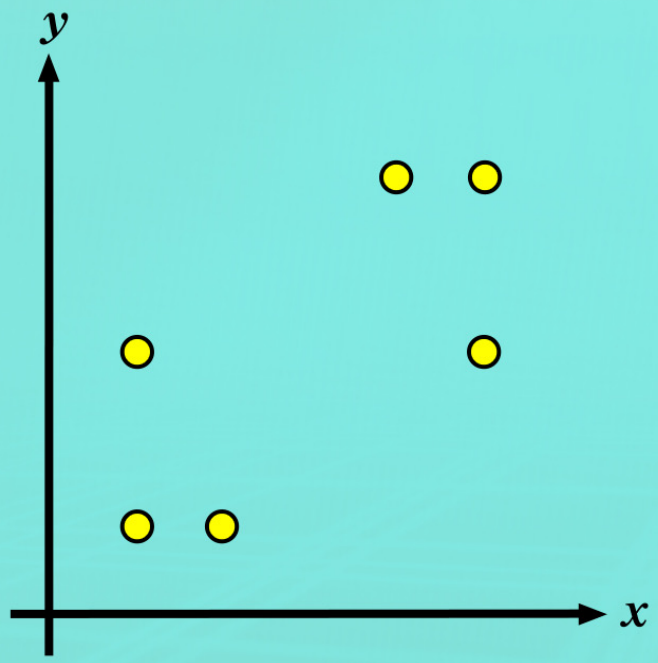
データの分析と知識発見

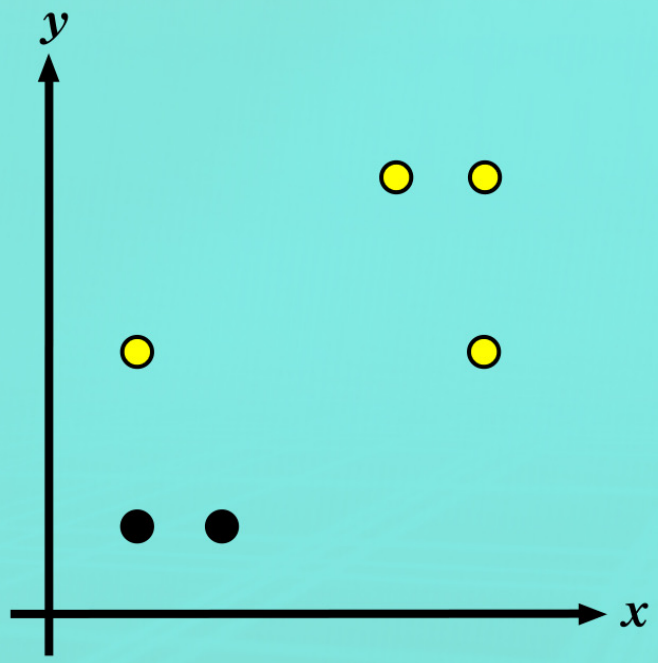
k 平均法

6個の点を2つのグループに分ける



k 平均法

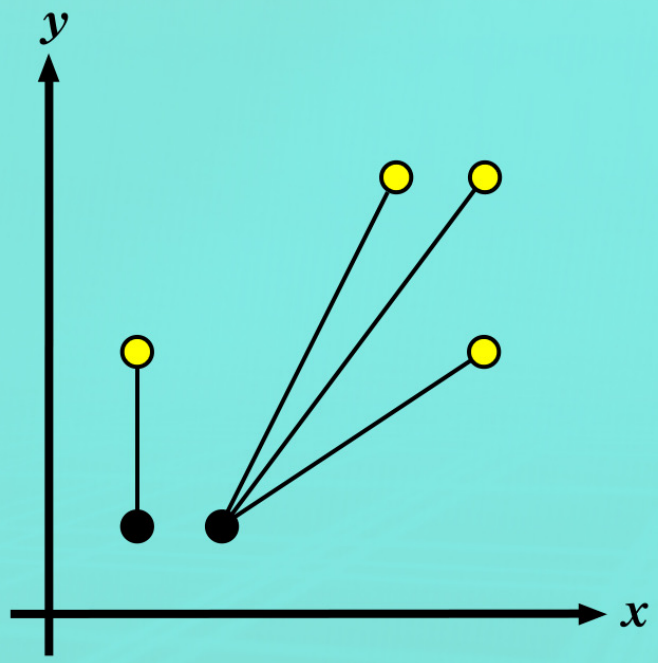
6個の点を2つのグループに分ける



◆2個の点を選び
その点を2つのグループの
代表点とする

k 平均法

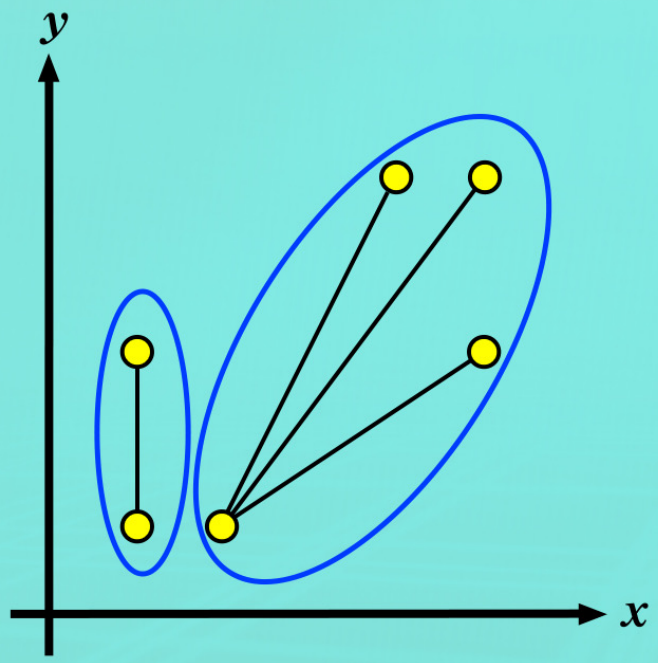
6個の点を2つのグループに分ける



◆その2点との距離を計算する

k 平均法

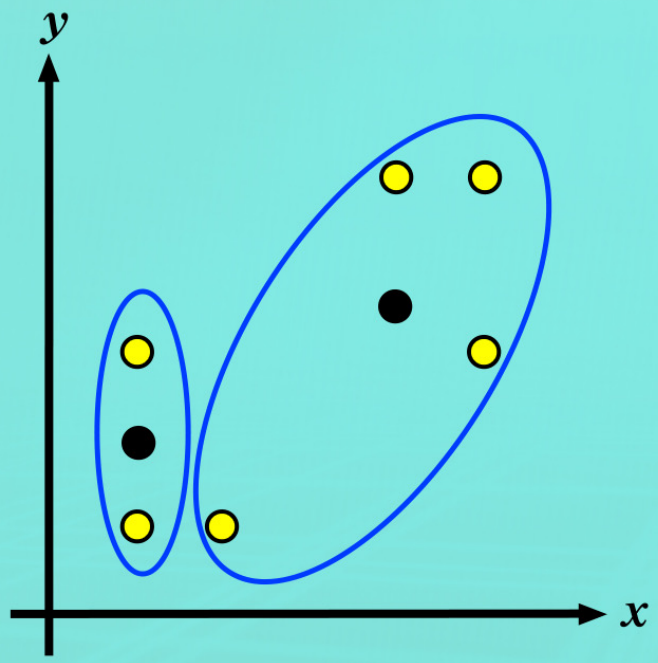
6個の点を2つのグループに分ける



- それぞれの点は
  近い方の代表点のグループに
  属することにする

k 平均法

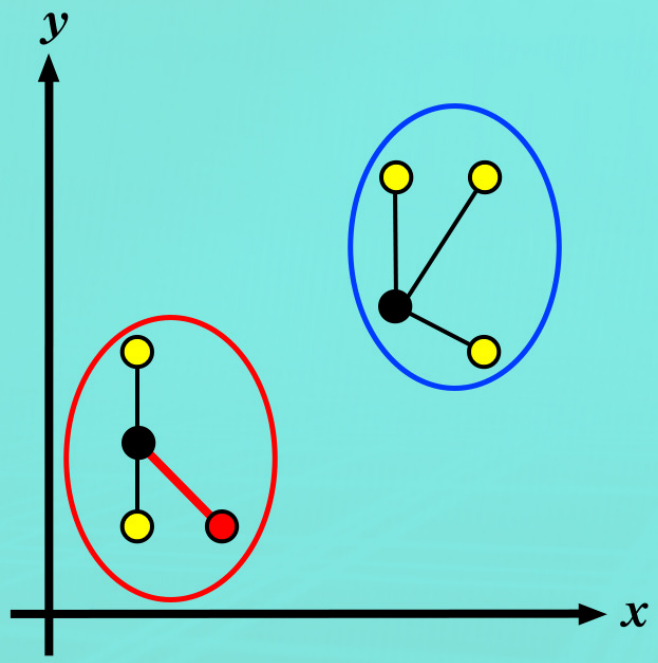
6個の点を2つのグループに分ける



- それぞれのグループの重心を新たな代表の点とする

k 平均法

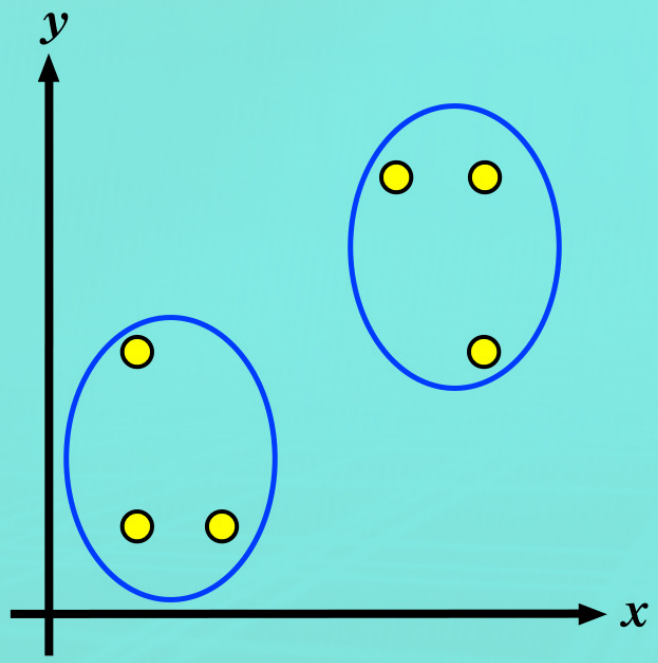
6個の点を2つのグループに分ける



- 先ほどと同様に
