データサイエンス講座
第3回 機械学習その２

- ロジスティクス回帰
- カーネル法とサポートベクターマシン
- アンサンブル学習

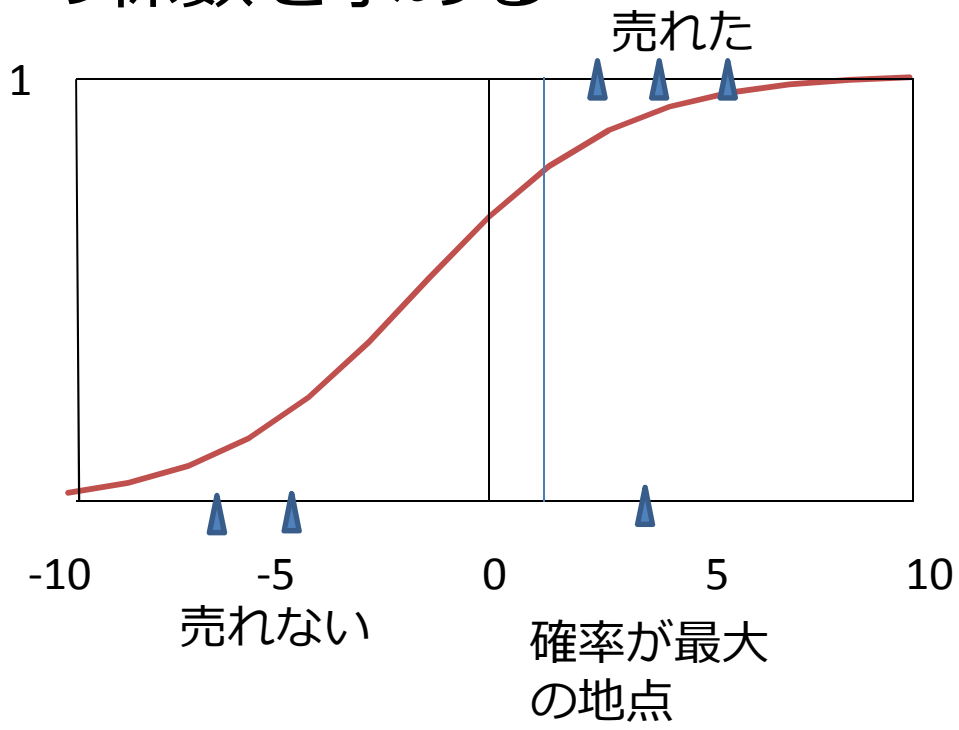
ロジスティクス回帰

- 基本的には重回帰分析のモデルと考え方は似ている

$$y = \frac{1}{1 + e^{-(a_1x_1 + a_2x_2 + a_3x_3 + a_px_p + b)}}$$

目的変数　＝　係数　×　説明変数　＋　定数

- この式をグラフ化するとyは0～1に収まる（シグモイド関数）
- トレーニングデータから確率を最大となる地点をもとめ、それぞれの係数を求める



ロジスティクス回帰

ロジスティクス回帰のメリット

- 結構、メリットが多い

1. カテゴリ変数（男性・女性、好き・嫌い）も説明変数として扱うことができる
2. 重回帰分析の一種なのでステップワイズ(AIC)によって、パラメータを削減して、説明力の高いモデルを作ることができる
3. 個々の説明変数をオッズ比（他と比較した確率の起こりやすさ）で示すことができるので、何が重要なパラメータか説明しやすい。

ランク	説明変数	P値 有意確率	オッズ比
1	40代_年収	0.1%	1.23
2	性別_男性	0.4%	1.12

- Excelではサポートしていないので、 R, Python, SPSSなどで実施するケースが多い、医療統計ではデフォルトで利用
- ぜひ、マスターして、戻ってから実践してください

