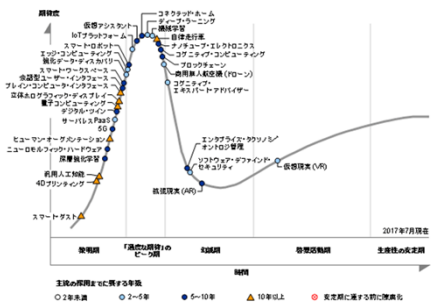
情報意味論（1）

慶應義塾大学理工学部
櫻井 彰人

この講義では

- 機械学習のいくつかの代表的な手法を知る
  - 基本原理
  - 基本アルゴリズム
  - 実際に使ってみよう
  - 少しアルゴリズムに触ってみる

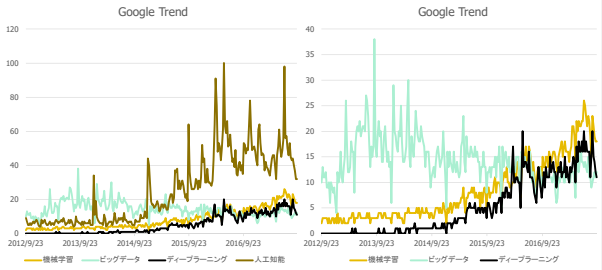
図1. 先進テクノロジのハイプ・サイクル：2017年



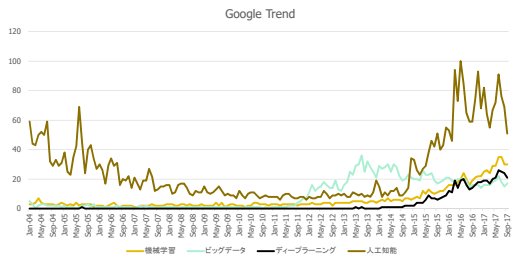
出典：ガートナー (2017年8月)

https://www.gartner.co.jp/press/html/pr20170823-01.html

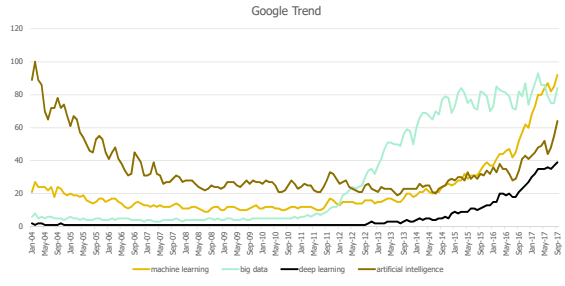
機械学習、ビッグデータ、他



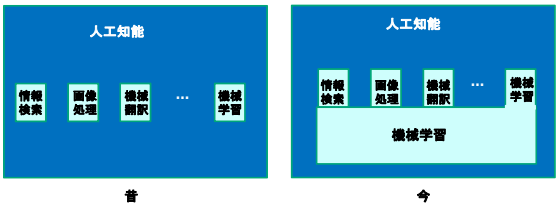
もう少し長い目でみると



英語でみると



機械学習の位置づけ



7

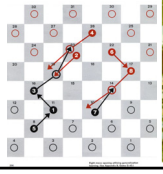
機械学習

機械学習は、コンピュータ科学の一分野であり、
機械学習をコンピュータに用いれば、
コンピュータは、明示的にプログラムされなくとも、
（自分で）学習する能力を持つ（Arthur Samuel）