  近い方の代表点と同じ
  グループとして分類する

k 平均法

6個の点を2つのグループに分ける



- ◆ 代表となる点を選び直す
- ◆ 距離を計算する
- ◆ 近い方のグループの点とする

という作業を繰り返す

グループに変化がなければ終了とする

Rによる演習（1）

```
> library(tidyverse)
> set.seed(38)
> sigma <- 0.25
> ca <- tibble( x=rnorm(50,1,sigma),    y=rnorm(50,1,sigma),   cl_t = “A”)
> cb <- tibble( x=rnorm(50,-1,sigma),   y=rnorm(50,1,sigma),   cl_t = “B”)
> cc <- tibble( x=rnorm(50,-1,sigma),   y=rnorm(50,-1,sigma),  cl_t = “C”)
> cd <- tibble( x=rnorm(50,1,sigma),    y=rnorm(50,-1,sigma),  cl_t = “D”)
> t0 <-rbind(ca,cb,cc,cd)
> ggplot(t0,aes(x=x, y=y, color=cl_t) ) + geom_point()
> k0 <- t0 %>% select(x , y) %>% kmeans(centers = 4)
> k0
> t0 %<>% mutate(cl_k=factor(k0$cluster, label=c(“A”,“B”,“C”,“D”), levels=c(4,2,1,3) ) )
> t0 %>% select(cl_t,cl_k) %>% table()
```

```
# (1,  1) を中心に
# (-1,  1) を中心に
# (-1, -1) を中心に
# ( 1, -1) を中心に
# 4つのグループ

# k平均法
```

```
# k平均法
# kmeans(訓練データの座標, 中心の数)
```

Rによる演習(2)

```
> min <- -2
> max <- 2
> N_test <- 100
> int <- (max-min) / (N_test-1)
> x <- rep(seq(min,max,int), each=N_test)
> y <- rep(seq(min,max,int),  N_test)
> test <-   tibble(x=x, y=y)
```

```
# rep は繰り返し, each とするとそれぞれ N_test個
# eachがない場合には N_test 周繰り返す
```

