人工知能A

4

Topic 4: 機械学習の分類と決定木

Topic 4: Machine learning and Decision trees

機械学習と決定木

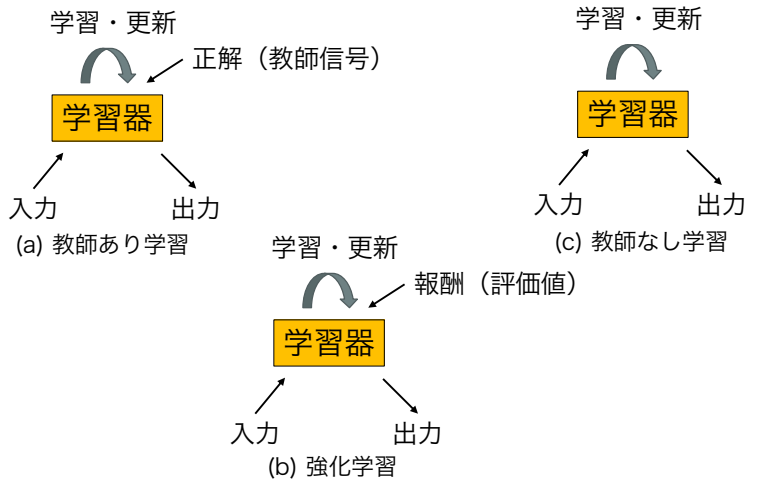
- 講義の内容
  - 機械学習の分類
  - 決定木 (ID3、CART)
  - ランダムフォレスト（時間があれば）
- 目標：
  - 機械学習の種類を理解する
  - エントロピー（復習？）
  - ID3アルゴリズムの理解

3 機械学習の種類と分類

学習とは

- うまい定義はなかなかない。哲学的で広い概念。人間の場合の例（現象面からの例示）を使って。
- 技能習得：
  - スキーができるようになった。自転車に乗れるようになった。
- 技能の向上（速くできる、正確にできるなど）
  - 100m、1分10秒台で泳げるようになった。
  - ゴルフのハンデが減った。キーボードの間違いが減った。
- 知識が増える （記憶した）
  - 英単語を覚えた。不定積分の公式を覚えた。
  - 不定積分の公式の理由を理解した。$\pi > 3.1$であることを理解した。
- 知識が増える（概念を創造した）
  - 虚数の概念を導入した。

学習の分類 (情報 -- 何から学ぶ?)



学習の分類 (情報 -- 何から学ぶ?)

- 教師有り学習 (supervised learning)
  - 訓練データとそれに対する正解が与えられる。この正解を教師信号という。データと正解（不正解）の関係を学習する。
- 強化学習 (reinforcement learning)
  - 出力結果に対し報酬が与えられ、この値を大きくする。
  - 教師信号と違い、報酬が(不)正解の判定にならない（もっと大きな報酬が得られるかもしれない。常に最大報酬との判断ができない
  - 行動の最後に成功 (の程度) の代価として報酬が得られ、行動一つ一つの正解情報はない
- 教師無し学習 (unsupervised learning)
  - 教師信号も報酬も存在しない。
  - クラスタリングや次元圧縮、概念学習が主。

学習の分類 (方法--どのように学ぶ?)

- 例からの帰納学習 (inductive learning, learning by example)
  - 赤いものをいくつか見て、赤という概念を学ぶ。
- 助言に基づく学習
  - 助言を与えられ、それに基づく行動を決める。
- 暗記学習 (rote learning)
  - 九九などの丸暗記もの。
- 効率化学習 （演繹推論など）
  - 経験を通し無駄な行為を省いたり、推論を一部省略する (EBLなど)
- 強化学習 (reinforcement learning)
  - 行為などに（正負の）報酬があり、それに基づき学習する。
- 発見的学習 (heuristic learning)
  - 現象を観測して、そこに潜む法則を発見する。
  - 他分野で起こる現象の類似点を考え、新しい法則を見つける (類推)

機械学習の基準 (何ができたら学習?)

