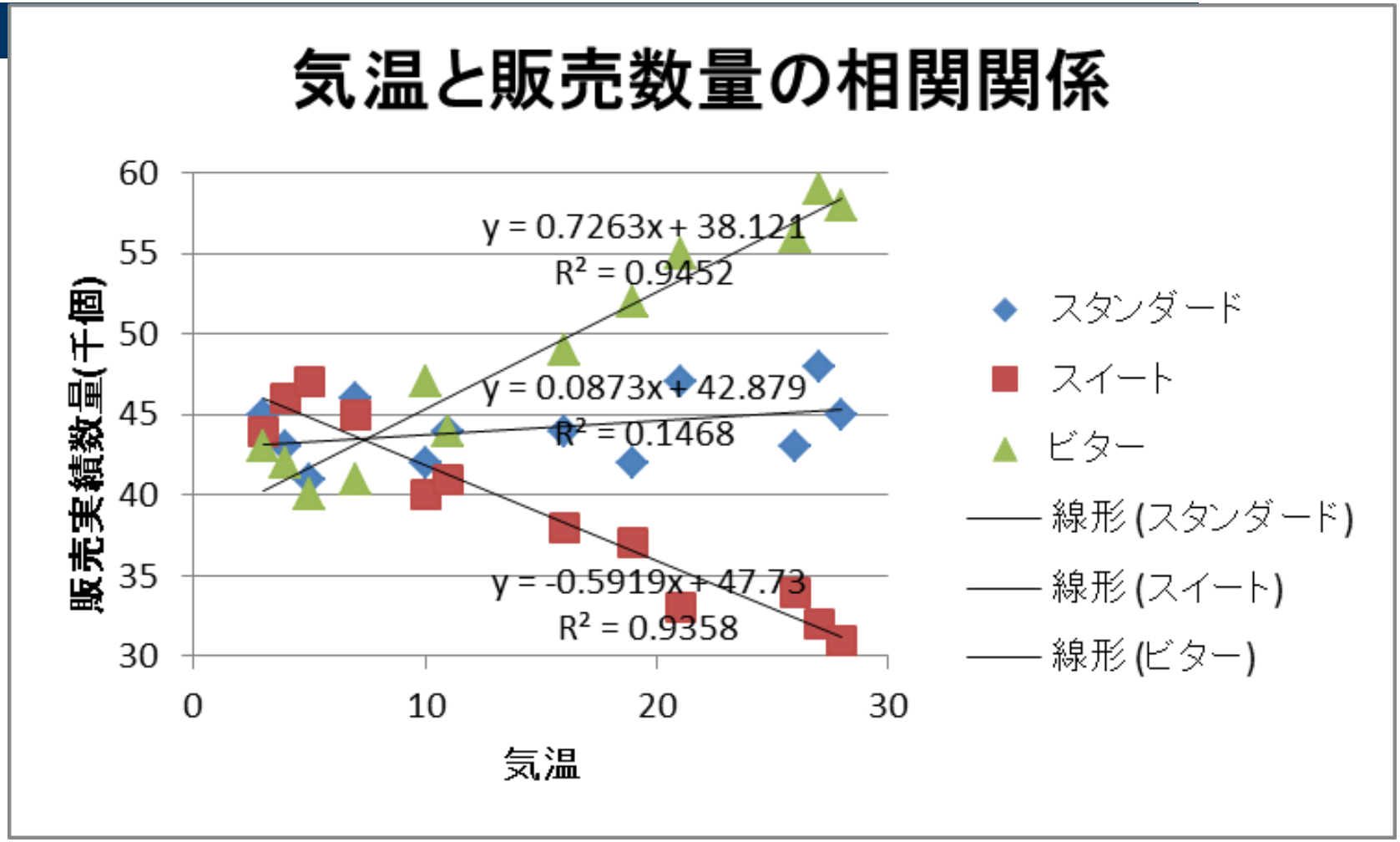
- データ
- オントロジー
- 理論
- データマイニング
- データセット例
- 分類例
- Python Scikit-learnパッケージ
- データテキストの分類器

EXAMPLE

データ例:コーヒのテストを分析する

月	昨年の実測気温	スタンダード	スイート	ビター
1	4	43	46	42
2	3	45	44	43
3	5	41	47	40
4	10	42	40	47
5	16	44	38	49
6	21	47	33	55
7	27	48	32	59
8	28	45	31	58
9	26	43	34	56
10	19	42	37	52
11	11	44	41	44
12	7	46	45	41

回帰分析の例


気温と販売数量の相関関係
販売実績数量(千個)
気温
60
55
50
45
40
35
30
0
10
20
30
y = 0.7263x + 38.121
R² = 0.9452
y = 0.0873x + 42.879
R² = 0.1468
y = -0.5919x + 47.73
R² = 0.9358
スタンダード
スイート
ビター
線形 (スタンダード)
線形 (スイート)
線形 (ビター)

気温予測に基づく販売数量		
本年の予測気温	スイート	ビター
8	42.99529	43.93098
10	41.8115	45.38353
9	42.4034	44.65725
14	39.44392	48.28863
16	38.26013	49.74118
23	34.11686	54.8251
28	31.15739	58.45647
29	30.56549	59.18275
26	32.34118	57.00392
20	35.89255	52.64627
9	42.4034	44.65725
6	44.17908	42.47843

	前年比	
月	スイート	ビター
1	−7%	5%
2	−5%	6%
3	−10%	12%
4	−1%	3%
5	1%	2%
6	3%	0%
7	−3%	−1%
8	−1%	2%
9	−5%	2%
10	−3%	1%
11	3%	1%
12	−2%	4%

1年間の販売数量の予測