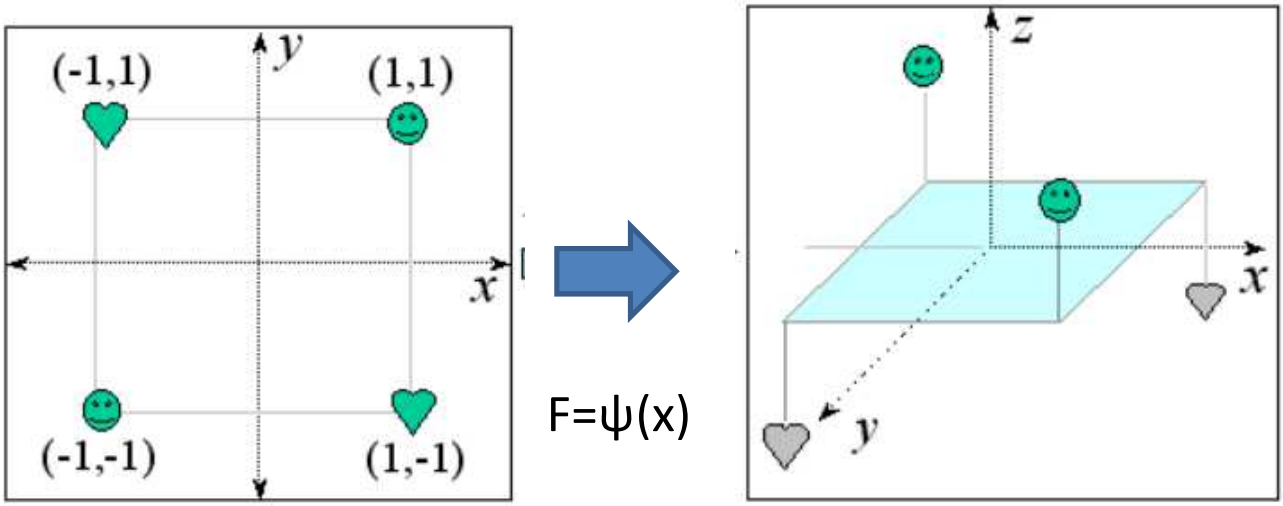
カーネル法

◘ カーネル法

- データを高次元の特徴空間に写像したときの主成分分析
- 具体的には、２次元平面座標(x,y)に、A1(1,1),A2(1,-1),A3(-1,-1),A4(-1,1)があるとする
- A1,A3が一つのクラスであるとすると、平面上にクラスの境界線を引けない
- 二次元平面(x,y)の４つの点を３次元空間(x,y,z)に射影すると、A1(1,1,1)、A2(1,-1,-1)、A3(-1,-1,1)、A4(-1,1,-1)になり、両クラスは平面で切り分けることが可能になる。
- 高次元の特徴変換をカーネル法（ψ(x))と呼ぶ

参考ビデオ
https://www.youtube.com/watch?v=3liCbRZPrZA&hl=ja&gl=JP

データ写像



明確に境界を線引きできない　　z=0の平面を境界面

カーネル法

- 高次元空間変換の問題点
  - 変換にあたって内積を計算する
  - ３次元の場合の内積、　(X,Y,Z) → (X, Y ,Z, X^2.Y^2.Z^2,XY,YZ,ZX)
  - 次元が増えれば増えるほど計算量は増える
  - どうやって計算量をすくなくするか？
  - カーネル法の登場　特徴空間への非線形写像

$$\varphi(x_1)^T \varphi(x_2) = K(x_1, x_2)$$

  - 再生核ヒルベルト空間（RKHS)の条件で、高次元（T)の内積をカーネル関数K(x1,x2)に変換することが可能（カーネルトリック）
  - いくつかのカーネル関数（K)
  - ガウスカーネル（RBFカーネル）　正規分布に従う

