大規模離散計算科学特論（11月8日）

白井 康之

（独）科学技術振興機構
ERATO 湊離散構造処理系プロジェクト

shirai@erato.ist.hokudai.ac.jp

講義内容

- 自己紹介その他（11月7日）

- 手法編（11月7～8日）
  - (1−1)はじめに(全体概要)
  - (1−2)頻出パタンマイニングとその周辺
  - (1−3)決定木分析法とその周辺

- 事例紹介 (11月8日)
  - クレジットカード自動審査, 行動履歴解析, 健康生活支援, アンケート分析など

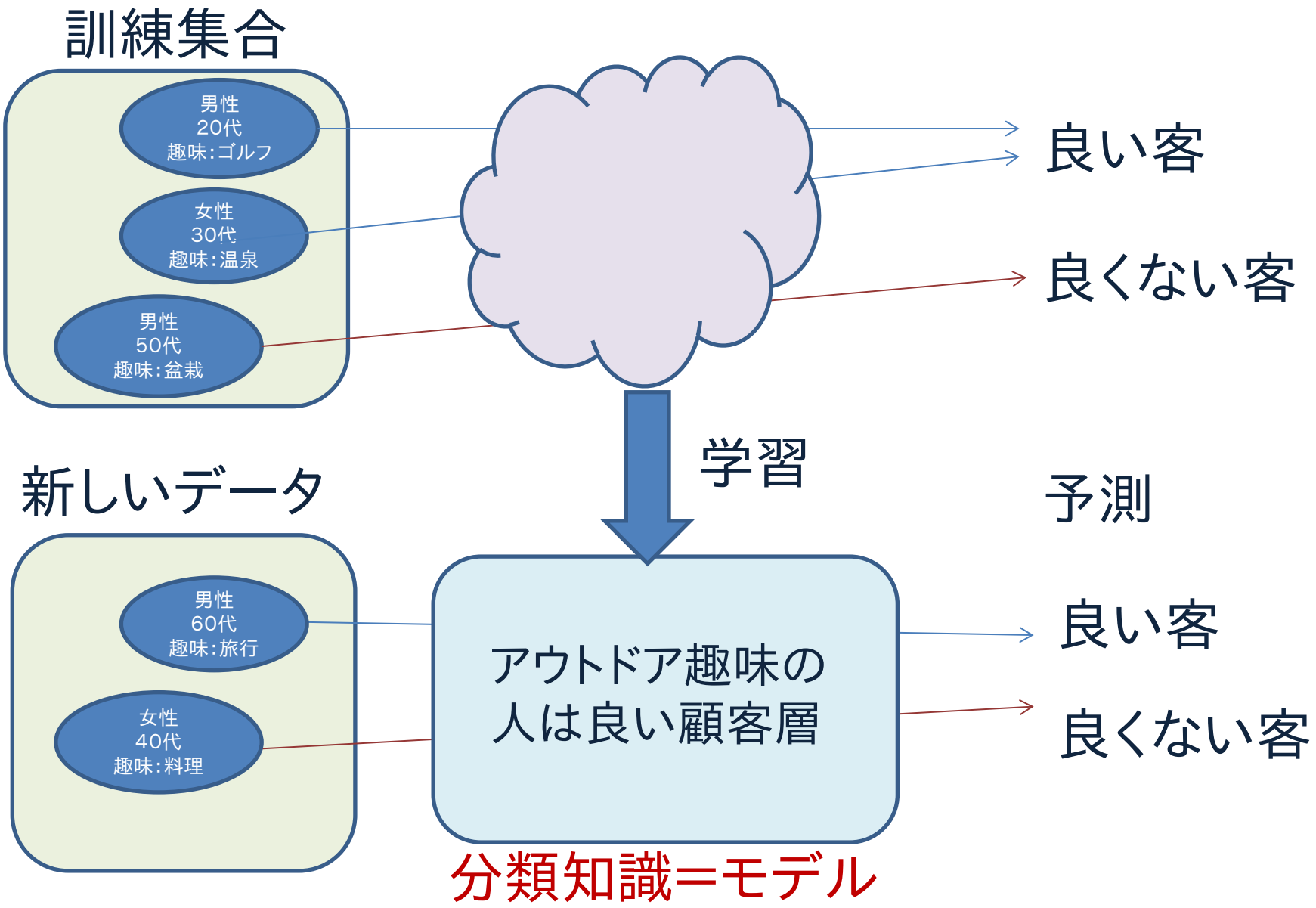
- ERATOセミナー（11月8日）
  人気感度と多様性に基づく顧客のセグメント化とその応用

（1−3）　決定木分析法とその周辺
(Decision Tree Analysis)

内 容

- 分類問題とは？
- 基本的指標の定義
  （再現率, 適合率, 情報エントロピーほか）
- 決定木分析（Decision Tree）
- 回帰木（Regression Tree）
- モデル木（Model Tree）
- ストリームデータに対する決定木

分類問題とは（例）：


訓練集合
男性
20代
趣味：ゴルフ
女性
30代
趣味：温泉
男性
50代
趣味：盆栽
良い客
良くない客
学習
新しいデータ
男性
60代
趣味：旅行
女性
40代
趣味：料理
アウトドア趣味の
人は良い顧客層
予測
良い客
良くない客
分類知識＝モデル

分類問題

分類知識を近似的な知識構造としてあらわすこと.

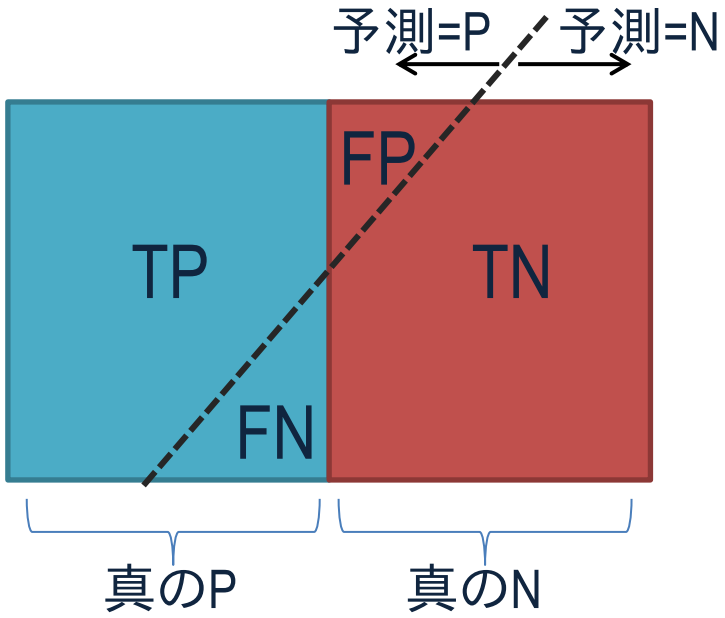
主な手法：線形近似, 決定木分析, k-NN, ニューラルネットワーク, SVM, . . .

