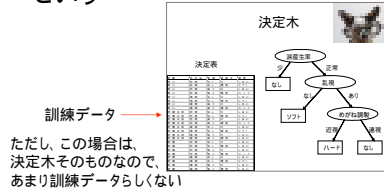


決定木 その2 決定木の作り方

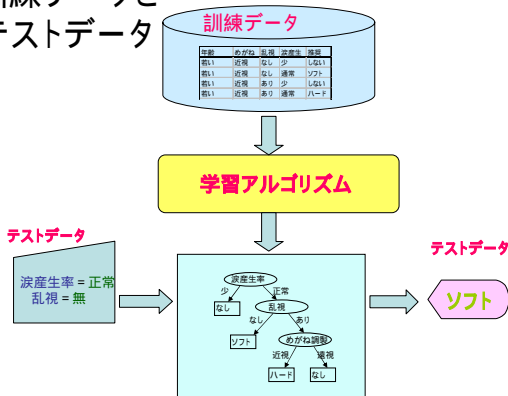
慶應義塾大学理工学部
櫻井彰人

決定木の作り方

- 決定木を作ることを考えよう
- 材料が必要
- 材料を、機械学習の分野では、「訓練データ」という



訓練データとテストデータ



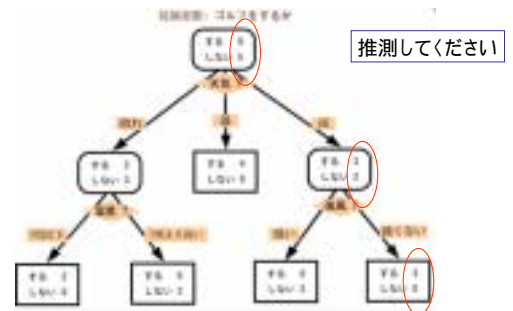
訓練データとテストデータ



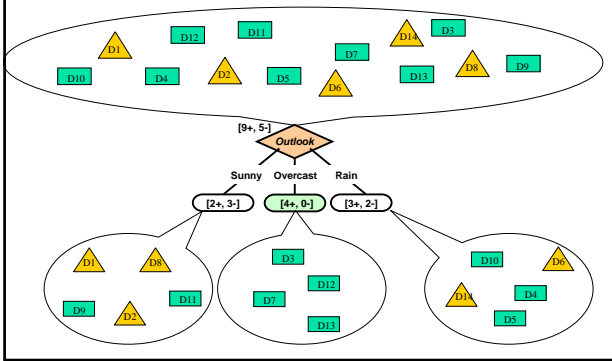
訓練データとテストデータ

- 訓練データとテストデータは、原則的には、別物
 - 丸暗記のテストをするときには、当然、同じもの。
 - 普通は、学習結果は、訓練データを一般化したもの。テストをするときには、いろいろなテストデータを用いてテストを行う。それゆえ、訓練データとテストデータは別物
 - 人間だって、自動車運転教習所で習った道路(訓練データ)と実際に走る道路(テストデータ)とは別物

