データサイエンスの諸相（決定木）

前田利之（まえだ　としゆき）
maechan@hannan-u.ac.jp

阪南大学　経営情報学部

(2024.11.27)

apriori【復習】

☆R を立ち上げ、必要に応じて arules パッケージをインストールした上で、以下を実行する

```
library(arules)
data(Groceries)
g0 <- Groceries
grule5 <- apriori(g0,
 parameter=list(support=0.005,confidence=0.7),
 appearance=list(rhs="whole milk",default="lhs"))
# ↑は本当は１行
inspect(grule5)
```

決定木とは

- 「もし…あるいは…」＝ [ ] の結合による [ ]
- 分類に利用することも、予測に活用することも、可能
- クラスター分析も [ ] を作るが、分類の分岐条件を提示しないため、人間がそれを解釈しなければならない。しかし決定木はこのような [ ] によって分類の分岐条件を説明してくれる

決定木とは

- 「もし…あるいは…」＝ IF-THEN(-ELSE) ルール の結合による
- 分類に利用することも、予測に活用することも、可能
- クラスター分析も を作るが、分類の分岐条件を提示しないため、人間がそれを解釈しなければならない。しかし決定木はこのような によって分類の分岐条件を説明してくれる

決定木とは

- 「もし…あるいは…」＝
  IF-THEN(-ELSE) ルール の結合による
  樹形図
- 分類に利用することも、予測に活用することも、可能
- クラスター分析も [ ] を作るが、分類の分岐条件を提示しないため、人間がそれを解釈しなければならない。しかし決定木はこのような [ ] によって分類の分岐条件を説明してくれる

決定木とは

- 「もし…あるいは…」＝ IF-THEN(-ELSE) ルール の結合による 樹形図
- 分類に利用することも、予測に活用することも、可能
- クラスター分析も 樹形図 を作るが、分類の分岐条件を提示しないため、人間がそれを解釈しなければならない。しかし決定木はこのような　　によって分類の分岐条件を説明してくれる

決定木とは

- 「もし…あるいは…」＝
  IF-THEN(-ELSE) ルール の結合による
  樹形図
- 分類に利用することも、予測に活用することも、可能
- クラスター分析も 樹形図 を作るが、分類の分岐条件を提示しないため、人間がそれを解釈しなければならない。しかし決定木はこのような IF-THEN ルール によって分類の分岐条件を説明してくれる

決定木とは（続き）

- 例：０から７までの数字をあてるゲームで、質問は「その数は△より小さいですか？」に限るものを考える（情報理論の例題とほぼ同じ）
- この条件で、どのように質問すれば［　　］かを考える
- 例えば「１より小さいですか？」と尋ねていって数字を大きくしていく作戦だと：
  - 「はい」なら０で確定する、が…
  - 「いいえ」なら１から７のどれかであるので、最悪［　　］質問が必要

決定木とは（続き）

- 例：０から７までの数字をあてるゲームで、質問は「その数は△より小さいですか？」に限るものを考える（情報理論の例題とほぼ同じ）
- この条件で、どのように質問すれば **効率的** かを考える
- 例えば「１より小さいですか？」と尋ねていって数字を大きくしていく作戦だと：
  - 「はい」なら０で確定する、が…
  - 「いいえ」なら１から７のどれかであるので、最悪 [　　] 質問が必要

決定木とは（続き）

- 例：０から７までの数字をあてるゲームで、質問は「その数は△より小さいですか？」に限るものを考える（情報理論の例題とほぼ同じ）
- この条件で、どのように質問すれば **効率的** かを考える
- 例えば「１より小さいですか？」と尋ねていって数字を大きくしていく作戦だと：
  - 「はい」なら０で確定する、が…
  - 「いいえ」なら１から７のどれかであるので、最悪 **７回** 質問が必要

決定木とは（続き）

- 一番効率的なのは：まず中央の「　　より小さいですか？」と質問する
- すると、「はい」なら　　、「いいえ」なら　　のどれかになる
- つまり、８つから４つに候補を　　ことになる
- 以下同様に候補を半分に絞っていくことで、　　の質問で確実に数を当てることが出来る　　

決定木とは（続き）

- 一番効率的なのは：まず中央の「 4 より小さいですか？」と質問する
- すると、「はい」なら [　　]、「いいえ」なら [　　] のどれかになる
- つまり、８つから４つに候補を [　　] ことになる
- 以下同様に候補を半分に絞っていくことで、[　　] の質問で確実に数を当てることが出来る [　　]

決定木とは（続き）

- 一番効率的なのは：まず中央の「 4 より小さいですか？」と質問する
- すると、「はい」なら 0から3 、「いいえ」なら [　　] のどれかになる
- つまり、８つから４つに候補を [　　] ことになる
- 以下同様に候補を半分に絞っていくことで、[　　] の質問で確実に数を当てることが出来る [　　]