ただし, 全データのクラス分類が可能であれば, 近似する必要はない. 学習の前に, まず変数をうまく集約することにより全列挙可能かどうかを検討するべき. データを編集することなく, 最初からマイニングにゆだねてしまう例が実は結構多い. 項目値がyes/no であれば, 10項目でも 1024通りの答をつくっておけば済む. ただし, 不完全なデータ（矛盾するデータ）の場合には, 学習結果を出力することは構造を把握するという意味もある.

ID	Attribute 1	Attribute 2	…	Attribute 10	Class
1	yes	no	…	no	P
2	no	yes	…	yes	N
3	yes		…	yes	P
…	…	…	…	…	…
1024	no	yes	…	yes	P

分類の指標　－　再現率（recall）と適合率（precision）

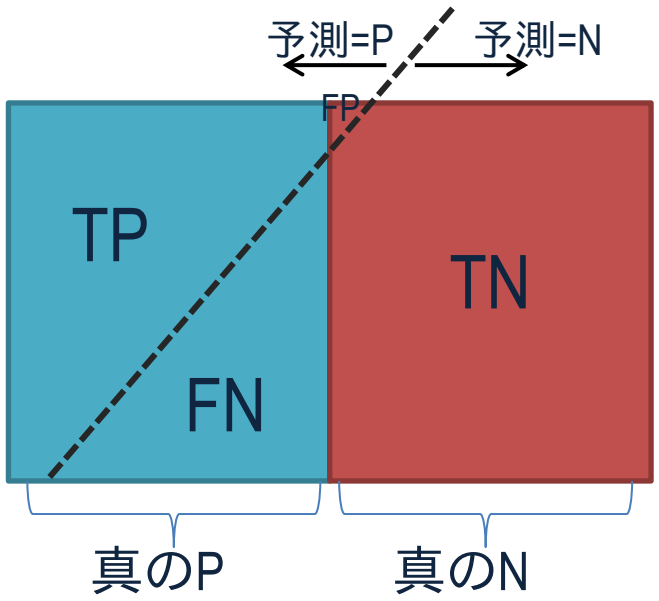
- {True, False}　×　{Positive, Negative}
  {正しい, 間違った} × {正例, 負例}



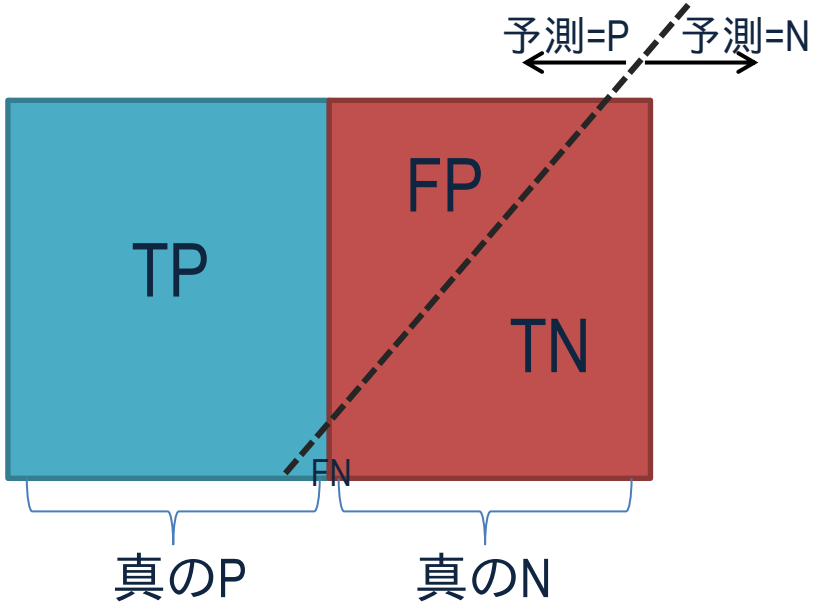
- 再現率（recall）：真のP のうち, 正しくP と判定される割合
  TP/(TP+FN)=TP/（真のP）
- 適合率（precision）：Pと判定したもののうち, 正しく分類できた割合
  TP/(TP+FP)=TP/（判定がP）

分類の指標　－　再現率（recall）と適合率（precision）

- 適合率TP/(TP+FP)を上げるには？
- FPをできるだけ小さくすると良い（ただし傾きが不変の場合）
- 例：スパムメールの分類など
- 疑わしきは罰せず



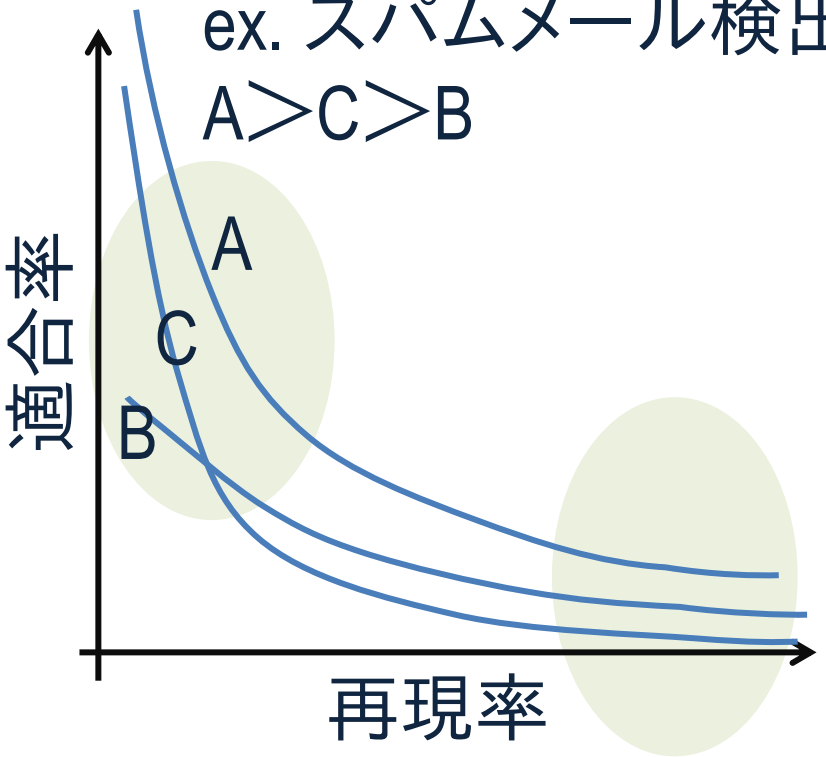
- 再現率TP/(TP+FN)を上げるには？
- FNをできるだけ小さくすると良い（ただし傾きが不変の場合）
- 例：不正アクセスの検出
- 疑わしきは罰する



