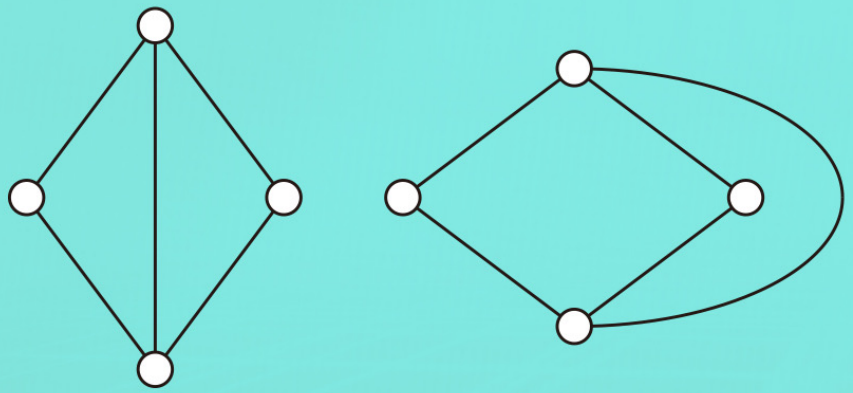
# データの分析と知識発見

# グラフと木

## グラフ

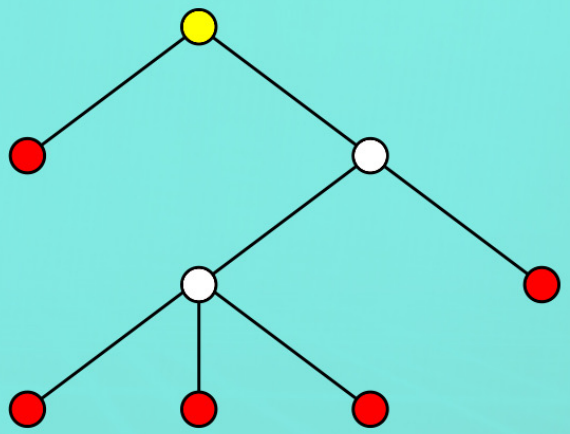
頂点（ノード）と枝（エッジ）の集まり



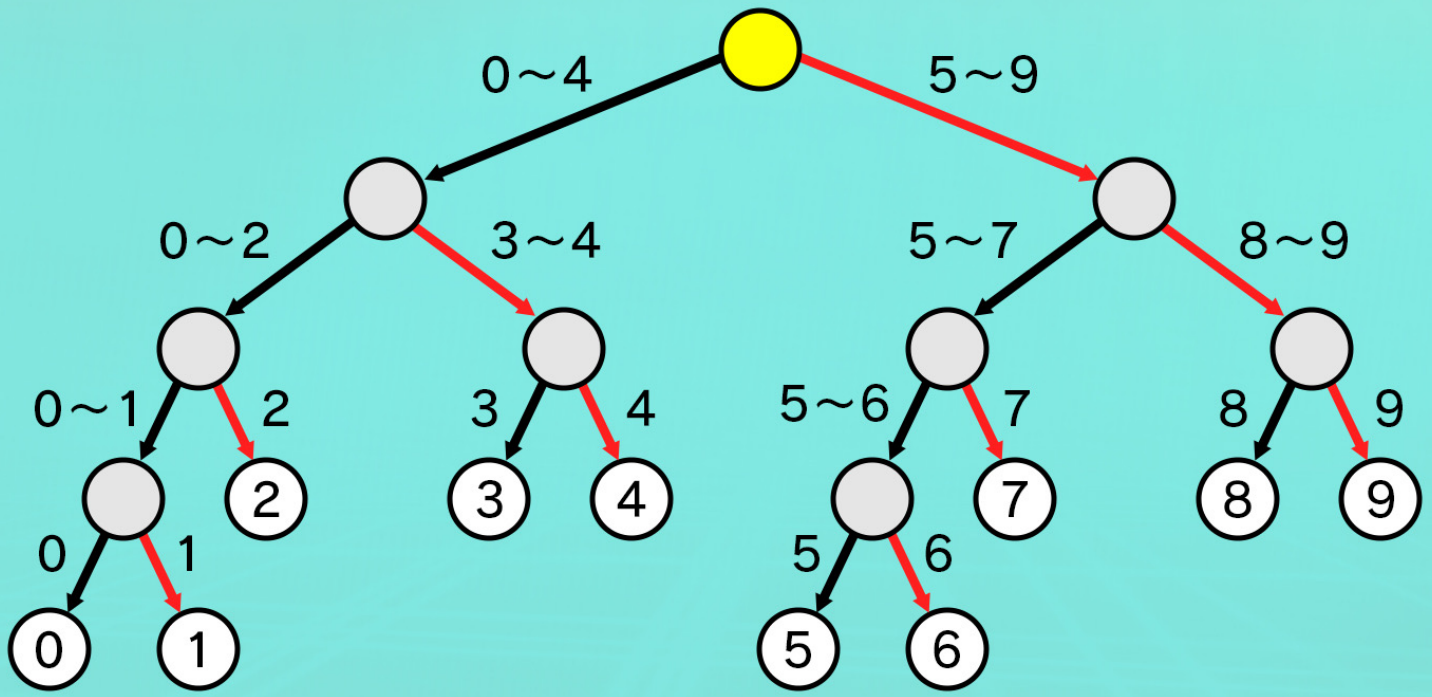
点のつながり方を表す

## 木（根つき木）

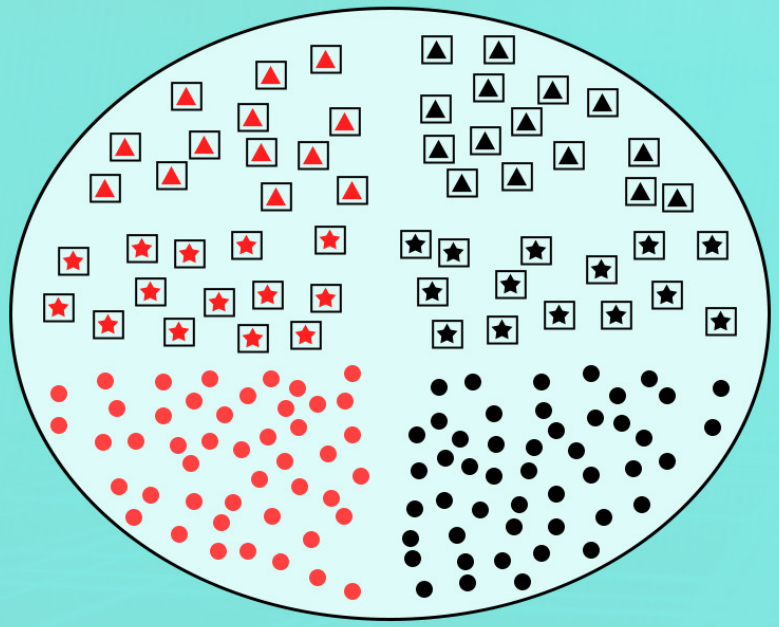
ループのないグラフ



# 数当てゲーム

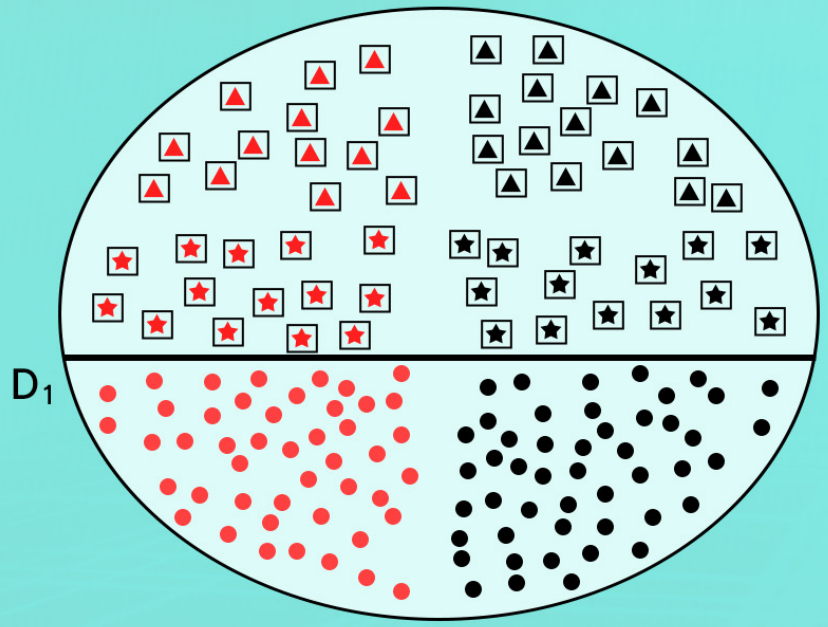