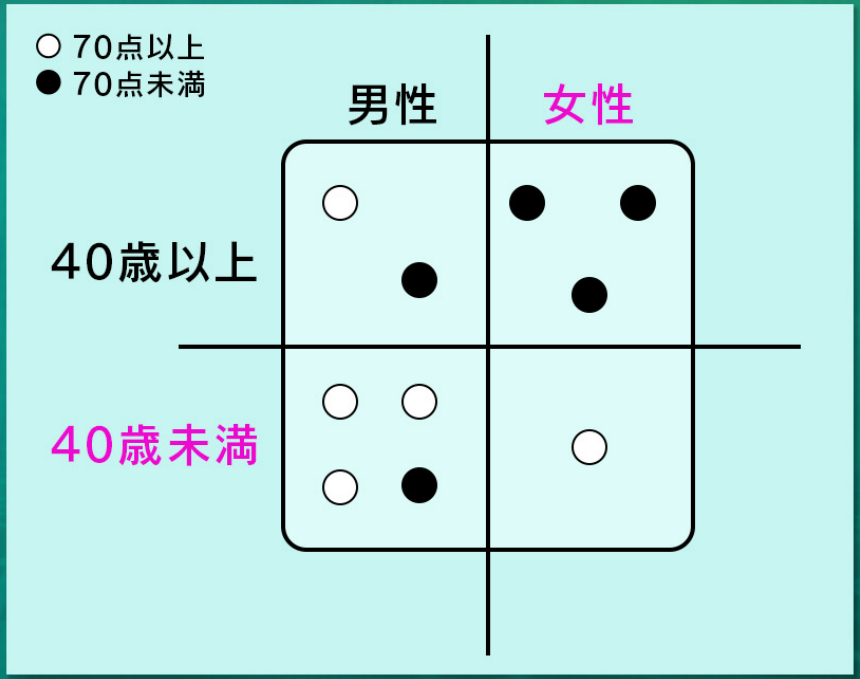
```
> library(class)
> cl_test <-  t0  %>% select(x, y)  %>%   knn(test, cl=t0$cl_t, k=5)
> ggplot(test,aes(x=x,y=y) )+  geom_point(aes(color=cl_test),size=0.5 )
```

```
# k近傍法
# cl_test の色で表示
```

```
# k近傍法
# knn(訓練データの座標, 検証データの座標, 訓練データのラベル, 近傍の数)
# 検証データの各座標はk個の近傍にある訓練データのラベルの多数決で決める
```

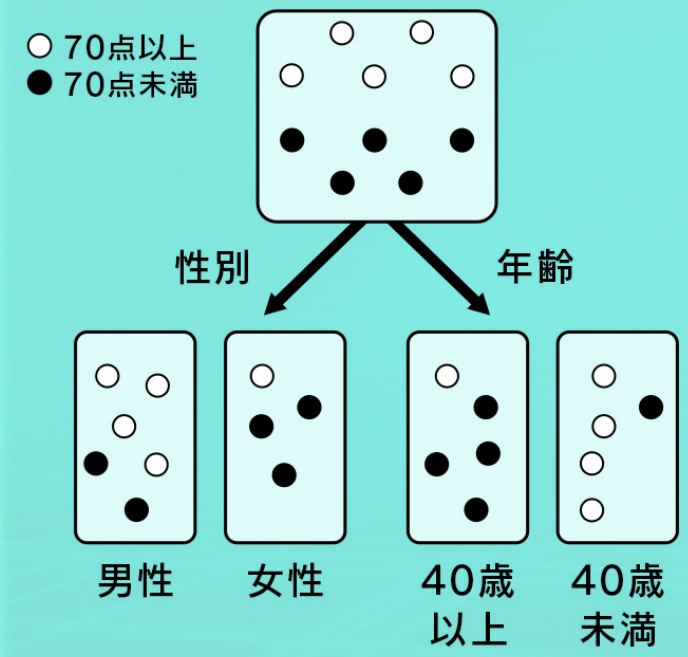
不純度とデータの分割

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

Rによる演習(3)