決定木とは（続き）

- 一番効率的なのは：まず中央の「 4 より小さいですか？」と質問する
- すると、「はい」なら 0から3 、「いいえ」なら 4から7 のどれかになる
- つまり、８つから４つに候補を
  ことになる
- 以下同様に候補を半分に絞っていくことで、
  の質問で確実に数を当てることが出来る

決定木とは（続き）

- 一番効率的なのは：まず中央の「 4 より小さいですか？」と質問する
- すると、「はい」なら 0から3 、「いいえ」なら 4から7 のどれかになる
- つまり、８つから４つに候補を 半分に絞れた ことになる
- 以下同様に候補を半分に絞っていくことで、 の質問で確実に数を当てることが出来る

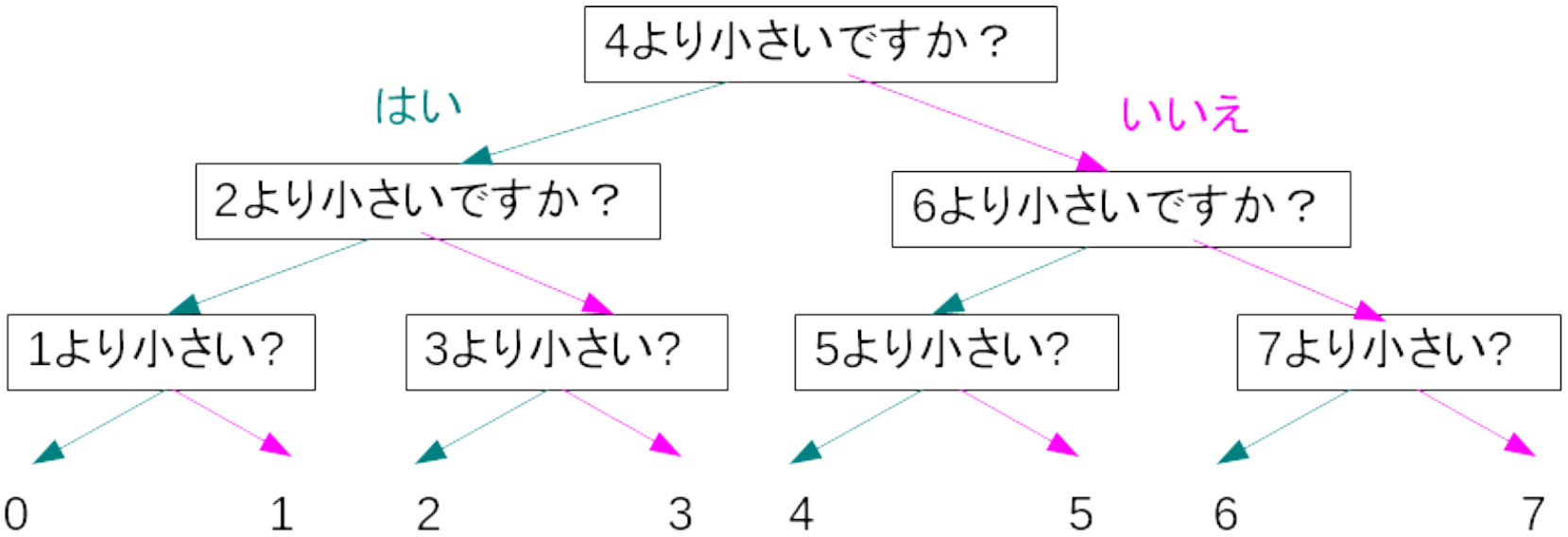
決定木とは（続き）

- 一番効率的なのは：まず中央の「 4 より小さいですか？」と質問する
- すると、「はい」なら 0から3 、「いいえ」なら 4から7 のどれかになる
- つまり、８つから４つに候補を 半分に絞れた ことになる
- 以下同様に候補を半分に絞っていくことで、 3回 の質問で確実に数を当てることが出来る

