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

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 は道具として使うが概説のみ

2008年度予定

1	9月28日	月	情報と意味と機械学習
2	10月5日	月	概念学習と決定木
3	10月19日	月	決定木と過学習
4	10月26日	月	コネクショニズム
5	11月2日	月	多層神経回路網
6	11月9日	月	ベイズ学習
7	11月16日	月	EMアルゴリズム
8	11月30日	月	ベイジアンネットワーク
9	12月7日	月	事例ベース学習
10	12月14日	月	support vector machine
11	12月21日	月	Boosting
12	1月15日	金	相関規則
13	1月18日	月	強化学習

他キャンパスで、時間がとれない方がいるときには、休講にし、以下（といっても一回だけ）、予定をずらします

機械学習

- データから意味を抽出する作業を、従来から、機械学習とよんできた
- 機械学習（machine learning）：
  - データ間の規則性（意味）の抽出（学習）を計算機に行なわせる
    - これは「学習」か？ yes!
  - 知識獲得ともいう
    - 規則性が知識だって？
  - 適応（adaptation）でもある。
- データを集めて情報となすことにはかわりない

学習



- 少しずつ異なった意味で用いられるが
  - 外界と自分があるときに、自分を少しずつ変化させて、外界に適応する（よりよいメリットを得る）こと
  - すなわち、対象とする系の表現・表出に基づき、最適行動を計画・実行する
    - そのために、ある系の振舞い（データ）をもとに、その系を表現する（本質をとらえた一般記述）ことが必要

学習

- もっと一般化して考えると、学習とは
- 具体例を知り、具体例を一般化すること
  - 丸暗記という学習もある。
- 具体例(instance)を一般化する。
  - りんご1が落ちた、栗2が落ちた、、、
    ⇒ 物体は支えがなくなれば落ちる
  - 叩いたら痛かった: 一週間前、昨日、今日、、、
    ⇒ 叩くと(いつでも)痛い
  - 隣のAさんはケイタイを持っている、会社のBさんも、、、、
    ⇒ みんなケイタイを持っている
- 特徴: 間違っているかもしれない
  - わずか(有限個の)具体例に基づくので当然。

機械学習

- 「機械学習」はこの「一般化」を理論化するにあたり、結果の正しさ(という評価基準は常に必要)を、
  - 具体例が無限個になれば、正解が得られる、すなわち、
  - 具体例が無限個になれば、モデルが同定できるような

  一般化を求めることにした。
  - 後に、この「モデル同定」でない、機械学習の特徴づけ(PAC)がなされ、機械学習のさらなる発展が起こることになる
- データ(対象とする系の動作の具体値(instance)をもとに、その系の記述を得る、その系を同定する。
  - 2, 4, 6, 8, 10, 12,,,, ⇒偶数
  - 1, 2, 4, 8, 16, 32,,,, ⇒ 2 の冪乗

機械学習

- ますます「学習」から離れるかもしれないが、「学習」の本質は捉えている

  学習: 経験(具体例)をもとにパフォーマンスを上げる

  (パフォーマンスを上げるには、未経験の事例に対しても、うまく動作する必要があるので)
  学習: 経験(具体例)をもとに未知の(類似の)事態に対応すること

  そのためには、相手(外界)を知ることが必要。知るとは記述できること。
  本質: 経験から(相手の)記述を帰納すること。未知事例に対して適用する。

学習の実例1 実世界

ロボットにペナルティキックをさせたい。もし関与するすべての物体の力学的性質が分かり、数値が測定可能かつ天候・芝の状態、キーパーの癖等がわかれば、最適なキック方法が選択できる。しかしそのようなことはない。どうするか？

自動清掃ロボットを作りたい。顧客ごとに部屋の配置を入力させるのは(入力するのは)大変だ。ロボット自身に「学習」させたい。どうしたらよいか？

学習の実例2 パターン認識

郵便番号(宛先)自動読み取り装置:
郵便番号・住所として書かれた文字のデータが10000組ある。これをもとに、宛先を読みとり分配するシステムを作るにはどうしたらよいか？

血液像自動分類装置: (診断に必要な)白血球・赤血球の画像とその分類例が10000枚ある。これをもとに、血球を自動分類しその個数を数える装置を作るにはどうしたらよいか。