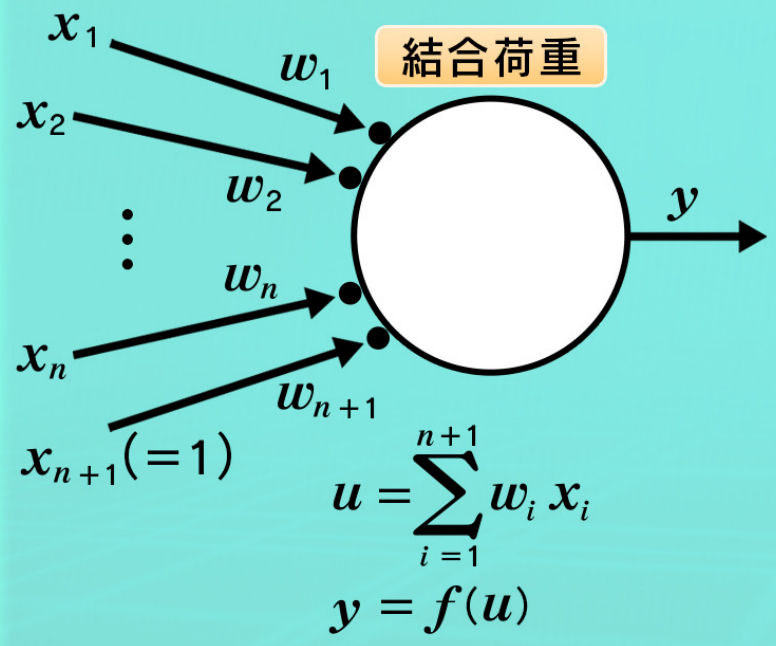
```
> library(rpart)
> rp_t <- rpart(data=t0, cl_t  ~  x + y,  method="class")
> library(rpart.plot)
> rpart.plot(rp_t)
> rp_test <- predict(rp_t, newdata=test,  type="class")

> ggplot(test,aes(x=x,y=y)) + geom_point(aes(color=rp_test), size=0.5 )
```

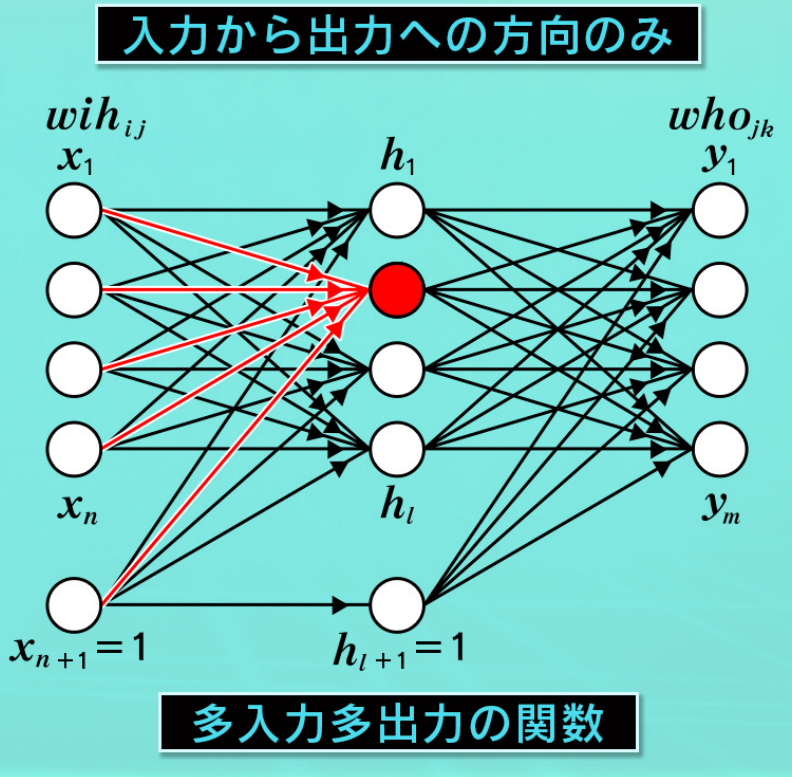
```
# 分類木


# predict で予測, type でclass を指定
```

```
> library(nnet)
> t1 <- t0 %>% mutate(cl_t = as.factor(cl_t) )
> nn_t <- nnet(data=t1, cl_t ~ x+y, size=5)
> nn_test <- predict(nn_t, newdata=test, type="class")

> ggplot(test,aes(x=x,y=y)) +geom_point(aes(color=nn_test), size=0.5 )
```

