数理統計 補助資料
～判別分析～

2023年度2学期：　月曜1限, 水曜3限
担当教員：　石垣　司

判別分析

新しく観測されたデータがどの群（グループやクラス）に属するのかを判別するための手法

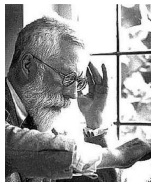
- Fisher の線形判別分析(Fisher 1936)
- 所属する群が既知のデータから群を区別する線形関数を構成
- 応用例：疾病の有無，製品の不良発見，優良顧客の判別など

多変量データ

ID	変数1	変数2	…	変数 P	群
1	x_{11}	x_{12}	…	x_{1P}	A
2	x_{21}	x_{21}	…	x_{2P}	B
⋮	⋮	⋮	⋱	⋮	
i	x_{i1}	x_{i2}	…	x_{iP}	A
⋮	⋮	⋮	⋱	⋮	
N	x_{N1}	x_{N2}	…	x_{NP}	?

線形判別関数 → ID番号 N のデータは群Aに所属

群が未知のデータ



Ronald Fisher 1890-1962

ビジネスにおける応用例

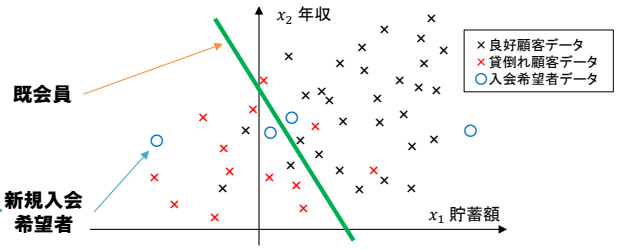
クレジットカードの入会審査

- 過去のクレジットカード会員の属性データと結果を用いて判別関数（線形関数：直線や平面）を構成
- 新規入会希望者の将来の結果を予測（判別）

過去のデータ

顧客 No.	貯金額	年収	群（良好or貸倒れ）
1	1200	600	良好
2	-100	250	貸倒れ
3	110	300	良好
4	300	400	良好
5	0	700	貸倒れ
6	770	250	良好
⋮	⋮	⋮	⋮
999	50	900	良好
1000	0	400	良好
1001	-200	1000	?
1002	1000	300	?

1. 過去のデータから判別関数を構成
2. 所属群が未知のデータの群を予測



判別関数により入会希望者の中から貸倒れ顧客を予測

Fisherの線形判別分析の基本事項

線形判別

- 判別関数は線形空間（直線や平面）

データの次元が P の時、判別関数の次元は $P-1$

- 例：2変数データの判別空間は、1次元の判別線
- 例：3変数データの判別空間は、2次元の判別面

2群の判別

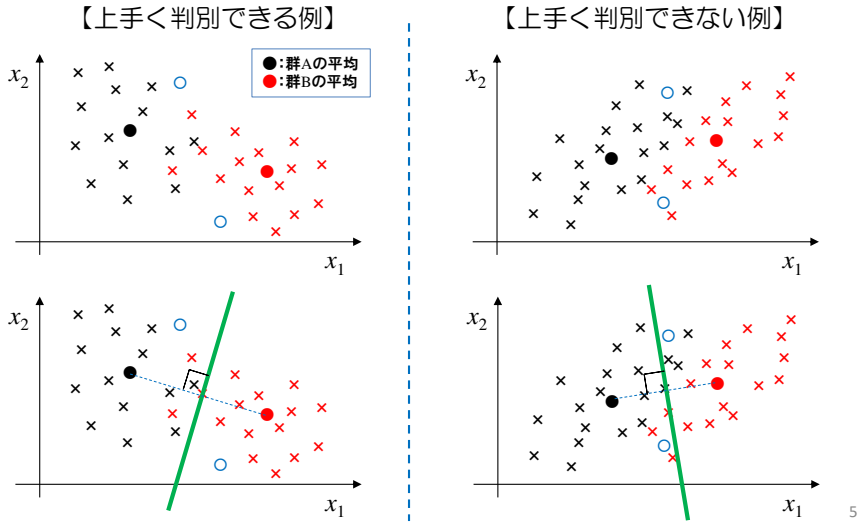
- 多群の判別は2群判別の拡張

教師あり学習

- 所属する群の正解（ラベル）が付与されたデータを用いて判別関数を構成

質問：簡単じゃないですか？

2群間の平均のど真ん中を判別関数



Fisherの線形判別分析

次の2つの尺度のバランスが良い射影軸を構成

- ① 両群の分離度（平均間の距離）を大きくする
  群間分散を大きく ⇒ 2群間のデータを判別しやすくする
- ② 射影後に両群の分布の交わりを小さくする
  群内分散を小さく ⇒ 2群間のデータの交わり部分を小さくする