学習の実例3 膨大なデータ

世界中にあるWWWページを自動的に収集・分類し、ユーザが指定した観点から自動的に類似性を判定し、関連性・塊を表示するシステムを作りたい。どうしたらよいか？

1GBある売上データから、売上増が見込める商品カテゴリを知りたい。売上増が見込める、販売戦略を練りたい。

学習の実例4 自動走行
ニューラルネットワーク

Autonomous Learning Vehicle In a Neural Net (ALVINN): Pomerleau *et al*

Navlab-5 に到り終了 (1995). 高速道路を 70mph で. "No Hands Across America"
http://www-2.cs.cmu.edu/afs/cs/user/tjochem/www/nhaa/nhaa_home_page.html




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

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

- 機械学習（machine learning）：
  - データ間の規則性（意味）の抽出（学習）を計算機に行なわせる
    - これは「学習」か？ yes!
  - 知識獲得ともいう
    - 規則性が知識だって？
  - 適応（adaptation）でもある。
    - 外界（自分以外の世界）の変化に自分を合わせる

ところで、何故情報意味論？

- もともとは、データと情報と意味を議論する講義であった（にしたかった）
  - データから意味・情報をとりだす
    - 考え方と方法
  - 取り出し方
    - 学習理論とアルゴリズム
    - 2つの方法：記号的な方法、統計的な方法
  - 応用
    - 様々な adaptation
    - データマイニング
- その中でも「学習」に重点を置くことにした

情報とは何か？

- 英語では information
  - Inform がもとの動詞。どう使う？
- 日本語：いつごろ訳したか？
  - 情とは
  - 報とは

（小野厚夫, "情報という言葉を尋ねて" (1)～(3), 情報処理(2005) を参照）

（インフヲルメーションで調べてみよ）

意味とは何か

- ①記号・表現によって表される内容またはメッセージ。② 物事が他との連関において持つ価値や重要さ。（広辞苑）
- 動作で考えてみよう。例えば、「意味がある」行動とは？
- 次に、表現と意味との関係を考えてみよう。
  - 現実世界における「表現」は常に、冗長である。では、徹底して冗長性を排除したらどうなるか？
  - なぜ、冗長なのかも考えてみよう

情報理論における情報

- データを生み出す「データ源」の記述
  - 例1：0は確率1/4で, 1は確率3/4でランダムに生成する
  - 例2: n番目には、n番目の素数の10進第一位を生成する

データ源の記述ができると

- データ源の記述ができると何がよいか？
  - 予測ができる
  - もしそれがノイズ源であれば、ノイズを効果的に低減することができる
  - （もっと一般的には）制御することができる

例えば、

- 一つの音源の音を正確に採取するために、複数のマイクを使う。
  - 2つのデータ中の相関の大きな成分が当該音源の音である（ノイズには相関がない）
- 経済予測： 株価予測、売上予測
  - 潜在需要の発見とその利用（刺激して新市場創造）
- 物理現象・化学現象・社会現象の記述と予測

つまり、機械学習

- 要は、
  - 目的、方法、評価方法は様々であれ、
  - データから意味（これって、目的によって変わります）をとりだすこと

  が機械学習

データマイニングとは？

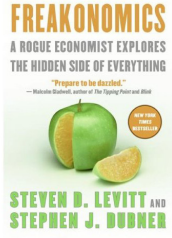
- データマイニング(データベースからの知識発見):
  - 興味深い(当たり前でない、潜在的、これまで知られていなかった、しかも、役に立つと思われる)情報あるいはパターンを大規模データベースから抽出すること
- データマイニングの別名
  - データマイニング：命名を間違えた？
  - データベースからの知識発見(Knowledge discovery in databases, KDD)、知識抽出、データ/パターン解析、データ考古学、情報収穫、ビジネスインテリジェンス、など
- データマイニングでないのは何か？
  - (演繹)質問応答処理
  - エキスパートシステムあるいは小規模な機械学習システム/統計パッケージ

データマイニングの応用例

- データベース解析と意思決定支援システム
  - マーケット分析とマネジメント
    - ターゲット・マーケッティング、CRM(customer relation management)、購入品目分析（market basket analysis）、マーケット区分（market segmentation）
  - 危機分析とマネジメント
    - 予測、顧客維持、保険の査定の改善、品質管理、競争力分析
  - 不正検知と管理：アクセスログ解析
- 他の応用
  - テキストマイニング(電子メール、webドキュメント、ブログ)
  - Web アクセスログ解析
  - 遺伝子解析（文献解析含む）