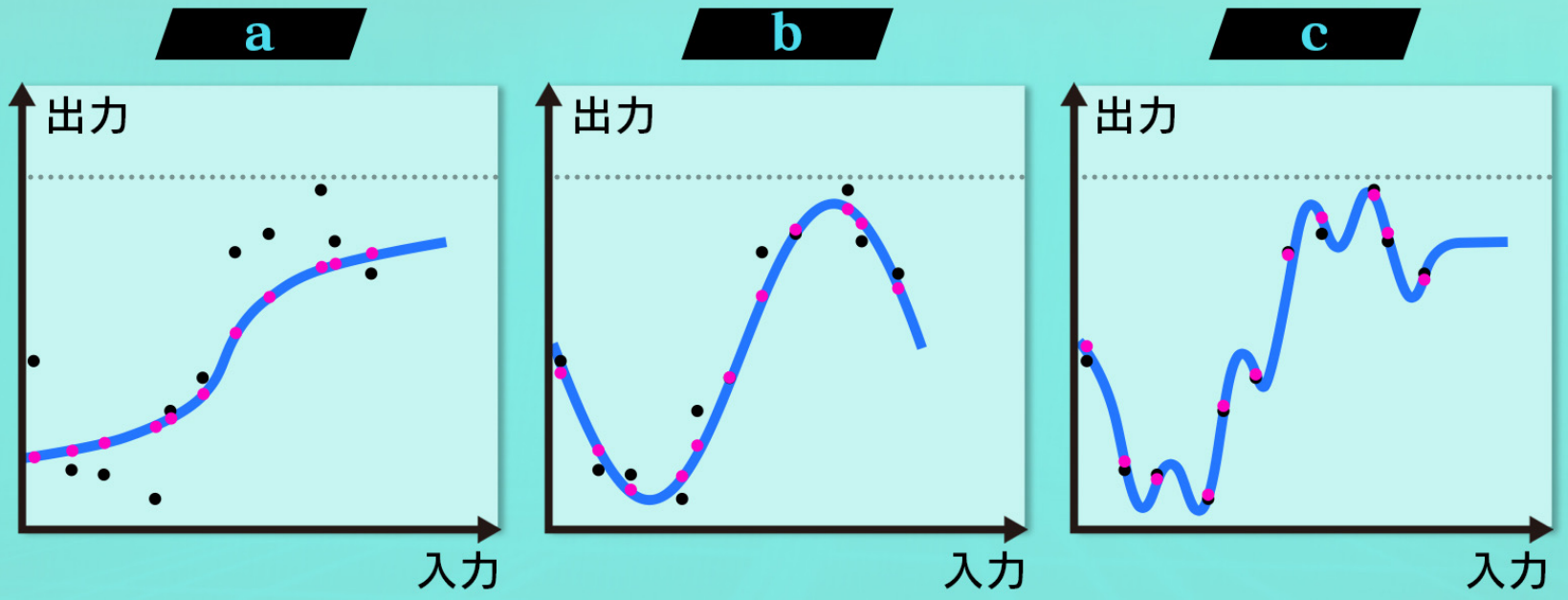
```
# nnet ではクラス分類をするには因子に変更

# predict で予測, type でclass を指定
```

階層型ネットワーク


x_1
x_2
x_n
$x_{n+1}(=1)$
w_1
w_2
w_n
w_{n+1}
結合荷重
y
$u = \sum_{i=1}^{n+1} w_i x_i$
$y = f(u)$


入力から出力への方向のみ
wih_{ij}
who_{jk}
x_1
h_1
y_1
x_n
h_l
y_m
$x_{n+1}=1$
$h_{l+1}=1$
多入力多出力の関数

汎化と過学習(3)



学習が不十分

過学習

Rによる演習(3)

```
> library(rpart)
> rp_t <- rpart(data=t0, cl_t  ~  x + y,  method="class")
> library(rpart.plot)
> rpart.plot(rp_t)
> rp_test <- predict(rp_t, newdata=test,  type="class")

> ggplot(test,aes(x=x,y=y)) + geom_point(aes(color=rp_test), size=0.5 )
```

```
# 分類木


# predict で予測, type でclass を指定
```

```
> library(nnet)
> t1 <- t0 %>% mutate(cl_t = as.factor(cl_t) )
> nn_t <- nnet(data=t1, cl_t ~ x+y, size=5)
> nn_test <- predict(nn_t, newdata=test, type="class")

> ggplot(test,aes(x=x,y=y)) +geom_point(aes(color=nn_test), size=0.5 )
```

```
# nnet ではクラス分類をするには因子に変更

# predict で予測, type でclass を指定
```