# 集合の分割と木構造

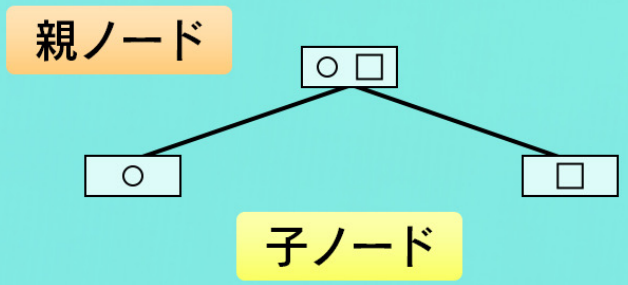


要素に応じて集合を分割

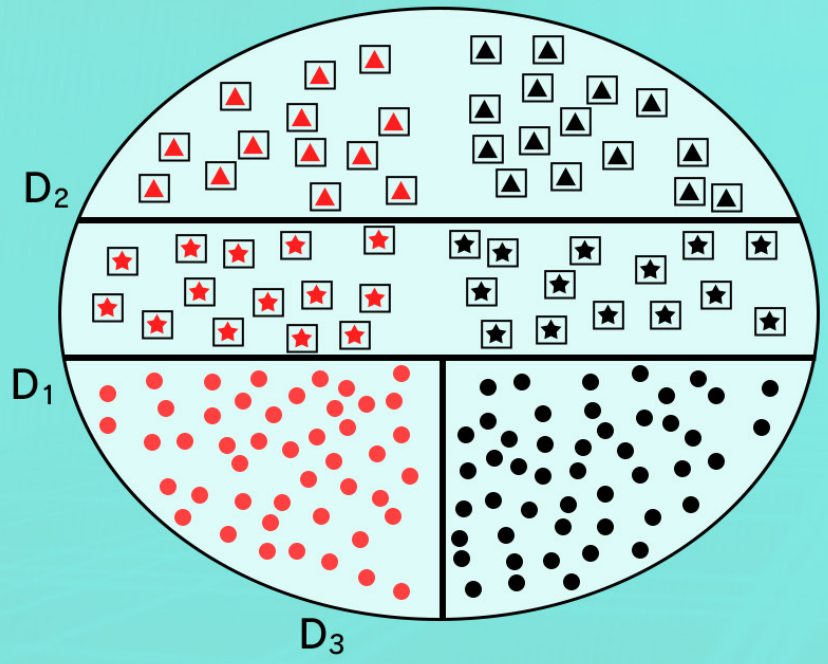
# 集合の分割と木構造



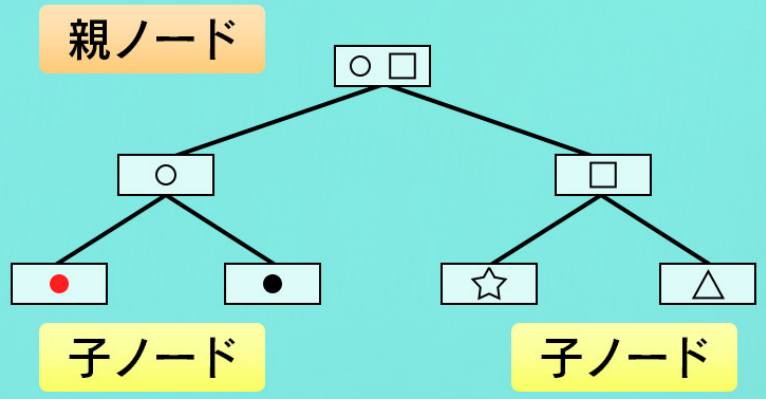

**要素に応じて集合を分割**



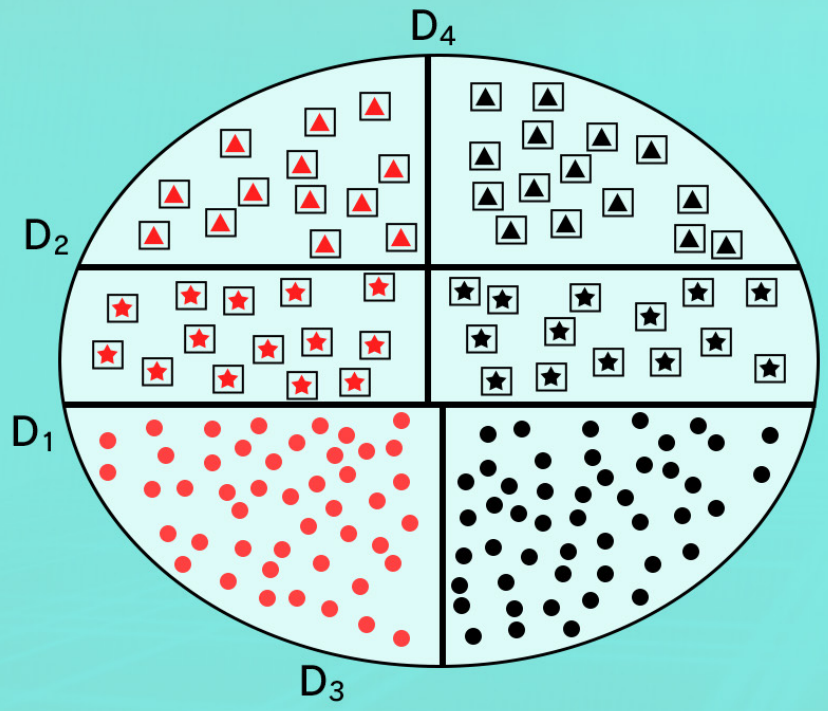
# 集合の分割と木構造



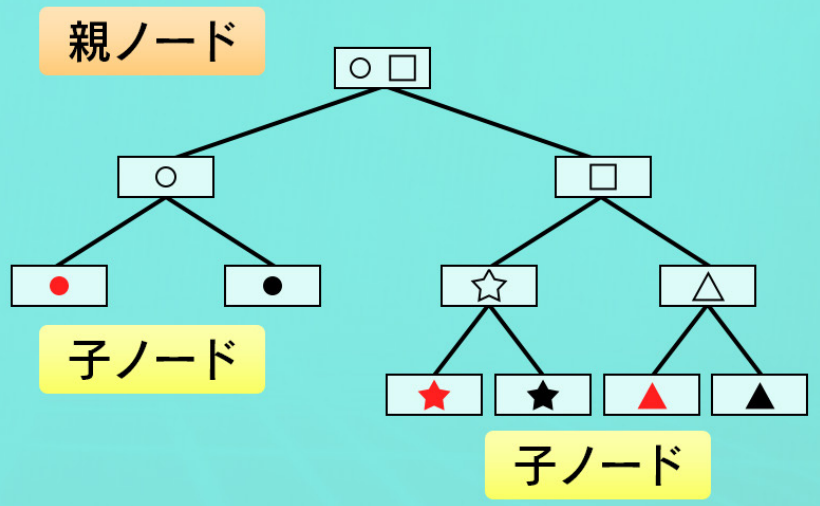

要素に応じて集合を分割



# 集合の分割と木構造

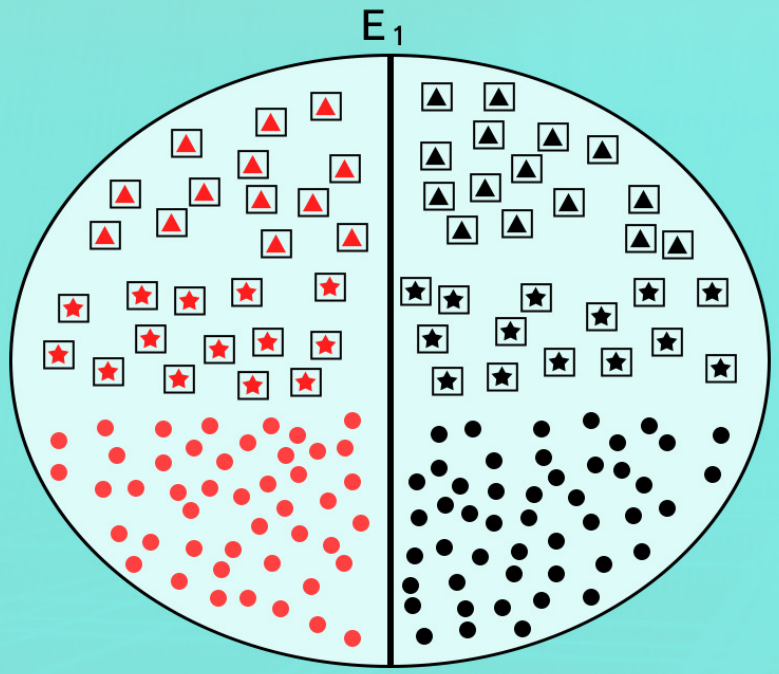



