

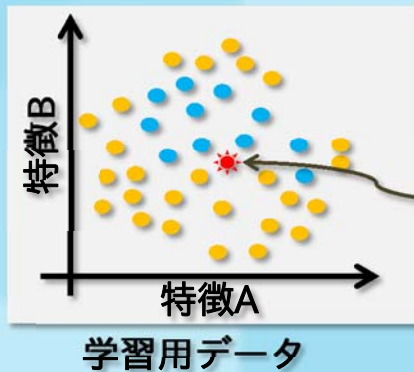
13回 データの分類

- k 近傍法による分類
- k 近傍法による数値の予測
- データの標準化
- k 近傍法の応用

k 近傍法 による分類


k 近傍法 (k Nearest Neighbor Algorithm; KNN)

学習用のデータに含まれる情報を用いて
問い合わせデータがどのようなカテゴリに属するのを予測



- ◆ データの特徴
- ◆ データが属するカテゴリ 

問い合わせデータ

問い合わせデータ  は、
● と ● のどちらのカテゴリか？

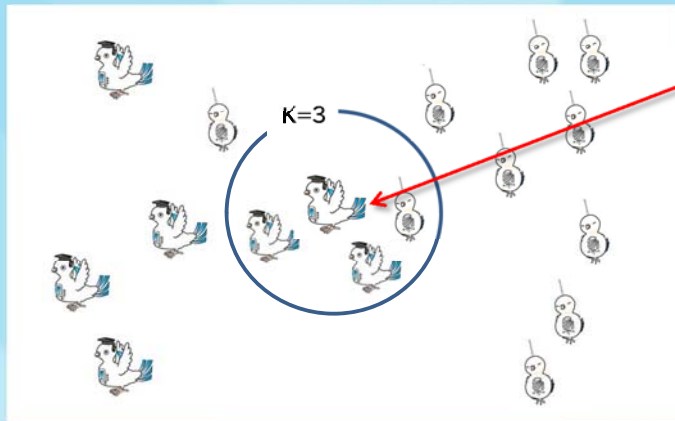
k 近傍法 (k Nearest Neighbor Algorithm)

● k 近傍法のアルゴリズム

1. パラメーター k の値を決定する
2. 問い合わせデータと学習用データとの類似度を計算する
3. 計算したデータの類似度にもとづいてデータを並べ替える
4. 決定した k の個数だけ、類似度に基づいて
問い合わせデータに類似するデータを選択し
選択されたデータのカテゴリの多数決投票を行う
5. 最多となるデータのカテゴリが、
問い合わせデータの推測されるカテゴリとなる

k 近傍法 (カテゴリ予測の事例 1-1)

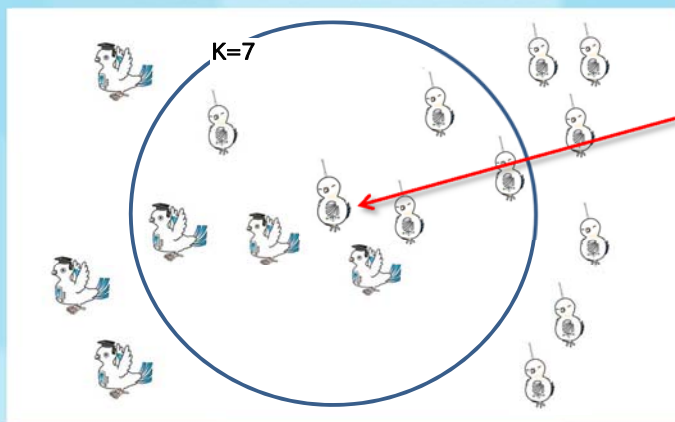
- k 近傍法で鳥の種類を予測する



1. $k = 3$ とする
2. 類似度 (距離) を計算
3. 距離で並替えて選択
4. 多数決投票
5. 最多となる種類を選ぶ

k 近傍法 (カテゴリ予測の事例 1-2)

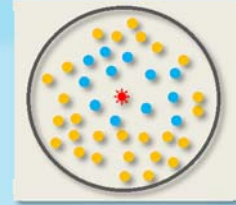
- k 近傍法で鳥の種類を予測する



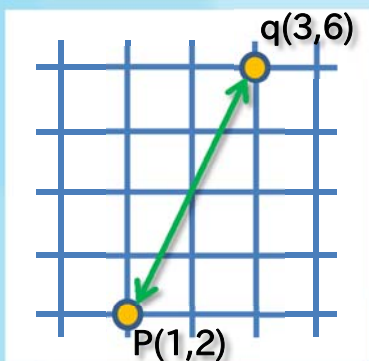
1. $k = 7$ とする
2. 類似度 (距離) を計算
3. 距離で並替えて選択
4. 多数決投票
5. 最多となる種類を選ぶ

k 近傍法 (k の値の選び方)