これは何だと思えますか？



決定木の学習の1ステップ

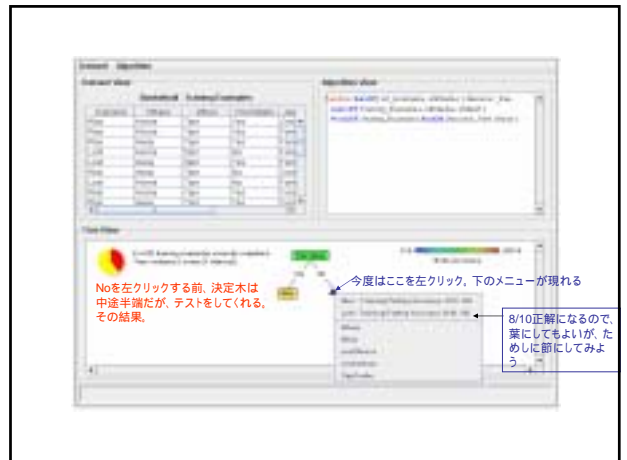
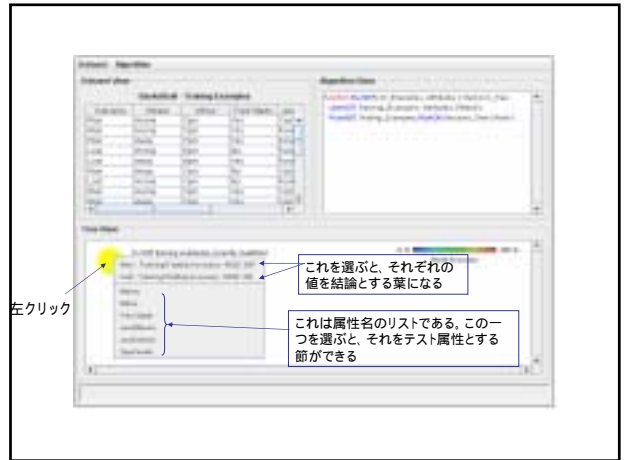


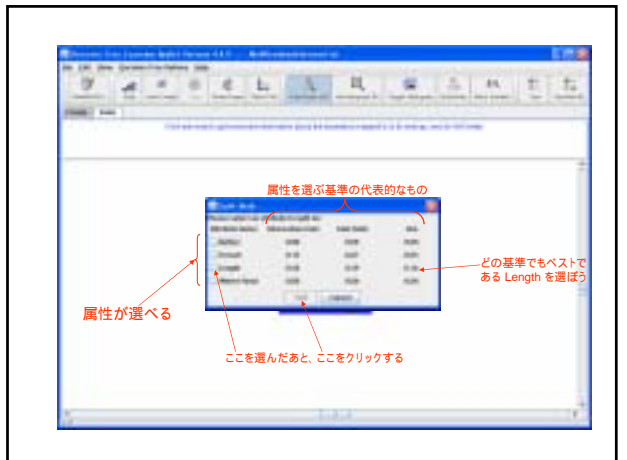
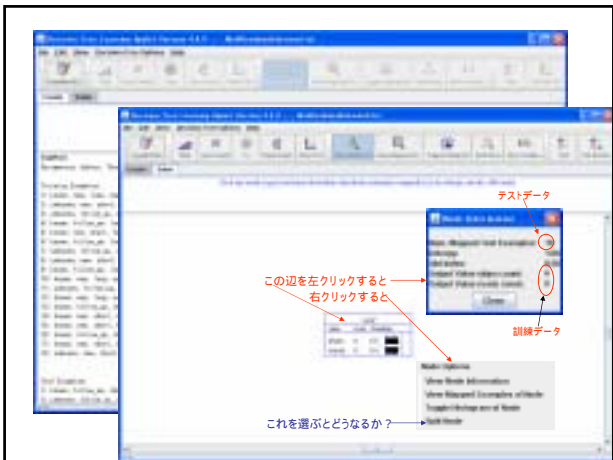
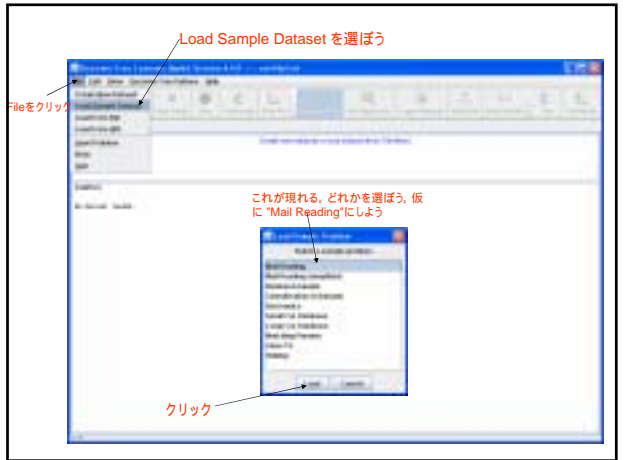
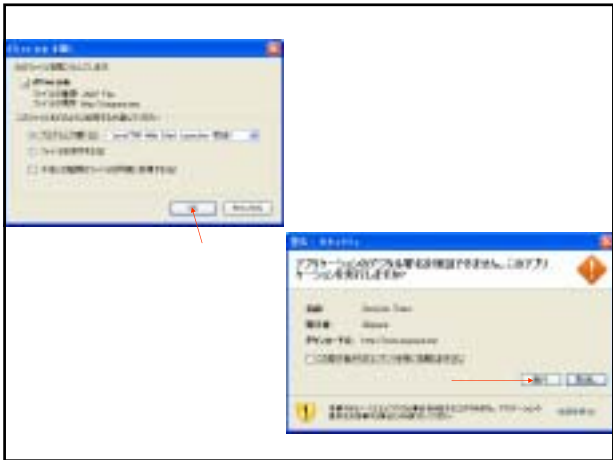
applet によるデモ

