データサイエンス特論
Data Science

パターン認識
Pattern recognition

1. パターン認識とは
2. 性能評価，トレーニングデータとテストデータ
3. クラス分類（教師あり学習）
   - サポートベクターマシーン（SVM）
   - k-近傍法
4. クラスタリング（教師なし分類）
   - 群平均法
   - k平均法

パターン認識とは

1. パターン
2. クラス分類問題

パターン認識という考え方は、紀元前のPlato, Aristotleの提唱が嚆矢と言われている。特に、Aristotleは、物理的世界を観察し、それらを一般化する概念を打ち出している。

パターンとは

パターン

- 右に示す葉を区別するために，「特徴量」を設ける
- 例えば，サイズ，色，形状など
  - とはいえ，これら三つを数値やラベルで分類するのはさらに多数の項目に分類
  - この特徴量をどう決めるかも厄介な話であるが，ここでは，簡単に考える
- 数字の「0」，「1」，・・・，「9」の特徴量は？

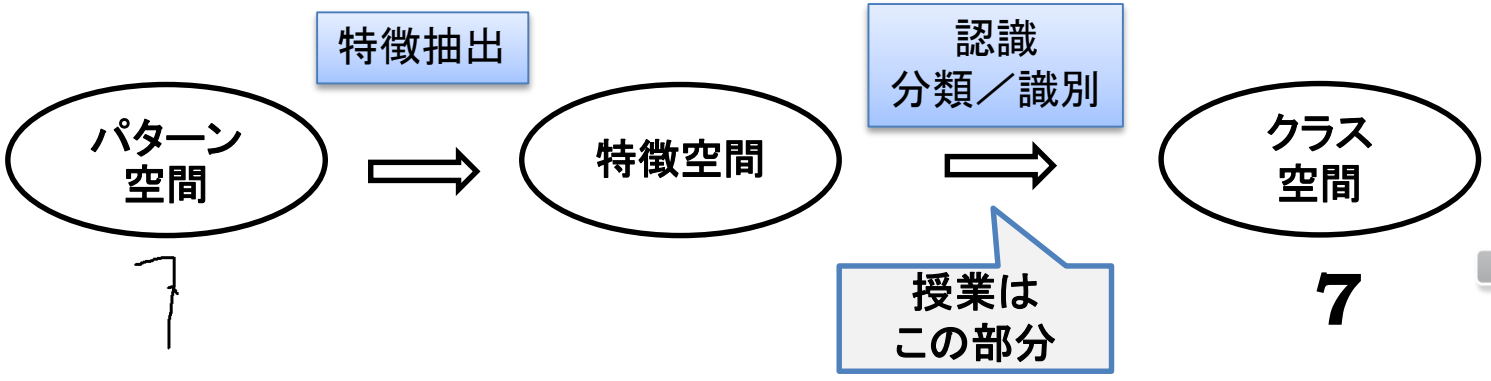
引用　https://sozaikoujou.com/category/clipart

パターン認識

- これらの特徴量を手掛かりに，クラス（class）に分類すること
- クラス分類（classification）とも呼ぶ

パターン認識として次の二つの問題を取り上げる

- クラス分類，クラスタリング
- 他にも，パターン認識の範疇に入る話題は幾つかあるが，この二つに限定，しかも，この二つにはそれぞれ多数の手法があるが，その中でも，限定して話を進める。



クラス分類

リンゴの分類

- バラ科に属するリンゴは多数の種類がある。この分類を考えたとき，手始めに，色，サイズ，重さ，形状（果肉，ツルもと），を設ける。
- 下記の表は，この考えに基づくものである。もちろん，専門的にはもっと深い分類が必要であるが，練習用であるので，この分類でご容赦願いたい。
- このとき，リンゴの名称をクラスとする。
- ある名称不明のリンゴを見たとき，下記の表に照らし合わせて，このリンゴの名称（クラス）がなにであるか，すなわち，どのクラスに属するかの分類を行う。これがクラス分類である。
- この種類の分類を行うとき，何らかの類似度を導入して，類似度が高ければ同じ種とするものとする。この類似度をどう考えるか？　これを考えるため，もう少し簡単な例を用いる。