再現率と適合率の関係

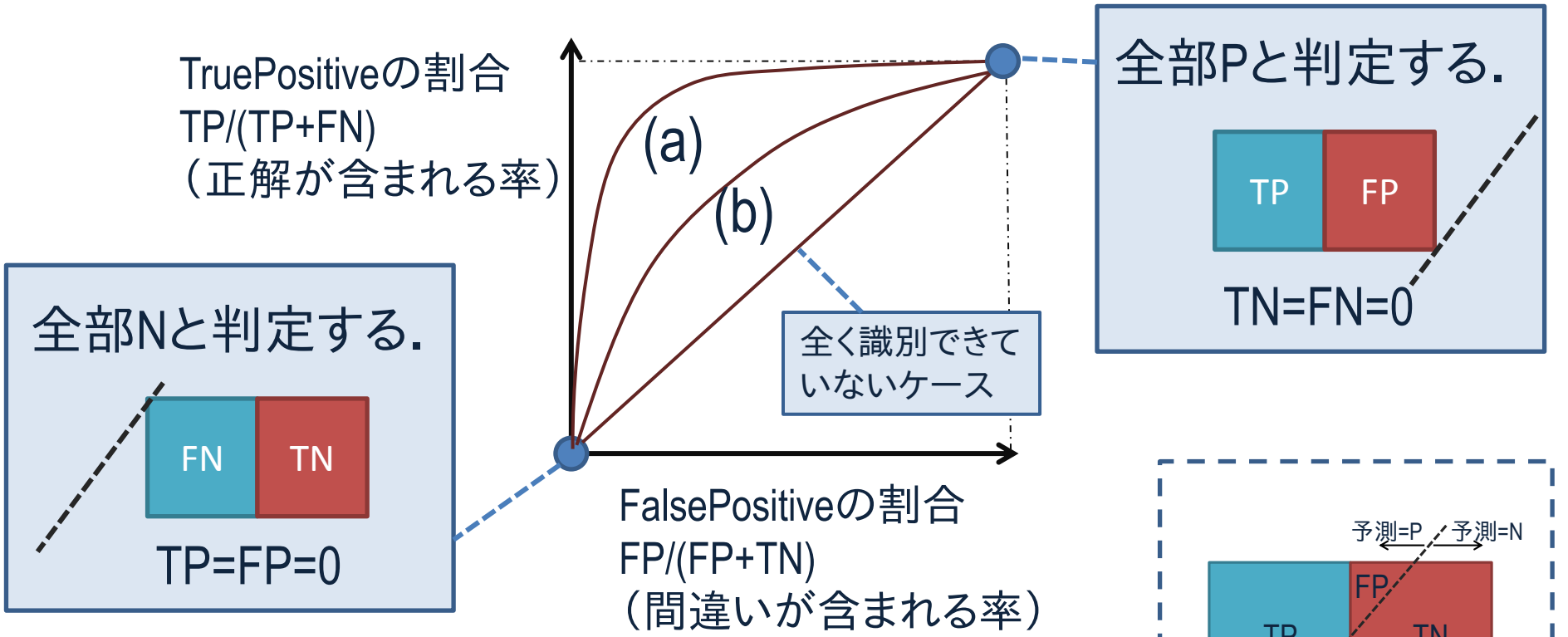
分類器 A,B,C があったとする.

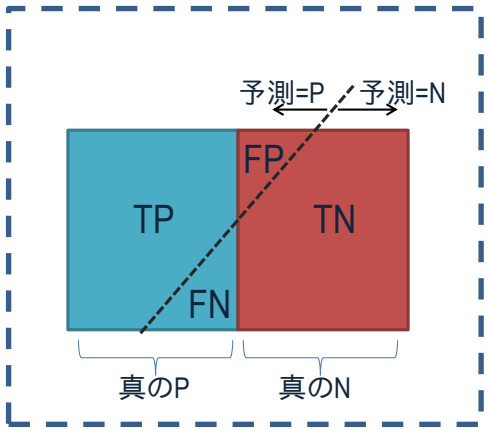
高い適合性を求めたい場合
ex. スパムメール検出
A＞C＞B



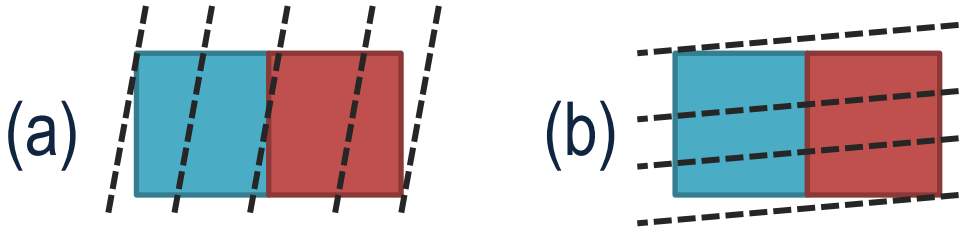
高い再現性を求めたい場合
ex. 不正アクセス検出
A＞B＞C

ROC（Receiver Operating Characteristic）曲線





ROC曲線（面積）により，分類器の精度がわかる.

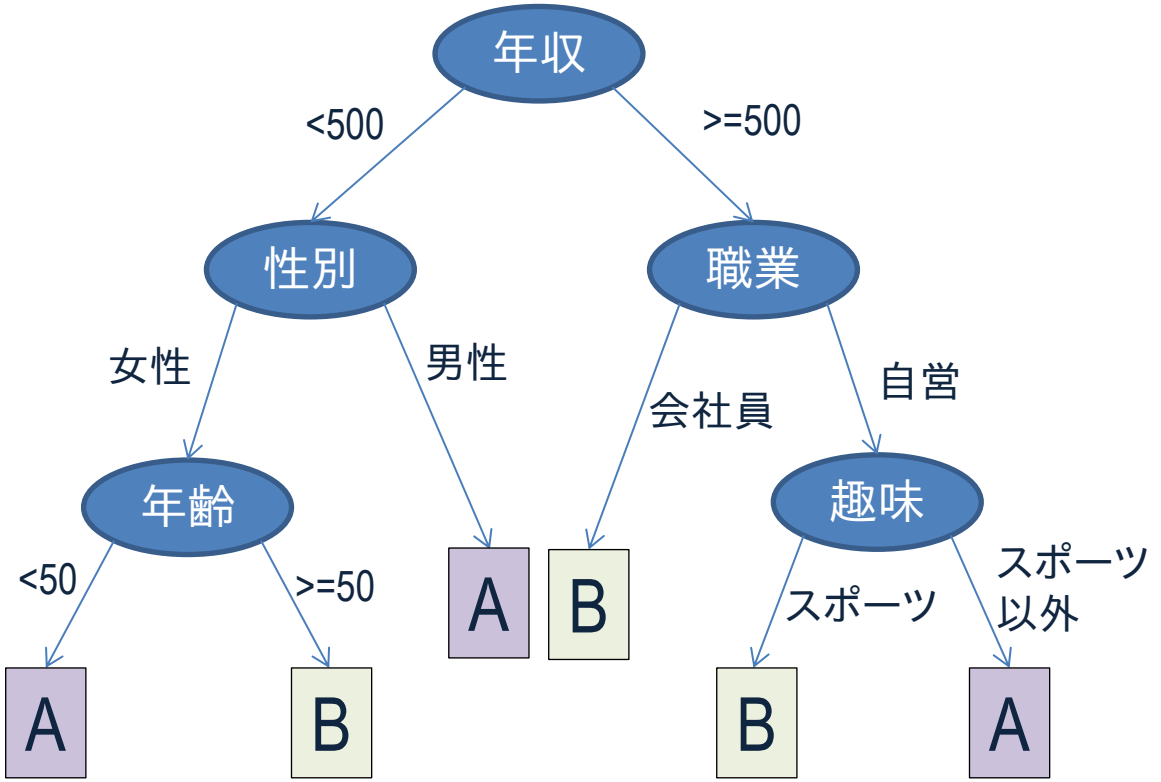


決定木分析法（Decision Tree Analysis）

- 分類ルールを木構造として表現したもの．
- J.R.Quinlan 1983 (ID3)，“AIによるデータ解析”（1995）
- その後 C4.5, C5.0/See5（商用）