**要素に応じて集合を分割**

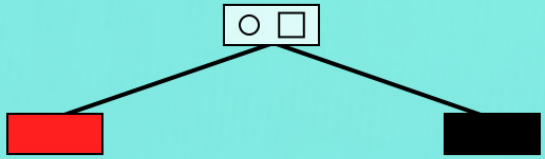



**葉は子ノードを持たない**

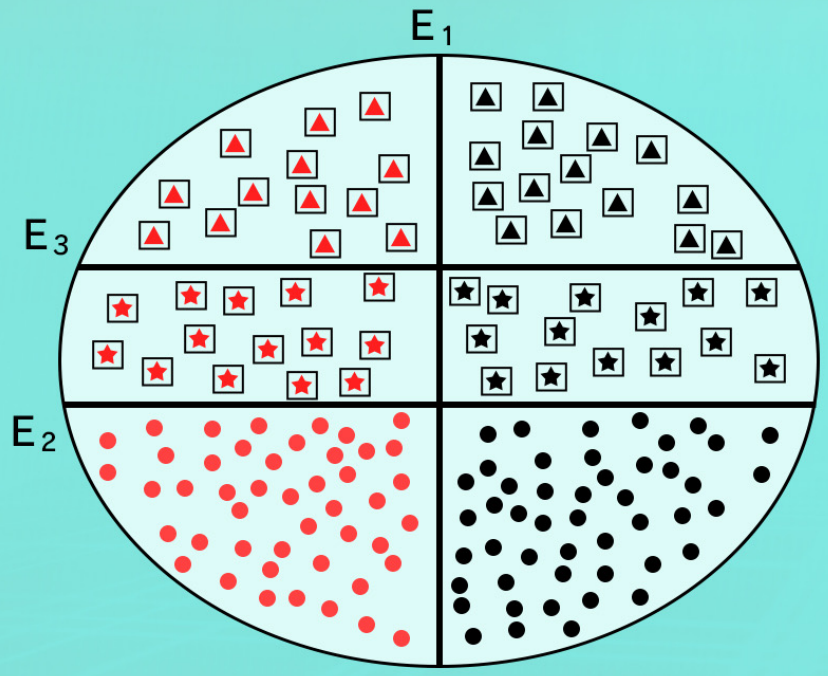
# 集合の分割と木構造



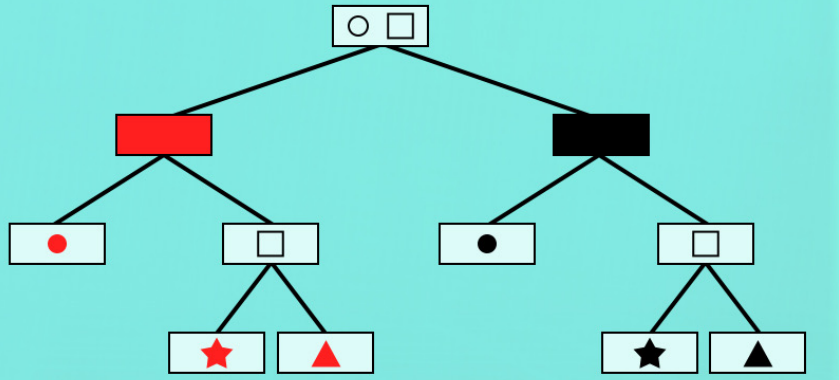

**構造は１種類とは限らない**



# 集合の分割と木構造




構造は分割方法に依存する



どんな目的で分割をするかに依る

# （例題）成績データ

ある科目の10人の試験結果

| 学生番号 | 年 齢 | 性 別 | 点 数 |
|---|---|---|---|
| 1 | 40歳未満 | 男性 | 70点以上 |