- 計算機に学習機能を実現する際に、下記の４つの基準を考え、システムの設計を行う。
  - 強化対象（何を学ぶか）
  - 外界からの情報
  - 事前知識との関係
- 注意：本講義での学習は、計算機が学習することに着目しており、人間が学習すること（教育、education）ではない。もちろん、計算機を使った教育でも機械学習は必要だが、その場合は計算機側の学習（たとえば生徒の能力などの把握）に着目している

機械学習の基準（１）

- 強化対象（何を学ぶか）：
  - どのような観点でエージェントの能力を向上させるか
    - 知識量の増加（前より解ける問題の範囲や種類が増える）
    - 応答速度の向上（前より速く解答をだせる）
    - 知識の記述量のコンパクトさ（よく使うルールはまとめて簡単に結論を導ける。前より効率になるだろう）
    - 正答率の向上
  - 余談：応答速度の向上。ならキャッシュなどは学習か？
    - 広い意味では学習といえる。よく使う、あるいは使うと予測されるときに、直ちに使えるように途中の情報を一時的に覚えておく。
    - ただし、ここでは学習とは言わないことにする。知識として整理した形で蓄え、推論過程を短縮したり、効率化することを念頭に置く。

機械学習の基準（２－１）

- 外界からの情報（前出 ― 外から得る情報は?）
  - 教師あり学習：前出
    - ただし、どのように与えられるかはいくつかのタイプがある。
    - 教師が一方的に情報を与える（教師主導型）データと（不）正解の組を提示することも含む。
    - 学習者（＝エージェント）が教師に質問する（学習者主導型）
    - 上記の混合（インタラクティブ型）
    - 一般化した知識を与える（一般的な指針を与える。たとえば、囲碁で右隅を攻撃せよなど）
    - 特殊化された知識を与える（例（囲碁では具体的な指し手）をいくつか示し、一般化された知識を習得させる）
    - 最終的な学習目標を直接与える。

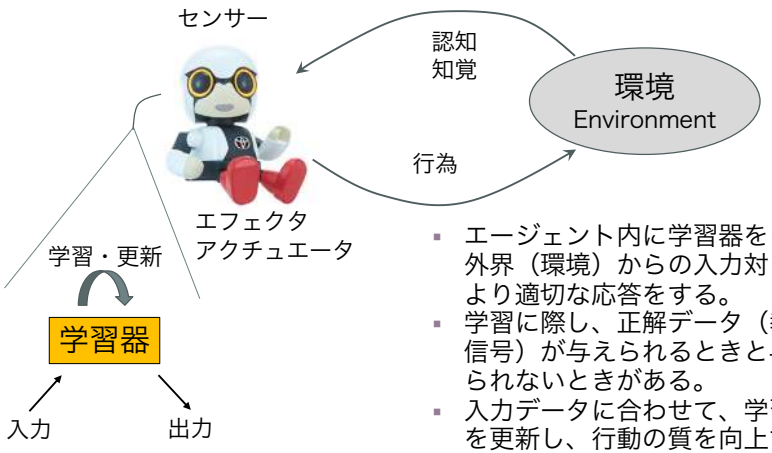
機械学習の基準（２－２）

- 外界からの情報（前出 ― 外から得る情報は?）
  - 強化学習
  - 教師なし学習(unsupervised learning)
    - これらは前出と同じ。特に後者では、特にデータのみ与えられ、正解データや報酬などはない。

機械学習の基準（３）

- 事前知識（との関係）
  - すでに知っていることでは（あと少しで）説明できなかったものを補う形で学習する。これを知識調節（knowledge accommodation）という。学習する項目を自然に制限するので、知識バイアスとも言う。
  - 新しく学習したことと、すでにある知識とは矛盾しない形で保存（記憶）されることが望ましい。このように矛盾を排除した形に学習を調整することを知識同化（knowledge assimilation）という。

学習



- エージェント内に学習器をもち、外界（環境）からの入力対してより適切な応答をする。
- 学習に際し、正解データ（教師信号）が与えられるときと与えられないときがある。
- 入力データに合わせて、学習器を更新し、行動の質を向上する