Arthur Samuel は、1959年IBM在籍時に machine learning という言葉を作った。

A. Samuel, Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development, 44:1.2 (1959).

IBM704 42Kops/s; 30s/move



機械学習

機械学習とは、明示的にプログラムをしなくても、（例えば碁の良い打ち方を）学習することができるメカニズムである

目的：（人間が調整しなくとも）経験しているうちに、段々と賢くなっていくプログラム（ロボット）を作ること

中間目標：（人間がプログラムしなくとも）動いているうちに、段々と性能が上がっていくプログラムを作る

中間目標：（経験を表現したものである）データが多ければ多いほど、性能が高いプログラムを作る

9

性能が上がるとは？

天気予報なら、予報がより正しくなる

翻訳機なら、より正しい翻訳ができるようになる

店員なら、店の売り上げがあがる

赤ちゃんなら、ハイハイができるようになる

機械学習後のプログラム（を持ったロボット）は、

より正しい天気予報を出力する

より正しい翻訳結果を出力する

より購入しそうな商品名を出力する

より適切な手足の動作を命令する

10

つまり、機械学習するとは

1日後の天気予報なら、
過去の「ある時点での気象状況と1日後の天気」というデータから
予測モデルを作り、
そのモデルに現時点での気象状況を与えて、1日後の天気を予測する

英仏翻訳なら、
「英文とその仏訳文」というデータから、
翻訳モデルを作り、
そのモデルに英文を与えて、仏訳文を得る

店員なら、
過去の経験である「この客はこれを買った」というデータから
（客属性から）購入しそうな商品を予測するモデルを作り、
そのモデルに客属性を与えて、購入しそうな商品を予測する

赤ちゃんなら、
過去の「こういう状況で手足をこう動かせばこう動く」というデータから
動作モデルを作り、
行きたい方向と現況から、手足の動きを得る

11

アルゴリズムのいろいろ

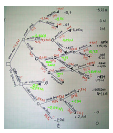
- 回帰
- 事例ベース
- 正則化
- 決定木
- 統計的分類
- カーネル法
- クラスタリング
- 相関規則
- ニューラルネットワーク
- ディープラーニング
- 次元圧縮
  - トピックモデリング
- アンサンブル法
- ブースティング

回帰 regression

- Regression: 回帰と訳すが
  - 後戻り, 復帰, 後退, 退歩, 退化, 退行
  - もともとは、今の意味とは異なる、「平均への回帰」の意味で使われた
- 説明変数のある関数で、被説明変数の値を近似する。次のものに依存する
  - 関数の形
  - 誤差の形
- 学習：訓練データで、回帰関数を作る
- 推測：未知データを回帰関数に入れ、出力値を予測値とする

決定木 decision tree

- 「木」を使って、学習結果を表現する
- 分類が主であるが、回帰もできる
- 学習：ヒューリスティックな構築方法
  - 各ノードには属性1個に関する値のテスト
- 推測： 未知データに決定木を適用する

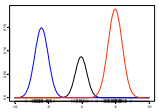


統計的分類

- 尤度最大化や事後確率最大化を図る。
  - その際、ベイズの定理を利用
- 学習：説明変数を確率変数と考え、その分布のモデルを作成する
  - モデルは、簡単化する。
    - Naïve Bayes
    - 判別分析
- 推測：非説明変数の値の分布を求める。

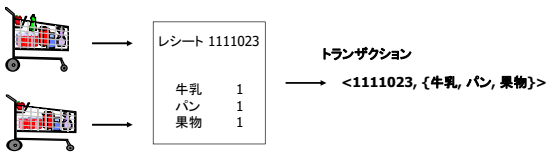
クラスタリング clustering

- 非説明変数に対する教師データはない。
  - 非説明変数はない、と言ってもよい
- 説明変数値の分布を用いて、各データをいくつかのグループ・塊り（クラスタ）に分ける
- 統計的には、隠れ変数のある統計モデルの推定問題として扱われる



相関規則 association rule

- 買い物籠1個がデータ1個
- 相関規則: If AとBを買う then Cも買う
- 発掘：大量の買い物籠データから、信頼性と精度が高い相関規則を抽出



事例ベース instance-based

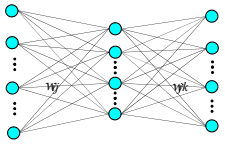
- 丸暗記＋類推
- 学習：事例をすべて記憶する
- 推測：新規データに最も近い事例を取り出す
  - 「近い、遠い」の決め方にいろいろ
  - 「近い、遠い」を学習する手法もある

