知的情報処理
1. 導入

櫻井彰人
慶應義塾大学理工学部

本講義の目的

- 機械学習の「応用」「理屈」「計算方法」の基本を理解する
  - 応用：小規模な実際のデータを用いてみる
  - 理屈：理論のさわり。
  - 計算方法：少しは、何をやっているかを。

機械学習って何？

- COM実験と少し重複します。ご勘弁を。
- データの背後にある構造を（直接的または間接的に）知って、爾後の行動の役に立てる
- 何で「機械」？
  - 「機械」とはコンピュータのこと。構造を探るのにコンピュータを活用するから（今では、コンピュータがないとできない）
- 何で「学習」？

何で「学習」か？

- 「学習」とは何か、という議論は大昔から行われています。
- 深い議論は知らなくとも、
  - 言葉で表現された知識の獲得
    - 理解して
    - 理解しないで。丸暗記
  - 体験としての知識の獲得
    - 頭に入ること
    - 体で覚えること

学習



- 少しずつ意味は異なるでしょうが、基本は
  - ある系の振舞い（データ）をもとに、その系の本質を表現する
  - 対象とする系の表現に基づき、最適行動を計画・実行する

機械（コンピュータによる）学習とは？

- Wikipedia に気の利いた表現あり
  - to extract information from data automatically, by computational and statistical methods.
- 要は、
  - データから情報を抽出する。ただし、
  - 自動的に（人間の介在なく）、そして、
  - コンピュータサイエンスと統計学の手法を用いる

データと情報

- どう違う？
- まぁ、ここにも意見・異見があると思いますが
  - データ：数字の羅列
  - 情報：意味の表現
- ただし、
  - 意味：我々の世界への対応
    - 例えば、「犬」という文字から生きている犬への対応
    - 犬のディジタル画像ファイル中の犬のイメージから生きている犬への対応
  - 対応先があると、もとのデータに意味があることになる。

機械学習

- もっとも、コンピュータは現実世界では生きていません。ですから意味が分かりません。
- それでも、データ間に何らかの関係（相関関係等）や構造があれば、それを見つけるようにプログラムすることはできます。
- 一方、人間は、一件何の脈絡もない大量のデータから規則性を見つけることはできません。それに、人間には、飽きやすい、間違いやすい、客観的になれない（自分に都合のよい解釈をしてしまう）という欠点があります。
- そこで、コンピュータの登場です。

現実1

- データは大量にあります
- web上には、テキストデータや数値データがありますし、言語データは米国を中心に集めていますし、各企業は売り上げ・在庫・人事・金融等々、国は経済・福祉・防衛等々、大量のデータを持っています。
- コンピュータも十分強力になりました。

現実2

- コンピュータ誕生の昔から、機械学習は興味深い研究テーマでした。
- 何しろ、人工知能という魅力的な研究テーマの中心部分ですから。

Arthur Samuel(1901-1990)：Playing checkers at SAIL, with teletype ~1970

A. L. Samuel (1959). Some Studies in Machine Learning Using the Game of Checkers.
A. L. Samule (1967). Some Studies in Machine Learning Using the Game of Checkers. II—Recent Progress.

注：人工知能

- 二つの立場
  - 人間の知能そのものをもつ機械を作ろう
  - 人間が知能を使ってすることを機械にさせよう
- 後者が普通。
- 機械学習の技術も使うが、使わなくてもよい
- ロボット（知能機械）の動作に、人工知能技術は必ずしも必要ない。機械学習技術も同様
- 一方、ロボット（知能機械）でなくても、機械学習技術が必要なところはある。人工知能技術も同様