| 2 | 40歳未満 | 女性 | 70点以上 |
| 3 | 40歳未満 | 男性 | 70点以上 |
| 4 | 40歳未満 | 男性 | 70点以上 |
| 5 | 40歳以上 | 男性 | 70点以上 |
| 6 | 40歳未満 | 男性 | 70点未満 |
| 7 | 40歳以上 | 女性 | 70点未満 |
| 8 | 40歳以上 | 女性 | 70点未満 |
| 9 | 40歳以上 | 男性 | 70点未満 |
| 10 | 40歳以上 | 女性 | 70点未満 |

カテゴリー変数
（数種類）

バイナリー変数
（2種類）

連続変数

# （例題）成績データ

ある科目の10人の試験結果

カテゴリー変数
（数種類）

バイナリー変数
（2種類）

連続変数

| 学生番号 | 年 齢 | 性 別 | 点 数 |
|---|---|---|---|
| 1 | 40歳未満 | 男性 | 70点以上 |
| 2 | 40歳未満 | 女性 | 70点以上 |
| 3 | 40歳未満 | 男性 | 70点以上 |
| 4 | 40歳未満 | 男性 | 70点以上 |
| 5 | 40歳以上 | 男性 | 70点以上 |
| 6 | 40歳未満 | 男性 | 70点未満 |
| 7 | 40歳以上 | 女性 | 70点未満 |
| 8 | 40歳以上 | 女性 | 70点未満 |
| 9 | 40歳以上 | 男性 | 70点未満 |
| 10 | 40歳以上 | 女性 | 70点未満 |