- k が小さすぎる
⇒ × ノイズがあるデータに弱い
- k が大きすぎる
⇒ × 他のクラスのデータを含む
- k を奇数にする
⇒ ◎ 多数決投票で同点をなくす



距離 (ユークリッド距離、L2)



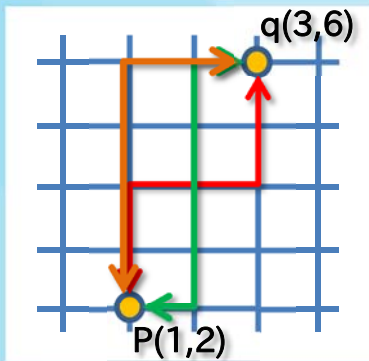
$$d(p, q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- $i = 2$ で2次元平面の点間の距離

計算例：

$$\begin{aligned} d(p, q) &= \sqrt{(1-3)^2 + (2-6)^2} \\ &= \sqrt{4 + 16} = \sqrt{20} \\ &\approx 4.47 \end{aligned}$$

距離 (マンハッタン距離、市街地距離、L1)



$$d(p, q) = \sum_i |p_i - q_i|$$

- $i = 2$ で2次元平面の点間の距離
- マンハッタン距離では、同じ距離となる複数の通り道が存在する

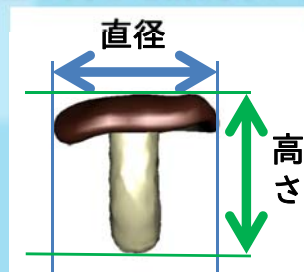
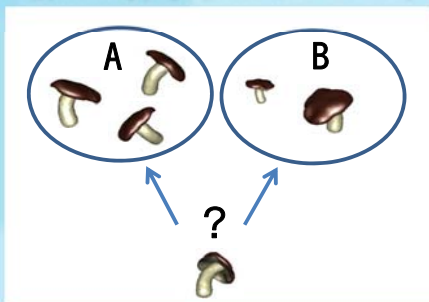
計算例：

$$\begin{aligned} d(p, q) &= |1 - 3| + |2 - 6| \\ &= |-2| + |-4| \\ &= 2 + 4 \\ &= 6 \end{aligned}$$

k 近傍法 (カテゴリ予測の事例 2-1)

k 近傍法でキノコの種類を予測する

- 2種類のキノコに関するデータがある
- データにはキノコの傘の直径、高さ、種類 (AかB)が記録されている
- 新たに採取したキノコの傘の直径と高さを計測し、種類を推測する



k 近傍法 (カテゴリ予測の事例 2-2)

k 近傍法でキノコの種類を予測する

#	キノコ	高さ	直径	種類
d1		8	9	A
d2		8	4	A
d3		4	4	B
d4		2	4	B
d5		7	7	A

5つのキノコのデータベース

#	キノコ	高さ	直径	種類
Q		4	8	?







新しいキノコのデータ

種類Aか
種類Bか
不明







k 近傍法 (カテゴリ予測の事例 2-3)

#	キノコ	高さ	直径	種類
d1		8	9	A
d2		8	4	A
d3		4	4	B
d4		2	4	B
d5		7	7	A
Q		4	8	

k 近傍法 (カテゴリ予測の事例 2-4)







#	キノコ	高さ	直径	種類	データQとの距離
d1		8	9	A	$\sqrt{(8-4)^2+(9-8)^2}=\sqrt{17}\approx 4.12$
d2		8	4	A	$\sqrt{(8-4)^2+(4-8)^2}=\sqrt{32}\approx 5.66$
d3		4	4	B	$\sqrt{(4-4)^2+(4-8)^2}=\sqrt{16}=4.00$
d4		2	4	B	$\sqrt{(2-4)^2+(4-8)^2}=\sqrt{20}\approx 4.47$
d5		7	7	A	$\sqrt{(7-4)^2+(7-8)^2}=\sqrt{10}\approx 3.16$
Q		4	8		

k 近傍法 (カテゴリ予測の事例 2-5)

