自然言語処理研究室

B3 Seminar

2013 年度　第2回

～機械学習と自然言語処理について～

長岡技術科学大学　B3 竹野 峻輔

機械学習とは？

- 計算機が
  データから規則性や法則性を見出し
  それ自身をアルゴリズムに反映させること

例えば…

広告（Facebook, Google…）
ロボットのバランス制御
天気予報、地震予測などなど

機械学習とデータマイニングの違い

- **機械学習**(Machine Learning)
  - **既知のデータから法則性を発見し**
    データの予測できるようにすること

- **データマイニング**(Data Mining)
  - **既存のデータから**
    有益な未知のデータの特徴を発掘すること

機械学習の種類

- 教師有あり学習（Supervised ML）
  - 予め用意されたサンプルから法則性を見つける
  - クラス分類
    - （ナイーブベイズ推定、SVM、ニューラルネットワーク）
- 教師なし学習(Unsupervised ML)
  - サンプルなしでデータから法則性を見つける。
  - クラスタ分析
    - (k-means法、EMアルゴリズム)
- 強化学習(Reinforcement ML)
  - 評価関数からアルゴリズムへフィードバックを行い改良を、独自に改善を図っていく。
  - 自動要約？
    - TD学習、Q学習

機械学習の種類

- 教師有あり学習（Supervised ML）（一番やりやすい）
  - 予め用意されたサンプルから法則性を見つける
  - クラス分類
    - （ナイーブベイズ推定、SVM、ニューラルネットワーク）
- 教師なし学習(Unsupervised ML)
  - サンプルなしでデータから法則性を見つける。
  - クラスタ分析
    - (k-means法、EMアルゴリズム)
- 強化学習(Reinforcement ML)
  - 評価関数からアルゴリズムへフィードバックを行い改良を、独自に改善を図っていく。
  - 自動要約？
    - TD学習、Q学習

どうやって自然言語処理に対応するか？

- 文書（自然言語）そのままでは処理しづらい

どうやって自然言語処理に対応するか？

- 文書（自然言語）そのままでは処理しづらい
  ⇒具体的な数値（素性抽出）を知る必要がある。

どうやって自然言語処理に対応するか？

- 文書（自然言語）そのままでは処理しづらい
  ⇒具体的な数値（素性抽出）を知る必要がある。
  ⇒どのようなことに気を付ければよいだろうか？

どうやって自然言語処理に対応するか？

- 文書（自然言語）そのままでは処理しづらい
  ⇒具体的な数値（素性抽出）を知る必要がある。
  ⇒どのようなことに気を付ければよいだろうか？

改めて、機械学習とは？

どうやって自然言語処理に対応するか？

- 文書（自然言語）そのままでは処理しづらい
  ⇒具体的な数値（素性抽出）を知る必要がある。
  ⇒どのようなことに気を付ければよいだろうか？

改めて、機械学習とは？

既知のデータから法則性を発見し

データの予測ができるようにすること

…未知のデータと既知のデータとの比較が必要

どうやって自然言語処理に対応するか？

- 文書（自然言語）そのままでは処理しづらい
  ⇒具体的な数値（素性抽出）を知る必要がある。
  ⇒どのようなことに気を付ければよいだろうか？

改めて、機械学習とは？

既知のデータから法則性を発見し

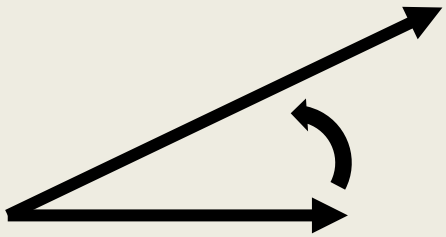
データの予測ができるようにすること

…未知のデータと既知のデータとの比較が必要

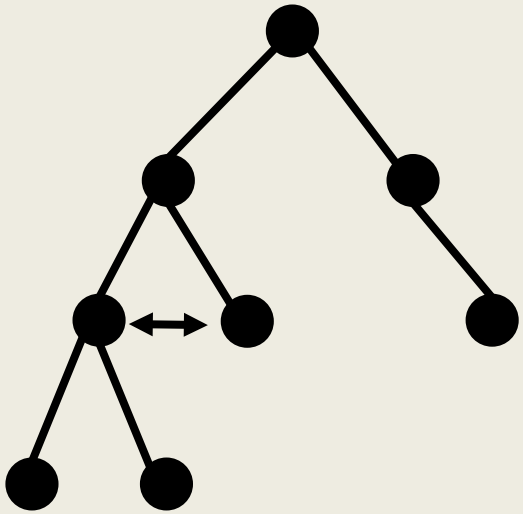
類似度の計算ができるような値を取り出す

類似度が計算できるもの（例）

- ベクトル
  - 内積

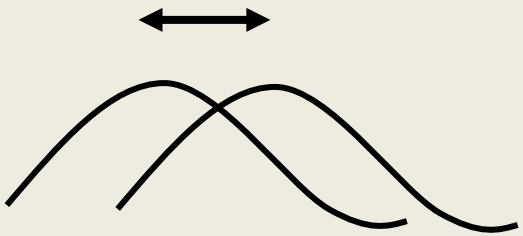


- 木構造（グラフ）データ
  - シソーラス



- 格フレーム

- 確率分布（関数）
  - 平均値、偏差、歪度、尖度（モーメント）



代表的な素性

- ベクトル：
  - Bag-of-words（文書、文比較）
    - ある単語（方向）の頻度（長さ）

    Ex) The pen is better than that pen!

    ⇒(pen, better, stick) = (2, 1 , 0)

  - 文脈ベクトル（単語の比較）
    - 空 高く 飛ぶ（名詞 副詞 動詞）

      ⇒（名詞, 形容詞, 副詞, 動詞,　形容動詞）= (1,0,0,0,1,0)

クラス分類(Classification)のための機械学習

- 訓練データからクラスの傾向を学習し、
  データがどのクラスに所属するか予測する。
  not クラスタ解析（≒クラス分析）

・ナイーブベイズ分類器
  -条件付き確率を学習 P(c|d) ≅ P(c)P(d|c)
  簡単、学習早い、精度それなり

・SVM(Support Vector Machine)
  -多次元の境界面を学習
  -解析的、学習時間かかる、精度高い

参考文献

- **奥村学 監修 「言語処理のための機械学習入門」, 高村大地著**
- **機械学習をはじめよう**, gihyo.jp, **http://gihyo.jp/dev/serial/01/machine-learning**