属性	何データ？
色	赤，黄，青
サイズ	S，M，L
重さ	S，　M，　L
形状（果肉）	円形，ハート型，楕円型
形状（ツルもと）	深い，浅い

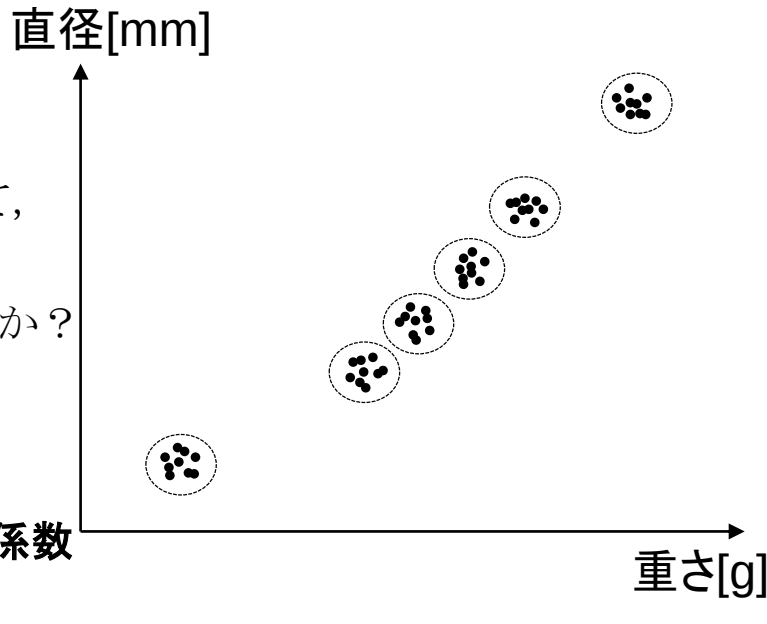
参考文献:

- リンゴミュージアム:http://www.ringomuseum.com
- リンゴ大学:http://www.ringodaigaku.com

クラス分類

コインの分類

- 100枚のコインがあるとして，この分類を考える。単純に考えて，直径[mm]と重さ[g]で分類することを考える。
- 100枚のコインの直径と重さの散布図
- （日本のコインはもっと精度が高い）
- 右図で，直径と重さが類似したコインは同じ種類として，破線で囲んだような分類を行いたい
- では，同じ破線内のコインが類似しているとどう言えるか？



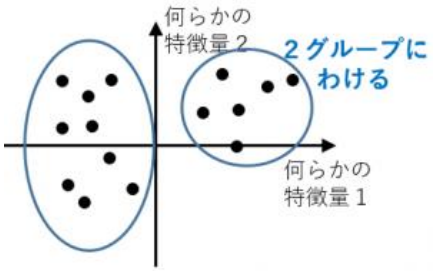
類似度

- 右図を見ると**距離**で測るというアイディアが浮かぶ。
- 距離以外には，**内積のコサイン**や**ピアソンの相関係数**を用いて類似度を見る考え方がある（他書参照）。
- 類似度を距離で測ることとする。この説明は付録に。
- 今は，距離は，ものさしで直線を測ることをイメージしてほしい。

クラスタリング

- **cluster は房，集団，群れのように塊りを表す。ブドウのクラスタ（下図）は，一粒一粒が集まって房（クラスタ）を形成し，そのクラスタが三つあることを示している。**
  - clusterの他の用例に，星団（star cluster），コンピュータクラスタ（computer cluster），水クラスタ（water cluster）などがある。
  - clusterに分類することをクラスタリング(clustering)という。
  - クラスタ解析，クラスタ分析とも称される。
- **目的変数のない（教師なしの）場合に用いられる。**
  - データには特徴量（性別，年齢など）があるが，どのように区別したらよいかがわからない（教師なし）。
  - これらの条件の下で，いくつかのグループ分けを行うことがクラスタリングである。