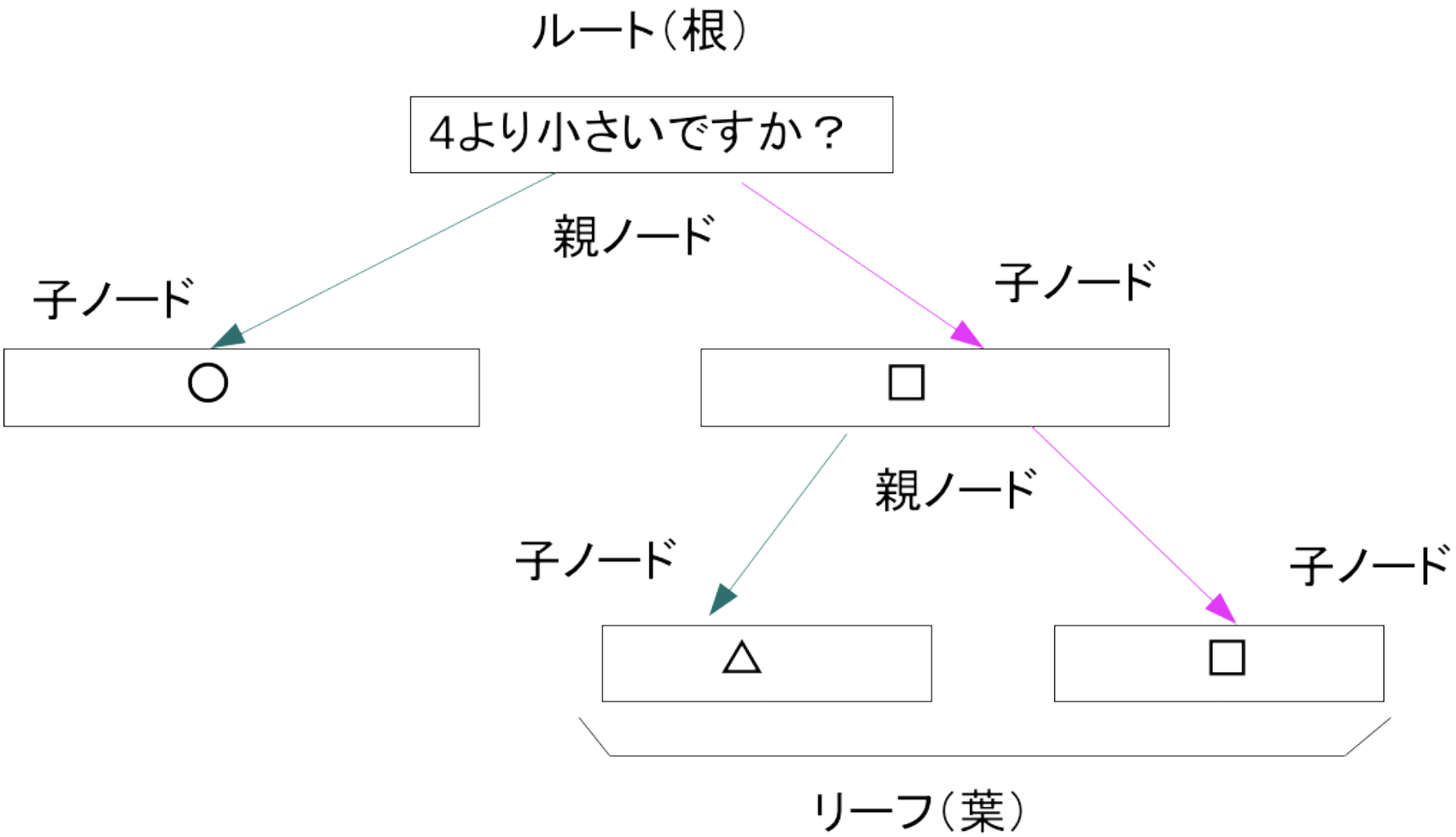
決定木とは（続き）

- 一番効率的なのは：まず中央の「 4 より小さいですか？」と質問する
- すると、「はい」なら 0から3 、「いいえ」なら 4から7 のどれかになる
- つまり、８つから４つに候補を 半分に絞れた ことになる
- 以下同様に候補を半分に絞っていくことで、 3回 の質問で確実に数を当てることが出来る （二分探索）

決定木とは（続き）


4より小さいですか？
はい
いいえ
2より小さいですか？
6より小さいですか？
1より小さい?
3より小さい?
5より小さい?
7より小さい?
0
1
2
3
4
5
6
7

決定木とは（要素名）



決定木の分析

- 分析：データを元に、このような　　こと
- 様々な特性のデータの集まりがあったとき、その　　して分割する
  - 例えば「丸かどうか」といった条件を考えることができる
- これを機械的に作成する→各成分の値に従って　　について考える

決定木の分析

- 分析：データを元に、このような木をつくること
- 様々な特性のデータの集まりがあったとき、その　　して分割する
  - 例えば「丸かどうか」といった条件を考えることができる
- これを機械的に作成する→各成分の値に従って　　について考える