#	キノコ	高さ	直径	種類	データQとの距離	順位
d1		8	9	A	$\sqrt{(8-4)^2+(9-8)^2}=\sqrt{17}\approx 4.12$	3
d2		8	4	A	$\sqrt{(8-4)^2+(4-8)^2}=\sqrt{32}\approx 5.66$	5
d3		4	4	B	$\sqrt{(4-4)^2+(4-8)^2}=\sqrt{16}=4.00$	2
d4		2	4	B	$\sqrt{(2-4)^2+(4-8)^2}=\sqrt{20}\approx 4.47$	4
d5		7	7	A	$\sqrt{(7-4)^2+(7-8)^2}=\sqrt{10}\approx 3.16$	1
Q		4	8			

昇順
「しょうじゅん」
(Ascending
Order)

k 近傍法 (カテゴリ予測の事例 2-6)

#	キノコ	高さ	直径	種類	データQとの距離	順位	K=3
d1		8	9	A	$\sqrt{(8-4)^2+(9-8)^2}=\sqrt{17}\approx 4.12$	3	○
d2		8	4	A	$\sqrt{(8-4)^2+(4-8)^2}=\sqrt{32}\approx 5.66$	5	×
d3		4	4	B	$\sqrt{(4-4)^2+(4-8)^2}=\sqrt{16}=4.00$	2	○
d4		2	4	B	$\sqrt{(2-4)^2+(4-8)^2}=\sqrt{20}\approx 4.47$	4	×
d5		7	7	A	$\sqrt{(7-4)^2+(7-8)^2}=\sqrt{10}\approx 3.16$	1	○
Q		4	8	A	← Aが2個、Bが1個。多数決でA種		

k 近傍法による 数値の予測

k 近傍法 (数値予測の事例 1-1)

k 近傍法でキノコの傘の直径数値を予測する

#	キノコ	高さ	直径
d1		8	9
d2		8	4
d3		4	4
d4		2	4
d5		7	7

5つのキノコのデータベース

#	キノコ	高さ	直径
Q		5	?

新しいキノコのデータ






キノコの傘の直径の大きさが不明

k 近傍法 (数値予測の事例 1-2)

#	キノコ	高さ	直径	距離 (高さ)
d1		8	9	$\sqrt{(8-5)^2} = \sqrt{9} = 3$
d2		8	4	$\sqrt{(8-5)^2} = \sqrt{9} = 3$
d3		4	4	$\sqrt{(4-5)^2} = \sqrt{1} = 1$
d4		2	4	$\sqrt{(2-5)^2} = \sqrt{9} = 3$
d5		7	7	$\sqrt{(7-5)^2} = \sqrt{4} = 2$




Q		5	?
---	---	---	---

k 近傍法 (数値予測の事例 1 - 3)

#	キノコ	高さ	直径	距離 (高さ)	順位
d1		8	9	$\sqrt{(8-5)^2} = \sqrt{9} = 3$	3
d2		8	4	$\sqrt{(8-5)^2} = \sqrt{9} = 3$	3
d3		4	4	$\sqrt{(4-5)^2} = \sqrt{1} = 1$	1
d4		2	4	$\sqrt{(2-5)^2} = \sqrt{9} = 3$	3
d5		7	7	$\sqrt{(7-5)^2} = \sqrt{4} = 2$	2

Q		5	?
---	---	---	---

k 近傍法 (数値予測の事例 1 - 4)

#	キノコ	高さ	直径	距離 (高さ)	順位	K=2
d1		8	9	$\sqrt{(8-5)^2} = \sqrt{9} = 3$	3	×
d2		8	4	$\sqrt{(8-5)^2} = \sqrt{9} = 3$	3	×
d3		4	4	$\sqrt{(4-5)^2} = \sqrt{1} = 1$	1	○
d4		2	4	$\sqrt{(2-5)^2} = \sqrt{9} = 3$	3	×
d5		7	7	$\sqrt{(7-5)^2} = \sqrt{4} = 2$	2	○

Q		5	?
---	---	---	---

k 近傍法 (数値予測の事例 1 - 5)

#	キノコ	高さ	直径	距離 (高さ)	順位	K=2	直径
d1		8	9	$\sqrt{(8-5)^2} = \sqrt{9} = 3$	3	×	
d2		8	4	$\sqrt{(8-5)^2} = \sqrt{9} = 3$	3	×	
d3		4	4	$\sqrt{(4-5)^2} = \sqrt{1} = 1$	1	○	4
d4		2	4	$\sqrt{(2-5)^2} = \sqrt{9} = 3$	3	×	
d5		7	7	$\sqrt{(7-5)^2} = \sqrt{4} = 2$	2	○	7

