デシジョンツリーモデルの発展

2011/4/21

データマインテック株式会社
代表取締役　谷岡日出男

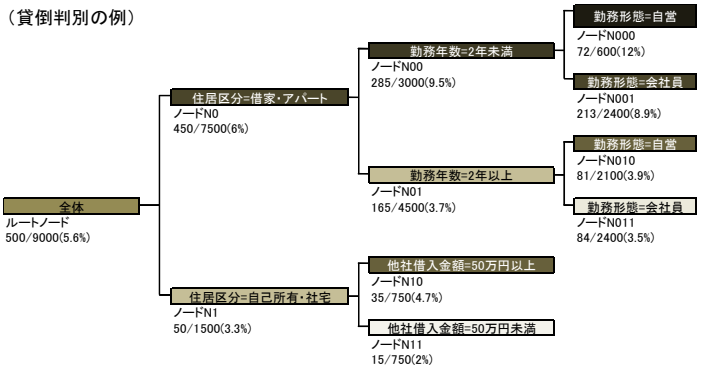
内容

1. はじめに
　1.1 デシジョンツリーモデルとは
　1.2 統計モデルとの比較
2. デシジョンツリーモデルの発展
　2.1 デシジョンツリーモデルの出現
　2.2 アンサンブル法の出現
　2.3 その他のデシジョンツリーモデル
3. まとめ（ホワイトボックスモデルとブラックボックスモデル）

1.はじめに
1.1 デシジョンツリーモデルとは

- **クラス判別**や**数値予測**に用いる
- 目的変数＝モデル＋誤差　の式におけるモデルが、説明変数の値による**逐次的データ分割**（**ノード分岐**）を繰り返し行うことにより得られる**木構造のモデル**

（貸倒判別の例）

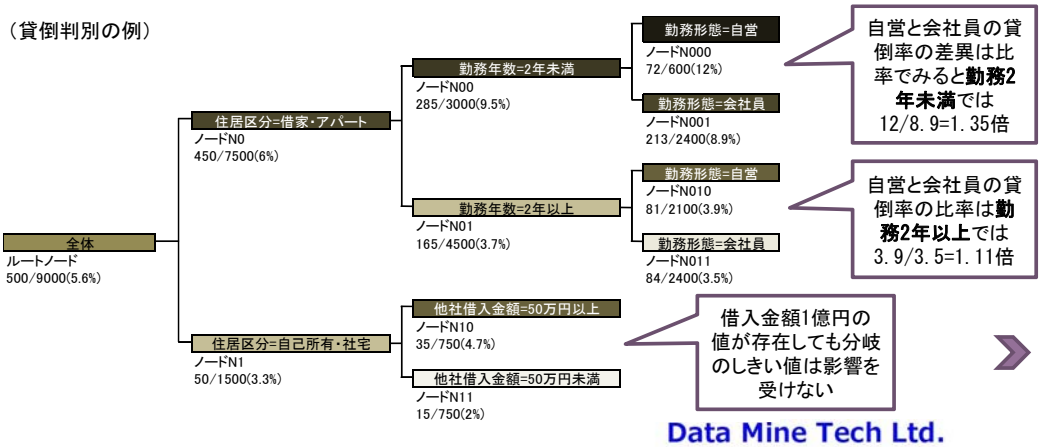


終端ノード	分岐条件1		分岐条件2		分岐条件3	ノード該当件数	貸倒件数	貸倒率
N000	借家またはアパート	かつ	勤務2年未満	かつ	自営	600	72	12.0%
N001					会社員	2,400	213	8.9%
N010			勤務2年以上		自営	2,100	81	3.9%
N011					会社員	2,400	84	3.5%
N10	自己所有または社宅		他社借入50万以上			750	35	4.7%
N11			他社借入50万未満			750	15	2.0%

デシジョンツリーモデルの長所

- **大規模データ**の分析が可能
- **交互作用を自動的にモデルに取り込む仕組み**（説明変数の値の変化がもたらす目的変数の変化の大きさが他の説明変数の値によって異なる状況）
- 説明変数の**異常値**に引きずられにくい
- 説明変数に**欠損値**があっても分析可能
- 予測や判別といった用途以外に、**要因分析**や**変数選択**に利用できる
- モデルの作り方やアウトプットが**直観的に理解できるホワイトボックスモデル**

（貸倒判別の例）

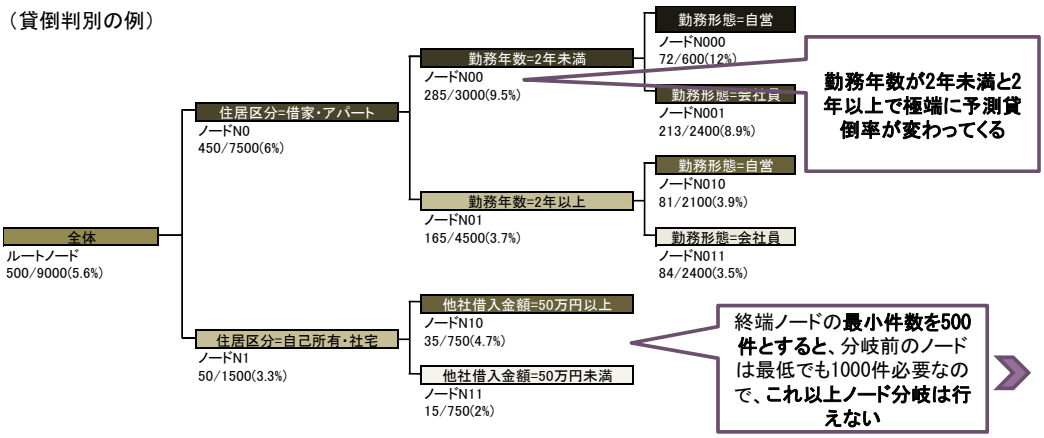


Data Mine Tech Ltd.
Data Bring New Insight to Your Business

デシジョンツリーモデルの短所

- **予測値の種類数が少ない**（予測値の種類数＝**終端ノード数**）
- **終端ノード数を増やすには十分なデータ件数が必要**（分岐数を増やすと予測値の種類数が増加する。その一方、分岐数を増やすとノード該当件数が減少するため、予測精度が悪化する）
- **数値説明変数の値の予測値への影響の仕方が極端**（あるしきい値を境としてその値以上/未満によってノード分岐を行う。**特に最初の方のノード分岐に採用された数値説明変数**の値がしきい値を超える/超えないで予測結果が大きく異なる場合がある）

（貸倒判別の例）



Data Mine Tech Ltd.
Data Bring New Insight to Your Business

1.2 統計モデルとの比較

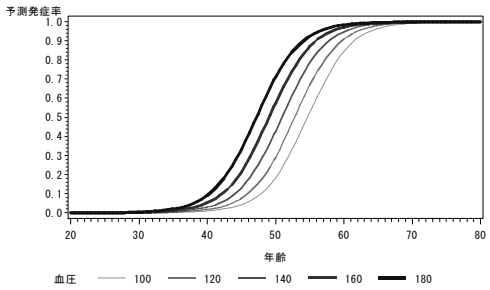
- クラス判別や数値予測を行うためのデータ解析手法としては、19世紀末～20世紀前半にかけて、以下のような統計解析手法が開発された
  - 回帰分析(1877 Galton,1921 Fisher)
  - ロジスティックモデル(1838 Verhulst, 1920 Peal,Reed)
  - 分散分析法(1925 Fisher)共分散分析法(1928 Fisher)
  - 判別関数法(1936 Fisher)
  - 正準判別分析(1962 Cooley,Lohnes)
  - 数量化I類,II類(1950年代 林)
- ロジスティックモデルの例

$$Z = -56.5 + 0.32 * \text{年齢} + 0.03 * \text{血圧} + 0.45 * \text{腹囲}$$

$$\text{成人病発症率} = \frac{\exp(Z)}{1+\exp(Z)}$$

$Z = 0$を境として予測出現率　は50%以上／以下となる

ロジスティックモデルによる年齢と血圧と成人病発症率の関係



- 統計モデルの長所:
  - **ホワイトボックスモデル**(モデル式が明確)
  - **多数の予測値**(説明変数の値による**連続的な予測値**)
  - 比較的**少数データ**でモデル構築が可能
- 統計モデルの短所:
  - **説明変数効果の独立性**(その説明変数が1単位増加したときの目的変数の変化の大きさ(重み係数)は**他の説明変数の値に無関係で一定**という前提)
  - **重み係数の解釈が困難**な場合がある(説明変数間に相関がある場合の係数の符号など)
  - **異常値・欠損値**に弱い

Data Mine Tech Ltd.
Data Bring New Insight to Your Business

統計モデルとデシジョンツリーモデルの比較

- 長所・短所をまとめると

比較項目	統計モデル		デシジョンツリーモデル	
	対応	比較	対応	比較
モデル作成に**必要な**データ件数	少数(**～数万件**)	○	多数(**数千～数百万件**)	×
モデル作成に用いることが**可能な**データ件数		×		○
分析できる**説明変数の数**	少数(**～数百**)	×	多数でも可(**～数千**)	○
説明変数間の**交互作用**	**事前知識**もしくはモデル構築過程で検討	×	**自動抽出**	○
説明変数の**異常値・欠損値**	異常値は分析結果に影響大。欠損値を含むオブザベーションは分析に用いることができない	×	異常値は分析結果に影響小。欠損値を含むオブザベーションを分析に用いることができる	○
数値説明変数の値の変化による**予測値の変化**	数値タイプ説明変数の値の変化は予測値の**連続的な変化**として現れる	○	数値説明変数は通常、モデル上ではある値をしきい値としてそれ以上とそれ未満の2つのカテゴリカル変数としてモデルにとりこまれるため、他の説明変数値を一定に保つと、しきい値を超えない変化は予測値に何ら変化をもたらさない。一方、**しきい値を跨ぐ変化は予測値に大きな変化**をもたらす可能性がある	×
予測値の**種類数**	連続的な予測値をとるため**無数**にある	○	**終端ノード数**になる	×
モデルの**作り方**(理論)の理解のしやすさ	モデル推計方法について理解をしておく必要があるが作り方は明確(**ホワイトボックスモデル**)	○	通常はロジックは簡単で理解しやすい**ホワイトボックスモデル**。(ただし、ブースティングツリーモデルなどブラックボックスモデルもある)	○
モデルの**アウトプット**(結果)の理解しやすさ	モデルは予測値を算出する**計算式**の形で出力され、非常に分かりやすい。ただし、説明変数間の相関があると、**説明変数にかかる重みパラメータの符号や値の大きさが解釈不能**となる場合がある	△	モデルは**ツリー図**(樹形図)の形で出力され、非常に分かりやすい。予測結果も説明変数の値の組合せによって定義された各ノード内の目的変数の分布に基づいており、解釈しやすい	○
モデルの**システムへの組込み**のしやすさ	計算式に含む**説明変数と重み係数の掛け算を説明変数分足しこむ**ことで予測値(予測スコア)が得られるという仕組みなので、システムへの組込みは簡単	○	**ツリー図の分岐を辿ってどの終端ノードに所属するかを決定**し終端ノードに付いた予測値を参照する仕組みなので、システムへの組込みは簡単	○