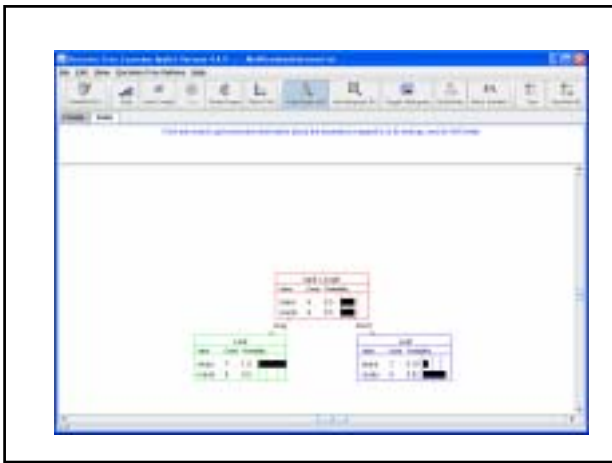
- 出来ていくプロセスがわかる、applet によるデモがあります。自動と手動モードがあります。

Aixploratorium - Decision Trees
<http://www.cs.ualberta.ca/~aixplore/learning/DecisionTrees/Applet/DecisionTreeApplet.html>

Alspace
<http://aispace.org/dTree/>







データセットを自分で作る場合 データセットの作り方

- 左上の「File」から「Create New Dataset」をクリック
- パラメータを入力
(カンマで区切る)
- 「OK」をクリック

データセットの作り方

- 左上の「ViewEdit Examples」をクリック
- 「Add New」より、各パラメータの値を入力
- 左側に学習用データ、右側にテスト用データを入力し、終わったら「Close Window」を選択

決定木を自動的に作る 決定木の作成

- 左上の「Solve」を選択
- 「Step」を選択すると、一段階ずつ木が作成される
- 各ノードの上で左クリックをすると、ノードの情報などを見ることができる

赤: 根ノード
青: 葉ノード(さらに分割可能)
緑: 葉ノード(これ以上分割不可)

決定木の作成

- 「Auto Create」を選択すると、最後まで木を生成する
- 「Reset Graph」を選択すると、木を作る前の状態に戻る
- 「Show Plot」を選択すると、error rate をグラフで見ることができる

テストデータへの適用

- 「Test」を選択すると、生成した決定木をテストデータへ適用した結果が表示される
- 「Test New Example」を選択すると、各要素を変更した際、結果がどう変わるかがわかる

決定木を作る時の課題

- 節(ノード)に置く属性を決めれば、自動的にできる(節を分割し、枝を伸ばすことができる)。

次のスライドの Hunt のアルゴリズムを参照

- 節(ノード)に置く属性の決め方が課題だ。
 - どうしよう?