- ID3 → C4.5
  - ✓ 連続値の取り扱い
  - ✓ 属性値の欠損への対応（?で表記）

- C5.0/See5
  - ✓ 商業利用
  - ✓ 速度向上



情報エントロピー

分割は（局所的に）最も効果が大きい属性を選択する.
このために，まずデータ集合に対するエントロピーを定義する.

$$Entropy \equiv \sum_{i=1}^{c} - p_i \log_2 p_i$$

クラスを決定するために必要なビット数．2クラスでポジデータ，ネガデータが均等に分布している状況（p=0.5）では1ビット必要．クラスが1つしかない場合は，クラスは確定しているので0ビット.

Information Gain（情報利得）

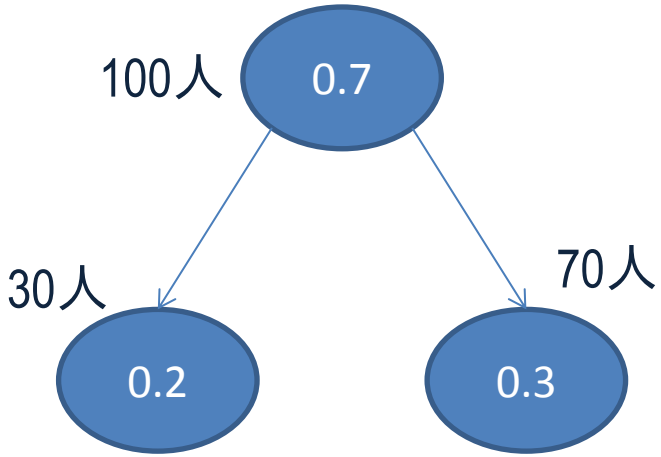
ある属性選択（Attribute）に対して，Information Gain を，エントロピーの差として以下のように定義する．

$$Gain(S, Attribute) = Entropy(S) - \sum_{v \in Attribute} \frac{|S_v|}{|S|} Entropy(S_v)$$

元のノードのエントロピー

属性で分割されたノードのエントロピーの重み付き和

最も大きい Information Gain が得られる分割を選択する．（hill-climbing）．



$$Gain = 0.7 - (\frac{30}{100} \times 0.2 + \frac{70}{100} 70/100 \times 0.3) = 0.43$$

分類の例

性別	身長	技量	クラス
M	High	Bad	P
F	High	Good	P
M	High	Bad	P
F	Low	Good	P
M	Low	Good	N
F	Low	Bad	N
M	Low	Good	N
F	High	Bad	N

Entropy(S)=1

性別で分類

$0.5\times Entropy(S_M)+0.5\times Entropy(S_F)$

$= 0.5\times(-0.5 \log 0.5-0.5 \log 0.5)$

$+ 0.5\times(-0.5 \log 0.5-0.5 \log 0.5) = 1$

身長で分類

$0.5\times Entropy(S_{high})+0.5\times Entropy(S_{low})$

$= 0.5\times(-0.75 \log 0.75 - 0.25 \log 0.25)$

$+ 0.5\times(-0.75 \log 0.75 - 0.25 \log 0.25)$

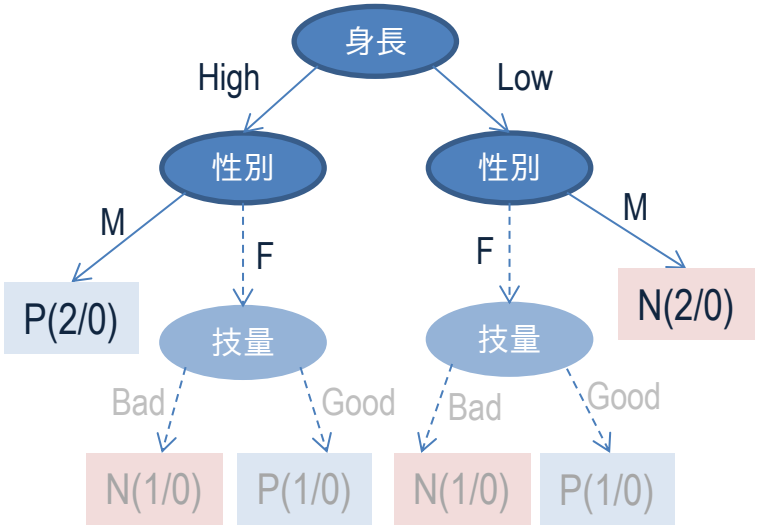
$\fallingdotseq 0.81$

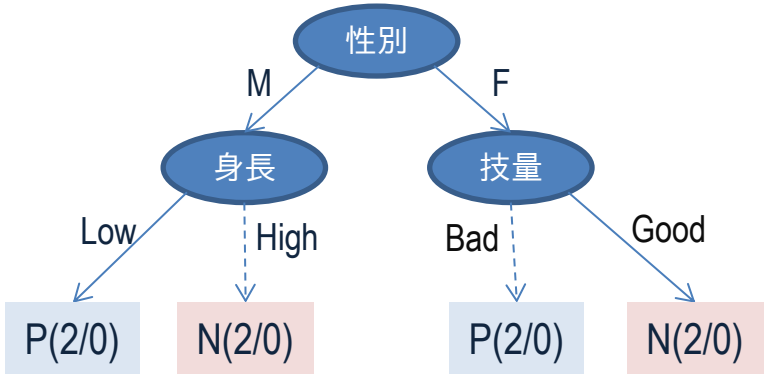
よって, 身長で分類したほうが
Information Gain は大きい.

決定木分析は山登り法

項目の選択方法は，山登り法であるため，局所解に陥ることがあるので注意．

性別	身長	技量	クラス
M	High	Bad	P
F	High	Good	P
M	High	Bad	P
F	Low	Good	P
M	Low	Good	N
F	Low	Bad	N
M	Low	Good	N
F	High	Bad	N