基本的な使い分け

- **データ件数が少ない**ときは統計モデル、
  データ件数が多いときはデシジョンツリーモデル

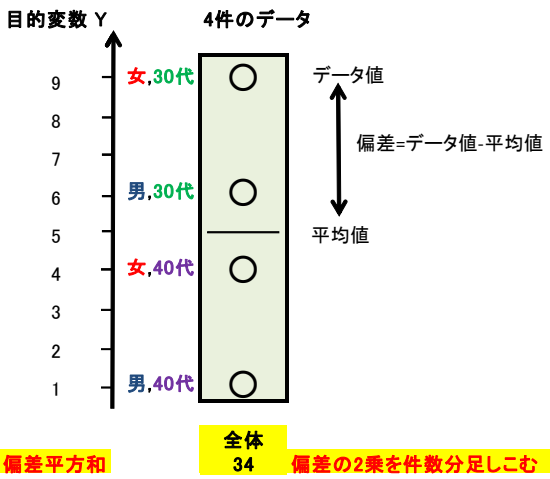
Data Mine Tech Ltd.
Data Bring New Insight to Your Business

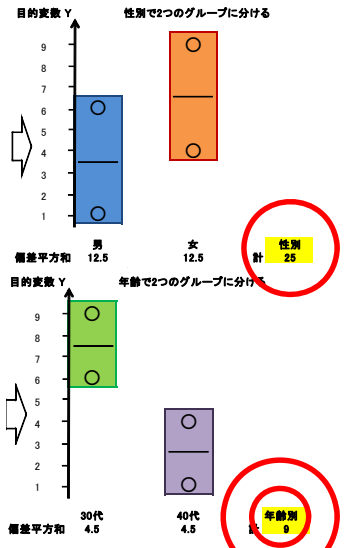
2.デシジョンツリーモデルの発展
2.1 デシジョンツリーモデルの出現

- 1990年代半ば以前に、統計学の応用分野の1つであったマーケティングリサーチから生まれたAID, CHAIDなどと、人工知能の一分野である機械学習から生まれたCLS, CART, ID3, C4.5などの2つの系統のデシジョンツリーモデルが誕生した
  - CLS(Concept Learning System 概念学習システム)(1966 Hunt, Martin, Stone)
  - AID(Automatic Interaction Detector 自動交互作用検知器)(1971 ミシガン大 Sonquist, Baker, Morgan)
  - CHAID(ChiSquare AID カイ2乗AID)(1980 ノースカロライナ大チャペルヒル校 Perreault, Hiram, Barksdale)
  - CART(Classification And Regression Trees 分類木と回帰木生成器)(1984 カリフォルニア大バークレー校 Breiman, Stone とスタンフォード大 Friedman, Olshen)
  - ID3(Iterative Dichotomizer 3 繰り返し2分木生成器)(1986 シドニー大 Quinlan)
  - C4.5(ID3の拡張版)(1993 Quinlan)

AID
数値予測

- **目的変数は数値**変数
- 説明変数は数値・カテゴリカルいずれでも良い
- 2分木(2つのグループに分岐)
- 分岐説明変数選択基準はその説明変数で分けた後の目的変数の**偏差平方和の合計**最小





- この場合は、性別で分けた場合より**年齢**で分けた後のYの偏差平方和の合計が小さくなっている

AID
数値予測

- 12件の事例データ
- 勤務、性別、年齢を説明変数、年収を目的変数とする

ID	勤務	性別	年齢	年収 Y	年収 - 年収平均 $Y - \overline{Y}$	(年収 - 年収平均)^2 $(Y-\overline{Y})^2$
0001	嘱託	男性	45	580	62.5	3906.3
0002	正社員	女性	36	390	-127.5	16256.3
0003	嘱託	男性	38	450	-67.5	4556.3
0004	嘱託	女性	41	480	-37.5	1406.3
0005	正社員	女性	25	380	-137.5	18906.3
0006	正社員	男性	35	620	102.5	10506.3
0007	嘱託	女性	22	300	-217.5	47306.3
0008	嘱託	男性	40	550	32.5	1056.3
0009	嘱託	男性	52	580	62.5	3906.3
0010	正社員	女性	40	550	32.5	1056.3
0011	嘱託	女性	50	580	62.5	3906.3
0012	正社員	男性	55	750	232.5	54056.3
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$	分岐前の平方和
				517.5		166825.0	**166825.0**

Data Mine Tech Ltd.
Data Bring New Insight to Your Business

AID
数値予測

- 例として性別と年齢でそれぞれ2つのノードに分けた場合の年収の偏差平方和の合計値を調べる
- 分けた後の2つのノードの偏差平方和の合計が小さい方の分け方を分岐方法として採用する

勤務形態で分けてみる

ID	勤務	性別	年齢	年収 Y
0001	嘱託	男性	45	580
0002	正社員	女性	36	390
0003	嘱託	男性	38	450
0004	嘱託	女性	41	480
0005	正社員	女性	25	380
0006	正社員	男性	35	620
0007	嘱託	女性	22	300
0008	嘱託	男性	40	550
0009	嘱託	男性	52	580
0010	正社員	女性	40	550
0011	嘱託	女性	50	580
0012	正社員	男性	55	750

ID	勤務	性別	年齢	年収 Y	年収 - 年収平均 $Y - \overline{Y}$	(年収 - 年収平均)^2
0002	正社員	女性	36	390	-148.0	21904.0
0005	正社員	女性	25	380	-158.0	24964.0
0006	正社員	男性	35	620	82.0	6724.0
0010	正社員	女性	40	550	12.0	144.0
0012	正社員	男性	55	750	212.0	44944.0
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$
				538.0		98680.0

ID	勤務	性別	年齢	年収 Y	年収 - 年収平均 $Y - \overline{Y}$	(年収 - 年収平均)^2
0001	嘱託	男性	45	580	77.1	5951.0
0003	嘱託	男性	38	450	-52.9	2793.9
0004	嘱託	女性	41	480	-22.9	522.4
0007	嘱託	女性	22	300	-202.9	41151.0
0008	嘱託	男性	40	550	47.1	2222.4
0009	嘱託	男性	52	580	77.1	5951.0
0011	嘱託	女性	50	580	77.1	5951.0
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$
				502.9		64542.9

分岐前の平方和 166825.0

勤務で分けた場合の平方和 163222.9

性別で分けてみる

ID	勤務	性別	年齢	年収 Y
0001	嘱託	男性	45	580
0003	嘱託	男性	38	450
0006	正社員	男性	35	620
0008	嘱託	男性	40	550
0009	嘱託	男性	52	580
0012	正社員	男性	55	750

ID	勤務	性別	年齢	年収 Y	年収 - 年収平均 $Y - \overline{Y}$	(年収 - 年収平均)^2
0001	嘱託	男性	45	580	-8.3	69.4
0003	嘱託	男性	38	450	-138.3	19136.1
0006	正社員	男性	35	620	31.7	1002.8
0008	嘱託	男性	40	550	-38.3	1469.4
0009	嘱託	男性	52	580	-8.3	69.4
0012	正社員	男性	55	750	161.7	26136.1
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$
				588.3		47883.3

ID	勤務	性別	年齢	年収 Y	年収 - 年収平均 $Y - \overline{Y}$	(年収 - 年収平均)^2
0002	正社員	女性	36	390	-56.7	3211.1
0004	嘱託	女性	41	480	33.3	1111.1
0005	正社員	女性	25	380	-66.7	4444.4
0007	嘱託	女性	22	300	-146.7	21511.1
0010	正社員	女性	40	550	103.3	10677.8
0011	嘱託	女性	50	580	133.3	17777.8
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$
				446.7		58733.3

分岐前の平方和 166825.0

性別で分けた場合の平方和 106616.7

年齢で分けてみる（22歳とそれを超える）

ID	勤務	性別	年齢	年収 Y	年収 - 年収平均 $Y - \overline{Y}$	(年収 - 年収平均)^2
0007	嘱託	女性	22	300	0.0	0.0
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$
				300.0		0.0

0005	正社員	女性	25	380	-157.3	24734.7
0006	正社員	男性	35	620	82.7	6843.8
0002	正社員	女性	36	390	-147.3	21689.3
0003	嘱託	男性	38	450	-87.3	7616.5
0008	嘱託	男性	40	550	12.7	162.0
0010	正社員	女性	40	550	12.7	162.0
0004	嘱託	女性	41	480	-57.3	3280.2
0001	嘱託	男性	45	580	42.7	1825.6
0011	嘱託	女性	50	580	42.7	1825.6
0009	嘱託	男性	52	580	42.7	1825.6
0012	正社員	男性	55	750	212.7	45252.9
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$
				537.3		115218.2

分岐前の平方和 166825.0

年齢22以下と22超で分けた場合の平方和 115218.2

年齢で分けてみる（25歳以下とそれを超える）

ID	勤務	性別	年齢	年収 Y	年収 - 年収平均 $Y - \overline{Y}$	(年収 - 年収平均)^2
0007	嘱託	女性	22	300	-40.0	1600.0
0005	正社員	女性	25	380	40.0	1600.0
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$
				340.0		3200.0

0006	正社員	男性	35	620	67.0	4489.0
0002	正社員	女性	36	390	-163.0	26569.0
0003	嘱託	男性	38	450	-103.0	10609.0
0008	嘱託	男性	40	550	-3.0	9.0
0010	正社員	女性	40	550	-3.0	9.0
0004	嘱託	女性	41	480	-73.0	5329.0
0001	嘱託	男性	45	580	27.0	729.0
0011	嘱託	女性	50	580	27.0	729.0
0009	嘱託	男性	52	580	27.0	729.0
0012	正社員	男性	55	750	197.0	38809.0
				年収平均 $\overline{Y}$		年収の偏差平方和 $\sum (Y - \overline{Y})^2$
				553.0		88010.0

分岐前の平方和 166825.0

年齢25以下と25超で分けた場合の平方和 91210.0

この中では**年齢25歳をしきい値**として2つに分けると分岐後の平方和の合計が最小となる

1.