これはデータマイニング？

- 経済学？ yes.
  - 経済的インセンティブを取り扱っている
- データマイニング？ yes
  - 多量データの分析結果に基づく



例えば

- 相撲の星取り表から
- 全米統一テストの結果から
- 中絶率と犯罪の発生件数から
- ベーグルの料金回収率から

蛇足: なぜ機械学習か？

- 様々な意味で「計算能力が向上」
  - データベースマイニング: データを知識に
  - 自動カスタマイズプログラム: ニュースのフィルタ、適応的な監視カメラ
  - 行動の学習: ロボットの計画, 制御の最適化, 決定支援
  - プログラム困難なアプリケーション: 自動運転, 音声認識
- 人間の学習や教育のよりよい理解を求めて
  - 認知科学: 知識獲得の理論 (e.g., 実践を通じて)
  - パフォーマンス向上: 推論・推測, 推薦システム
- 時は今、、、
  - 学習アルゴリズムや理論の最近の進歩は目覚しい
  - 様々なソースから大量のオンラインデータが提供される
  - 計算機は安価・高速
  - 機械学習を用いた事業が発生・成長 (e.g., データマイニング/KDD)

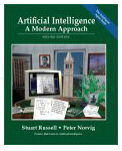
関連領域

- 認知科学: 言語獲得、推論の学習
- 統計学: バイアス vs. 分散, 信頼区間, 仮説検定
- ベイズの方法: ベイズの定理、欠測値の推定
- 人工知能: 記号表現、計画、知識を用いた学習
- 計算の複雑さの理論: PAC 学習、VC次元、誤差限界
- 制御理論: 最適化、動的計画、予測の学習
- 情報理論: エントロピー, MDL, 情報源符号化
- 神経科学: 人工神経回路網、脳(大脳、小脳、視床下部)
- 哲学: オッカムの剃刀, 帰納的一般化
- 心理学: 練習の冪法則(Power Law of Practice) 発見的学習

機械学習環境

- Weka: Waikato大学開発
  - http://www.cs.waikato.ac.nz/ml/weka/
- RapidMiner:
  - http://rapid-i.com/content/blogcategory/10/69/
  - 旧名: Yale: yet another learning environment
    - http://www-ai.cs.uni-dortmund.de/SOFTWARE/YALE/index.html
- 掲示板
  - http://www.kdkeys.net/forums/

参考書等



- **Thomas Mitchell, Machine Learning, McGraw-Hill.**
- **Stuart Russell, Peter Norvig, エージェントアプローチ人工知能, 共立出版**
  - Artificial Intelligence: A Modern Approach (2nd edition), Prentice Hall
- http://www.sakurai.comp.ae.keio.ac.jp/

時間が余ったときの雑談

Induction (帰納)

- OED (Oxford English Dictionary) によれば
  - the process of inferring a general law or printciple from the observations of particular instances
  - これは、inductive inference のこととする
  - inductive reasoning は: the process of reassigning a probability (or credibility) to a law or proposition from the observation of particular events

答え(学習結果)の個数

- 一般には複数
  - 隣のAちゃんはたまごっちを持っている、向かいのBちゃんも、、、、
    ⇒ みんなたまごっちを持っている
    ⇒ 4歳から10歳はみなたまごっちを持っている
    ⇒ 10歳未満の男性の50%はたまごっちを持っている
    ⇒ Aちゃん、Bちゃん、、、はたまごっちを持っている

説明誤りなしとしても複数

- 6, 12, 18, 24, 30,,,,,
  - 6 の倍数
  - 3 の倍数の倍
  - 2 の倍数でありかつ3の倍数
  - 6 と 12 と18以上の6の倍数
  - 12の倍数の半分

エピクロスの多説明原理

- ギリシャの哲学者 Epicurus
  - If more than one theory is consistent with the observations, keep all theories (Principle of Multiple Explanations).
- その一つの理由：一つを他から選び出す理由がない

Occam の剃刀

- 人口に膾炙しているのは
  - Entities should not be multiplied beyond necessity.
- Bertrand Russell によれば
  - It is vain to do with more what can be done with fewer.
- 最も普通の解釈
  - Among the theories that are consistent with the observed phenomena, one should select the simplest theory.

Isaac Newton の言葉

- We are to admit no more causes of natural things than such as are both true and sufficient to explain the appearances. To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.