目的変数

# （例題）成績データ

ある科目の10人の試験結果

カテゴリー変数
（数種類）

バイナリー変数
（2種類）

連続変数

| 学生番号 | 年 齢 | 性 別 | 点 数 |
|---|---|---|---|
| 1 | 40歳未満 | 男性 | 70点以上 |
| 2 | 40歳未満 | 女性 | 70点以上 |
| 3 | 40歳未満 | 男性 | 70点以上 |
| 4 | 40歳未満 | 男性 | 70点以上 |
| 5 | 40歳以上 | 男性 | 70点以上 |
| 6 | 40歳未満 | 男性 | 70点未満 |
| 7 | 40歳以上 | 女性 | 70点未満 |
| 8 | 40歳以上 | 女性 | 70点未満 |
| 9 | 40歳以上 | 男性 | 70点未満 |
| 10 | 40歳以上 | 女性 | 70点未満 |

説明変数　　目的変数

# （例題）成績データ

ある科目の10人の試験結果

カテゴリー変数
（数種類）

バイナリー変数
（2種類）

連続変数

| 学生番号 | 年 齢 | 性 別 | 点 数 |
|---|---|---|---|
| 1 | 40歳未満 | 男性 | 70点以上 |
| 2 | 40歳未満 | 女性 | 70点以上 |
| 3 | 40歳未満 | 男性 | 70点以上 |
| 4 | 40歳未満 | 男性 | 70点以上 |
| 5 | 40歳以上 | 男性 | 70点以上 |
| 6 | 40歳未満 | 男性 | 70点未満 |
| 7 | 40歳以上 | 女性 | 70点未満 |
| 8 | 40歳以上 | 女性 | 70点未満 |
| 9 | 40歳以上 | 男性 | 70点未満 |
| 10 | 40歳以上 | 女性 | 70点未満 |