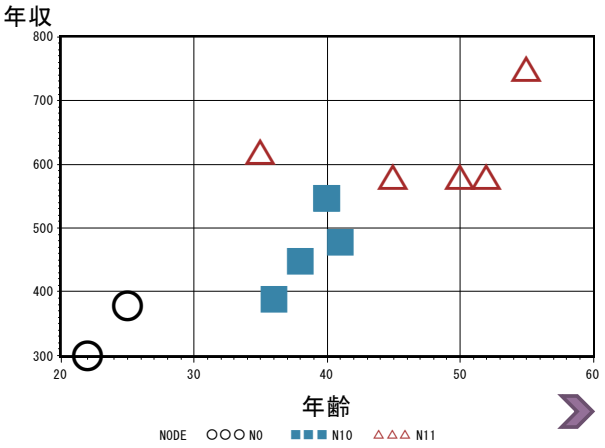
Data Bring New Insight to Your Business

AID
数値予測

- 年収を目的変数、勤務・性別・年齢を説明変数、最小ノード件数=2として2階層までツリーを作成すると年齢によって以下の3つのグループに分かれたツリーモデルができる

LVL0	LVL1	LVL2
全体(ルートノード):件数=12,年収平均=517.5,標準偏差=117.91	ノードN0:(件数=2,平均年収=340,標準偏差=40)年齢=LOW~25	
	ノードN1:(件数=10,平均年収=553,標準偏差=93.8)年齢=25~HIGH	ノードN10:(件数=5,年収平均=484,標準偏差=61.2)年齢=36~41
		ノードN11:(件数=5,年収平均=622,標準偏差=65.8)年齢=LOW~36,41~HIGH

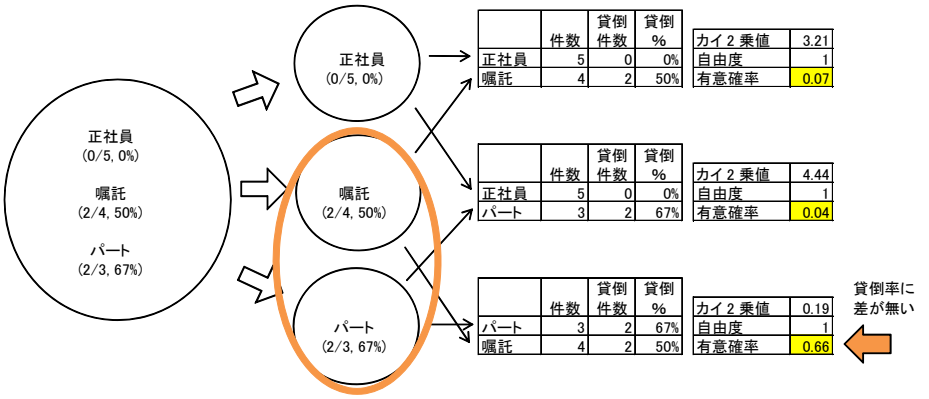
No.	ノード	条件	件数	年収平均値	年収標準偏差
1	N0	年齢25以下	2	340	40
2	N10	年齢36以上41以下	5	484	61
3	N11	年齢25超36未満または年齢41超	5	622	66



Data Mine Tech Ltd.
Data Bring New Insight to Your Business

CHAID
最適多分岐ツリー生成法

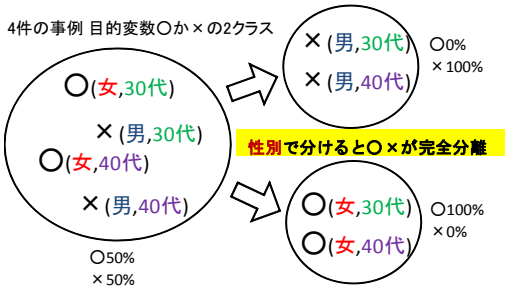
- 元来の目的変数は**クラス変数（多クラス可）**
- 説明変数は数値・カテゴリカルいずれでも良い
- 説明変数は**カイ2乗検定**を用いた**目的クラス出現率分布の一様性**に基づき、数値の場合は順序カテゴリとして、カテゴリカルの場合は名義カテゴリとして**自動併合**
- 結果として**分岐数が最適とみなされる多分岐ツリーモデル**が得られる点に特徴がある
- 数値予測を可能とするため、**F検定**に基づく説明変数カテゴリ併合法も開発された

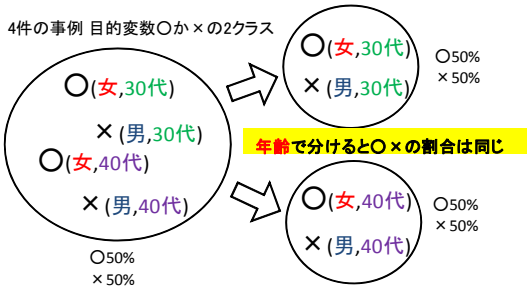


Data Mine Tech Ltd.
Data Bring New Insight to Your Business

C4.5
クラス判別

- **目的変数はクラス**変数（2クラス以上可）
- 説明変数は数値・カテゴリカルいずれでも良い
- 2分木～多分木（説明変数がカテゴリの場合にはカテゴリごとにノード分岐）
- 分岐説明変数選択基準はクラス事例混在度合いを表す**エントロピー減少差（利得）**または**利得をノード分岐数で補正した利得比**（2つに分ける場合、利得比＝利得）



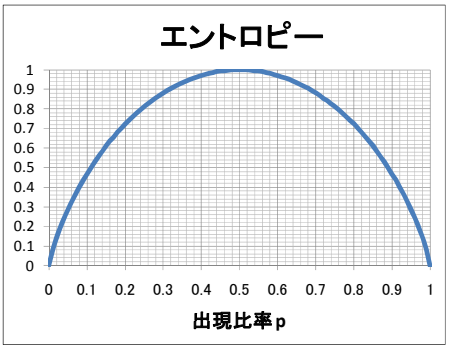


- この場合は性別で分ける方が年齢で分けるよりクラス分けがうまくできることがわかる
- クラスの混在度を計測する一般的な指標として**エントロピー**がある

C4.5
クラス判別

- **エントロピー**
  クラスの混在度を評価する指標。 ID3, C4.5の分岐説明変数選択基準。 分かれた後のエントロピーが小さい状態になるような分け方を探索する
  2クラスの場合は、一方のクラスの出現率を pとすると、

$$\text{エントロピー} = -p\log_2(p) - (1-p)\log_2(1-p)$$



- pが0か1に近く一方のクラスが優勢なほどエントロピーは小さくなり、p=0.5で2つのクラスが拮抗している状態のとき最も大きい値をとる
- k個のクラスの場合、エントロピー値を0～1に基準化する場合、

$$\text{エントロピー} = \sum_{i}^{k} -p_i \log_k(p_i) \quad \text{ただし} k\text{はクラス数}$$

- たとえばk=3の場合、p1=p2=p3=1/3（拮抗状態）のときエントロピーは最大値1、p1, p2, p3のいずれかが1のとき最小値0をとる

C4.5
クラス判別

- **12件の事例データ**
- 勤務、性別、年齢、年収を説明変数、貸付1年後の状態Yの貸倒か非貸倒の2クラスを目的変数とする

ID	勤務	性別	年齢	年収	貸付1年後の状態 Y
0001	嘱託	男性	45	580	正常残なし
0002	正社員	女性	36	390	正常残あり
0003	嘱託	男性	38	450	貸倒
0004	嘱託	女性	41	480	貸倒
0005	正社員	女性	25	380	正常残なし
0006	正社員	男性	35	620	正常残あり
0007	嘱託	女性	22	300	正常残あり
0008	嘱託	男性	40	550	貸倒
0009	嘱託	男性	52	580	貸倒
0010	正社員	女性	40	550	正常残なし
0011	嘱託	女性	50	580	正常残あり
0012	正社員	男性	55	750	正常残あり

貸倒出現率	p	貸倒/非貸倒のエントロピー	$E(p) = -p*\log_2(p) - (1-p)*\log_2(1-p)$		分岐前のエントロピー	
	0.3333		0.9183	⇨		0.9183

C4.5
クラス判別

- 勤務形態と性別でそれぞれデータを2つに分けた場合のエントロピー値を調べる
- 分けた後の件数の**重み付き平均エントロピー**が最小となる分け方を採用する

性別で分けてみる

ID	勤務	性別	年齢	年収	貸付1年後の状態 Y
0001	嘱託	男性	45	580	正常残なし
0003	嘱託	男性	38	450	貸倒
0006	正社員	男性	35	620	正常残あり
0008	嘱託	男性	40	550	貸倒
0009	嘱託	男性	52	580	貸倒
0012	正社員	男性	55	750	正常残あり