機械学習のアプローチ

- 学習のアプローチの流派
  - 記号主義（シンボリズム）
    - 述語論理、推論規則、様相論理など記号に基づく。
    - 本講義では、最後に少し触れる程度（人工知能論Bで。多分）
  - 統計的学習
    - 大量の学習データから分類、傾向、相関関係を抽出
    - 回帰分析、推定や仮説検定
    - クラスタリングなど
  - コネクショニズム
    - 神経回路網（ニューラルネット。神経系の構造をまね、ニューロンの結合やその強さを変えて学習する。深層学習など）
    - 通常は大量の教師データが必須

15 決定木

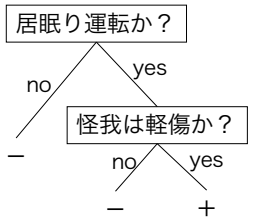
決定木とは

- 概念を区別するための分類木のこと。
  - 分類に有効な概念を見いだすことでもあり、概念学習の一部とも考えられる。
  - たとえば、以下の事故の分類で、どのような時に軽過失となるかを分類（区分）する（通常はもっとデータが多い）。

ID	場所	原因	事故内容	負傷	分類
E1	交差点	居眠り	追突	軽症	軽過失
E2	一般道	居眠り	接触	軽症	軽過失
E3	一般道	酒酔い	追突	重症	重過失
E4	交差点	酒酔い	接触	軽症	重過失
E5	交差点	居眠り	接触	軽症	軽過失
E6	一般道	居眠り	追突	重症	重過失
E7	一般道	酒酔い	接触	重症	重過失

決定木とは

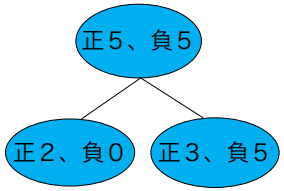
- たとえば右のような分類
- このような決定木は一般に一通りではない。なるべくよい決定木が望ましい（たとえば、右の決定木の上に、交差点での事故か？という判定を入れても、意味がない。右の木が二つぶら下がるだけ）
- 与えられた事例から、なるべくよい決定木を作りたい。よい決定木とは？



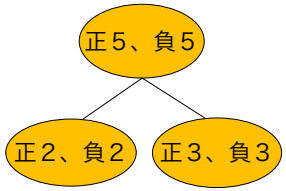
+が軽過失

よい決定木とは？

- なるべく少ない情報で（つまり決定木の深さが浅い）分類できるものであろう。
  - 同じ分類精度なら単純な方がよい。
  - データを整理したいので、分類したときにより均一（多様でない）データに分かれること
  - 仮に10 個の事例（正の例が5、負の例が5）がある。ある2値の情報で分類したとき、それが綺麗に分かれるとよい。以下の例。



右は正負のばらつきが変わらず、情報量はない。左は部分的ではあるが、正の事例のみを分類できている。



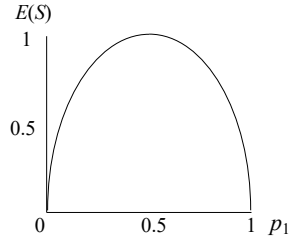
情報量（エントロピー）の考え方が有効

エントロピー

- 観測データ（正例、負例など）の集合S、その要素数をnとする。観測データに関するある情報がm個の値を持ち、それぞれのデータの個数をC_i個 ($i=1,\cdots,m$)とする。
- p_iをi番目の値となるデータの個数の出現確率 (つまり、C_i/n) とすると、この情報に関するエントロピーを以下のように定義する。