一気に現代に

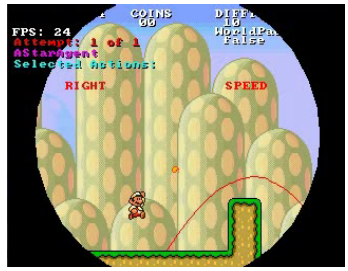
- 身近なところで使われて・使われようとしています。

- コンピュータ将棋ってご存じですか？

これは人工知能？

- 多分、人工知能でも、機械学習でもない
  - 勿論、使うことは可能であろうが、、、

これは？



- 人工知能です。探索技術を使っている
  - 機械学習はしていない

http://www.doc.ic.ac.uk/~rb1006/projects:marioai

ではこれは？

Michael Schmidt and Hod Lipson, "Distilling Free-Form Natural Laws from Experimental Data, Science, Vol. 324, April 3, 2009.

機械学習と関連するもの

- 映画推薦
  - Bayesian network を使用
- 併売分析
  - association rule mining
- コンピュータ将棋
  - 探索＋予測と最適化

他の例

- 画像識別・認識
  - 文字認識は、COM実験で
  - ふる～い論文の例
- 時系列予測（というより、、、、）
  - GA（これもCOM実験で）を用いた予測の論文

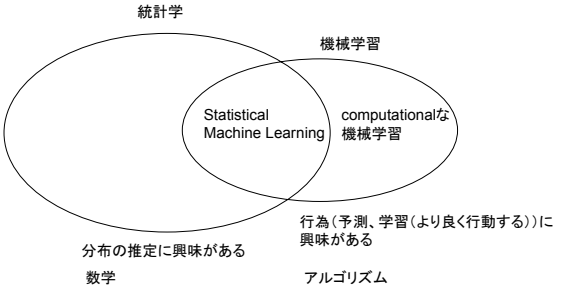
統計学との違い

- 統計では、分布の推定や（ほぼ同じことであるが）パラメータの推定を目指す。
- 機械学習では、予測を目指す。

- 統計で目的とする分布は、数式で書ける分布が多い。
  - モデルを単純にし、理論的に正確に。ノンパラメトリックという手法はある
- 機械学習では、数式で書けないような分布を対象とする
  - モデルは複雑（かどうかは分からないが）に、予測結果は正確に。
  - 機械学習でも「分布の推定」ということを行うし、研究ではその評価式がたくさん出てくるのですが、実用上は、その推定精度は大したことはない。結果としての予測精度が重要。
  - 分布の推定精度が測れるほどのデータ量が、実は、ないのが原因。
    - モデルパラメータが多すぎるのが原因。

- 融合が進んでいますが、相変わらず、違いがある

統計学と機械学習



この講義の目的

- 知的な情報処理を実現する技術の一つである「機械学習・データマイニング」技術の基礎を知る

評価他

- レポート(3回ほど)と試験、またはレポート(4回ほど)に基づく
  - 講義の進行状況に依存して決める。
  - レポート採点は、考察重視
  - 出席はとらない予定
  - ただし、簡単な即レポで代替することもあるので、ご注意ください
- 講義資料は、櫻井研究室 website に掲載予定

  http://www.sakurai.comp.ae.keio.ac.jp/

予定

1	9月29日	火	知的さと学習 – 人間とコンピュータの学習
2	10月6日	火	最近傍法 – 似ていれば似ているか？
3	10月13日	火	決定木と過学習 – 育成法と剪定法
4	10月20日	火	確率的に考えよう – ナイーブに考えよう
5	10月27日	火	ベイジアンネット – ナイーブばかりではないベイズ
6	11月10日	火	生成と判別 – モデルに頼るか頼らないか
7	11月17日	火	オッカムの剃刀の切れ味
8	12月1日	火	ニューラルネットワーク – 夢と限界と広がり
9	12月8日	火	SVMとその不思議 – 確率を無視して大成功. でも?
10	12月15日	火	カーネルトリックとPAC学習 – 過学習を手なずける
11	12月22日	火	自然言語とコンピュータ – 最も苦手なもの
12	1月12日	火	友達の友達はちょっと友達 – クラスタリング
13	1月19日	火	Open notebook test