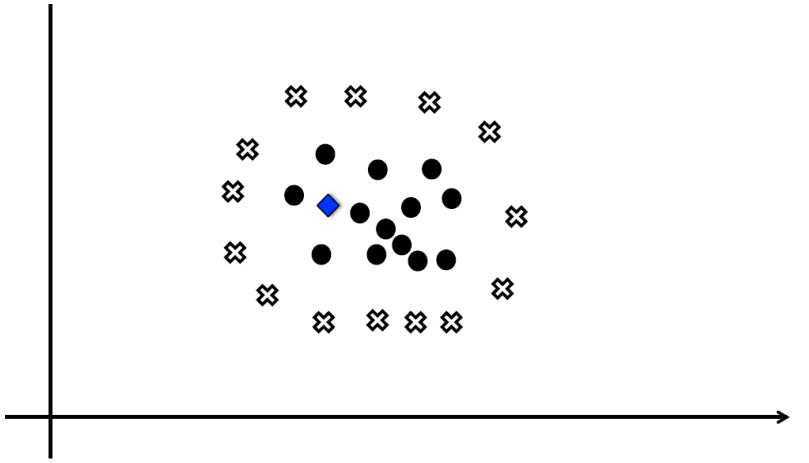
$$K(x_1, x_2) = \exp\left(\frac{-\|x_1 - x_2\|}{2\sigma^2}\right)^2$$

$$K(x_1, x_2) = \tanh(a\, x1^T - b)$$

  - 無限次元から写像できるガウスカーネルがカーネル法のカーネル関数としてよく利用される

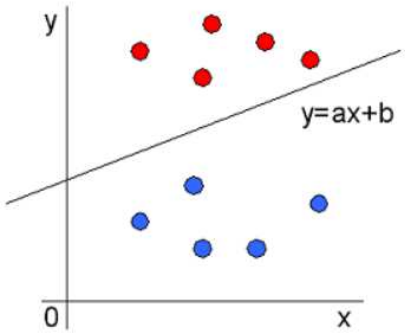
サポートベクターマシン

- カーネル法　≠　機械学習
- サポートベクター　＝　機械学習
- サポートベクターマシンの目的：
  - 教師あり学習
  - あるデータを教師ありデータに基づいて分類したい
  - 分類　→　回帰式（ y = ax + b)で説明したい
  - ただし、単純な回帰式では、◆を説明できない
  - カーネル法で、２次元座標を多次元特徴空間に変換
  - 特徴空間で、回帰分析をすることで、分類する
  - 境界線のギリギリ（マージン最大化）で線を引く
  - サポートベクターマシン＝カーネル法＋回帰分析