貸倒出現率	p	貸倒/非貸倒のエントロピー	$E(p) = -p*\log_2(p) - (1-p)*\log_2(1-p)$
	0.5000		1.0000

ID	勤務	性別	年齢	年収	貸付1年後の状態 Y
0002	正社員	女性	36	390	正常残あり
0004	嘱託	女性	41	480	貸倒
0005	正社員	女性	25	380	正常残なし
0007	嘱託	女性	22	300	正常残あり
0010	正社員	女性	40	550	正常残なし
0011	嘱託	女性	50	580	正常残あり

貸倒出現率	p	貸倒/非貸倒のエントロピー	$E(p) = -p*\log_2(p) - (1-p)*\log_2(1-p)$
	0.1667		0.6500

$E = \frac{n1*E1 + n2*E2}{n1+n2}$ = 0.8250

勤務形態で分けてみる

ID	勤務	性別	年齢	年収	貸付1年後の状態 Y
0002	正社員	女性	36	390	正常残あり
0005	正社員	女性	25	380	正常残なし
0006	正社員	男性	35	620	正常残あり
0010	正社員	女性	40	550	正常残なし
0012	正社員	男性	55	750	正常残あり

貸倒出現率	p	貸倒/非貸倒のエントロピー	$E(p) = -p*\log_2(p) - (1-p)*\log_2(1-p)$
	0.0000		0.0000

$E = \frac{n1*E1 + n2*E2}{n1+n2}$ = 0.5747

ID	勤務	性別	年齢	年収	貸付1年後の状態 Y
0001	嘱託	男性	45	580	正常残なし
0003	嘱託	男性	38	450	貸倒
0004	嘱託	女性	41	480	貸倒
0007	嘱託	女性	22	300	正常残あり
0008	嘱託	男性	40	550	貸倒
0009	嘱託	男性	52	580	貸倒
0011	嘱託	女性	50	580	正常残あり

貸倒出現率	p	貸倒/非貸倒のエントロピー	$E(p) = -p*\log_2(p) - (1-p)*\log_2(1-p)$
	0.5714		0.9852

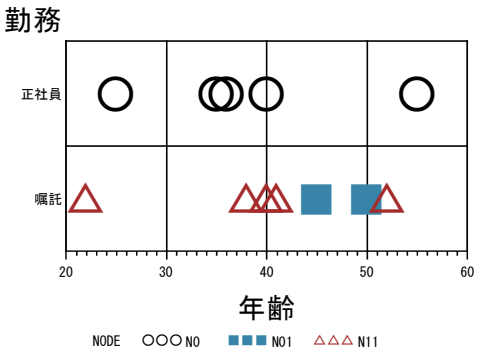
- **勤務形態**で分ける(エントロピー0.5747)方が性別で分ける(同0.8250)より良い

C4.5
クラス判別

- エントロピー最小基準分岐アルゴリズムで最小ノード件数=2として2階層までツリーを作成すると勤務と年齢によって以下の3つのグループに分かれる
- なお、2分木の場合のエントロピー最小基準は、C4.5の利得比基準と同じ

LVL0	LVL1	LVL2
全体(ルートノード):件数=12,貸倒率=33.33%	ノードN0:(件数=5,貸倒率=0%)分岐条件:勤務="正社員"	
	ノードN1:(件数=7件,貸倒率=57.14%)分岐条件:勤務="嘱託"	ノードN10:(件数=2,貸倒率=0%)分岐条件:年齢45以上50以下
		ノードN11:(件数=5,貸倒率=80.00%)分岐条件:年齢45未満または50超

No.	ノード	条件	件数	貸倒率
1	N0	正社員	5	0%
2	N10	嘱託かつ年齢45以上50以下	2	0%
3	N11	嘱託かつ(年齢45未満または50超)	5	80%



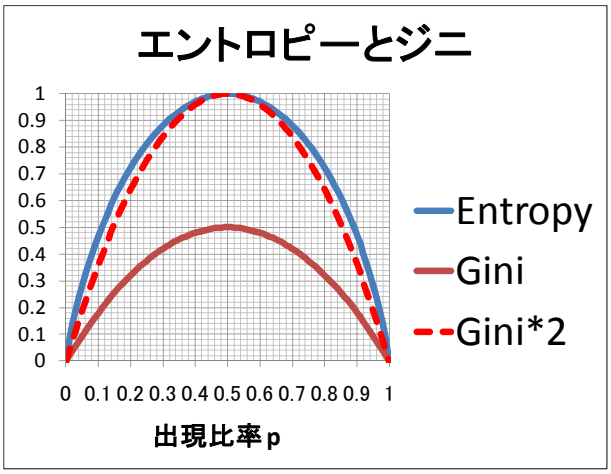
Data Mine Tech Ltd.
Data Bring New Insight to Your Business

CART
デシジョンツリーモデルの"老舗"

- **数値予測**と**クラス判別**両方に対応
- クラス判別には、**ジニ最小基準**を提案

$$E(p) = -p * \log_2(p) - (1-p) * \log_2(1-p)$$

$$Gini\ (p) = 1 - \{p^2 + (1-p)^2\}$$



- エントロピーとジニは良く似た分岐基準！

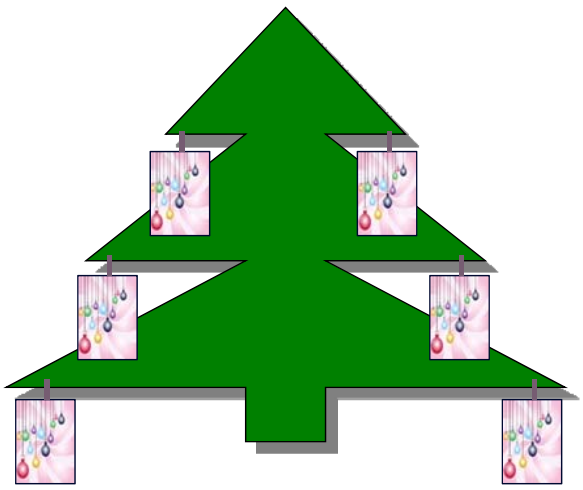
Data Mine Tech Ltd.
Data Bring New Insight to Your Business

2.2 ハイブリッド法とアンサンブル法

- 90年代半ば以降、データマイニングの手法としてAID、CHAID、ID3、C4.5、CARTなど盛んに用いられるようになった。また、様々なデシジョンツリー手法が開発され発表された。中でも、Bagging, Boostingなどのアンサンブル法が注目され、既存のデシジョンツリーモデルはこれらのベースモデルとして用いられた
  - NBTree(ナイーブベイズと決定木のハイブリッド法)(1996 シリコングラフィクス社 Kohavi)
  - CART&Logit(CARTとロジットモデルのハイブリッド法)(2002 サルフォードシステム社 SteinbergとCardell)
  - Bagging (ブートストラップデータに対する複数のモデル合成)(1996 カリフォルニア大バークレー校 Breiman)
  - Random Forests(多数のBaggingツリーの合成)(2001 カリフォルニア大バークレー校 Breiman)
  - Boosting (当てはまりの悪い事例に重みを加えたデータに対する複数の"弱い"モデルの合成)(1988 Kearns, 1996 AT&Tシャノン研究所 Freund, Schapire)
  - MART または Stochastic Gradient Boosting Tree または TreeNet(多重加算的回帰木;確率的勾配ブースティングツリー;多数の小ツリーの直列結合モデル)(1999 スタンフォード大 Friedman)

NBTree
ツリー終端ノードにナイーブベイズモデルを追加

- 目的:**クラス判別の精度アップ**
- C4.5などの標準ツリーモデルの**各終端ノードにナイーブ(単純)ベイズモデルを追加**したモデル
- 単一のツリーモデルでは予測数を増やすために分岐数を多くすると、予測値のバラツキが大きくなるというジレンマがあるが、NBTreeは**分岐数を多くせずに予測数を増やすことができる**
- アンサンブル法とは異なり、モデルが理解しやすい**ホワイトボックスモデル**という利点もある



NBTree

ツリー終端ノードにナイーブベイズモデルを追加

- ベイズの定理とナイーブ(単純)ベイズモデルの考え方

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)} \qquad P(Y|X)=\frac{P(X\text{かつ}Y)}{P(X)},\ P(X|Y)=\frac{P(Y\text{かつ}X)}{P(Y)} \text{ から導かれる}$$

① $P(\text{貸倒}|\{\text{男かつ30代}\})=\frac{P(\text{男かつ30代}|\text{貸倒})P(\text{貸倒})}{P(\text{男かつ30代})}\cong\frac{P(\text{男}|\text{貸倒})P(30\text{代}|\text{貸倒})P(\text{貸倒})}{P(\text{男かつ30代})}$

② $P(\text{非貸倒}|\{\text{男かつ30代}\})=\frac{P(\text{男かつ30代}|\text{非貸倒})P(\text{非貸倒})}{P(\text{男かつ30代})}\cong\frac{P(\text{男}|\text{非貸倒})P(30\text{代}|\text{非貸倒})P(\text{非貸倒})}{P(\text{男かつ30代})}$