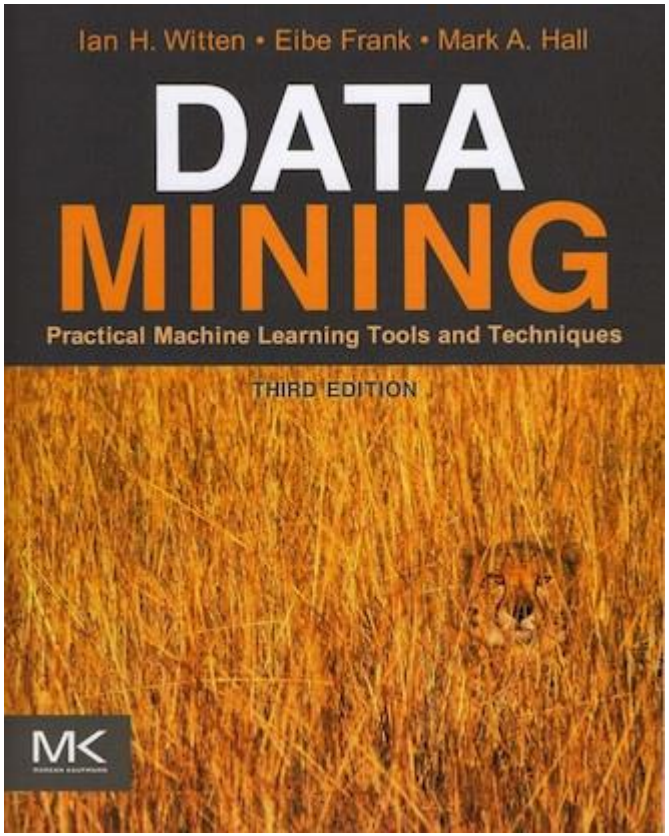


Weka

http://www.cs.waikato.ac.nz/ml/weka/index.html

- Univ. of Waikato, New Zealand
- Weka（データマイニングツールのパッケージ）
  (Waikato Environment for Knowledge Analysis)
- 1993年にプロジェクトを開始.
- 当初は, さまざまな言語と Tcl/Tk で書かれたインタフェースからなっていた.
- 1997年から Java 化
  （フリーソフトとして広く流通）
- その後, C で各ライブラリを統一して記述.
- データ加工, クラスタリング, 分類, 回帰, 可視化等をサポート.

- 小規模データの実験には向いているが, 大規模データに対する効率性は重視されていない.



一般的なDMの教科書のようでいて,
実はマニュアル（に近い）

weka: 実装されているアルゴリズムと問題記述例

- 分類手法
  ナイーブベイズ, ベイジアンネットワーク, NN（Nearest Neighbor）, 決定木（C4,5, M5, etc）, 線形回帰, SVM, その他もろもろ
- クラスタリング
  EM アルゴリズム, k-means, 各種階層的クラスタリング手法, その他
- 相関ルール分析
  Apriori など
- 可視化

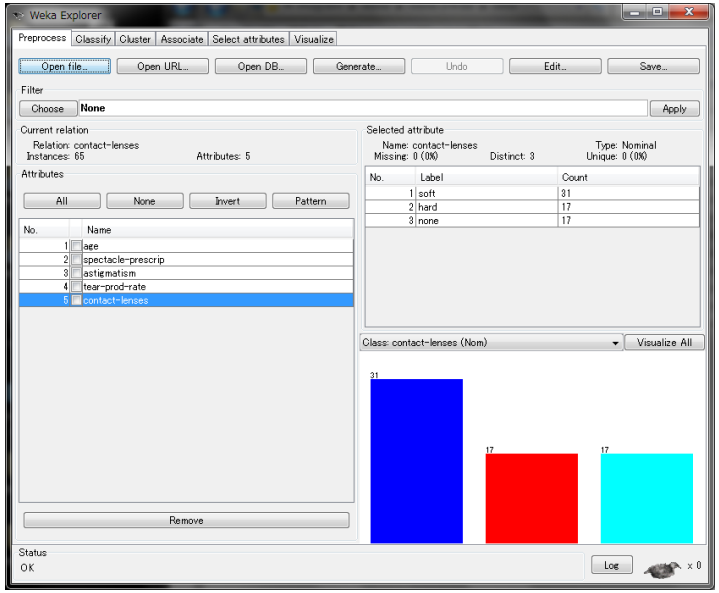
```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,hot,high,FALSE,yes
overcast,cool,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,cool,normal,TRUE,no
rainy,mild,high,TRUE,no
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
```