ブドウのクラスタ



クラスタリングのイメージ

https://jp.123rf.com/photo_42376154_白い背景の上のつるブドウのクラスター.html

クラスタリング

特徴

- 統計分野では多変量解析の一種
- 教師なしデータを対象とする
- アルゴリズムは比較的簡単

用途

- 顧客，店舗，製品，動植物の特徴データに対して，特徴が近いものをグループ化することで，どのグループがどのような特徴を有するかを見る（留意点を見よ）。

アルゴリズム

- 階層型クラスタリング
  - 群平均法(group average method). UPGMA (unweighted pair-group using arithmetic averages)., ウォード法（Ward‘s method）など
- 非階層型クラスタリング
  - k平均法（ k-means）など

留意点

- クラスタリングは探索的 (exploratory) なデータ解析手法であって，分割は必ず何らかの主観や視点に基づいているということである. よって，クラスタリングした結果は，データの要約などの知見を得るために用い，客観的な証拠として用いてはならない.（神嶌敏弘；データマイニング分野のクラスタリング手法 (1)，人工知能学会誌，vol.18,no.1,pp.59-65, 2003）
- すなわち，利用目的に応じて，クラスタリングの結果を検証・評価する必要があります。また，得られたクラスタは絶対的・客観的なものではなく，あくまで一つのある視点から見た結果であることに留意してください。

性能評価、
トレーニングデータとテストデータ

距離と類似度

距離が測れるとは，差が定義できること

- 数学的な距離の説明
- 量的データは距離が測れる。質的データの距離を測ることは意味が無い。ただし，数学の位相論を導入すると質的データを位相の意味で測ることはできるが，数値表現とはならないので，ここでは扱わない。
- アンケート結果やランキング（順位）データの距離を測るのは，少しおかしなことになるので，距離で測るのではなく，他のmeasure（数学の意味での）を導入されたい。

量的データ（数値データ）

- 比率データ（比率尺度，四則演算可）
  - ✓ 比率に意味がある。例：スカイツリーの高さ（634m）は東京タワー（333m）の1.904倍である（この例は加減も可）。
- 間隔データ（間隔尺度）
  - ✓ 加減に意味がある。例：水温20度は，水温10度よりも10度高い（温度が2倍とはあまり言わない）

質的データ

- 順位データ（順位尺度）
  - ✓ 順序に意味がある。例：成績　S＞A＞B＞C＞D，アンケート結果（5：非常に良い，4：よい，3：ふつう，2：悪い，1：非常に悪い）
- カテゴリデータ（名義尺度）
  - ✓ 分類，区分に意味がある。例：サイズL，M，S。電話番号。形式的に数字を当てはめることもある，1：はい，2：いいえ。

距離と類似度

類似度を距離で測る

- 類似度を数値で表現しましょう，という意味。
- この考え方では，先のリンゴの分類も，複数の特徴量を数値で表すこととなる。
- 先のコインの分類では，距離の近いものをグルーピングして，分類を行う方法を考える。

余談

- コインの例を考える。コインの種類を人間は視覚的に分類できる。この行為は視知覚におけるゲシュタルトの法則（Gestalt Principle）の近接の法則（Principal of Proximity）に基づくとされている。コインのクラス分類は，この自動化を行っていることになるが，ゲシュタルト則とは異なる方法で分類している。自動化とは，人間と異なる考え方や方法を用いることといえる。

分類器の性能評価

2レベル識別器の場合

混同行列 (Confution Matrix)

		分類クラス	
		Positive	Negative
真のクラス	Positive	True Positive (TP)	False Negative (FN)
	Nega	False Positive (FP)	True Negative (TN)

正確度 (Accuracy) $= \dfrac{TP + TN}{TP + FP + FN + TN}$

適合率 (Precision) $= \dfrac{TP}{TP + FP}$

再現率 (Recall) $= \dfrac{TP}{TP + FN}$

正確度 (Accuracy)

- 推定した値と真の値が一致した割合

適合率 (Precision)

- モデルが陽性と判断したものの中に、どれだけ本当に陽性なものが含まれていたかを示す指標

再現率 (感度、Recall)