カーネル法

- 特徴量を、ある非線形関数を用いて高次元空間に写像し、そこで、線形関数を用いた分類や回帰を行う
  - 元になる手法（線形関数を用いる手法）が、カーネルトリックが有効となるような手法であるべき
  - 例: SVM
- 学習： 学習データでパラメータを推定。
  - カーネル関数は事前知識に基づいて選ぶ。ただし、情報量基準やCVを用いて選択するも可
- 推測： 未知データを入力
  - カーネルトリックを用いる故、計算量は（次元を高くしても）多くならない

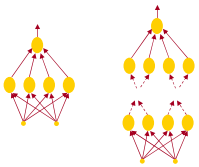
ニューラルネットワーク

- 単純な機能を持った素子（神経素子の単純なモデル）を多数結合したもの
- 学習: コスト（誤差等）が最小となるよう素子間の結合荷重を調節する
- 推測： 説明変数値を入力し、出力値を推定値とする



ディープラーニング

- 中間層数が多い（２以上）のニューラルネットワーク　2ではなかなかDLNとは認めてくれない
- 基本的にはニューラルネットワーク
- 学習アルゴリズムに本質的な工夫がある



正則化 regularization

- 過学習を抑えるため、最小化すべきコストに、モデルが複雑になるほど大きくなるペナルティ項を加える
  - コスト関数＝本来のコスト＋λ ペナルティ項
  - λの決め方に恣意性が残る

$$\min_f \sum_{i=1}^{n} |Y_i - f(X_i)|^2$$

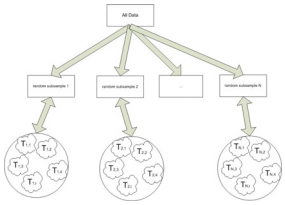
$$\min_f \sum_{i=1}^{n} |Y_i - f(X_i)|^2 + \lambda \|f\|^2$$

次元圧縮

- 説明変数の個数を減らす
  - 被説明変数がある場合、ない場合
  - 手法は多数あり
    - 主成分分析（PCA）
    - 因子分析
    - 多次元尺度法（MDS）
    - 潜在意味分析（LSA, LSI）
    - 確率的潜在意味分析（pLSA, pLSI）
    - Latent Dirichlet Analysis
    - 非負行列分解（non-negative matrix factorization）
    - LASSO（least absolute shrinkage and selection operator）
    - word2vec

アンサンブル法

- 複数の（多数の）学習器を組み合わせる
- 多数
  - ブースティング
  - バッギング
  - AdaBoost
  - Random Forest



The top 10 algorithms in DM

- the IEEE International Conference on Data Mining (ICDM) in December 2006 で決めたもの
  - C4.5
  - k-means
  - SVM
  - A priori
  - EM
  - PageRank
  - AdaBoost
  - k-Nearest Neighbor
  - Naïve Bayes
  - CART

Deep learning がない….

講義形態

- 普通の講義形態
- できるだけ、動作例を見ててもらう
- シラバスから順序等多少変更あるかも
- 確率・統計の基礎はできるだけ省略
- Weka と R は道具として使うが概説のみ

評価方法

- 3回～4回のレポートに基づく

2017年度予定

1	9月25日	月	情報と意味と機械学習
	10月2日	月	休講
3	10月16日	月	コネクショニズム
4	10月23日	月	多層神経回路網
5,2	10月30日	月	ベイズ学習, 過学習
6	11月6日	月	モデル選択
7	11月13日	月	EMアルゴリズム
8	11月20日	月	ベイジアンネットワーク
9	11月27日	月	トピックモデル
10	12月4日	月	word2vec
11	12月11日	月	SVM
12	12月18日	月	Boosting
13	12月25日	月	事例ベース学習/相関規則
14	1月9日	火	Deep Learning, 強化学習
15	1月15日	月	予備

10/2 休講のため変更