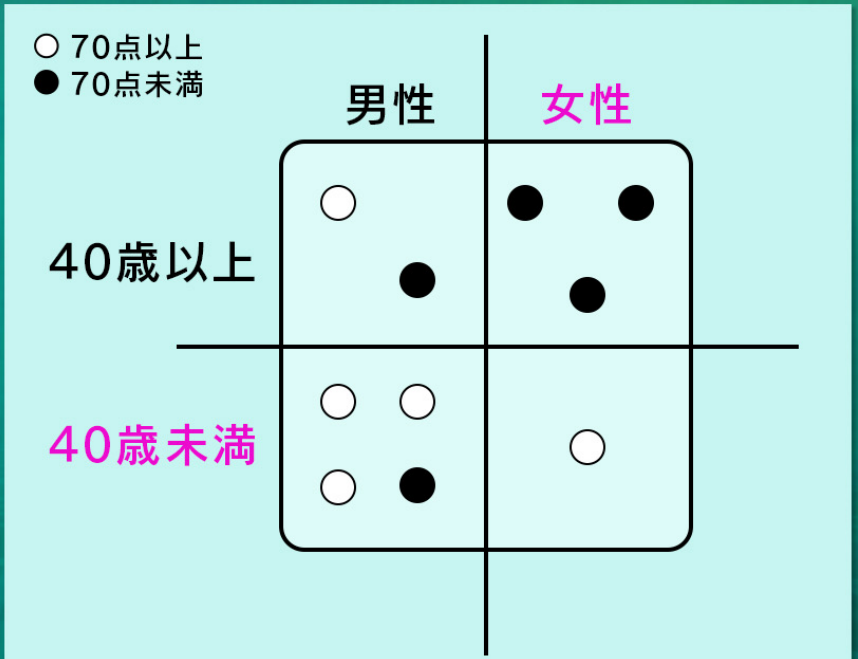
説明変数　　目的変数

目的変数が

- ◇ カテゴリー変数
  - ➡ 分類木
- ◇ 連続変数
  - ➡ 回帰木

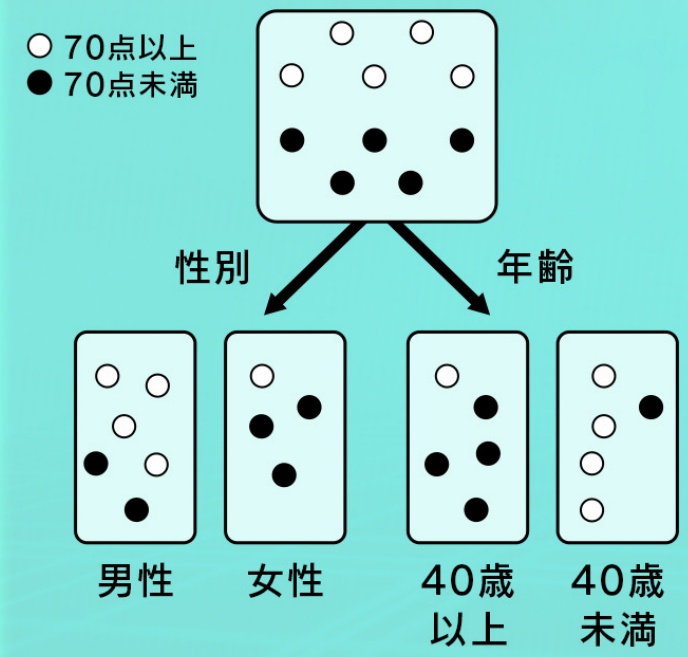
# 不純度とデータの分割

| 学生番号 | 年 齢 | 性 別 | 点 数 |
|---|---|---|---|
| 1 | 40歳未満 | 男性 | 70点以上 |
| 2 | 40歳未満 | 女性 | 70点以上 |
| 3 | 40歳未満 | 男性 | 70点以上 |
| 4 | 40歳未満 | 男性 | 70点以上 |
| 5 | 40歳以上 | 男性 | 70点以上 |
| 6 | 40歳未満 | 男性 | 70点未満 |
| 7 | 40歳以上 | 女性 | 70点未満 |
| 8 | 40歳以上 | 女性 | 70点未満 |
| 9 | 40歳以上 | 男性 | 70点未満 |
| 10 | 40歳以上 | 女性 | 70点未満 |



# 不純度とデータの分割

## 集合の分割




## ジニ係数

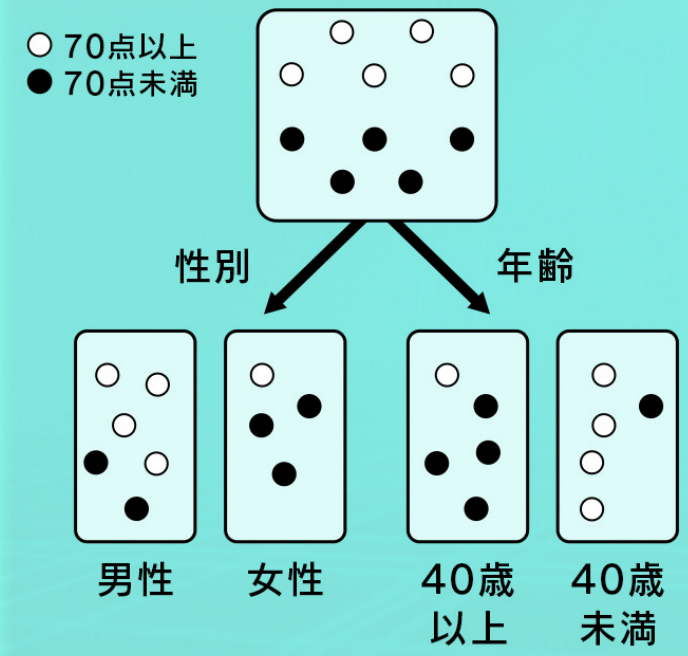
不純度が低くなるように分割

$$I = 1 - p_A^2 - p_B^2$$
$$= 2p_A(1 - p_A)$$

AとBという 2種類のボール：
1つ取り出して 元に戻して
もう1つ取り出すとする
不純度が低い
⇔ 別のものを取り出す割合が小さい

# 不純度とデータの分割

## 集合の分割




## 分割前

$$I_P = 1 - \left(\left(\frac{5}{10}\right)^2 + \left(\frac{5}{10}\right)^2\right) = 0.5$$

## 性 別

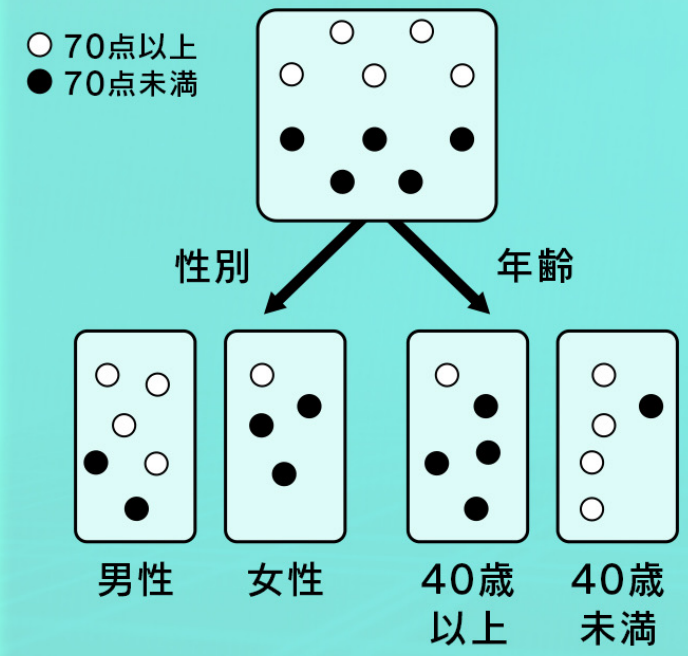
$$I_S(\text{男性}) = 1 - \left(\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right) = \frac{4}{9} = 0.444\cdots$$

$$I_S(\text{女性}) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) = \frac{3}{8} = 0.375$$

$$I_S = \frac{6}{10} \times \frac{4}{9} + \frac{4}{10} \times \frac{3}{8} = \frac{5}{12} = 0.416\cdots$$