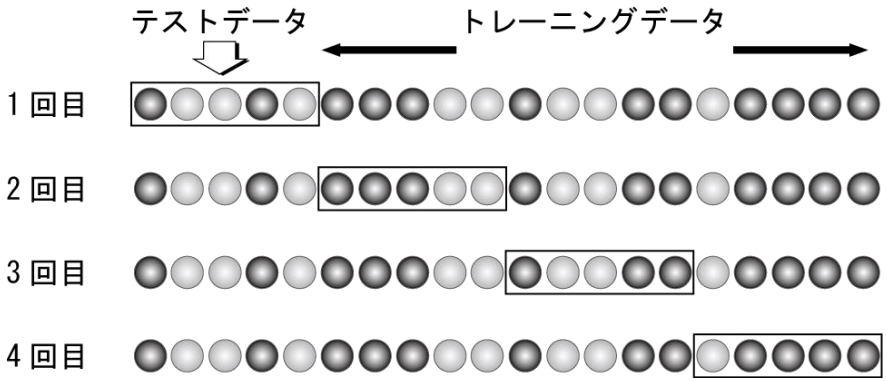
- 実際に陽性だったもののうちモデルが陽性と判断したものの割合

F-値 (F-score, F-measure)

- 適合率と再現率の調和平均，2つの指標を総合的に見るときに用いる。F検定のFとは異なる。

トレーニングデータとテストデータ

- **トレーニングデータ（training data）**
  - クラス分類を試みるために用いられ，クラス分類器を作成する
- **テストデータ（test data）**
  - クラス分類器が他のデータでも有効であるかを確認するためのデータ
- **ホールドアウトと交差検証法（holdout and cv: cross validation）**
  - データの数をそれほど多くは持てない
  - ホールドアウト
    - 1つのデータセットを，例えば，トレーニングデータに7割，テストデータに3割分割する方法
    - 簡易であるが，分離した片方に偏りのデータがあった場合に，クラス分類器の性能が保証できない
  - 交差検証法
    - データセットをk個に分割し，そのうちの(k-1)個をトレーニングデータ，残りの1つをテストデータに割り当てる。
    - この割り当てを，順繰りk回繰り返し，計算結果の評価を行う。



扱うパターン認識

クラス分類（教師あり学習）

- SVM（Support Vector Machine）：有力な分類性能を有する，非線形にも対応
- kNN（k-Nearest Neighbors）：アルゴリズムは単純だが，比較的性能は良い。ただし，ノンパラメトリックのため，数式でクラス分類できない。

クラスタリング（教師なし学習）

- 階層型
  - 類似度の高いもの（距離の近い）から順にまとまり（クラスタ）を作成する方法で，この途中の過程が階層のように表せる。このため，樹形図（デンドログラム，dendrogram）で表すことができる。
  - 凝集型 (agglomerative)と分割型 (divisive) がある。凝集型のみを説明。
  - 凝集：細かいものが散らばっていて，それが一つに集まり，固まること。「勢力を凝集させる」。
- 非階層型
  - 分割最適化手法の他，partitional や optimization など多くの呼び方がある。
  - 分割の良さの評価関数に基づいて分割を探索する。
  - 異なる性質のものが混じり合ったデータに対して，予めクラスタ数を与え，クラスタを順に作成する手法で，その過程が階層的でないことから非階層型と名付けられた。
  - 階層型クラスター分析と比較して，サンプル数が大きいビッグデータを分析するときに適している。ただし、予め与えるクラスタ数を幾つにするかの決め方に課題を残す。
  - k-meanを説明する。

scikit learn

全体の説明

- User Guide
- http://scikit-learn.org/stable/user_guide.html

- 1. Supervised learning（教師付き学習）
  - http://scikit-learn.org/stable/supervised_learning.html
  - 1.4. Support Vector Machines (SVM) http://scikit-learn.org/stable/modules/svm.html
  - 1.6. Nearest Neighbors (kNN) http://scikit-learn.org/stable/modules/neighbors.html

- 2. Unsupervised learning（教師無し学習）
  - http://scikit-learn.org/stable/unsupervised_learning.html
  - 2.3. Clustering http://scikit-learn.org/stable/modules/clustering.html

scikit-learn cheat-sheet