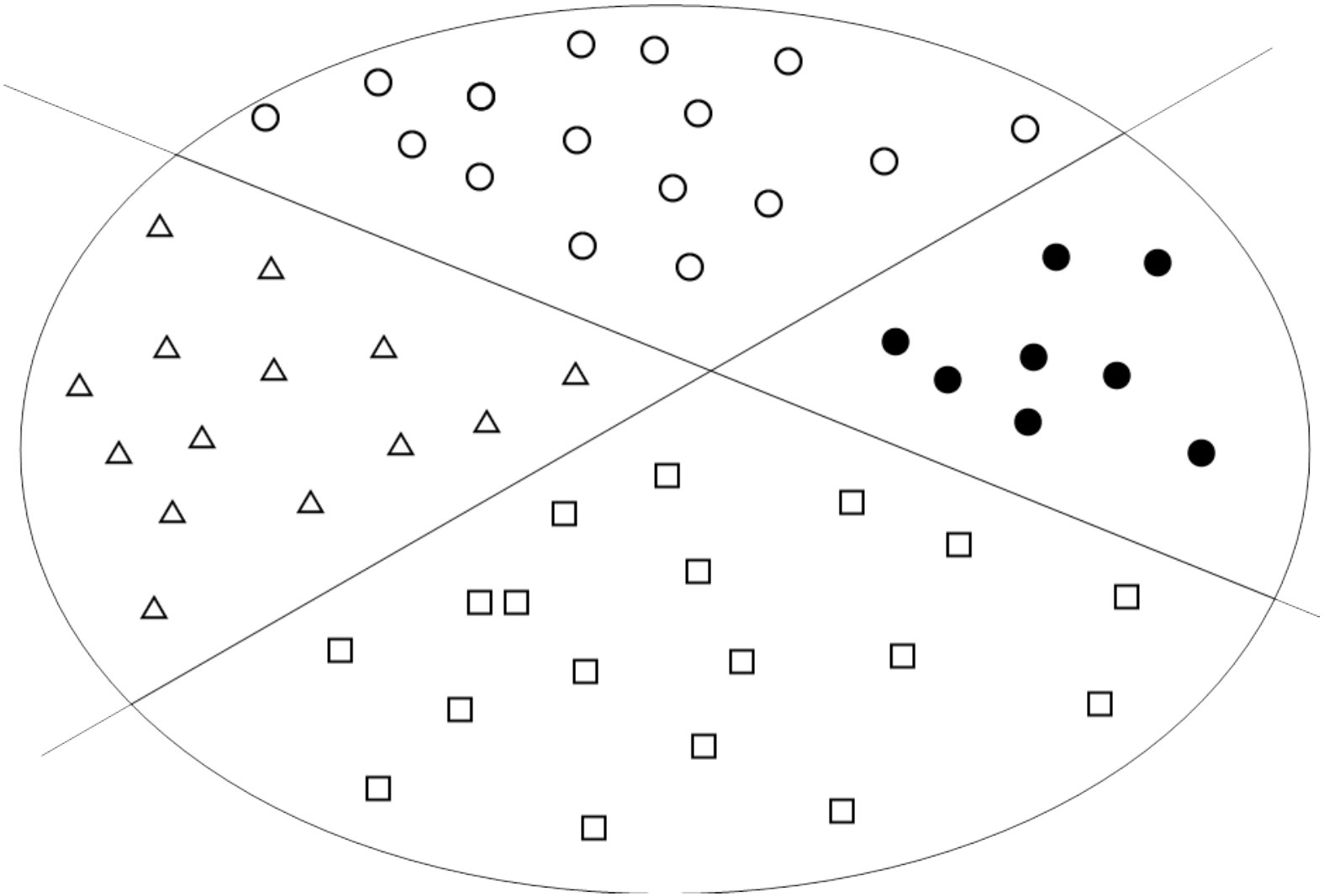
決定木の分析

- 分析：データを元に、このような
  木をつくること
- 様々な特性のデータの集まりがあったとき、その
  特徴によって線引き して分割する
  - 例えば「丸かどうか」といった条件を考えることができる
- これを機械的に作成する→各成分の値に従って
  について考える

決定木の分析

- 分析：データを元に、このような**木をつくる**こと
- 様々な特性のデータの集まりがあったとき、その**特徴によって線引き**して分割する
  - 例えば「丸かどうか」といった条件を考えることができる
- これを機械的に作成する→各成分の値に従って**分岐の条件**について考える

決定木の分析（続き）



不純度とジニ係数

- 木構造を作るためには何等かの基準でグループを分け、さらにそれぞれのグループを分けて…を繰り返す
- 一般に、分析対象の変数（目的変数）は事前にきめておく
- 出来た木については以下のように呼ぶこともある
  - 目的変数が [　　] 変数の場合: [　　]
  - 目的変数が [　　] 変数の場合: [　　]

不純度とジニ係数

- 木構造を作るためには何等かの基準でグループを分け、さらにそれぞれのグループを分けて…を繰り返す
- 一般に、分析対象の変数（目的変数）は事前にきめておく
- 出来た木については以下のように呼ぶこともある
  - 目的変数が カテゴリー 変数の場合:
  - 目的変数が 変数の場合:

不純度とジニ係数

- 木構造を作るためには何等かの基準でグループを分け、さらにそれぞれのグループを分けて…を繰り返す
- 一般に、分析対象の変数（目的変数）は事前にきめておく
- 出来た木については以下のように呼ぶこともある
  - 目的変数が カテゴリー 変数の場合: 分類木
  - 目的変数が [　　] 変数の場合: [　　]