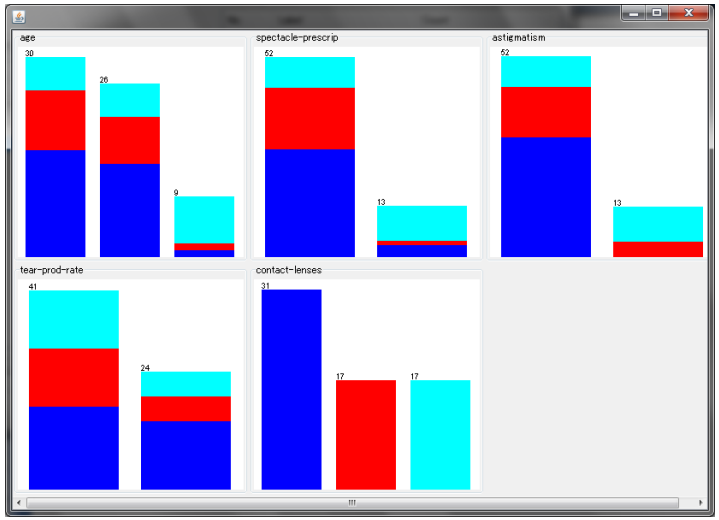
（weka サンプルデータより）

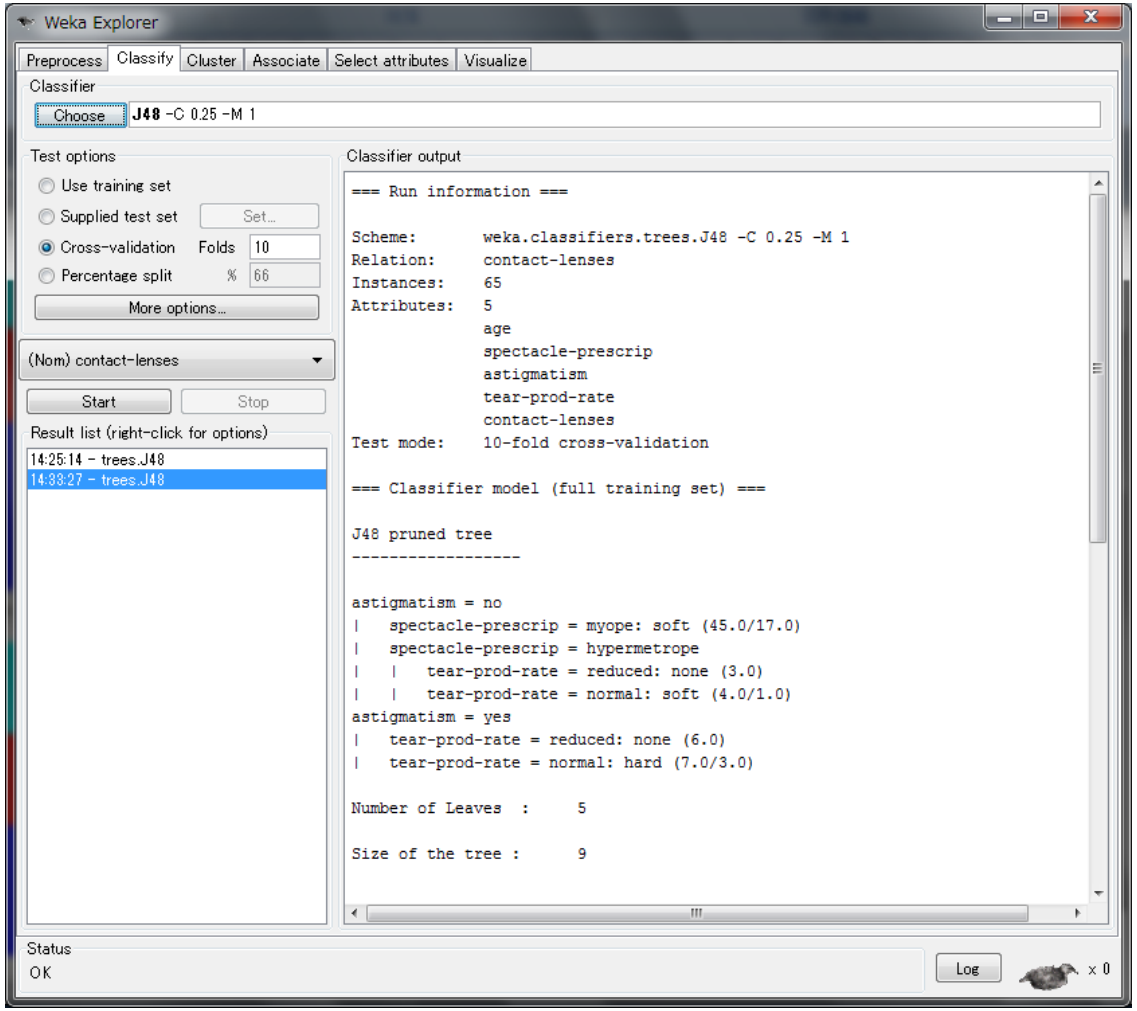
weka: 実行例（データファイル）

```
% Title: Database for fitting contact lenses
%    -- 3 Classes (soft/hard/none)
%   1. age : (1) young, (2) pre-presbyopic, (3) presbyopic（老眼）
%   2. spectacle prescription 処方:  (1) myope（近視）, (2) hypermetrope（遠視）
%   3. astigmatic 乱視:     (1) no, (2) yes
%   4. tear production rate 涙の分泌量:  (1) reduced, (2) normal

@relation contact-lenses
@attribute age                          {young, pre-presbyopic, presbyopic}
@attribute spectacle-prescrip           {myope, hypermetrope}
@attribute astigmatism                  {no, yes}
@attribute tear-prod-rate       {reduced, normal}
@attribute contact-lenses       {soft, hard, none}

@data
young,myope,no,reduced,hard
pre-presbyopic,myope,no,reduced,hard
young,myope,no,reduced,hard
pre-presbyopic,myope,no,reduced,hard
young,myope,no,reduced,hard
```

weka: 実行例

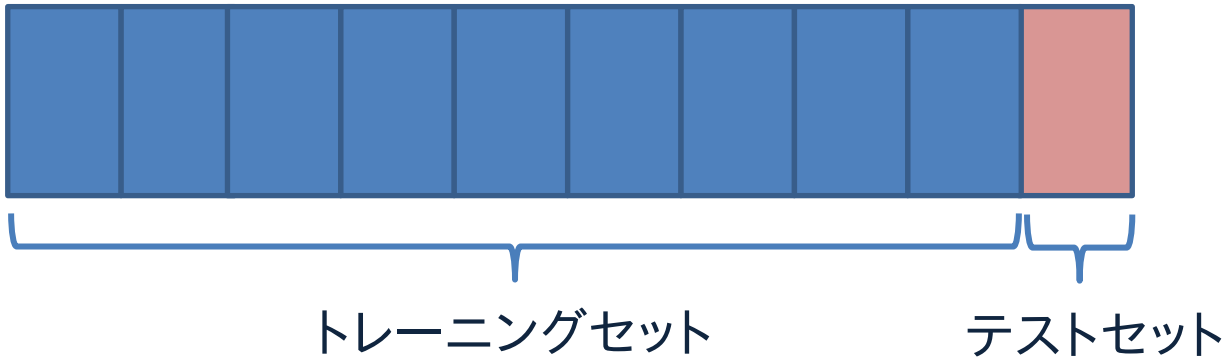

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: contact-lenses
Instances: 65
Attributes: 5
Attributes
All
None
Invert
Pattern
No.
Name
1 age
2 spectacle-prescrip
3 astigmatism
4 tear-prod-rate
5 contact-lenses
Remove
Selected attribute
Name: contact-lenses
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
1 soft 31
2 hard 17
3 none 17
Class: contact-lenses (Nom)
Visualize All
31
17
17
Status
OK
Log
x 0


age
spectacle-prescrip
astigmatism
tear-prod-rate
contact-lenses


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) contact-lenses
Start
Stop
Result list (right-click for options)
14:25:14 - trees.J48
14:33:27 - trees.J48
Classifier output
=== Run information ===
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 1
Relation: contact-lenses
Instances: 65
Attributes: 5
age
spectacle-prescrip
astigmatism
tear-prod-rate
contact-lenses
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
astigmatism = no
| spectacle-prescrip = myope: soft (45.0/17.0)
| spectacle-prescrip = hypermetrope
| | tear-prod-rate = reduced: none (3.0)
| | tear-prod-rate = normal: soft (4.0/1.0)
astigmatism = yes
| tear-prod-rate = reduced: none (6.0)
| tear-prod-rate = normal: hard (7.0/3.0)
Number of Leaves : 5
Size of the tree : 9
Status
OK
Log
x 0

k 分割交差検定（k-hold cross-validation）

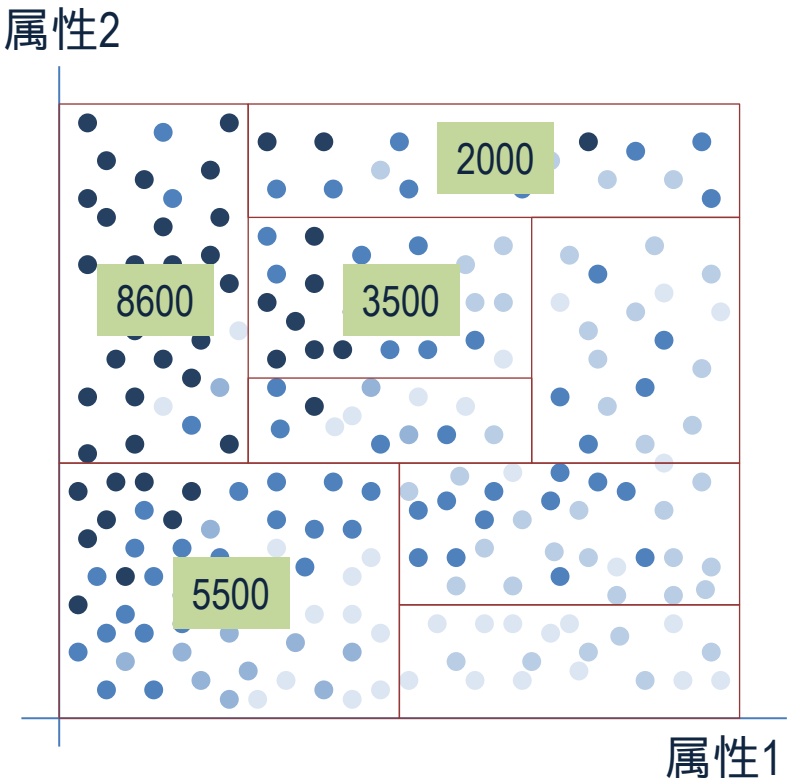
- 分類結果の精度評価.
- 過度な適合を避けることが目的.
- 標本データをトレーニングセットとテストセットに分割し，トレーニングセットでの解析結果をテストセットで検定する.
- k-分割交差検定では，標本群をk個に分割し，k-1 個のトレーニングセットで学習した結果を残りの1つのテストセットで検証する（これをk回繰り返して，平均をとり，分類方法の評価とする）.