分かった！しかし

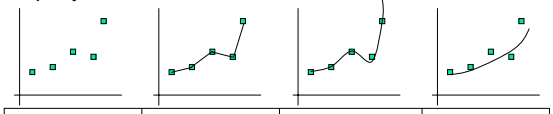
- 一般には複数
  - 隣のAちゃんはたまごっちを持っている、向かいのBちゃんも、、、、
    ⇒ みんなたまごっちを持っている
    ⇒ 4歳から10歳はみなたまごっちを持っている
    ⇒ 10歳未満の男性の50%はたまごっちを持っている
    ⇒ Aちゃん、Bちゃん、、、はたまごっちを持っている

注目：残余誤差と複雑さの二律背反

- 観測値には測定誤差がある
- 残余誤差 0 となる理論は複雑過ぎる
  - 丸暗記（⇒役に立たない）に相当
- 簡単過ぎる理論は残余誤差が多い
  - 過剰な一般化：すぐに「皆が持っている」
- 理論の複雑さと残余誤差を両立させればよい
- でも、どうやって？

関数近似の例

データ



	区分線形	全点を通る4次多項式	2次多項式
パラメータ数	2×4+3=11	5	3+ノイズ

妥協の基準

- 妥協の基準はいくつかある
  - AIC: an information criterion (Akaike's ….)
  - MDL: minimum description length

雑談のまとめ

- (機械)学習の仕事
  - データ
    ↓
  - 規則・情報・知識 + 冗長性 + ノイズ
- 可能解はたくさん(無限個)
  - 解の持ち方も、また、機械学習の研究目標
  - 解の絞り方も、また、機械学習の研究目標
    - 何らかの先験的知識で tradeoff の解消

マーケッティング分析とマネジメント (1)

- 分析のためのデータソースは何処か?
  - クレジットカード利用履歴、顧客カード、割引クーポン券、消費者苦情電話、生活スタイル調査
- ターゲット・マーケッティング
  - 同一特徴(趣味、収入、生活習慣など)をもつモデル顧客の発見
  - 顧客の購入時系列パターンの抽出
    - 銀行口座の単一名義から共有名義への変更:結婚、など
- クロスマーケット分析
  - 購入製品間の関連性
  - 関連性情報に基づく購入予測

マーケッティング分析とマネジメント(2)

- ターゲット・マーケッティング
  - 同一特徴(趣味、収入、生活習慣など)をもつモデル顧客の発見
  - 顧客の購入時系列パターンの抽出
    - 銀行口座の単一名義から共有名義への変更:結婚、など
- 顧客のプロファイリング
  - どんなタイプの顧客がどのような製品を購入するか (クラスタリングあるいは分類)
- 顧客の要求の同定
  - それぞれの顧客に対して最適な製品を同定する
  - どのような因子が新規顧客にとって魅力があるのかを予測する

企業分析とリスク管理

- 財務計画と資産評価
  - キャッシュフロー分析と予測
  - 資産評価のためのcontingent claim analysis
  - セクション間、時系列分析 (financial-ratio, trend analysis, など)
- リソースプラニング:
  - リソースのサマリーと比較
- 競合:
  - 競争相手とマーケットの方向性の観測
  - 顧客のクラス化と各クラス毎の価格ポリシーの策定
  - 競合性の非常に高いマーケットにおける価格戦略の決定

不正検出と管理

適用範囲

- 広い範囲で用いられている: 健康保険、小売、クレジットカード、電話(携帯、カード電話)
- 手法
  - 蓄積されたデータを用いて不正行為のモデルを作る。類似事例の発見にデータマイニングを使用
- 例
  - 自動車保険: 保険金を騙し取るために事故の偽装を起こす犯人達を探し出す
  - マネーロンダリング: 疑わしい現金の送金を検出する (US Treasury's Financial Crimes Enforcement Network)
  - 医療保険: 不正請求(病院、患者(米国の場合))
  - オーストラリア健康保険局 (節約Australian $1m/yr)

その他の応用

- スポーツ
  - IBM Advanced Scout: NBAの試合の統計データ (ブロックしたショット数、アシスト数、ファウル数) を分析して、New York Knicks と Miami Heatの強さの要因を明らかにした。
- 天文学
  - ジェット推進研究所(JPL)とパロマ-天文台(Palomar Observatory)は、データマイニングを用いて22のクエーサーを発見した。
- Internet Web Surf-Aid (IBM)
  - データマイニングアルゴリズムを用いて、マーケット関連のページのアクセルログを分析し、顧客がどのようなページを好むかを発見した。そして、Webマーケッティングの効果について分析し、Webサイトの構成の改良法を得た。