Q		5	5.5	← $(4+7) \div 2 = 5.5$
---	---	---	------------	------------------------

データの標準化

数値の標準化 (例 1 - 1)

ID	重さ (グラム)	高さ (cm)	金額 (円)
1	100	9	20
2	200	8	30
3	270	4	50
4	150	2	40
5	300	5	90
6	180	8	10

キノコのデータ (重さ、高さ、金額)

データに異なる単位や尺度が
使用されている



データの標準化が必要

数値の標準化 (例 1 - 2)

ID	重さ (グラム)	高さ (cm)	金額 (円)
1	100	9	20
2	200	8	30
3	270	4	50
4	150	2	40
5	300	5	90
6	180	8	10

最大	300	9	90
最小	100	2	10

標準化の式 :

$$x_{std} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

↖
↖
↖

↖
↖
↖

↖
↖
↖

計算例 :

$$x_{std} = \frac{50 - 10}{90 - 10} = \frac{40}{80} = 0.5$$

数値の標準化 (例 1-3)

ID	重さ (グラム)	高さ (cm)	金額 (円)
1	100	9	20
2	200	8	30
3	270	4	50
4	150	2	40
5	300	5	90
6	180	8	10

最大	300	9	90
最小	100	2	10

$$x_{std} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad \text{標準化後(0.00~1.00)}$$

ID	重さ	高さ	金額
1	$\frac{100-100}{300-100} = 0.00$	$\frac{9-2}{9-2} = 1.00$	$\frac{20-10}{90-10} = 0.13$
2	$\frac{200-100}{300-100} = 0.50$	$\frac{8-2}{9-2} = 0.86$	$\frac{30-10}{90-10} = 0.25$
3	$\frac{270-100}{300-100} = 0.85$	$\frac{4-2}{9-2} = 0.29$	$\frac{50-10}{90-10} = 0.50$
4	$\frac{150-100}{300-100} = 0.25$	$\frac{2-2}{9-2} = 0.00$	$\frac{40-10}{90-10} = 0.38$
5	$\frac{300-100}{300-100} = 1.00$	$\frac{5-2}{9-2} = 0.43$	$\frac{90-10}{90-10} = 1.00$
6	$\frac{180-100}{300-100} = 0.40$	$\frac{8-2}{9-2} = 0.86$	$\frac{10-10}{90-10} = 0.00$

数値の標準化 (例 1-4)

ID	重さ (グラム)
1	100
2	200
3	270
4	150
5	300
6	180

最大	300
最小	100

ID	重さ
1	$\frac{100-100}{300-100} = 0.00$
2	$\frac{200-100}{300-100} = 0.50$
3	$\frac{270-100}{300-100} = 0.85$
4	$\frac{150-100}{300-100} = 0.25$
5	$\frac{300-100}{300-100} = 1.00$
6	$\frac{180-100}{300-100} = 0.40$

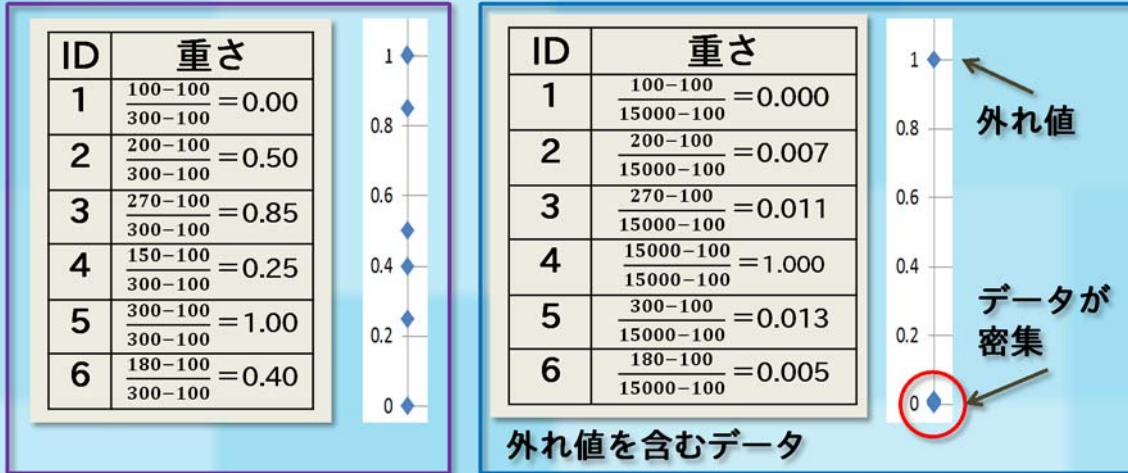
ID	重さ (グラム)
1	100
2	200
3	270
4	15000
5	300
6	180