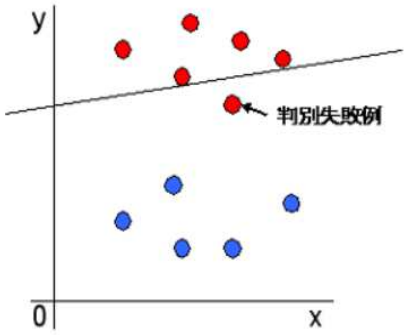


サポートベクターマシン

- サポートベクターマシンの目的
  - きちんと線引きをしたい

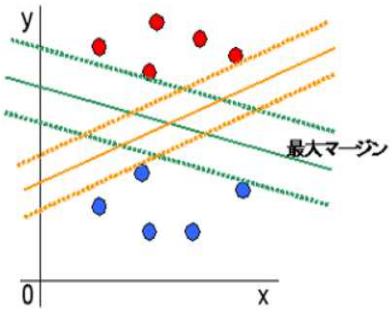


  - 場合によっては、判別に失敗するケースも

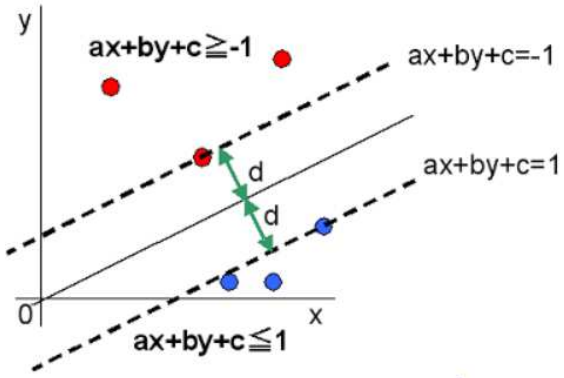


サポートベクターマシン

- サポートベクターマシンの考え方



  - 緑とオレンジ、2つの線を引く
  - 緑とオレンジと較べて、緑の方が赤と青の隔てる幅が広い（マージンが大きい）
  - 赤と青との最大マージンを取る線を選ぶ

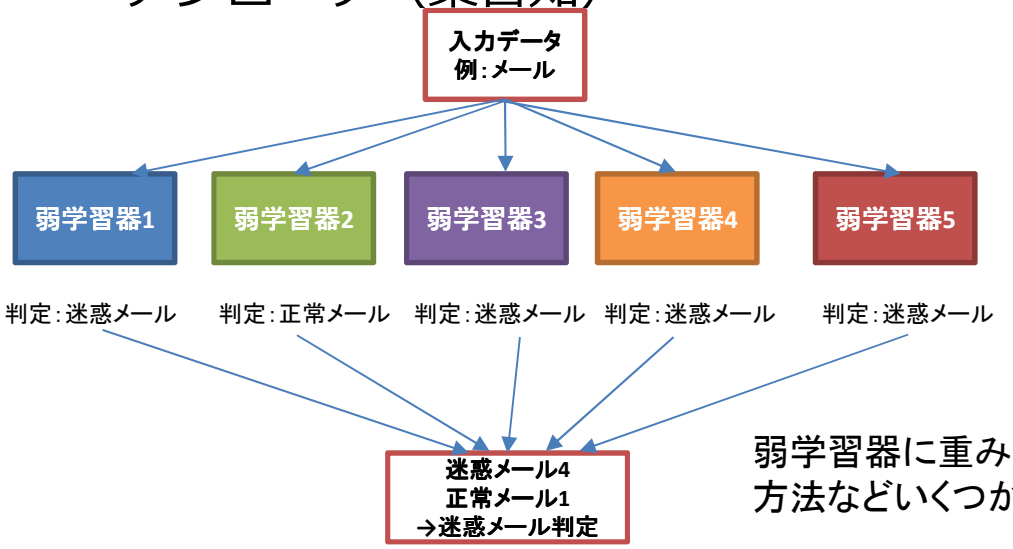


  - 赤と青のそれぞれの端っこのデータに注目（サポートベクター）
  - 赤であれば、-1となるax+by+c= -1, 青であれば、1となるax+by+c=1 となる端っこに線を引く
  - その赤と青の端っこが一番遠い距離（最大マージン）を数学的な手法を計算し、教師データを作る

アンサンブル学習

- アンサンブル　＝　“集合”
  - 例：　音楽　→　2人以上で演奏する

- アンサンブル学習の考え方
  - モデリングの難しいところ　→　過学習
  - 正確なモデルにしようとすればするほど、汎用性は失われる
  - アンサンブル学習の考え方　→　たくさんのモデルを作成して、そこから多数決・平均をとるアプローチ（集合知）



弱学習器に重みをつける
方法などいくつかパターンあり

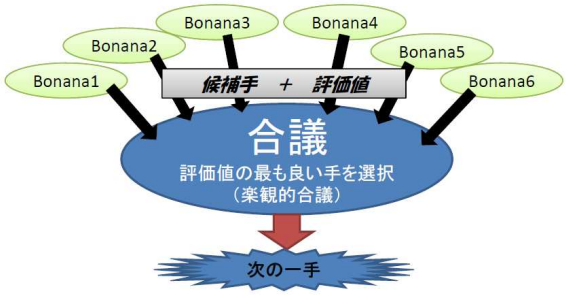
- 弱学習器のパターン
  - バギング　– 判定に際して単純に平均、多数決
  - ブースティング - 誤検知率に応じて重みをつける
  - ランダムフォレスト – 複数の決定木から平均多数決をとる

アンサンブル学習

- アンサンブル学習は精度は高いか？
  - 厳密に証明はできないものの、データ数が多くなればなるほど、トレーニングエラーが減る傾向があり、ビックデータ解析では広く利用