{男かつ30代}のデータについて①＞②　なら貸倒、そうで無いなら非貸倒と推計する

ノードN1 嘱託

ID	性別	年齢	貸付1年後の状態 Y	貸倒事前確率 P(貸倒)	貸倒尤度 P(X|貸倒)=P(性別|貸倒)*P(年齢|貸倒)	貸倒尤度*貸倒事前確率 P(性別|貸倒)*P(年齢|貸倒)*P(貸倒)	非貸倒事前確率 P(非貸倒)	非貸倒尤度 P(X|非貸倒)=P(性別|非貸倒)*P(年齢|非貸倒)	非貸倒尤度*非貸倒事前確率 P(性別|非貸倒)*P(年齢|非貸倒)*P(非貸倒)	ナイーブベイズ予測結果	予測判定結果
0001	男性	40超	正常残なし	0.571	0.375	0.214	0.429	0.222	0.095	貸倒	×
0003	男性	40以下	貸倒	0.571	0.375	0.214	0.429	0.111	0.048	貸倒	○
0004	女性	40超	貸倒	0.571	0.125	0.071	0.429	0.444	0.190	非貸倒	×
0007	女性	40以下	正常残あり	0.571	0.125	0.071	0.429	0.222	0.095	非貸倒	○
0008	男性	40以下	貸倒	0.571	0.375	0.214	0.429	0.111	0.048	貸倒	○
0009	男性	40超	貸倒	0.571	0.375	0.214	0.429	0.222	0.095	貸倒	○
0011	女性	40超	正常残あり	0.571	0.125	0.071	0.429	0.444	0.190	非貸倒	○

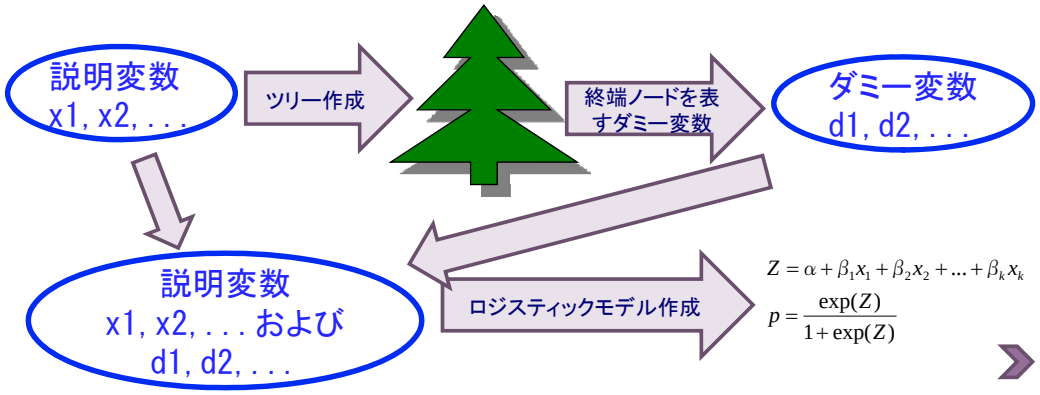
各説明変数 Xiごとの P(Xi|Y)

説明変数	値	貸倒	非貸倒
性別	男性	0.75	0.333
	女性	0.25	0.667
年齢	40以下	0.5	0.333
	40超	0.5	0.667

CART&Logit

説明変数にツリー終端ノードを表すダミー変数を追加したロジスティックモデル

- 目的:**クラス判別の精度アップ**
- **まず CARTツリーモデルを作成**し、分析データに**各終端ノードを表すダミー変数**を追加し、元の説明変数と終端ノード数分のダミー変数を説明変数とする**ロジスティックモデル**を作成する方法
- **ダミー変数部分が CARTツリーモデルと同等の説明力**を持ち、さらに元の説明変数を追加することで CARTで説明できなかったクラス判別説明力を高める仕組み
- NBTreeと同じように**終端ノードごとにロジスティックモデルを適用する方法**(初期のハイブリッド法)も提案していたが、これは**あまり効果が無かった**としている



CART&Logit

説明変数にツリー終端ノードを表すダミー変数を追加したロジスティックモデル

ID	勤務	性別	年齢	年収	貸付1年後の状態 Y	貸倒判別ツリー終端ノード	N0	N10	N11	貸倒予測出現率	予測判定結果
0001	嘱託	男性	45	580	正常残なし	N10	0	1	0	0.00033	○
0002	正社員	女性	36	390	正常残あり	N0	1	0	0	0	○
0003	嘱託	男性	38	450	貸倒	N11	0	0	1	0.99998	○
0004	嘱託	女性	41	480	貸倒	N11	0	0	1	0.99994	○
0005	正社員	女性	25	380	正常残なし	N0	1	0	0	0	○
0006	正社員	男性	35	620	正常残あり	N0	1	0	0	0	○
0007	嘱託	女性	22	300	正常残あり	N11	0	0	1	0.00372	○
0008	嘱託	男性	40	550	貸倒	N11	0	0	1	0.99995	○
0009	嘱託	男性	52	580	貸倒	N11	0	0	1	1	○
0010	正社員	女性	40	550	正常残なし	N0	1	0	0	0	○
0011	嘱託	女性	50	580	正常残あり	N10	0	1	0	0.00238	○
0012	正社員	男性	55	750	正常残あり	N0	1	0	0	0.02343	○

LVL0	LVL1	LVL2
全体(ルートノード):件数=12,貸倒率=33.33%	ノードN0:(件数=5,貸倒率=0%)分岐条件:勤務="正社員"	
	ノードN1:(件数=7件,貸倒率=57.14%)分岐条件:勤務="嘱託"	ノードN10:(件数=2,貸倒率=0%)分岐条件:年齢45以上50以下
		ノードN11:(件数=5,貸倒率=80.00%)分岐条件:年齢45未満または50超

$$Z = -41.18 + 24.08*(\text{kinmu} = \text{"嘱託"}) - 3.47*(\text{seibetu} = \text{"女性"}) + 1.09*\text{nenrei} - 0.03*\text{nenshu} - 22.57*N10$$

$$p = \frac{\exp(Z)}{1+\exp(Z)}$$

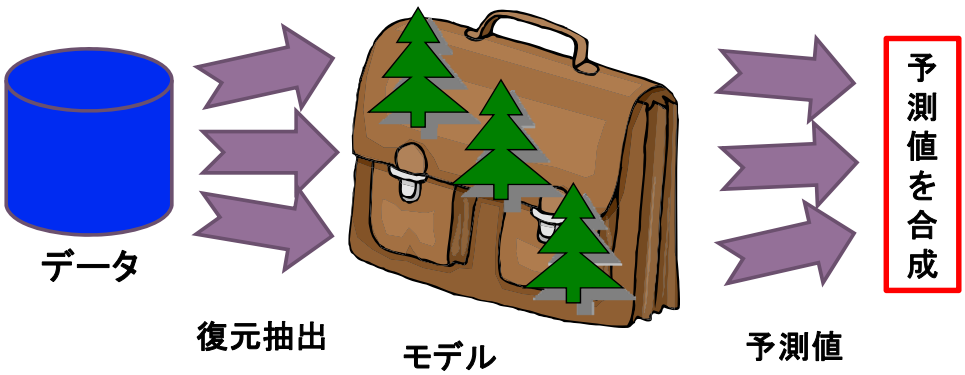
なお、N0およびN11のパラメータは0

Data Mine Tech Ltd.
Data Bring New Insight to Your Business

Bagging

ブートストラップデータに対する複数のモデル合成

- 目的:**モデル予測値のバラツキの低減**
- **ブートストラップデータ**とは元のデータから**復元抽出**法(重複を許すランダム抽出法)により同じ件数をランダム抽出したデータのこと
- ブートストラップデータは**母集団からの多数のランダム抽出データを得ることに相当**し、これを用いたモデルの合成によって精度を上げるという仕組み
- 合成する各モデルの**重みはすべて1**
- **多数(数百)のツリーモデルを合成**するため、モデルの内容はもはや理解できない**ブラックボックスモデル**



- 注:バッグに詰めた複数のツリーモデルの図は正しいBaggingの意味を表しておらず、単なる"しゃれ"です

Data Mine Tech Ltd.
Data Bring New Insight to Your Business

Bagging
ブートストラップデータに対する複数のモデル合成

元の分析データ

ID	勤務	性別	年齢	年収	貸付1年後の状態	目的変数 Y
0001	嘱託	男性	45	580	正常残なし	0
0002	正社員	女性	36	390	正常残あり	0
0003	嘱託	男性	38	450	貸倒	1
0004	嘱託	女性	41	480	貸倒	1
0005	正社員	女性	25	380	正常残なし	0
0006	正社員	男性	35	620	正常残あり	0
0007	嘱託	女性	22	300	正常残あり	0
0008	嘱託	男性	40	550	貸倒	1
0009	嘱託	男性	52	580	貸倒	1
0010	正社員	女性	40	550	正常残なし	0
0011	嘱託	女性	50	580	正常残あり	0
0012	正社員	男性	55	750	正常残あり	0

1回目のブートストラップデータ

ID	勤務	性別	年齢	年収	貸付1年後の状態	目的変数 Y
0003	嘱託	男性	38	450	貸倒	1
0012	正社員	男性	55	750	正常残あり	0
0005	正社員	女性	25	380	正常残なし	0
0004	嘱託	女性	41	480	貸倒	1
0012	正社員	男性	55	750	正常残あり	0
0012	正社員	男性	55	750	正常残あり	0
0007	嘱託	女性	22	300		
0007	嘱託	女性	22	300		
0001	嘱託	男性	45	580		
0001	嘱託	男性	45	580		
0010	正社員	女性	40	550		
0007	嘱託	女性	22	300		

2回目のブートストラップデータ

ID	勤務	性別	年齢	年収	貸付1年後の状態	目的変数 Y
0007	嘱託	女性	22	300	正常残あり	0
0011	嘱託	女性	50	580	正常残あり	0
0003	嘱託	男性	38	450	貸倒	1
0010	正社員	女性	40	550	正常残なし	0
0010	正社員	女性	40	550	正常残なし	0
0011	嘱託	女性	50	580	正常残あり	0
0008	嘱託	男性	40	550	貸倒	1
0008	嘱託	男性	40	550	貸倒	1
0002	正社員	女性	36	390	正常残あり	0
0003	嘱託	男性	38	450	貸倒	1
0006	正社員	男性	35	620	正常残あり	0
0007	嘱託	女性	22	300	正常残あり	0

ブートストラップデータ 1 からモデル作成

LVL0	LVL1
全体:16.67%(2/12)	N0:66.67%(2/3)年収=380~480
	N1:0.00%(0/9)年収=LOW~<380,480<~HIGH