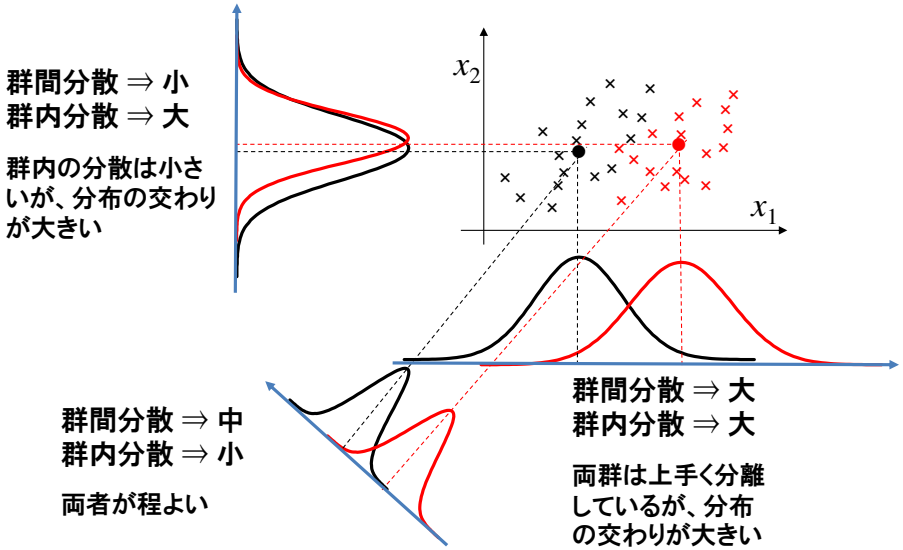
群間分散と群内分散の比 r の最大化

- ラグランジュ未定乗数法と固有値問題へ帰結

$$r = \frac{\text{群間分散}}{\text{群内分散}} \rightarrow \text{最大化}$$

群間分散と群内分散

例：3つの線形軸（→）へのデータの射影



線形判別関数の求め方 #1

Notation

- 群Aの i 番目データ：$\boldsymbol{x}_{Ai} = [x_{Ai1}, \ldots, x_{AiP}]^T \ (i = 1, \ldots N_A)$
- 群Bの i 番目データ：$\boldsymbol{x}_{Bi} = [x_{Bi1}, \ldots, x_{BiP}]^T \ (i = 1, \ldots N_B)$
  N_A: 群Aのデータ数, N_B: 群Bのデータ数, P: データの次元
- 射影関数の重み：$\boldsymbol{w} = [w_1, \ldots, w_P]^T$
- 群Aのデータ $\boldsymbol{x}_{Ai}$ を y 軸へ射影した値：$y_{Ai} = \boldsymbol{w}^T \boldsymbol{x}_{Ai}$
- 群Bのデータ $\boldsymbol{x}_{Bi}$ を y 軸へ射影した値：$y_{Bi} = \boldsymbol{w}^T \boldsymbol{x}_{Bi}$
- 群Aのデータの標本平均ベクトル：$\bar{\boldsymbol{x}}_A = [\bar{x}_{A1}, \ldots, \bar{x}_{AP}]^T$
- 群Bのデータの標本平均ベクトル：$\bar{\boldsymbol{x}}_B = [\bar{x}_{B1}, \ldots, \bar{x}_{BP}]^T$

線形判別関数の求め方 #2

群間分散

- y_{Ai} と y_{Bi} の平均

$$\bar{y}_A = \frac{1}{N_A}\sum_{i=1}^{N_A} y_{Ai} = \sum_{k=1}^{P} w_k \bar{x}_{Ak} = \boldsymbol{w}\bar{\boldsymbol{x}}_A,$$

$$\bar{y}_B = \frac{1}{N_B}\sum_{i=1}^{N_B} y_{Bi} = \sum_{k=1}^{P} w_k \bar{x}_{Bk} = \boldsymbol{w}\bar{\boldsymbol{x}}_B.$$

#メモ：導出計算は前回の講義と同様

- 群間分散の定義

$$(\bar{y}_A - \bar{y}_B)^2 = \{\boldsymbol{w}^T(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)\}^2$$

線形判別関数の求め方 #3

群内分散

- 群Aと群Bのデータの y 軸上での標本分散

S_A と S_B はそれぞれの群の標本分散共分散行列

$$\frac{1}{N_A - 1}\sum_{i=1}^{N_A}(y_{Ai} - \bar{y}_A)^2 = \boldsymbol{w}^T S_A \boldsymbol{w}$$

$$\frac{1}{N_B - 1}\sum_{i=1}^{N_B}(y_{Bi} - \bar{y}_B)^2 = \boldsymbol{w}^T S_B \boldsymbol{w}$$

#メモ：導出計算は前回の講義と同様

- 群内分散の定義

群Aと群Bのy軸上での標本分散の重み付け和

$$\frac{1}{N_A + N_B - 2}\{(N_A - 1)\boldsymbol{w}^T S_A \boldsymbol{w} + (N_B - 1)\boldsymbol{w}^T S_B \boldsymbol{w}\} = \boldsymbol{w}^T S \boldsymbol{w}$$

ただし、$S = \frac{1}{N_A + N_B - 2}\{(N_A - 1)S_A + (N_B - 1)S_B\}$

線形判別関数の求め方 #4

群間分散と群内分散の比 r の最大化

$$r = \frac{\{\boldsymbol{w}^T(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)\}^2}{\boldsymbol{w}^T S \boldsymbol{w}}$$