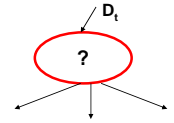
講義は、
 (1) Huntの方法
 (2) 属性の決め方と属性値のグループ化
 (3) やめ方

- 実は、もう2つあります。
- 一つは、「いつやめるか?」
- もう一つは、属性値がたくさんあるとき、属性値は、実はグループ分けした方がよい。それをどうするか?

Hunt のアルゴリズムの構造

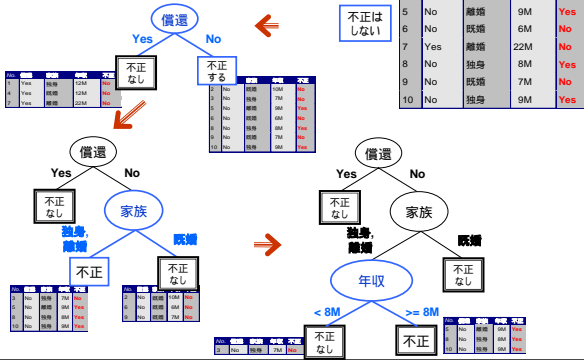
No.	職業	家族	年収	不正
1	Yes	独身	12M	No
2	No	既婚	10M	No
3	No	独身	7M	No
4	Yes	既婚	12M	No
5	No	離婚	9M	Yes
6	No	既婚	6M	No
7	Yes	離婚	22M	No
8	No	独身	8M	Yes
9	No	既婚	7M	No
10	No	独身	9M	Yes

- D_t をノード t にたどりついた訓練データの集合とする
- 手続きの概要:
 - もし D_t に含まれるデータがすべて同じクラス y_t に属するなら、 t は葉ノードであってそのラベルは y_t .
 - もし D_t が空集合なら、 t は葉ノードであって、そのラベルは予め決めておくデフォルトラベル y_d となる
 - もし D_t に含まれるデータは複数個のクラスに属するとき、 t は属性値のチェックを入れ、訓練データ D_t をより小さな集合(部分集合)に分割する。この手続きを再帰的に、当該部分集合に適用する。



Hunt のアルゴリズム

No.	職業	家族	年収	不正
1	Yes	独身	12M	No
2	No	既婚	10M	No
3	No	独身	7M	No
4	Yes	既婚	12M	No
5	No	離婚	9M	Yes
6	No	既婚	6M	No
7	Yes	離婚	22M	No
8	No	独身	8M	Yes
9	No	既婚	7M	No
10	No	独身	9M	Yes

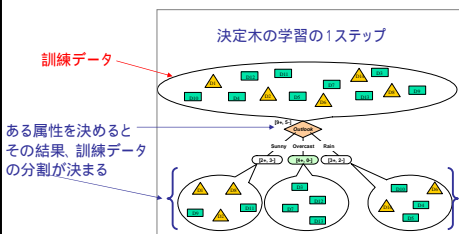


節に置く属性の決定

- グリーディな方略をとる。
 - いったん決めたら、心変わりしない
 - 迷路を進むときに、後戻りしない。一度掘んだら離さない。
 - 最適ではないが、後戻りしない分、速い。
 - つまり、ある節に置く属性(と使う属性値グループ化)を決めたら、撤回しない。
 - 一度分割してある枝を作ったら、それを取りやめることはない
 - その結果、最適な決定木となることは保証出来ない(一般には最適ではない)
- 課題
 - できるだけよい
 - 属性の選び方は?
 - 属性値のグループ化の仕方は?

考え方

- 結局、「訓練データの分割方法として、どれがいいか」を考えればよいと気がつく



宿題

Alxploratorium - Decision Trees
<http://www.cs.ualberta.ca/~aixplore/learning/DecisionTrees/Applet/DecisionTreeApplet.html>

を使って、決定木を作ってください。
 自動と手動で作り、その違いを比較してください。
 説明は続きのスライドにあります。