- アンサンブル学習の応用例１
  - コンピュータ将棋

「文殊 with Bonanza (2009.11.7.ver)」
・・・評価関数を用いた合議



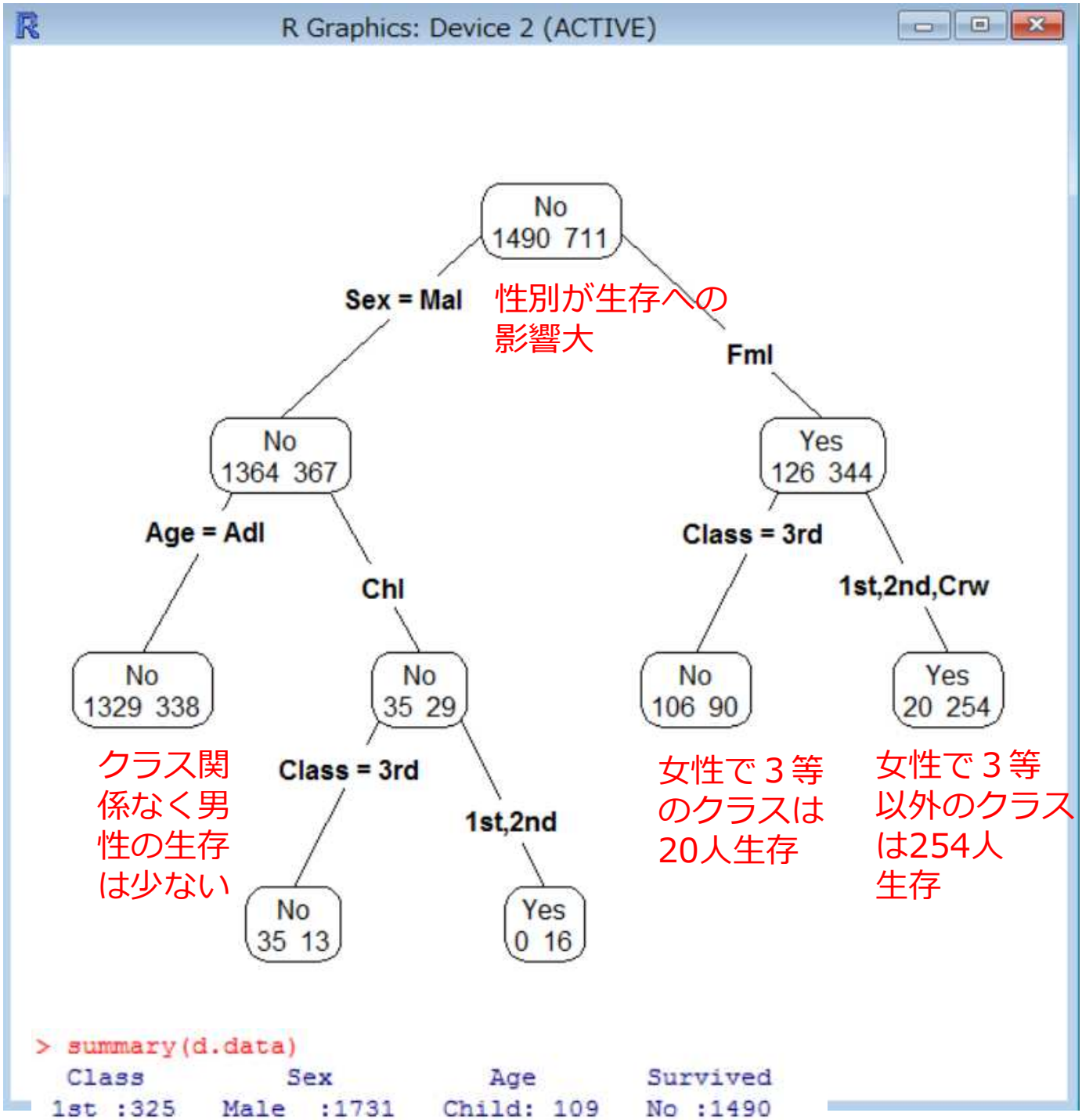
評価関数に乱数を加えた6個のBonanzaを疎結合で並列に6台繋いで、6つの候補手とその評価値を出させ、その中で、
最も評価値の高い手を選択する。
↑ *これを「楽観的合議」と呼ぶことにする*