不純度とジニ係数

- 木構造を作るためには何等かの基準でグループを分け、さらにそれぞれのグループを分けて…を繰り返す
- 一般に、分析対象の変数（目的変数）は事前にきめておく
- 出来た木については以下のように呼ぶこともある
  - 目的変数が カテゴリー 変数の場合: 分類木
  - 目的変数が 連続 変数の場合:

不純度とジニ係数

- 木構造を作るためには何等かの基準でグループを分け、さらにそれぞれのグループを分けて…を繰り返す
- 一般に、分析対象の変数（目的変数）は事前にきめておく
- 出来た木については以下のように呼ぶこともある
  - 目的変数が **カテゴリー** 変数の場合: **分類木**
  - 目的変数が **連続** 変数の場合: **回帰木**

不純度とジニ係数（データ例）

number	age	sex	score
1	Under19	Man	Over60
2	Under19	Woman	Over60
3	Under19	Man	Over60
4	Under19	Man	Over60
5	Over20	Man	Over60
6	Under19	Man	Under59
7	Over20	Woman	Under59
8	Over20	Woman	Under59
9	Over20	Man	Under59
10	Over20	Woman	Under59

不純度とジニ係数（CART）

- [ ]: L. Breiman et. al. によって提案されたアルゴリズム
- 前の例で目的変数を score (成績）として、分割する場合を考えるとき、綺麗に＝混ざりが少なく分離してくれるほうが望ましい
- →データの中の混ざり具合＝ [ ]を求める
- そのために、 [ ] という指標を考える
  - 別のアルゴリズムでは [ ]を使う場合もある（詳細略）

不純度とジニ係数（CART）

- CART: L. Breiman et. al. によって提案されたアルゴリズム
- 前の例で目的変数を score (成績）として、分割する場合を考えるとき、綺麗に＝混ざりが少なく分離してくれるほうが望ましい
- →データの中の混ざり具合＝ [ ] を求める
- そのために、 [ ] という指標を考える
  - 別のアルゴリズムでは [ ] を使う場合もある（詳細略）

不純度とジニ係数（CART）

- CART: L. Breiman et. al. によって提案されたアルゴリズム
- 前の例で目的変数を score (成績）として、分割する場合を考えるとき、綺麗に＝混ざりが少なく分離してくれるほうが望ましい
- →データの中の混ざり具合＝ 不純度 を求める
- そのために、 [ ] という指標を考える
  - 別のアルゴリズムでは [ ] を使う場合もある（詳細略）

不純度とジニ係数（CART）

- CART: L. Breiman et. al. によって提案されたアルゴリズム
- 前の例で目的変数を score (成績）として、分割する場合を考えるとき、綺麗に＝混ざりが少なく分離してくれるほうが望ましい
- →データの中の混ざり具合＝ 不純度 を求める
- そのために、 ジニ係数 という指標を考える
  - 別のアルゴリズムでは [　　] を使う場合もある（詳細略）

不純度とジニ係数（CART）

- CART: L. Breiman et. al. によって提案されたアルゴリズム
- 前の例で目的変数を score (成績）として、分割する場合を考えるとき、綺麗に＝混ざりが少なく分離してくれるほうが望ましい
- →データの中の混ざり具合＝ 不純度 を求める
- そのために、 ジニ係数 という指標を考える
  - 別のアルゴリズムでは 情報量 を使う場合もある（詳細略）