最大	15000
最小	100

ID	重さ
1	$\frac{100-100}{15000-100} = 0.000$
2	$\frac{200-100}{15000-100} = 0.007$
3	$\frac{270-100}{15000-100} = 0.011$
4	$\frac{15000-100}{15000-100} = 1.000$
5	$\frac{300-100}{15000-100} = 0.013$
6	$\frac{180-100}{15000-100} = 0.005$

外れ値

数値の標準化 (例1-5)



数値の標準化 (例2-1)

ID	重さ (グラム)	高さ (cm)	金額 (円)
1	100	9	20
2	200	8	30
3	270	4	50
4	150	2	40
5	300	5	90
6	180	8	10

	重さ	高さ	金額
合計	1200	36	240
平均	200	6	40
標準偏差	68.07	2.52	25.82

標準化の式: 平均

$$x_{norm} = \frac{x - x_{avg}}{\sigma}$$

標準偏差

偏差: 各データと平均値の差

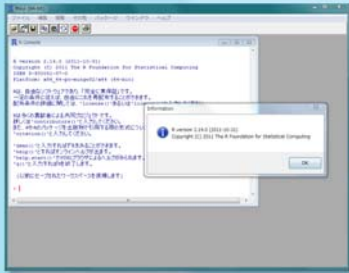
分散: 偏差の2乗の平均値

標準偏差: 分散の平方根

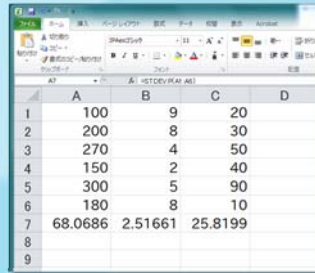
※標本分散 (Sample Variance) で計算

数値の標準化 (例2-2)

- 標準偏差の計算には、統計ソフトウェア、表計算ソフトウェア、関数電卓などの利用が便利



統計ソフトウェア



表計算ソフトウェア



関数電卓

数値の標準化 (例2-3)