判定を図る４つの指標

真の値 ＼ 判定結果	陽 性	陰 性
陽 性	真陽性 (True Positive) TP	偽陰性 (False Negative) FN
陰 性	偽陽性 (False Positive) FP	真陰性 (True Negative) TN

正解率（*Accuracy*）$= \dfrac{TP+TN}{TP+FP+FN+TN}$

適合率（*Precision*）$= \dfrac{TP}{TP+FP}$

再現率（*Recall*）$= \dfrac{TP}{TP+FN}$

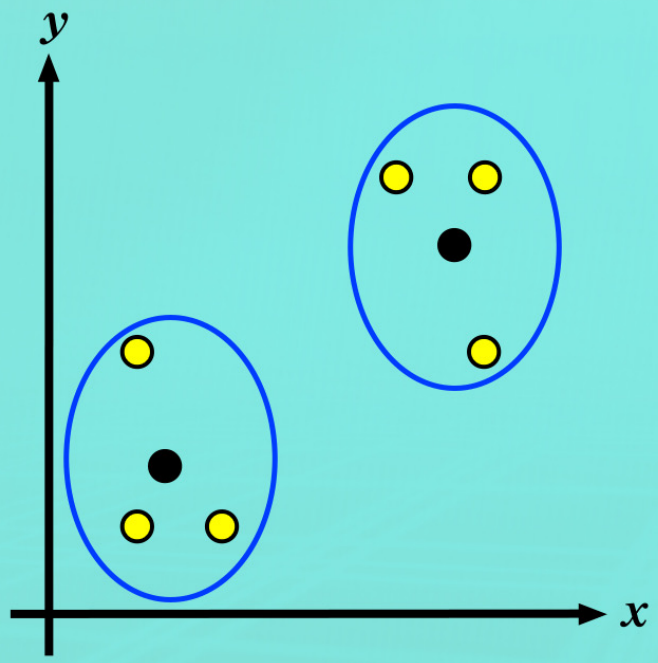
F値（*F measure*）$= \dfrac{2}{\dfrac{1}{Precision}+\dfrac{1}{Recall}}$

$= \dfrac{2\,(Precision \cdot Recall)}{Precision + Recall}$

データの分析と知識発見

k 平均法

6個の点を2つのグループに分ける

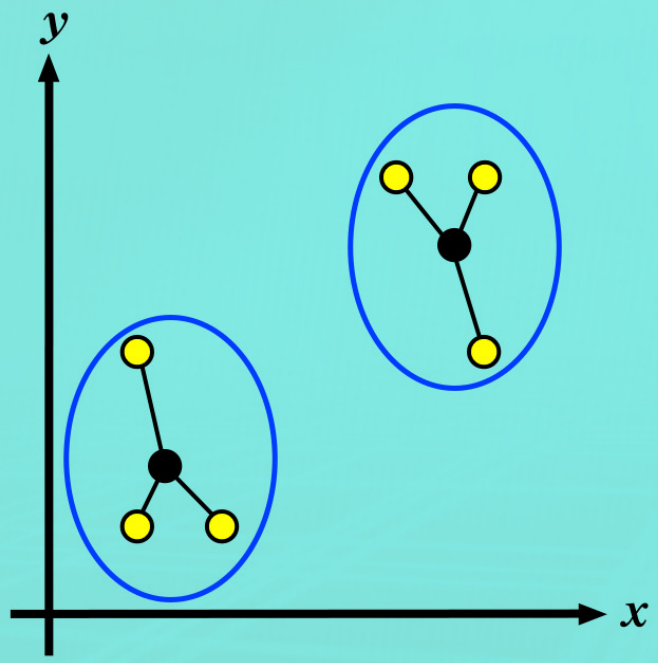


- ◆ 代表となる点を選び直す
- ◆ 距離を計算する
- ◆ 近い方のグループの点とする

という作業を繰り返す

k 平均法

6個の点を2つのグループに分ける



- ◆ 代表となる点を選び直す
- ◆ 距離を計算する
- ◆ 近い方のグループの点とする

という作業を繰り返す