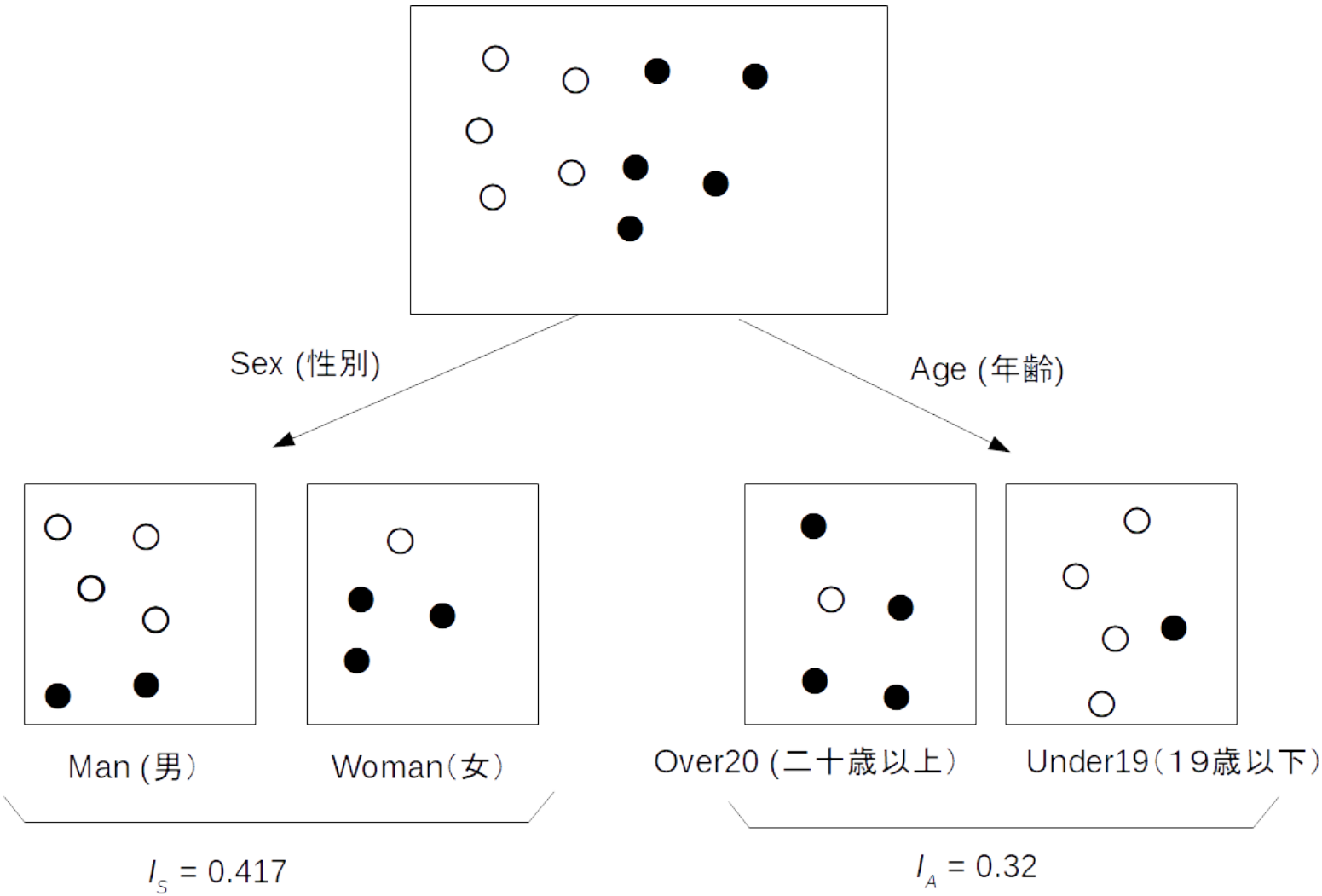
不純度とジニ係数（ジニ係数）

- C. Gini が考案した指数
- データから　　に２つの要素を抜き出したときに、その２つが別のクラスに属する確率
- クラス A に属する確率を P_A、B に属する確率を $P_B = (1 - P_A)$ とすると、ジニ係数 I は以下になる
  $I = 1 - {P_A}^2 - {P_B}^2 = 2P_A(1 - P_A)$
- 一般には $1 - \sum_{i=1}^{n} P_i^2$

不純度とジニ係数（ジニ係数）

- C. Gini が考案した指数
- データから **ランダム** に２つの要素を抜き出したときに、その２つが別のクラスに属する確率
- クラス A に属する確率を P_A、B に属する確率を $P_B = (1 - P_A)$ とすると、ジニ係数 I は以下になる
  $I = 1 - {P_A}^2 - {P_B}^2 = 2P_A(1 - P_A)$
- 一般には $1 - \sum_{i=1}^{n} P_i^2$

分割の候補


Sex (性別)
Age (年齢)
Man (男)
Woman（女）
Over20 (二十歳以上)
Under19（19歳以下）
I_S = 0.417
I_A = 0.32

不純度とジニ係数（続き）

- 分割前のジニ係数は $I_P = 1 - ((\frac{5}{10})^2 + (\frac{5}{10})^2) = 0.5$
- 性別で分割した場合のジニ係数は
  $I_S(Man) = 1 - ((\frac{4}{6})^2 + (\frac{2}{6})^2) = \frac{4}{9} = 0.444..$
  $I_S(Woman) = 1 - ((\frac{1}{4})^2 + (\frac{3}{4})^2) = \frac{3}{8} = 0.375$
  個数の割合をかけて $I_S = \frac{6}{10} \times \frac{4}{9} + \frac{4}{10} \times \frac{3}{8} = 0.417...$

不純度とジニ係数（続き）

- 年齢で分割した場合のジニ係数は
  $I_A(Over20) = 1 - ((\frac{1}{5})^2 + (\frac{4}{5})^2) = \frac{8}{25} = 0.32$
  $I_A(Under19) = 1 - ((\frac{4}{5})^2 + (\frac{1}{5})^2) = \frac{8}{25} = 0.32$
  個数の割合をかけて $I_A = \frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{8}{25} = 0.32$
- 不純度が高いほどジニ係数は［　　］（図参照）
- この例では年齢で分割するほうが［　　］なっている
- このように決定木はジニ係数の［　　］となる分岐を見つけ、それを繰り返す