$$E(S) = \sum_{i=1}^{m}(-p_i \times \log_2 p_i)$$

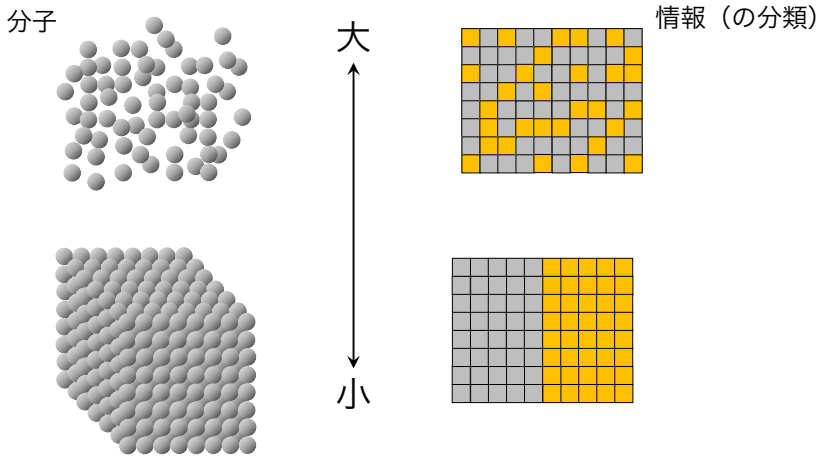
- エントロピーは情報のばらつき具合。従ってエントロピーの差分を情報量と考える。
  - 情報を与え、より綺麗に分類したい。
  - たとえば、$m = 2$とすると、$p_1 = 1 - p_2$に注意し、右図を参照。



エントロピー

- エントロピーが1とは、分類されておらずばらばらの状態（高温）。エントロピーが0（低温）とは、きれいに分類され、ばらつきのない状態に相当（ → 次ページ）
  - 前ページのエントロピーはシャノンのエントロピーと言われる。
- 決定木の生成の基本的考え方：
  - 決定木は、要素を分類すること。つまりエントロピーを下げる方向に要素を分類できるほうがよいという考え。エントロピーを最も下げる情報に基づいて要素を分類すると決定木を簡単にできる。
  - 上記の決定木は、最も単純な構造になると考えられる。単純なものが良いとする指針をオッカムのかみそりという。(Occam's razor, Ockham's razor. Ockhamは14世紀の哲学者)
  - このような決定木を求めるアルゴリズムをID3と呼ぶ
  - 大きくエントロピーを下げた分類は概念として保持するのは有効

エントロピー（アナロジー）



エントロピー：計算例

- 分類前： $E(S) = -\left(\frac{5}{10}\right)\log\frac{5}{10} - \left(\frac{5}{10}\right)\log\frac{5}{10} = 1$
- 左側：$E(S_1) = -\left(\frac{2}{2}\right)\log\frac{2}{2} - \left(\frac{0}{2}\right)\log\frac{0}{2} = 0$
  - $E(S_2) = -\left(\frac{3}{8}\right)\log\frac{3}{8} - \left(\frac{5}{8}\right)\log\frac{5}{8} = 0.9544$
  - 分類による重み付け。合わせて、 $\frac{2}{10}E(S_1) + \frac{8}{10}E(S_2) = 0.7635$
  - 情報量は $1 - 0.7635 = 0.2365$
- 右側：$E(S_1) = -\left(\frac{2}{4}\right)\log\frac{2}{4} - \left(\frac{2}{4}\right)\log\frac{2}{4} = 1,\ E(S_2) = 1$
  - 重み付け　＝1（変化無し。情報量 = 0）

エントロピー：計算例

- **分類前**： $E(S) = -(5/10)\cdot\log 5/10 - (5/10)\cdot\log 5/10 = 1$
- **左側**：
  - $E(S_1) = -(2/3)\cdot\log 2/3 - (1/3)\cdot\log 1/3 = 0.6365$
  - $E(S_2) = -(4/9)\cdot\log 4/9 - (5/9)\cdot\log 5/9 = 0.6870$
  - **分類による重み付けを行う。合わせて、** $3/12\cdot E(S_1) + 9/12\cdot E(S_2) = 0.6744$
- **右側**：
  - $E(S_1) = -(3/5)\cdot\log 3/5 - (2/5)\cdot\log 2/5 = 0.6730$
  - $E(S_2) = -(3/7)\cdot\log 3/7 - (4/7)\cdot\log 4/7 = 0.6829$
  - **重み付け**　$5/12\cdot E(S_1) + 7/12\cdot E(S_2) = 0.6788$ (若干、情報量は少ない)

エントロピーの例 (阪神ファン)