その他の応用(続き)

- スポーツ
  - IBM Advanced Scout: NBAの試合の統計データ (ブロックしたショット数、アシスト数、ファウル数) を分析して、New York Knicks と Miami Heatの強さの要因を明らかにした。
- 電子メールの自動分類
  - F社への電子メールによる各種の問い合わせを集積した事例を基に、86のクラスに自動分類する規則を抽出した.
- 古典和歌の歌集からの類似歌の抽出
  - 古今集, 新古今集などの二つの和歌集の間のすべての対に対して類似度を算出し、類似歌を抽出した. これまでに知られていなかった『本歌取り』の例を発見した.
- 脳の活動パターンの抽出
  - 足し算の暗算や、英会話などのタスクを実行するときに, 脳のどの部位を使っているかをf-MRI画像から、特定した. とくに、計算に小脳が関与していることを発見した.

ここまでのまとめ

- 学習: 具体例からその意味(一般的記述、説明)を得る、帰納する。
- 意味: 具体例から帰納した、系に関する一般的な記述
- 同程度の説明誤差なら、簡単なほうがよい
- 誤差があるときは、誤差との勘案

事前知識(prior knowledge)

- 背景知識(background knowledge)ともいう
- 学習理論では、多く、事前知識を仮定しない
- しかし、人間の学習の場合、大量の事前知識(常識もある)を前提として、学習の速度と精度を確保している
- 理論化は難しいが、例はある(EBL 他)

知識

- 辞典によれば:
  - 物事について、明確にあるいはいろいろと知っている事柄(小学館日本語大辞典)
- (ここでは)情報を集大成したもの

知識の獲得

- 読書百遍義自ずから通ず
  - 自己組織化、、、
- 知識を利用する側から見ると
  - 知識の獲得(人間が持つ知識を移植する)
  - 知識の抽出(データから機械的に;バッチ的)
  - 知識の学習(同上;少しづつ)、、、
  - acquisition, learning, data-mining

知識の種類(言語との対応)

- 形式知
  - 言語化した、または言語化可能な知識
- 暗黙知
  - 言語化できない、または言語化前の知識

知識の種類(抽象度)

- 事実の記述
  - 100% 客観的な記述はありえない、ことに注意
    - すべてを記述することはできない、
    - 解釈が入る、、、、
  - 体験による学習;丸暗記
- 一般化した規則・法則
  - 体験・事実からの帰納;規則自体の学習
- 「知識」はどちらかというと、後者

知識とは

- 頻繁には「経験・データを抽象・一般化して得られる、事実に対する記述」といった意味で用いられる

もう一つのポイント

- 知識表現の基礎
  - 知識の種類
    - 宣言的知識と手続き的知識
  - 知識の表現 (記号表現の場合)
    - 宣言的表現と手続き的表現
    - 宣言的表現の代表(手続き的理解も可)
      - 意味ネットワーク
      - フレーム表現
      - 論理表現

知識の種類

- 宣言的知識
  - 事物・状態・エージェント間の静的関係に関わる知識
    - ペンギンは鳥である（分類学的知識）
    - 殆どの鳥は飛ぶことができる（事物の属性）
    - ロボットAは時刻T_1 に場所 P_1 にいる（状態の表現）
  - 手続き的知識
    - 事物・状態・エージェントの時間変化に関わる知識
      - 受話器を取り上げ電話番号を回せば電話がかかる
      - 場所 P_1 にいる時に南下すれば場所 P_2 にいたる

実は

- 宣言的表現と手続き的表現の境界は曖昧
  - （表現したい知識は多くの場合プログラムほど手順が複雑ではないため）単位手順を組み合わせて表現できてしまうため
  - 単位手順を一言で表現したら、それは宣言的？

宣言的/手続き的＊知識/表現

	宣言的知識	手続き的知識
宣言的表現	（素直）	状態変化を2時点の状態とオペレータ（又は推論）で表現
手続き的表現	ex. 演繹的DB father(...,..)….. を事実とし、ancestor を father の合成関係で定義し、推論によって ancestor を得る	（素直）

（問題：推論は手続きか否か）

知識と学習

- どんな知識が学習できるか？
- 行われているもの：
  - パラメータ値の決定という形で書かれるもの
  - 値の予測
  - 分類（例題あり、例題なし）
  - 行動（ロボット等）