不純度とジニ係数（続き）

- 年齢で分割した場合のジニ係数は
  $I_A(Over20) = 1 - ((\frac{1}{5})^2 + (\frac{4}{5})^2) = \frac{8}{25} = 0.32$
  $I_A(Under19) = 1 - ((\frac{4}{5})^2 + (\frac{1}{5})^2) = \frac{8}{25} = 0.32$
  個数の割合をかけて $I_A = \frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{8}{25} = 0.32$
- 不純度が高いほどジニ係数は 大きい （図参照）
- この例では年齢で分割するほうが
  [　　] なっている
- このように決定木はジニ係数の [　　] となる分岐を見つけ、それを繰り返す

不純度とジニ係数（続き）

- 年齢で分割した場合のジニ係数は
  $I_A(Over20) = 1 - ((\frac{1}{5})^2 + (\frac{4}{5})^2) = \frac{8}{25} = 0.32$
  $I_A(Under19) = 1 - ((\frac{4}{5})^2 + (\frac{1}{5})^2) = \frac{8}{25} = 0.32$
  個数の割合をかけて $I_A = \frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{8}{25} = 0.32$
- 不純度が高いほどジニ係数は 大きい （図参照）
- この例では年齢で分割するほうが
  不純度が小さく なっている
- このように決定木はジニ係数の [　　] となる分岐を見つけ、それを繰り返す

不純度とジニ係数（続き）

- 年齢で分割した場合のジニ係数は
  $I_A(Over20) = 1 - ((\frac{1}{5})^2 + (\frac{4}{5})^2) = \frac{8}{25} = 0.32$
  $I_A(Under19) = 1 - ((\frac{4}{5})^2 + (\frac{1}{5})^2) = \frac{8}{25} = 0.32$
  個数の割合をかけて $I_A = \frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{8}{25} = 0.32$
- 不純度が高いほどジニ係数は **大きい** （図参照）
- この例では年齢で分割するほうが **不純度が小さく** なっている
- このように決定木はジニ係数の **差が最大** となる分岐を見つけ、それを繰り返す

決定木（演習）

1. まず dec-exam.csv をダウンロードして ”I:￥rdata” に保存する
2. パッケージ ”rpart” をインストールする。rpart() 関数＝ CART アルゴリズムが使えるようになる

```
install.packages("rpart")
```

決定木（演習・続き）

3. 以下を実行する。確認のため s2 の中身をみる

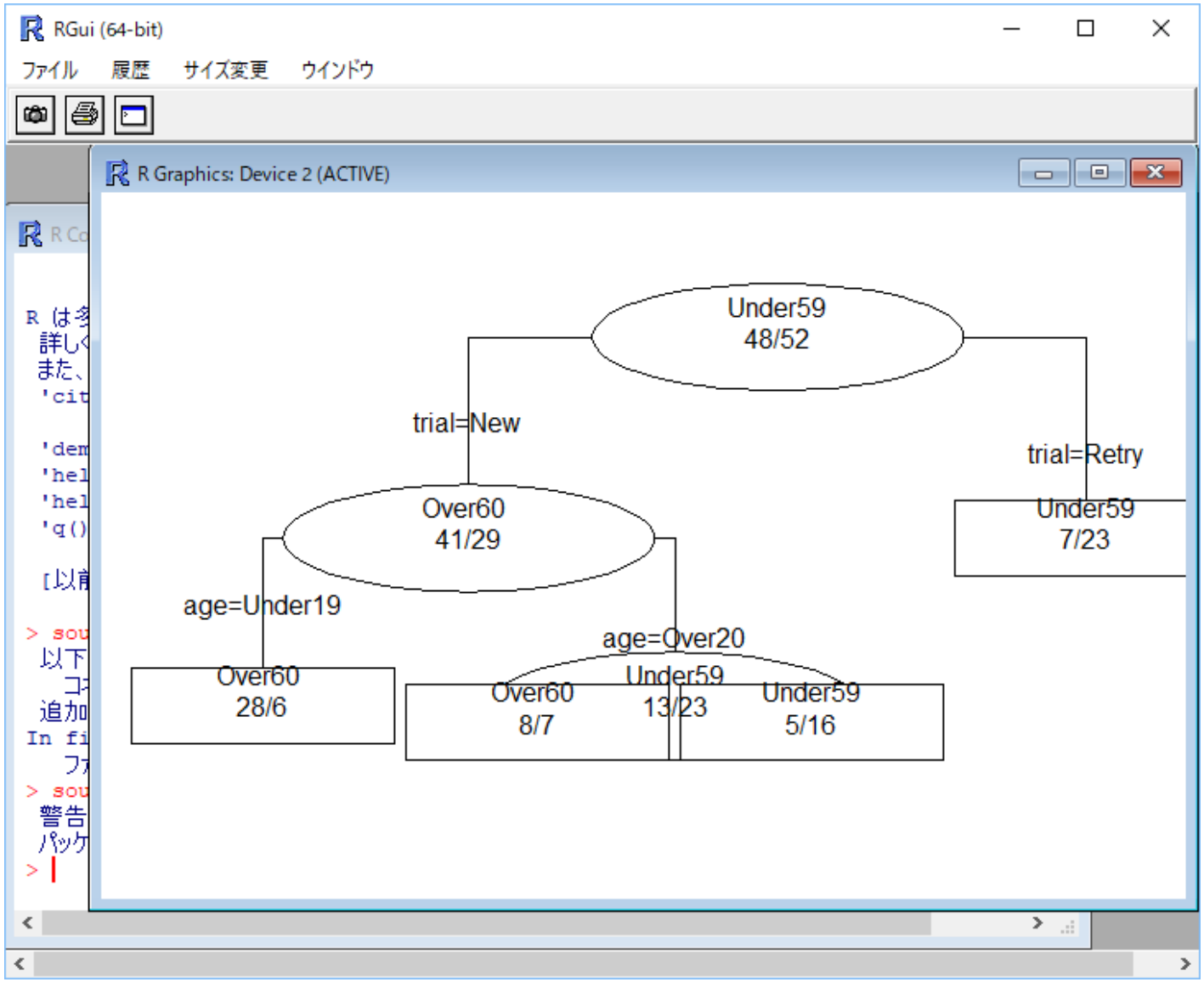
```
library(rpart)
s1 <- read.csv("I:/rdata/dec-exam.csv",
  header=T,row.names=1) # 本当は１行
s2 <- rpart(score~.,data=s1,
  method="class") # 本当は１行
# method="class" で分類木であることを指示
```