名前 (id)	性別	出身	学部	趣味	阪神ファン
A	男性	関東	工学部	スポーツ	×
B	女性	関東	工学部	スポーツ	×
C	男性	関西	工学部	スポーツ	○
D	男性	中部	理学部	スポーツ	○
E	男性	中部	法学部	旅行	○
F	女性	関西	法学部	旅行	○
G	男性	関東	法学部	旅行	○
H	男性	中部	理学部	旅行	○
I	女性	関東	理学部	旅行	○
J	男性	関西	工学部	旅行	○
K	女性	中部	理学部	スポーツ	×

エントロピーの例

- 阪神ファンかどうかを決定する決定木を作るとする。
- 最初の集合（これをSとする）
  - $E(S) = -(8/11) \cdot \log 8/11 - (3/11) \cdot \log 3/11 = 0.8453$
- 出身地で分ける。

  $S_{関東} = \{A, B, G, I\}, S_{中部} = \{D, E, H, K\}, S_{関西} = \{C, F, J\}$

  $E(S_{関東}) = -(2/4) \cdot \log 2/4 - (2/4) \cdot \log 2/4 = 1.0$

  $E(S_{中部}) = 0.8113, E(S_{関西}) = 0$
  - これをデータの個数で重みをつけると、

    $(4/11) \cdot E(S_{関東}) + (4/11) \cdot E(S_{中部}) + (3/11) \cdot E(S_{関西}) = 0.6037$
  - したがって、$0.8453 - 0.6037 = 0.2416$が出身地で分類したときに得られる情報量である。

ID3 (Iterative dichotomizer 3)

- Sを観測データの集合、Aを属性の集合としたとき、ID3アルゴリズムID3(S, A)は、下記の通り。
  - Step1: ルートノードを作る。
  - Step2: Sの要素すべてが同一カテゴリに属するなら（分類は完了しているので）、Sをそのままルートノードとして終了。
  - Step3: Aの各属性について、エントロピーによる情報効率（情報量）を計算し、もっとも効率の高い属性をcとして選択する。このテストを新しいルートノードの値とする。
  - Step4: Sの各要素を上記の属性に基づく分類で部分集合S_i ($i = 1, \ldots, m$)に分割する（$S = \cup S_i$ でありdisjointである）。各S_iについて、ID3(S_i, $A - \{c\}$)を再帰的に行う。
  - これらの出力を、このルートノードの下につけ出力とする。

課題4-1

- 阪神ファンかどうかを決定する決定木を以下のステップにしたがい完成させよ。
  - 例にしたがい、Sに対し出身地の他、性別、学部、趣味で分類した時のエントロピーに基づく情報量を計算せよ。その結果、決定木の最初のルート部分で適切な (情報量がもっとも多い）属性cを求めよ。
  - 上記で求めた属性に基づいてSを分割し、各S_iを求めよ。
  - 各S_iについて同様に残りの各属性に関する情報量を計算し、分割に適切な属性c_i を求めよ。
  - 上記を繰り返し、決定木を求めよ。

ID3の拡張

- 基本はID3と同じであるが、これを拡張したアルゴリズムや、それを利用したパッケージがある。
  - C4.5: ID3の機能を拡張したアルゴリズム
    - たとえば連続量データの扱いを加えたなど。
  - C5.0/See5
    - C4.5を高速化し、初期に商用化したパッケージ
    - C5.0 は各種統計処理パッケージ（たとえばR, SPSSなど）でも利用できる。
    - そのほかJ4.8など。

CART

- (エントロピーの代わりに)GINI係数に基づいて決定木を作る。
  - $E(S) = -\sum_{i=1}^{m}(p_i \times \log_2 p_i)$ …エントロピー
  - $G(S) = 1 - \sum_{i=1}^{m} p_i^2$ …ジニ係数 (Gini coefficient)
  - Gini係数は、データの（不）純度のようなもの（ベースは所得の統計処理）。
- CART (Classification and Regression Trees)
  - 不純度を下げる方向で分類をすすめる。
  - 2分木を生成
  - R, pythonのパッケージなどでも利用可