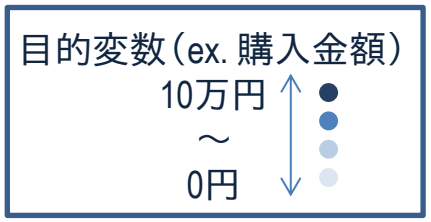


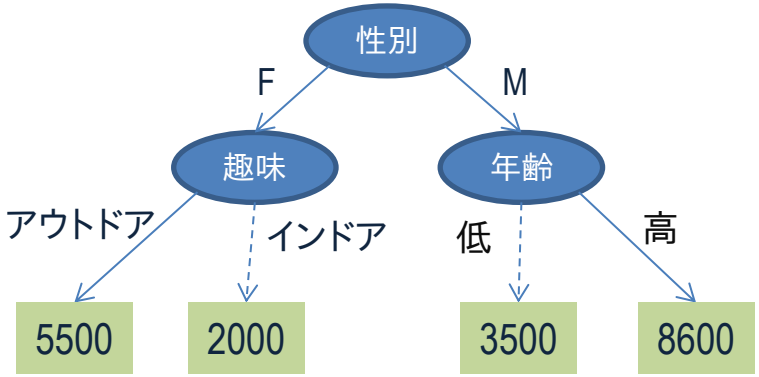
- 10-hold cross-validation あたりが一般的.

決定木分析の拡張（Regression Tree：回帰木）

[Breiman 1984] Classification and Regression Trees
「CART」（商用ソフトウェア）等に実装
目的変数が数値（i.e., 順序に意味がある）







決定木分析の拡張（Regression Tree：回帰木）

項目の選択（離散項目の場合）

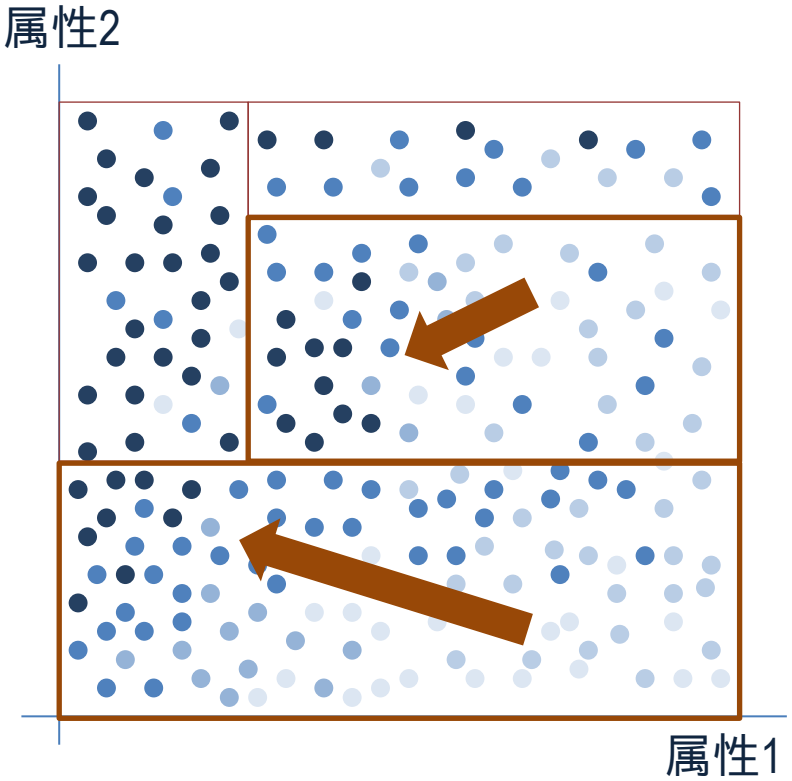
$$Gain(S, Attribute) = SquaredError(S) - \sum_{v \in Attribute} SquaredError(S_v)$$

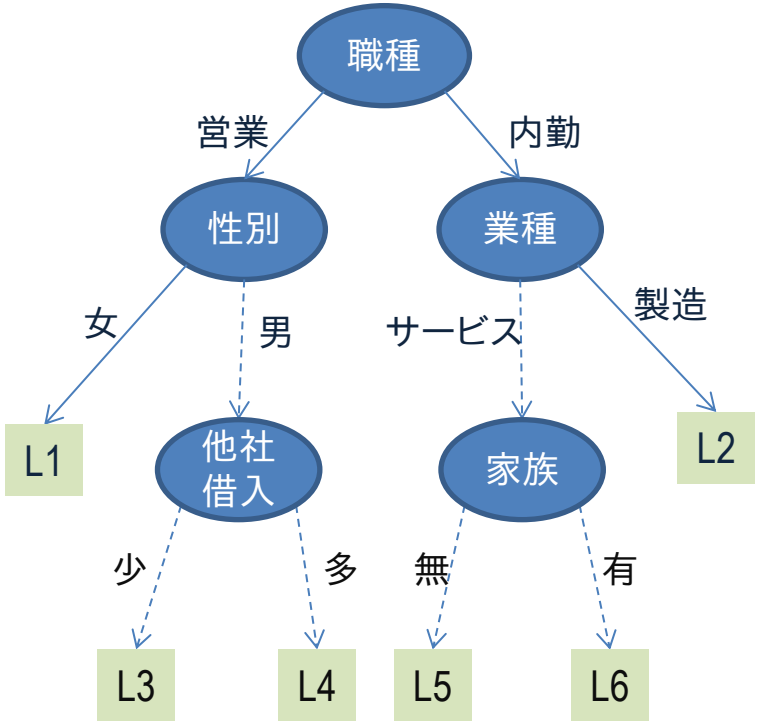
$$SquaredError(S) = \sum_{x \in S} (x - \overline{x_S})^2$$

連続値の場合には，最も二乗誤差が小さくなるような閾値を選択.