- アンサンブル学習の応用例２
  - IBM Watson

Watsonに質問が入力されると、100を超えるアルゴリズムがさまざまな方法で質問を分析し、ふさわしい答えをたくさん見つけます。そしてこの分析はすべて同時に行われるのです。別のアルゴリズム・セットが答えをランク付けし、スコアを与えます。答えの候補それぞれについて、Watsonはその裏付けとなる根拠、またはそれを否定する反証を見つけます。数百の候補それぞれについて、数百の根拠を見つけ、数百のアルゴリズム・スコアを使って、その根拠で導き出される確信度を評価します。根拠の評価が最も高い候補が、最も高い確信度を得ます。そのうちトップの候補が、その問題の答えになります。しかし、「Jeopardy!」の対戦中、トップの候補でも、十分な確信度に達していないと判断すると、Watsonはボタンを押しませんでした。答えを間違えて、賞金を失うことを避けるためです。Watsonコンピューターはこのすべてのプロセスをおよそ3秒の間に行います。

アンサンブル学習

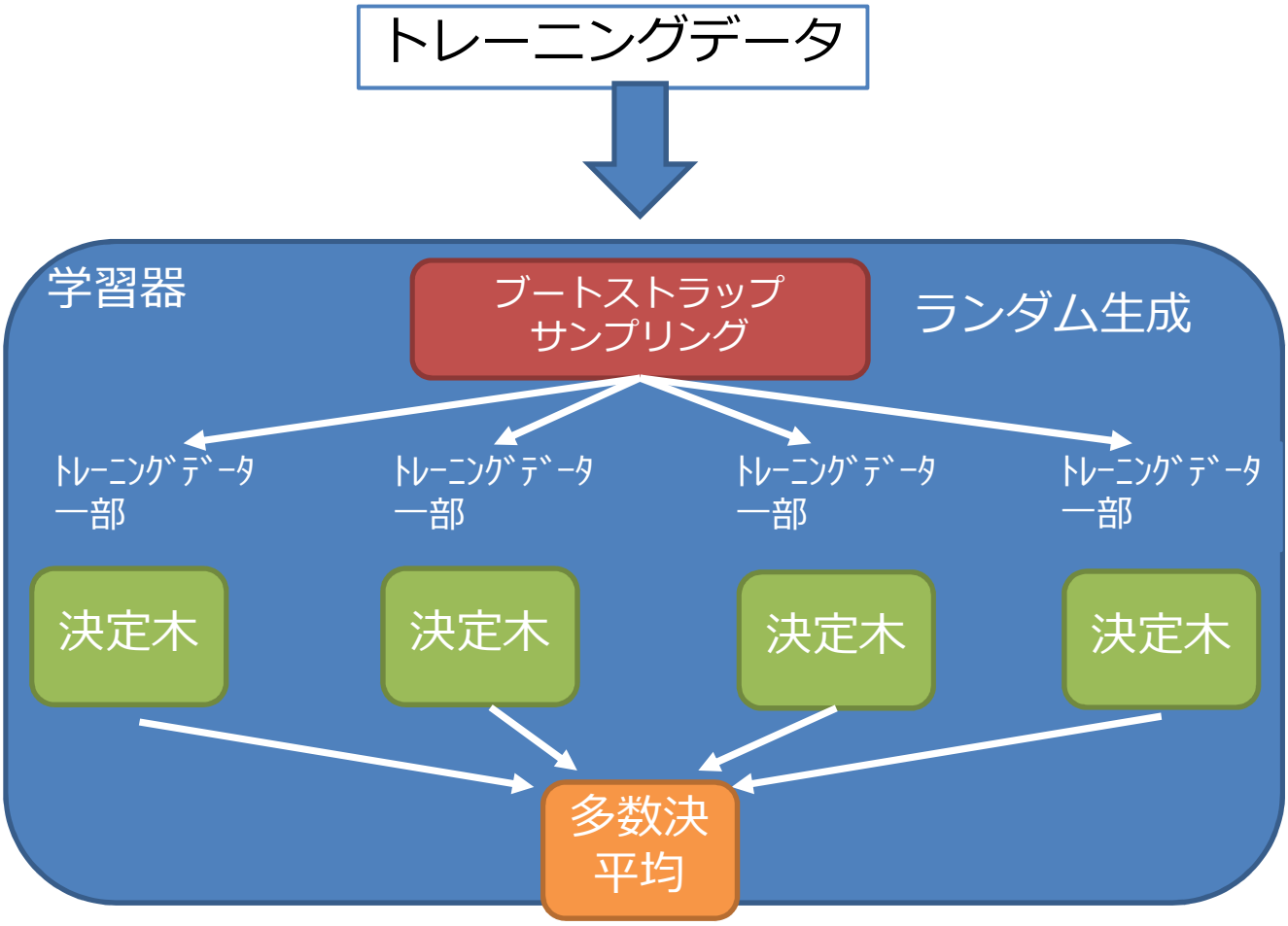
- 決定木　→　女性の生存率がもっとも高い→ロジスティクス回帰のオッズと同じ結論




```
> summary(d.data)
  Class          Sex          Age        Survived
1st :325    Male   :1731   Child: 109   No :1490
2nd :285    Female: 470    Adult:2092   Yes: 711
3rd :706
Crew:885
```