	重さ	高さ	金額
合計	1200	36	240
平均	200	6	40
標準偏差	68.07	2.52	25.82

$$x_{norm} = \frac{x - x_{avg}}{\sigma}$$

↙ 平均
↘ 標準偏差

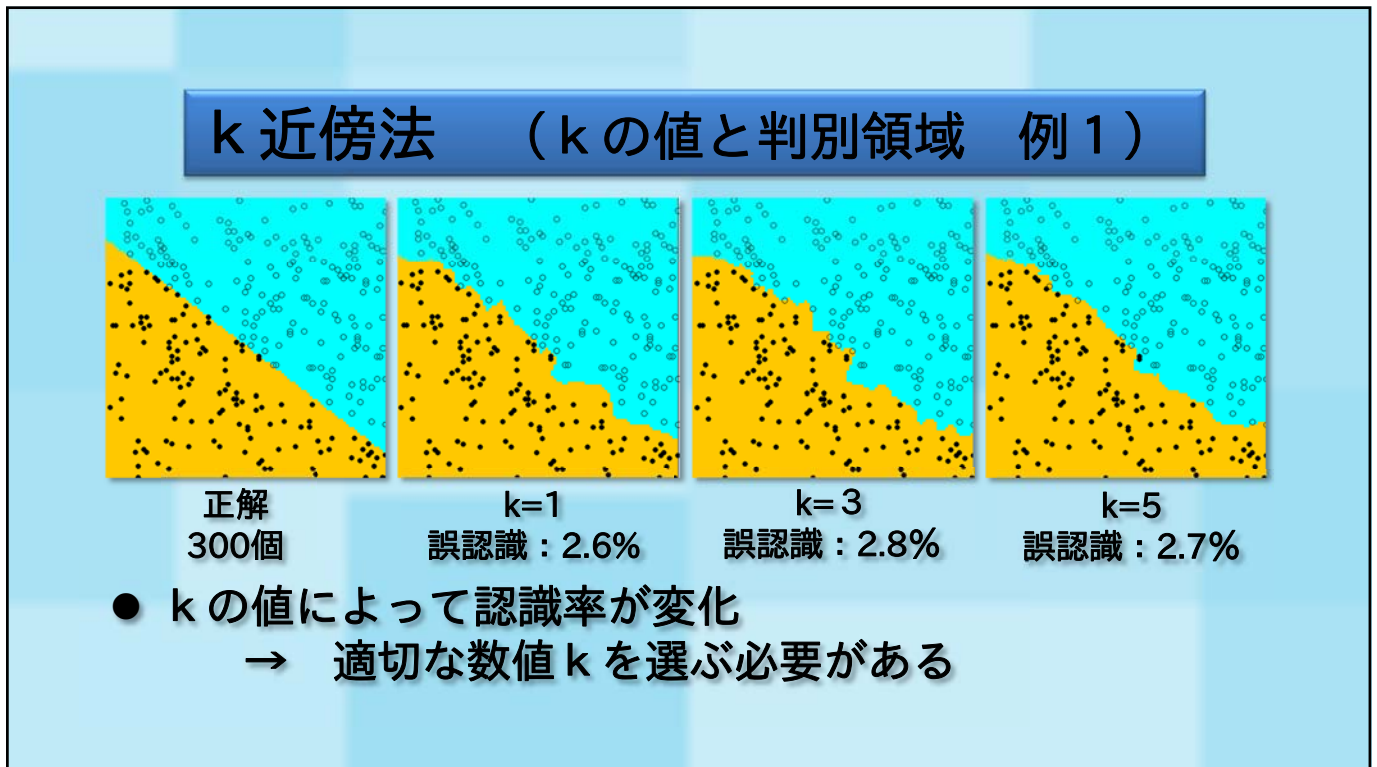
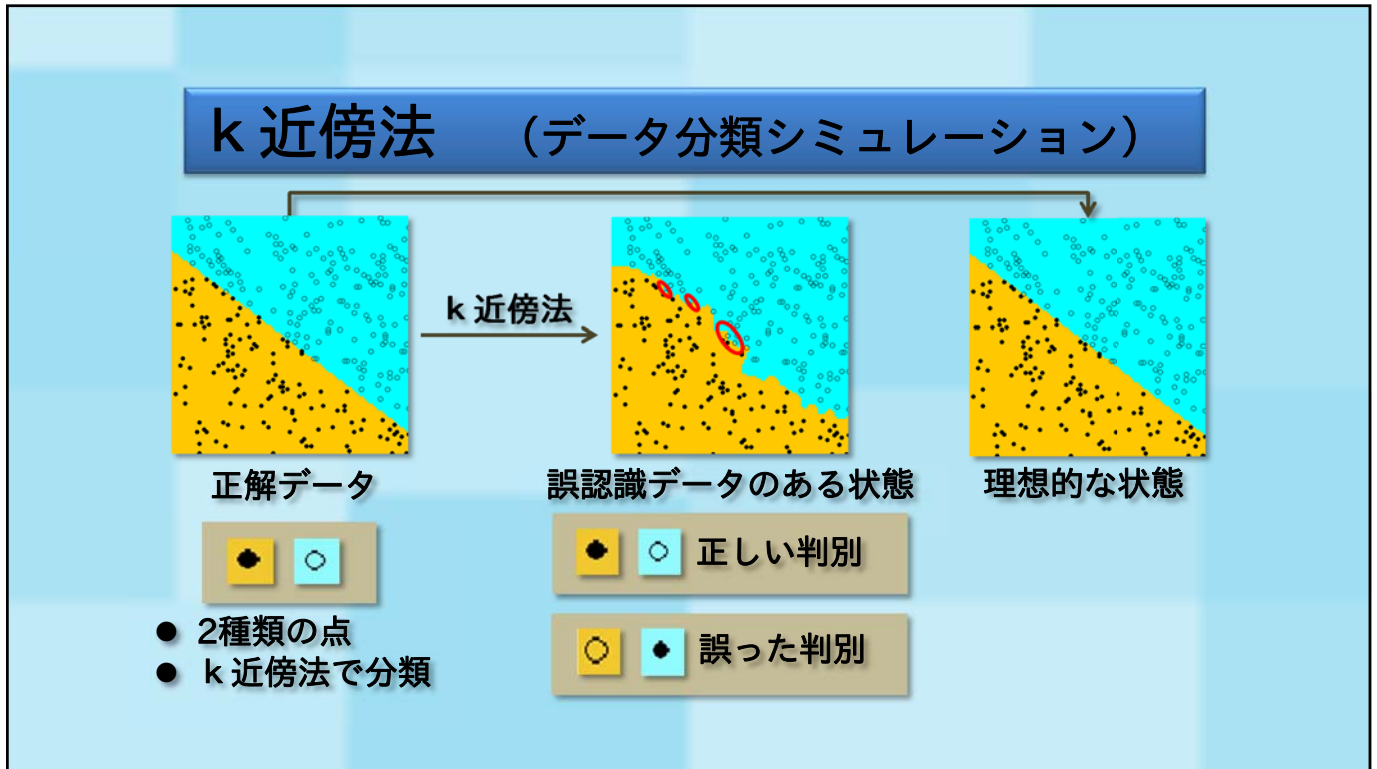
変数を平均0、分散1に標準化