例題

- 文系コースの学生の集合Aと理系コースの学生の集合B。それぞれのコースは100人。
- 英語の点数70点以上(未満)で分割したときのジニ係数
  - $1 - \left(\frac{60}{90}\right)^2 - \left(\frac{30}{90}\right)^2 = 0.4444$
  - $1 - \left(\frac{40}{110}\right)^2 - \left(\frac{70}{110}\right)^2 = 0.4628$
  - 分割前のジニ係数は、$1 - \left(\frac{100}{200}\right)^2 - \left(\frac{100}{200}\right)^2 = 0.5$
  - $0.5 - \frac{90}{200} \times 0.4444 - \frac{110}{200} \times 0.4628 = 0.04548$ ジニ係数が減少
  - 同様にジニ係数の減少を計算し、それがもっとも大きい指標で分割。

	A	B
英語の点数70点以上	60人	30人
英語の点数50点以上	90人	60人
数学の点数80点以上	15人	75人
数学の点数60点以上	40人	95人

課題4-2

- 前の例題で、その他のジニ係数の減少量を求め、どの指標で分類すると良いかを考えよ。
- 前の例題で「英語の点数70点以上(未満)」で分割したときのエントロピーの減少量（情報量）を求めよ。
- 同様に、エントロピーを使って、最初にどの指標で分類すると良いかを求めよ。

Random Forest

- 集団学習アルゴリズムの一つ
  - 集団学習とは弱学習器（たとえば精度が不十分、大量のデータには向かないなど）を組み合わせて、精度のよい学習器とすること。

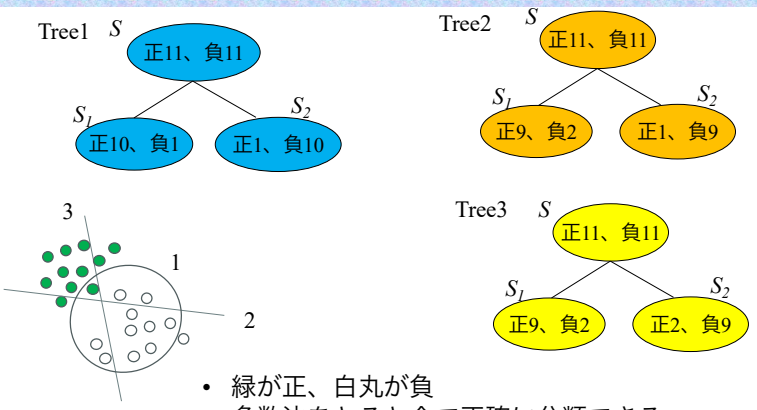
一つ一つの精度は十分ではないかもしれない
--- 特に大規模データのとき---

みんなで協力して考えよう。
そして多数決など。

Random Forest

- 方法（大量のデータを想定）：概略のみ
  - N個のランダムに選択したサンプル集合を作る
    - ブートストラップサンプルという
    - Nの値やサンプル集合の要素数は、適宜、問題に合わせて決める（ある程度の試行錯誤が必要）。
  - 各サンプル集合を訓練データと考え、決定木を作成する。ただし、あまり深い木は作らないのが一般。（N本の木ができる）
    - サンプルから生成している、木の深さも制限しているという意味でそれぞれの木は弱学習器
  - 木の利用
    - 各決定木で分類し、多数決などで決定。

Random Forest (効果のイメージ)



- 緑が正、白丸が負
- 多数決をとると全て正確に分類できる

Random Forest

- 利点など
  - 決定木の生成部分や、生成後に決定木で分類するところは、並列化が可能で、非常に高速となる。このため複雑な一つの決定木を作るより、効率化可能。
  - 比較的かつ一般的に分類精度はよい。過学習 (overfitting, あとで説明) も起こりにくい。弱学習器が功を奏す？
  - パラメータ設定や調整もほとんど無い。
  - 分類に使った木を解析することで、分類結果の理由を知ることができる。ただし多数あるので、決定木よりはわかりにくい。
  - しかし、何本の木をつくるか、木のサイズの指定などは、問題ごとに依存する（試行錯誤が必要）
  - 詳細は学部３年のレベルを超えるので、ここまで。Topic 8でも若干、触れます。