アンサンブル学習

- もどって、ランダムフォレスト
- ランダムフォレストの仕組み



アンサンブル学習

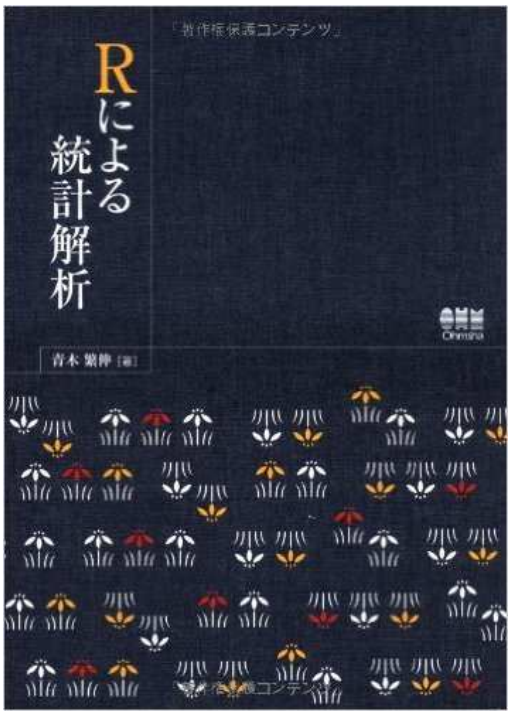
- ランダムフォレストのアルゴリズム
  1. トレーニングデータからブートストラップサンプリングを作成する
  2. ブートストラップサンプルから決定木Ti を構築、指定したノード数になるまで以下を繰り返す
     a. p個の説明変数からm個の変数をランダムに選択する
     b. mこの説明変数から最も説明しやすい変数を分岐ノードとする
  3. B個の決定木Ti を用いて学習器を構築
  4. 最終的に判別問題は多数決、回帰問題は平均で答えを出す
- 決めるべきパラメータとして、指定したノードの数、m個の変数
  - 判別問題：ノード1 $m=\sqrt{p}$
  - 回帰問題：ノード5 $m=p/3$

アンサンブル学習

- バイアス・バリアンス理論
  - 汎化誤差=バイアス+バリアンス+削除不能誤差
  - バイアス（Bias) → トレーニングデータから統計モデルを学習するアルゴリズムの良し悪し
  - バリアンス（Variance) → トレーニングデータに由来する誤差

- 単純なモデル（回帰式など)
  - 単純ゆえに、バイアスは大きい → 線形なので、十分モデルを説明できないこともある
  - 一方で、単純ゆえに、トレーニングデータに対する誤差は少ない、バリアンスは低い
- 複雑なモデル（ニューラルネットなど）
  - 複雑ゆえに、バイアスは小さい
  - 一方で、複雑ゆえに、トレーニングデータに対する誤差は大きい、バリアンスは大きい
- バイアス・バリアンスはトレードオフの関係
- ランダムフォレストは、弱学習器を多様なサンプルから学習してバリアンスを減らすアプローチ

おススメ書籍



「Rによる統計解析」

2009年4月

青木　繁伸（著）

オーム社

　統計解析の大部分について網羅的に掲載されている。Rでどうするかわからなくなったとき参照すると便利。

「データサイエンティスト養成読本」

2015年9月

技術評論社

　入門といいながらも、結構、高度なところまで言及。Pythonでの応用、画像認識など踏み込んで取り上げている。本講座の次のステップとしておススメ。