ID	重さ	高さ	金額
1	(100-200)/68.07= -1.47	(9-6)/2.52= +1.19	(20-40)/25.82= -0.77
2	(200-200)/68.07= 0.00	(8-6)/2.52= +0.79	(30-40)/25.82= -0.39
3	(270-200)/68.07= +1.03	(4-6)/2.52= -0.79	(50-40)/25.82= +0.39
4	(150-200)/68.07= -0.73	(2-6)/2.52= -1.59	(40-40)/25.82= 0.00
5	(300-200)/68.07= +1.47	(5-6)/2.52= -0.40	(90-40)/25.82= +1.94
6	(180-200)/68.07= -0.29	(8-6)/2.52= +0.79	(10-40)/25.82= -1.16

k 近傍法の応用

k 近傍法の特徴

- 適切な数値 k を選ぶ必要がある
 - ✓ k の値を設定 → クロスバリデーション等を利用
- 適切な距離尺度を選択する必要がある
- 学習に使うデータを選定する必要がある
- 大きなデータベースに対しては計算量が多くなる
 - ✓ 検索データと学習データ全てに対して距離計算をするため
 - ✓ 高速化 → kd木 (k Dimensional Tree)
→ Locality Sensitive Hashing (LSH)

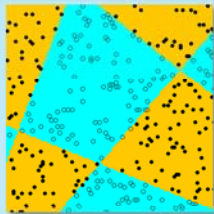


k 近傍法 (k の値と判別領域 例 2)

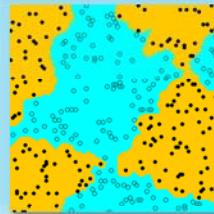


正解領域

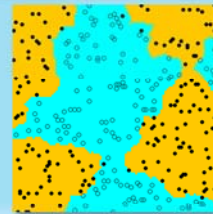
判別する領域が
やや複雑な例



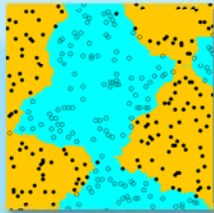
正解 (300個)



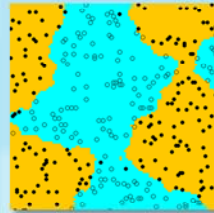
k=1, 誤認識 : 7.0%



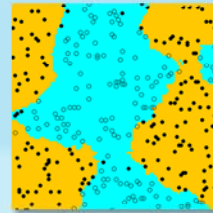
k=3, 誤認識 : 8.1%



k=5, 誤認識 : 7.6%

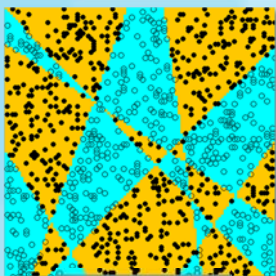


k=7, 誤認識 : 7.8%

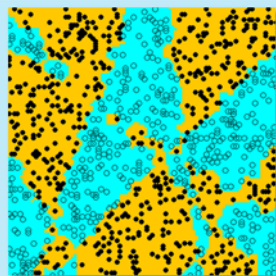


k=11, 誤認識 : 8.9%

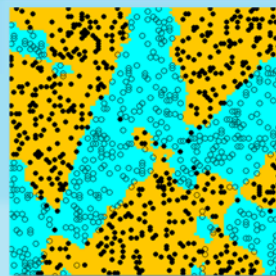
k 近傍法 (k の値と判別領域 例 3)



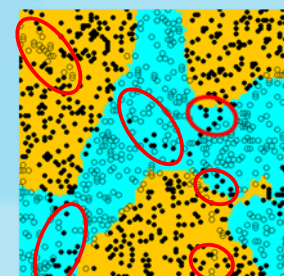
正解
1000個



k=1
誤認識 : 6.9%



k=5
誤認識 : 7.6%

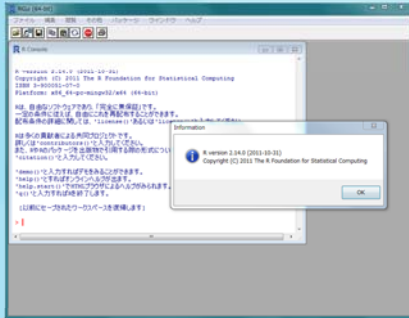


k=51
誤認識 : 15.6%

- k が大きすぎると
他のクラスのデータを含んでしまう

非常に大きな k の値

k 近傍法 (統計ソフトウェア 1-1)