ブートストラップデータ 2 からモデル作成

LVL0	LVL1	LVL2
全体:33.33%(4/12)	N0:0.00%(0/7)性別=″女性″	
	N1:80.00%(4/5)性別=″男性″	N10:66.67%(2/3)年齢=LOW~38
		N11:100.00%(2/2)年齢=38<~HIGH

Bagging
ブートストラップデータに対する複数のモデル合成

- ブートストラップモデルを増やしていくと、バギングモデルは精度が良くなっていく

ブートストラップデータ 1 から作成されたモデル

LVL0	LVL1
全体:16.67%(2/12)	N0:66.67%(2/3)年収=380~480
	N1:0.00%(0/9)年収=LOW~<380,480<~HIGH

ブートストラップデータ 2 から作成されたモデル

LVL0	LVL1	LVL2
全体:33.33%(4/12)	N0:0.00%(0/7)性別=″女性″	
	N1:80.00%(4/5)性別=″男性″	N10:66.67%(2/3)年齢=LOW~38
		N11:100.00%(2/2)年齢=38<~HIGH

バギングモデル予測結果

ID	勤務	性別	年齢	年収	貸付1年後の状態	目的変数 Y	モデル予測結果: ブートストラップデータ1のモデル1	ブートストラップデータ2のモデル	2個のバギングモデル	10個のバギングモデル	20個のバギングモデル
0001	嘱託	男性	45	580	正常残なし	0	0	1	0.5	0.5	0.431
0002	正社員	女性	36	390	正常残あり	0	0.667	0	0.334	0.067	0.040
0003	嘱託	男性	38	450	貸倒	1	0	0.667	0.334	0.908	0.710
0004	嘱託	女性	41	480	貸倒	1	0.667	0	0.334	0.617	0.506
0005	正社員	女性	25	380	正常残なし	0	0.667	0	0.334	0.067	0.040
0006	正社員	男性	35	620	正常残あり	0	0	0.667	0.334	0.142	0.127
0007	嘱託	女性	22	300	正常残あり	0	0	0	0	0.350	0.265
0008	嘱託	男性	40	550	貸倒	1	0	1	0.5	0.775	0.771
0009	嘱託	男性	52	580	貸倒	1	0	1	0.5	0.675	0.594
0010	正社員	女性	40	550	正常残なし	0	0	0	0	0	0.100
0011	嘱託	女性	50	580	正常残あり	0	0	0	0	0.250	0.240
0012	正社員	男性	55	750	正常残あり	0	0	1	0.5	0.175	0.094
					正答数		7	8	8	11	12

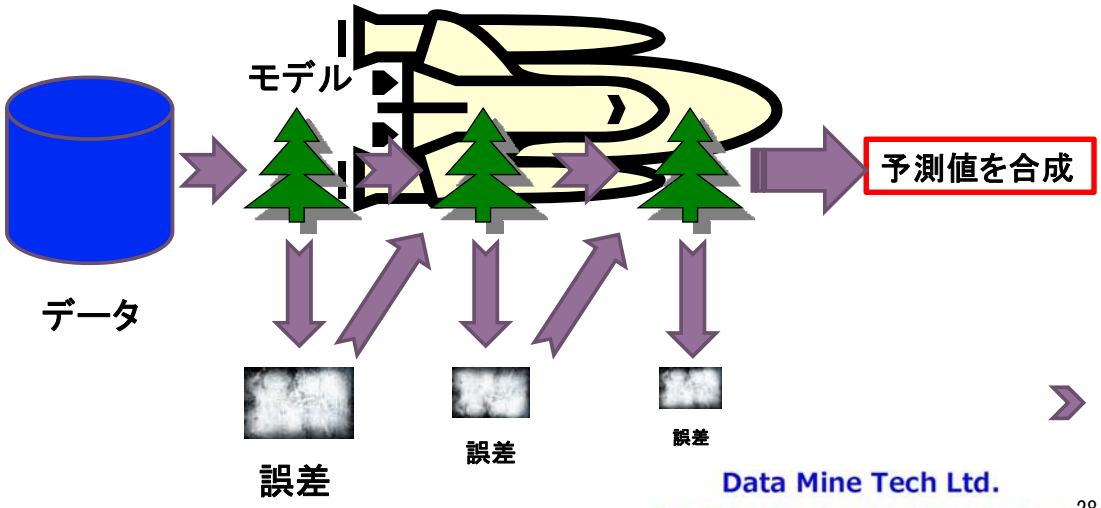
- ブートストラップデータごとに作成された全モデルの集合が1セットの Baggingモデルとなる
- 数百程度のブートストラップデータを用いると十分と言われている

Random Forests
多数のツリーモデルを作成し予測値を合成

- 目的変数は**数値変数でもクラス変数でも可**
- 分析データから**ブートストラップデータ**を抽出し、同時に**説明変数からランダム選択**（通常、候補説明変数の数を k個とすると **√k 個**の数の説明変数を選択）
- Random Forestsでは**大きなツリーモデルを並行的に多数作成し等しい重みで合成**する（バギングの一種）

Boosting
複数の"弱い"学習器による強化学習

- 目的：モデルの**訓練データに対する誤差の低減**
- **"弱い"学習器**とはランダムモデルよりは精度が高いが、それほど精度が高く無いモデル
- デシジョンツリーでは**ランダムに説明変数セットを選択**したり、**最良の説明変数を故意に除外**して2番目に良い説明変数を選択したり、**分岐数を絞る**などの方法で作る
- **最初のモデルはデータの重みはすべて同じ** 1として訓練データそのものから作成する
- **2番目以降のモデル**は、直前に作成したモデルを元のデータに適用した場合の**事例ごとに誤差の大きさと関連した重みによるリサンプル**データから作成し、最後に合成する



Boosting
複数の"弱い"学習器による強化学習

- 最初のモデル

ID	勤務	性別	年齢	年収	貸付1年後の状態	目的変数 Y	重み $w_1 = \frac{1}{N}$	予測値	予測クラス $\hat{Y}$	重み付き絶対誤差 $w_1 \lvert Y - \hat{Y} \rvert$	次回の重み $w_2 = \frac{w_1 \beta_1^{1-\lvert Y-\hat{Y} \rvert}}{\sum w_1 \beta_1^{1-\lvert Y-\hat{Y} \rvert}}$
0001	嘱託	男性	45	580	正常残なし	0	0.0833	0.5714	1	0.0833	0.1667
0002	正社員	女性	36	390	正常残あり	0	0.0833	0.0000	0	0	0.0556
0003	嘱託	男性	38	450	貸倒	1	0.0833	0.5714	1	0	0.0556
0004	嘱託	女性	41	480	貸倒	1	0.0833	0.5714	1	0	0.0556
0005	正社員	女性	25	380	正常残なし	0	0.0833	0.0000	0	0	0.0556
0006	正社員	男性	35	620	正常残あり	0	0.0833	0.0000	0	0	0.0556
0007	嘱託	女性	22	300	正常残あり	0	0.0833	0.0000	0	0	0.0556
0008	嘱託	男性	40	550	貸倒	1	0.0833	0.5714	1	0	0.0556
0009	嘱託	男性	52	580	貸倒	1	0.0833	0.5714	1	0	0.0556
0010	正社員	女性	40	550	正常残なし	0	0.0833	0.5714	1	0.0833	0.1667
0011	嘱託	女性	50	580	正常残あり	0	0.0833	0.5714	1	0.0833	0.1667
0012	正社員	男性	55	750	正常残あり	0	0.0833	0.0000	0	0	0.0556
							重み合計			誤差合計 ε_1	重み合計
							1			0.2500	1
										β パラメータ $\beta_1 = \frac{\varepsilon_1}{1-\varepsilon_1}$	
										0.3333	

最初のモデル作成（2番目に効く説明変数を選択）

LVL0	LVL1
全体:33.33%(4/12)	N0:57.14%(4/7)年収=450以上580以下
	N1:0.00%(0/5)年収=LOW~<450,580<~HIGH

- モデルを分析データにあてはめた場合の誤差の大きさに応じた重みを付与する

Boosting
複数の"弱い"学習器による強化学習

- 2番目のモデル
- **誤差の大きさに応じた重みを選択確率**として同じ件数を得るまでランダム抽出したデータから次のモデルを作成する
- 作成されたモデルを元データに適用し**誤差を再計算し、次回の重みを更新**する。何度も間違った事例にはより大きな重み（**累積的重み**）が付与されていく

ID	勤務	性別	年齢	年収	貸付1年後の状態	目的変数 Y	重み w_2	予測値	予測クラス $\hat{Y}$	重み付き絶対誤差 $w_2 \lvert Y - \hat{Y} \rvert$	次回の重み $w_3 = \frac{w_2 \beta_2^{1-\lvert Y-\hat{Y} \rvert}}{\sum w_2 \beta_2^{1-\lvert Y-\hat{Y} \rvert}}$
0001	嘱託	男性	45	580	正常残なし	0	0.1667	0.1250	0	0	0.1071
0002	正社員	女性	36	390	正常残あり	0	0.0556	0.1250	0	0	0.0357
0003	嘱託	男性	38	450	貸倒	1	0.0556	0.5000	1	0	0.0357
0004	嘱託	女性	41	480	貸倒	1	0.0556	0.5000	1	0	0.0357
0005	正社員	女性	25	380	正常残なし	0	0.0556	0.1250	0	0	0.0357
0006	正社員	男性	35	620	正常残あり	0	0.0556	0.1250	0	0	0.0357
0007	嘱託	女性	22	300	正常残あり	0	0.0556	0.1250	0	0	0.0357
0008	嘱託	男性	40	550	貸倒	1	0.0556	0.5000	1	0	0.0357
0009	嘱託	男性	52	580	貸倒	1	0.0556	0.1250	0	0.0556	0.1250
0010	正社員	女性	40	550	正常残なし	0	0.1667	0.5000	1	0.1667	0.3750
0011	嘱託	女性	50	580	正常残あり	0	0.1667	0.1250	0	0	0.1071
0012	正社員	男性	55	750	正常残あり	0	0.0556	0.1250	0	0	0.0357
							重み合計			誤差合計 ε_2	重み合計
							1			0.2222	1
										β パラメータ $\beta_2 = \frac{\varepsilon_2}{1-\varepsilon_2}$	
										0.2857	