決定木（演習・続き）

4. 引き続いて以下を実行し、図示する

```
par(xpd=TRUE)
# 文字が描画領域からはみ出しても切れないようにする
plot(s2) # 木構造を描画
text(s2,pretty=0,all=T,use.n=T,fancy=T)
# ラベルを追加
# pretty=0 : 各条件で変数の値を省略せず表示させる
# all=T :すべての分岐点でラベルを表示
# use.n=T :分類結果の数値を表示
# fancy=T :各ノードを楕円や四角で表示
```

決定木（演習・続き）


RGui (64-bit)
ファイル 履歴 サイズ変更 ウインドウ
R Graphics: Device 2 (ACTIVE)
Under59
48/52
trial=New
trial=Retry
Over60
41/29
Under59
7/23
age=Under19
age=Over20
Over60
28/6
Over60
8/7
Under59
13/23
Under59
5/16

決定木（演習の木の見方）

- 楕円内の、例えば根の ”Under59 48/52” は、全体が $48 + 52 = 100$ で、そのうち 52 人が Under59（多数派）であることを示す
- ”trial=Retry” のほうの子ノードで ”Under59 7/23” とは、Retry=再試験で受けた人は $7 + 23 = 30$ 人いて、そのうちの 23 人が Under59（多数派）であることを示す
- …以下同様に読み解け、再試を受けてる人はやはり 59 点以下の人が多かったり、１回目の受講で 19 歳以下の場合に 60 点以上のことが多かったり、等が読み取れる。

決定木（演習・続き）

5. 引き続いて以下を実行し、図示する

```
install.packages("rpart.plot")
library(rpart.plot) # 木構造を描画
# ライブラリ追加
rpart.plot(s2, extra = 4)
# ちょっと綺麗？に表示
```

決定木での予測

- 既存データから学習をおこない、データにはどのような傾向があるのか（どの属性が重要な決定要因となっているのか）を知ることができた
- この分析結果は予測モデルとして利用できる
- predict() 関数を用いる

決定木（演習・続き）

6. まずdec-test.csv をダウンロードして ”I:￥rdata” に保存する
7. 以下を実行する

```
t1 <- read.csv("I:/rdata/dec-test.csv",
  header=T,row.names=1) # 本当は１行
# 予測してみる
prd <- predict(s2,t1)
```

決定木（演習・続き）

8. prd の中身をチェックする

```
> prd
     Over60    Under59
1 0.8235294 0.1764706
2 0.2333333 0.7666667
3 0.2380952 0.7619048
```

決定木（演習・続き）

9 dec-test は

```
1,Under19,Man,New,UNK
2,Over20,Woman,Retry,UNK
3,Over20,Woman,New,UNK
```

だったので、

- １は 0.8235294 の確率で合格、
- ２は 0.7666667 の確率で不合格、
- ３は 0.7619048 の確率で不合格、

となると予測される

おしまい

質問があれば、お気軽に！