- k 近傍法の計算は統計ソフトウェア等を使うことができる

例：R <http://www.r-project.org/>

オープンソースでフリーソフトウェアの統計解析向けプログラミング言語及びその開発実行環境

k 近傍法 (統計ソフトウェア 1-2)

k 近傍法でキノコの種類を予測する

#	キノコ	高さ	直径	種類
d1		8	9	A
d2		8	4	A
d3		4	4	B
d4		2	4	B
d5		7	7	A

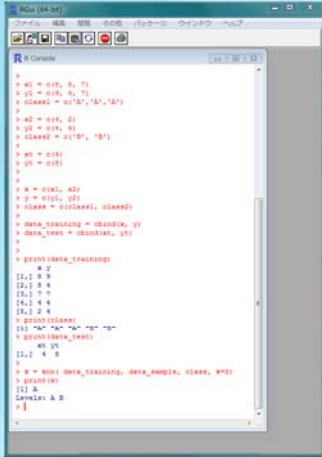
5つのキノコのデータベース

#	キノコ	高さ	直径	種類
Q		4	8	?

新しいキノコのデータ

種類Aか
種類Bか
不明

k 近傍法 (統計ソフトウェア 1-3)



Rの画面

R言語のプログラム

```
library(class)
x1 = c(8, 8, 7)
y1 = c(9, 4, 7)
class1 = c('A','A','A')
x2 = c(4, 2)
y2 = c(4, 4)
class2 = c('B', 'B')
xt = c(4)
yt = c(8)
x = c(x1, x2)
y = c(y1, y2)
class = c(class1, class2)
data_training = cbind(x, y)
data_test = cbind(xt, yt)
print(data_training)
print(class)
print(data_test)
k = knn( data_training, data_sample, class, k=3)
print(k)
```

k 近傍法 (統計ソフトウェア 1-4)

>library(class)

>x1 = c(8, 8, 7)

>y1 = c(9, 4, 7)

>class1 = c('A','A','A')

>x2 = c(4, 2)

>y2 = c(4, 4)

>class2 = c('B', 'B')

...

#	キノコ	高さ	直径	種類
d1		8	9	A
d2		8	4	A
d3		4	4	B
d4		2	4	B
d5		7	7	A

5つのキノコのデータベース

k 近傍法 (統計ソフトウェア 1-5)

```
...
>xt = c(4)
>yt = c(8)
...
```

#	キノコ	高さ	直径	種類
Q		4	8	?

新しいキノコのデータ
(種類を問い合わせるキノコのデータ)

k 近傍法 (統計ソフトウェア 1-6)

```
...
>x = c(x1, x2)
>y = c(y1, y2)
>class = c(class1, class2)
>data_training = cbind(x, y)
>data_test = cbind(xt, yt)

>print(data_training)
>print(class)
>print(data_test)
...
```

学習キノコのデータ結合 (A種とB種)

学習キノコのデータを横に並べて結合

問い合わせキノコQのデータを
横に並べて結合

データを表示して確認

k 近傍法 (統計ソフトウェア 1-7)

```
> print(data_training)
  x y
[1,] 8 9
[2,] 8 4
[3,] 7 7
[4,] 4 4
[5,] 2 4

> print(class)
[1] "A" "A" "A" "B" "B"

> print(data_test)
  xt yt
[1,] 4 8
```

学習用データ (高さ、直径) を表示

学習用データ (A種、B種) を表示

問い合わせ用キノコQのデータを表示

k 近傍法 (統計ソフトウェア 1-8)

K 近傍法 (K-Nearest Neighbor)

結果

学習データ

問い合わせデータQ

学習データ (キノコ種類)

Kの値

```
> k = knn( data_training, data_test, class, k=3 )
```

K 近傍法

```
> print(k)
[1] A
Levels: A B
```

結果
キノコは
A種と推測



まとめ (第13回 データの分類)

- **k 近傍法による分類**
 - Kの値、距離、類似度、多数決投票
- **k 近傍法による数値の予測**
 - 多数決投票と平均による数値予測
- **データの標準化**
 - 数値の標準化、標準偏差、ソフトウェアの利用
- **k 近傍法の応用**
 - k 近傍法のシミュレーション、統計ソフトウェアの利用