2番目のモデル（2番目に効く説明変数を選択）

LVL0	LVL1
全体:25%(3/12)	N0:50%(2/4)年齢=38以上41以下
	N1:12.5%(1/8)年齢=38未満または41超

Boosting
複数の"弱い"学習器による強化学習

- 最終的には**各モデルの誤差率を考慮した予測値の平均**をブースティング予測値とする
- ブースティングモデルは**モデル作成データに過剰適合**し、**汎化誤差が大きくなると予想**されるが、実際には汎化誤差も小さくなったという報告もある

1番目のモデル

LVL0	LVL1
全体:33.33%(4/12)	N0:57.14%(4/7)年収=450以上580以下
	N1:0.00%(0/5)年収=LOW~<450,580<~HIGH

2番目のモデル

LVL0	LVL1
全体:25%(3/12)	N0:50%(2/4)年齢=38以上41以下
	N1:12.5%(1/8)年齢=38未満または41超

ブースティングモデル予測結果

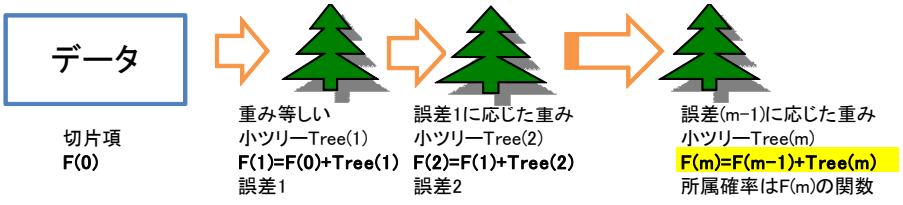
ID	勤務	性別	年齢	年収	貸付1年後の状態	目的変数 Y	1回目パラメータ β_1	1回目の予測クラス $\hat{Y}_1$	2回目パラメータ β_2	2回目の予測クラス $\hat{Y}_2$	2つの予測クラスのBoosting合成値 $Boost\_\hat{Y}$	Boostiong予測クラス
0001	嘱託	男性	45	580	正常残なし	0	0.3333	1	0.2857	0	0.4615	0
0002	正社員	女性	36	390	正常残あり	0		0		0	0.0870	0
0003	嘱託	男性	38	450	**貸倒**	**1**		**1**		**1**	**0.9130**	**1**
0004	嘱託	女性	41	480	**貸倒**	**1**		**1**		**1**	**0.9130**	**1**
0005	正社員	女性	25	380	正常残なし	0		0		0	0.0870	0
0006	正社員	男性	35	620	正常残あり	0		0		0	0.0870	0
0007	嘱託	女性	22	300	正常残あり	0		0		0	0.0870	0
0008	嘱託	男性	40	550	**貸倒**	**1**		**1**		**1**	**0.9130**	**1**
0009	嘱託	男性	52	580	**貸倒**	**1**		**1**		0	0.4615	0
0010	正社員	女性	40	550	正常残なし	0		1		**1**	**0.9130**	**1**
0011	嘱託	女性	50	580	正常残あり	0		1		0	0.4615	0
0012	正社員	男性	55	750	正常残あり	0		0		0	0.0870	0
							正答率	0.75	正答率	0.83	正答率	**0.83**

- この例示ではブースティングモデルのモデル作成データに対する正答率はサブモデルの1つと同じとなっているが、一般に精度は高くなることが証明されている

$$Boost\_\hat{Y} = \frac{1}{1+\beta_1^{2r-1}+\beta_2^{2r-1}} \quad , r = \frac{\hat{Y}_1 \log(\frac{1}{\beta_1}) + \hat{Y}_2 \log(\frac{1}{\beta_2})}{\log(\frac{1}{\beta_1}) + \log(\frac{1}{\beta_2})}$$

MART(TreeNet)
多数の小ツリーの直列結合モデル

- 目的変数は**数値変数でもクラス変数でも可**
- **終端ノード数が数個の小さなツリー**(弱い学習器)を**数百個程度直列に連結**したモデル
- 説明変数間の**交互作用を小ツリーモデルにより表現**し、小ツリーモデルを説明アイテムとして多数用いた**線形重回帰モデル**または**線形ロジスティックモデル**とみなすこともできる
- **多数の数値説明変数**を含むモデル構築に特に威力を発揮

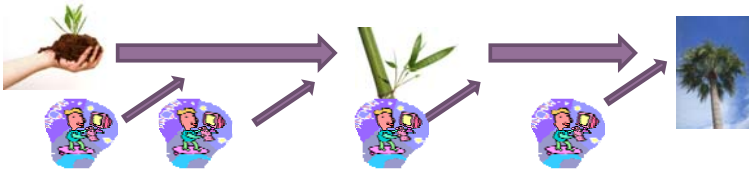


$$\text{予測スコア} = F(0) + \text{Tree}(1) + \text{Tree}(2) + \ldots + \text{Tree}(m)$$

2.3 その他のデシジョンツリーモデル

- ネットワークが普及し、大量の分析データが時系列でリアルタイムに入手できる状況の中で、ストリームデータのデシジョンツリー分析手法が提案された。また時系列データや生存時間データに対して適用する既存分析手法のパラメータが異なるグループを自動発見し、データ分割を行うツリーモデルなども開発された。
  - VFDT または Hoeffding Tree(データストリームからの高速決定木生成法;ホフディングツリー)(2000 ワシントン大 DomingosとHulten)
  - Survival Tree(生存ツリー)(1985 GordonとOlshenなど)
  - Autoregressive Tree(自己回帰ツリー)(2002 マイクロソフト社 MeekとChickeringとHeckerman)

Hoeffding Tree(VFDT)
ストリームデータから要因変化を自動検出



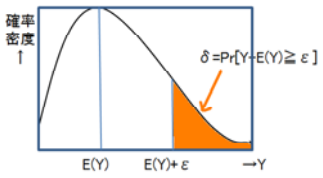
- 最初は**ルートノード**のみからスタート
- **ストリームデータ**を読んで、説明変数と目的変数の**クロス集計値を更新**していく(ストリームデータそのものは蓄積しない)**大規模なストリームデータ**を想定している
- 最適な説明変数による分岐と次善の分岐の評価値の差が**ホフディング範囲以上**になると最適な説明変数による**分岐**が発生する
- **新たな分岐**や、**分岐のやり直し**を実行しながら木が成長していく

原理(ホフディングの不等式)

- **nをノード内のデータ件数**、Yを最善の分岐説明変数の評価値Gaと次善の分岐説明変数の評価値Gbの差Ga-Gbの観測値(Ga>Gbとする)、E(Y)を真の Y(母集団期待値)とすると、Y>=E(Y)+**ε**(**εはホフディング範囲(正の値)**)となる確率δは以下の式で表現できる
- これから、Y>εのとき、1-δの確率で E(Y)>=Y-ε>0となり、これは母集団において真にGa>Gbであることを意味する

$$\delta = \mathrm{Prob}[Y - E(Y) \geq \varepsilon] \leq \exp\left(-2n\varepsilon^2\right)$$

$$\varepsilon \geq \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}$$



Hoeffding Tree(VFDT)
ストリームデータから要因変化を自動検出

分析データ（1件ずつ出現）

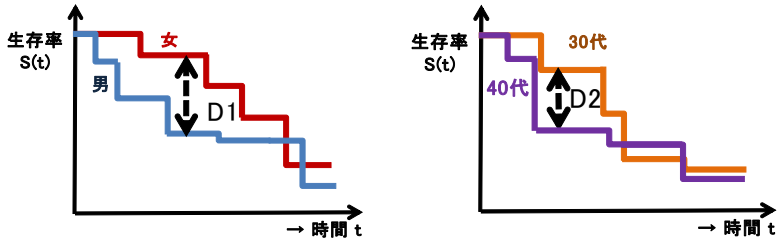
ID	勤務	性別	貸付1年後の状態 Y
0001	嘱託	男性	正常残なし
0002	正社員	女性	正常残あり
0003	嘱託	男性	貸倒
0004	嘱託	女性	貸倒
0005	正社員	女性	正常残なし
0006	正社員	男性	正常残あり
0007	嘱託	女性	正常残あり

ストリームNO	勤務						性別						分割後のエントロピー差(評価値の差)	ε（δ=0.05）
	カテゴリ	貸倒	非貸倒	貸倒率	エントロピー	分割後のエントロピー	カテゴリ	貸倒	非貸倒	貸倒率	エントロピー	分割後のエントロピー		
1	嘱託	0	1	0	0	0	男性	0	1	0	0	0	0	1.224
	正社員	0	0	0	0		女性	0	0	0	0			
2	嘱託	0	1	0	0	0	男性	0	1	0	0	0	0	0.865
	正社員	0	1	0	0		女性	0	1	0	0			
3	嘱託	1	1	0.5	1	0.6667	男性	1	1	0.5	1	0.6667	0	0.707
	正社員	0	1	0	0		女性	0	1	0	0			
4	嘱託	2	1	0.667	0.9183	0.6887	男性	1	1	0.5	1	0.75	0.061278	0.612
	正社員	0	1	0	0		女性	1	1	0.5	1			
5	嘱託	2	1	0.667	0.9183	0.551	男性	1	1	0.5	1	0.6	0.049022	0.547
	正社員	0	2	0	0		女性	1	2	0.333	0.9183			
6	嘱託	2	1	0.667	0.9183	0.4591	男性	1	2	0.333	0.9183	0.6258	0.166667	0.5
	正社員	0	3	0	0		女性	1	2	0.333	0.9183			
7	嘱託	2	2	0.5	1	0.5714	男性	1	2	0.333	0.9183	0.5364	0.035016	0.463
	正社員	0	3	0	0		女性	1	3	0.25	0.8113			