決定木分析の拡張（Model Tree：モデル木）

[Quinlan 1992] Learning with Continuous Classes
しばしばM5と呼ばれる.
リーフに線形回帰式



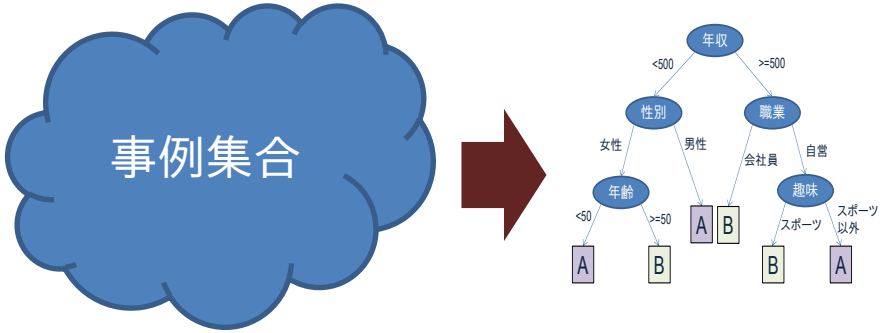


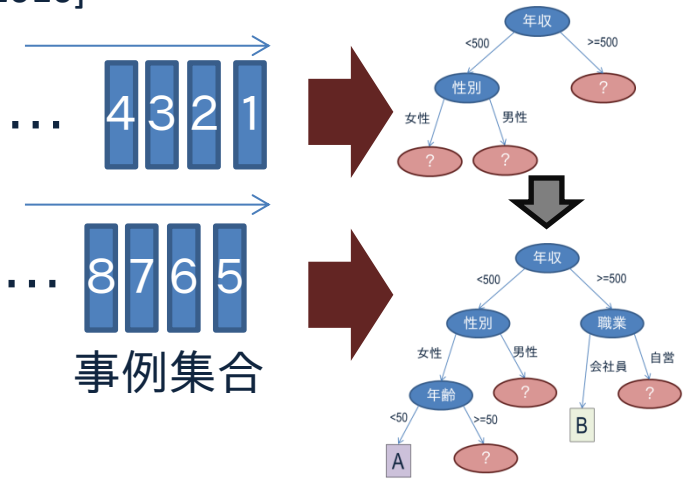
L4：c+w1*借入額+w2*年齢+w3*{家族=有}

木が深くなるとほとんど意味をなさない. 常識的には, リーフ数は数個程度が適当（な例が多い）. 単一の線形回帰では無理があるような場合.

ストリームデータマイニング（分類器）

- すべての事例をもとに決定木を構築する⇒オフライン型決定木
- ストリームデータに対して決定木を拡張していく⇒オンライン型決定木
- 特にリアルタイム性を要求されるアプリケーションでは，オンライン型決定木が有用
  （クレジットカードの不正使用，ネットワークの不正アクセス，金融取引など）
- 一般にはオフライン型と比べ，オンライン型は大容量のデータを扱う．
- VFDT(Very Fast Decision Tree learner)
  [P. Domingos et al, “Mining High-Speed Data Streams”, 2010]
  データストリームに対応した決定木構築アルゴリズム





オフライン型分類器　　オンライン型分類器

直感的には：確度の高いところから分岐させる．そうでないところは必要なデータのみを保持しておく．確度が高くなったところで分岐させ，不要な情報は消去する．

ストリームデータマイニング（VFDT）

- VFDT(Very Fast Decision Tree learner)
  [P. Domingos et al, “Mining High-Speed Data Streams”, 2010]
  データストリームに対応した決定木構築アルゴリズム

- ノードを分割するタイミングはいつか？
  → Hoeffding Bounds（ノード分割のための情報量基準値）

  一番情報利得の大きな属性分割 a と二番目に大きな属性分割 b に対して, 以下が成り立つとき, 属性 X_a でノードを分割することが $1-\delta$ の確率で正しい.

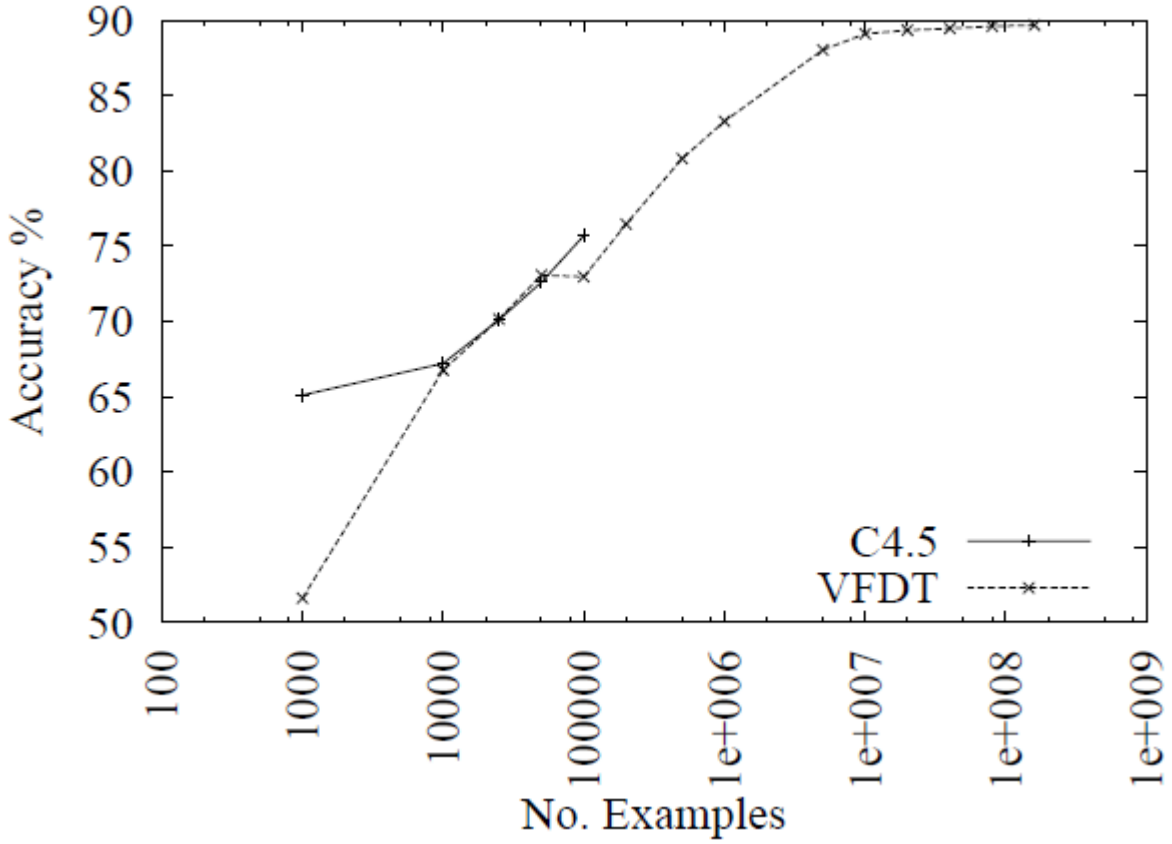
$$G(X_a) - G(X_b) > \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}$$

R: 変数の値域
n:観測回数

δを小さくすると信頼度が上がるが（当然）条件は厳しくなる.
n を大きくすると小さな差でも分割できる.

ストリームデータマイニング（VFDTとC4.5の比較）

精度（Accuracy）はほとんどC4.5 と遜色ない（当然）.
しかし, C4.5 では大量データをまとめて処理することができない.



[P. Domingos et al, “Mining High-Speed Data Streams”, 2010]

頻出パタンや決定木を使ったマイニング事例については, このあと紹介します.