# 不純度とデータの分割

## 集合の分割




## 分割前

$$I_P = 1 - \left(\left(\frac{5}{10}\right)^2 + \left(\frac{5}{10}\right)^2\right) = 0.5$$

## 年 齢

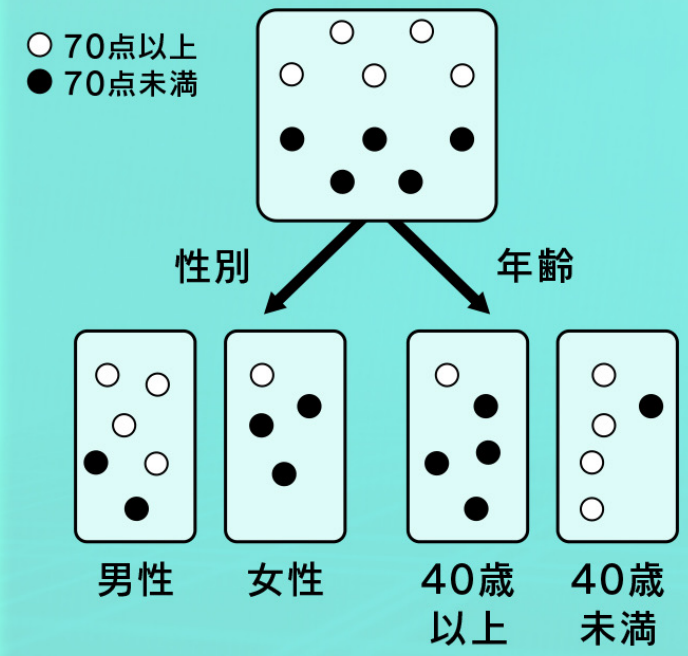
$$I_Y(\text{40歳以上}) = 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right) = \frac{8}{25} = 0.32$$

$$I_Y(\text{40歳未満}) = 1 - \left(\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right) = \frac{8}{25} = 0.32$$

$$I_Y = \frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{8}{25} = \frac{8}{25} = 0.32$$

# 不純度とデータの分割

## 集合の分割




$$I_P = 1 - \left(\left(\frac{5}{10}\right)^2 + \left(\frac{5}{10}\right)^2\right) = 0.5$$

## 性 別

$$I_S = \frac{6}{10} \times \frac{4}{9} + \frac{4}{10} \times \frac{3}{8} = \frac{5}{12} = 0.416\cdots$$

$$\Delta I_{PS} = I_P - I_S = 0.5 - 0.416\cdots = 0.083\cdots$$

## 年 齢

$$I_Y = \frac{5}{10} \times \frac{8}{25} + \frac{5}{10} \times \frac{8}{25} = \frac{8}{25} = 0.32$$

$$\Delta I_{PY} = I_P - I_Y = 0.5 - 0.32 = 0.18$$

# データについて

| number | age | gender | trial | score |
|---|---|---|---|---|
| 1, | Under40, | Man, | New, | Under70 |
| 2, | Under40, | Woman, | New, | Under70 |
| 3, | Over40, | Woman, | Retry, | Under70 |
| 4, | Over40, | Man, | New, | Over70 |
| 5, | Over40, | Woman, | New, | Over70 |
| 6, | Under40, | Man, | New, | Over70 |
| 7, | Under40, | Woman, | New, | Under70 |
| 8, | Under40, | Man, | New, | Over70 |
| 9, | Under40, | Woman, | Retry, | Under70 |
| 10, | Under40, | Man, | New, | Over70 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

# Rによる演習(rpart)

```
> library(rpart)
> s1 <- read_csv("rpart.csv")
> head(s1)
```

```
# install.packages(rpart)
# install.packages(rpart.plot)
```

```
> s2 <- rpart(score ~ age + gender + trial,
        data=s1,method="class")
> s2
```

```
# 目的変数 ~　説明変数1 + 説明変数2 + …
#  method=“class” 分類木
```

```
> library(rpart.plot)
> rpart.plot(s2,　box.palette = "Greys")
```