- ストリームデータが到着するたびに、ノードごと（最初は全体ノードのみ）に目的変数と各説明変数の度数集計表を更新していく
- ストリームデータ数（n）が増えると ホフディング範囲（ε）は小さくなっていく。
- 勤務で分けた場合の分割後のエントロピーと、住所で分けた場合の分割後のエントロピーの差を評価値の差とする
- この評価値の差は変動しているが、ストリームデータ件数が増えていくと、ホフディング範囲（ε）超える効果の高い説明変数が出てくると考えられる

Survival Tree
生存曲線の違いを表すツリーモデル

- 目的：一定期間内の**イベント出現率の時間推移**の違いにより対象を分類
- 目的変数は**生存率の違い**（差の最大値や面積の差）、また特定のイベントの**ハザード**や**累積出現率**とすることもできる
  ハザード＝単位時間内出現数／時間区間開始時点の残存件数
  累積出現率＝その時点までの累積出現数／時点0における残存件数（分析対象件数）
- **層別生存曲線やイベント累積出現率の差**が最大になる説明変数で分岐を行う



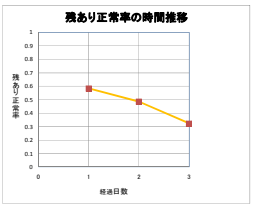
- 上の例では、性別で分ける方が年齢で分けるより**グループ間の生存率の差**が大きい（D1>D2）ので、性別で分ける方が良いという判断になる
- また、与信分野やマーケティング分野での利用においては、生存曲線の下側面積はローンの残あり日数や顧客として残る日数を表すので、**面積の差**を分岐変数選択基準とすることが考えられる

Survival Tree
生存曲線の違いを表すツリーモデル

分析データ

ID	勤務	性別	状態	貸付日	貸倒日	完済日	現在日	経過日数
0001	嘱託	男性	正常残なし	1		2	4	3
0003	嘱託	男性	貸倒	1	3		4	3
0002	正社員	女性	正常残あり	1			4	3
0005	正社員	女性	正常残なし	1		4	4	3
0004	嘱託	女性	貸倒	2	3		4	2
0006	正社員	男性	正常残あり	2			4	2
0007	嘱託	女性	正常残あり	2			4	2
0008	嘱託	男性	貸倒	3	4		4	1
0010	正社員	女性	正常残なし	3		4	4	1
0011	嘱託	女性	正常残あり	3			4	1
0012	正社員	男性	正常残あり	3			4	1
0009	嘱託	男性	貸倒	3	4		4	1

生存分析累積出現率				
	貸倒率	0.25	0.35	0.35
	完済率	0.1667	0.1667	0.3519
	正常残あり率	0.5833	0.4861	0.3241
	計	1.00	1.00	1.03



分析データ

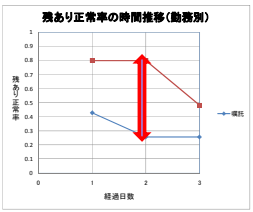
ID	勤務	性別	状態	貸付日	貸倒日	完済日	現在日	経過日数
0001	嘱託	男性	正常残なし	1		2	4	3
0003	嘱託	男性	貸倒	1	3		4	3
0004	嘱託	女性	貸倒	2	3		4	2
0007	嘱託	女性	正常残あり	2			4	2
0008	嘱託	男性	貸倒	3	4		4	1
0011	嘱託	女性	正常残あり	3			4	1
0009	嘱託	男性	貸倒	3	4		4	1
0002	正社員	女性	正常残あり	1			4	3
0005	正社員	女性	正常残なし	1		4	4	3
0006	正社員	男性	正常残あり	2			4	2
0010	正社員	女性	正常残なし	3		4	4	1
0012	正社員	男性	正常残あり	3			4	1

嘱託

生存分析累積出現率				
	貸倒率	0.4286	0.5918	0.5918
	完済率	0.1429	0.1429	0.1429
	正常残あり率	0.4286	0.2571	0.2571
	計	1.00	0.99	0.99

正社員

生存分析累積出現率				
	貸倒率	0	0	0
	完済率	0.2	0.2	0.52
	正常残あり率	0.8	0.8	0.48
	計	1.00	1.00	1.00



分析データ

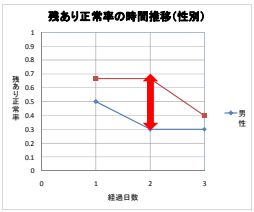
ID	勤務	性別	状態	貸付日	貸倒日	完済日	現在日	経過日数
0001	嘱託	男性	正常残なし	1		2	4	3
0003	嘱託	男性	貸倒	1	3		4	3
0006	正社員	男性	正常残あり	2			4	2
0008	嘱託	男性	貸倒	3	4		4	1
0012	正社員	男性	正常残あり	3			4	1
0009	嘱託	男性	貸倒	3	4		4	1
0002	正社員	女性	正常残あり	1			4	3
0005	正社員	女性	正常残なし	1		4	4	3
0004	嘱託	女性	貸倒	2	3		4	2
0007	嘱託	女性	正常残あり	2			4	2
0010	正社員	女性	正常残なし	3		4	4	1
0011	嘱託	女性	正常残あり	3			4	1

男性

累積出現率				
	貸倒率	0.3333	0.5238	0.5238
	完済率	0.1667	0.1667	0.1667
	正常残あり率	0.5	0.3	0.3
	計	1.00	0.99	0.99

女性

生存分析累積出現率				
	貸倒率	0.1667	0.1667	0.1667
	完済率	0.1667	0.1667	0.4444
	正常残あり率	0.6667	0.6667	0.4
	計	1.00	1.00	1.01



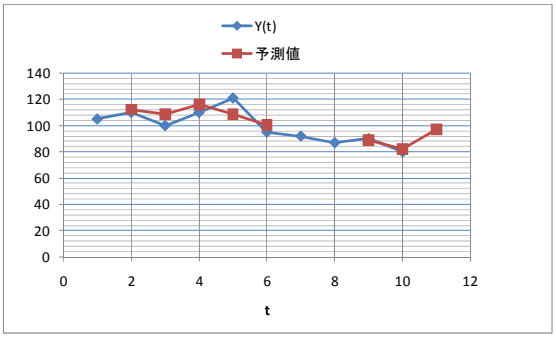
- この例では性別で分けるより勤務で分ける方が正常残あり率の差が大きくなっている

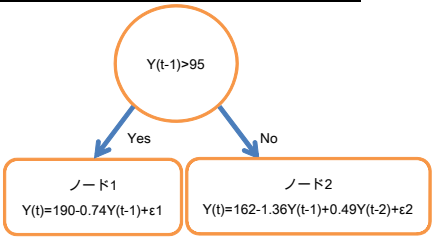
Data Mine Tech Ltd.
Data Bring New Insight to Your Business

自己回帰ツリー
時系列モデルの時間範囲を自動的に区切る

- 目的：時系列モデルの**パラメータが相互に異なる時間範囲**を区切り、別々のパラメータを推計することにより、時系列モデルの予測性能の向上を図る
- 目的変数は単一時系列
- 説明変数は目的時系列変数の**ラグ**をとった変数、および**他の時系列変数**

t	Y(t)	Y(t-1)	Y(t-2)	Y(t-3)	ノード	予測値
1	105					
2	110	105			1	112.3
3	100	110	105		1	108.6
4	110	100	110	105	1	116.0
5	121	110	100	110	1	108.6
6	95	121	110	100	1	100.5
7	92				2	
8	87	92			2	
9	90	87	92		2	88.8
10	80	90	87	92	2	82.2
11		80	90	87	2	97.3
12			80	90	2	
13				80	2	





- 全体のデータで1つの自己回帰モデルを作るより当てはまりが良くなる

Data Mine Tech Ltd.
Data Bring New Insight to Your Business

3. まとめ（ホワイトボックスモデルとブラックボックスモデル）

- ホワイトボックスモデル
  - モデルによる判別結果や予測結果の理由づけが明確に行えることは、データに基づく人間の判断を"モデルが代理で行う"という観点から非常に重要である。このような応用分野の代表として**与信（審査）モデル**がある
  - また、ホワイトボックスモデルは、モデルの適合性が悪くなった場合に、その原因を突き止めたり、部分調整を行うなどの対応を行うことが比較的容易という利点もある。ブラックボックスモデルの場合は、データを入れ替えたりして全体を作り替えるといった処置が必要になる
- ブラックボックスモデル
  - モデルの中身は理解しづらく、モデル予測結果に対する疑問を突き止めたり、調整したりすることが難しいという欠点がある。しかし、ブラックボックスモデルは予測精度を上げることを最大の目標としており、精度が優先的な応用分野に向いている。このような応用分野の代表として**不正検知モデル**や**株価予測モデル**などがある
- 一概にブラックボックスモデルが悪いというわけではなく、用途に応じて選択することが重要

参考文献

- Ritschard(2010) CHAID and Earlier Supervised Tree
- Perreault,Hiram,Barksdale(1980) A Model-Free Approach for Analysis of Complex Contingency Data in Survey Research
- Quinlan(1993) C4.5: Programs for Machine Learning
（邦訳:古川(1995)AIによるデータ解析）
- Breiman(1996) Bagging Predictors
- Freund,Schapire(1999) A Short Introdction to Boosting
- Kohavi(1996) Scaling Up the Accuracy of Naive-Bayes Classifiers:a Decision-Tree Hybrid
- Friedman(1999) Greedy Function Approximation: A Gradient Boosting Machine
- Domingos,Hulten(2000) Mining High-Speed Data Streams
- Meek,Chickering,Heckerman(2002) Forecasting Using Autoregressive Tree Models
- Berry,Linoff(2004) Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd Edition
（邦訳:江原他(2005) データマイニング手法　営業、マーケティング、CRMのための顧客分析）