- 目的関数 r, 制約条件 $\boldsymbol{w}^T S \boldsymbol{w} = 1$
- 制約条件付き最適化問題
  ⇒ ラグランジュ未定乗数法と固有値問題

- 最適解 check!

$$\boldsymbol{w}^* = S^{-1}(\bar{\boldsymbol{x}}_A - \bar{\boldsymbol{x}}_B)$$

#メモ：マハラノビス距離を用いる定式化も有名。その解は $\boldsymbol{w}^*$ に一致する

線形判別関数の求め方 #5

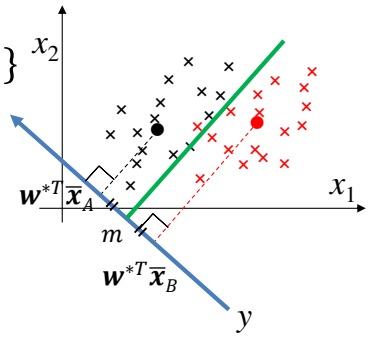
Fisherの線形判別関数

- 群Aと群Bのデータの y 軸上での平均の中点で定義

y 軸上での群Aの平均: $\boldsymbol{w}^{*T}\bar{\boldsymbol{x}}_A$

y 軸上での群Bの平均: $\boldsymbol{w}^{*T}\bar{\boldsymbol{x}}_B$

- 中点 $m = \frac{1}{2}\{\boldsymbol{w}^{*T}\bar{\boldsymbol{x}}_A + \boldsymbol{w}^{*T}\bar{\boldsymbol{x}}_B\}$



- 判別関数 $f(\boldsymbol{x})$

$$f(\boldsymbol{x}) = \boldsymbol{w}^{*T}\boldsymbol{x} - m$$

$f(\boldsymbol{x}) \geq 0$ならば未知の $\boldsymbol{x}$ を群Aに所属
$f(\boldsymbol{x}) < 0$ならば未知の $\boldsymbol{x}$ を群Bに所属
判別線ー：$f(\boldsymbol{x}) = 0$

様々な判別率の指標

正答率(Accuracy)
- **テストデータを正しく判別できた割合**

$$\frac{TP+TN}{TP+FN+FP+FN}$$

適合率(precision)
- **陽性と判別したデータの内, 実際に陽性である割合**

$$\frac{TP}{TP+FP}$$

再現率(recall)
- **実際に陽性であるデータの内, 正しく判別できた割合**

$$\frac{TP}{TP+FN}$$

F-**尺度**(F-measure)
- **精度と再現率のバランス(調和平均)**

#メモ: 適合率と再現率はトレードオフ

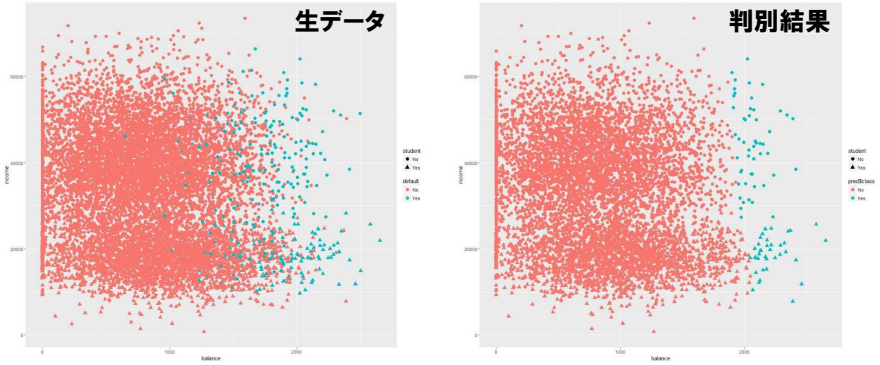
例: 癌の検査や工場の不良品検出に適した指標は?

例: 迷惑メールの検出に適した指標は?

		判別結果	
		群1(例:陽性)	群2(例:陰性)
正答	群1	True Positive (TP)	False Negative (FN)
	群2	False Positive (FP)	True Negative (TN)

Rによる判別分析の実行 #1

クレジットカードの貸倒れデータ
- Default Data (ISLR**パッケージ内の**10,000**人分の**simulated data)
  - **被説明変数:** default (Yes or No)
  - **説明変数:** student, balance, income



G. James, et al. "*An Introduction to Statistical Learning with Application in R*" Springer (2013)

Rによる判別分析の実行 #2

クレジットカードの貸倒れデータ

```
library(MASS)
library(ISLR)
library(ggplot2)
data(Default)
summary(Default)
TrainD = Default[1:3000,] #訓練データ3000人分
TestD = Default[3001:10000,] #テストデータ7000人分
lda = lda(default~student+balance+income,data=TrainD)
lda
pred = predict(lda,TestD)
pYes = pred$class=="Yes"
tYes = TestD[,"default"]=="Yes"
#正答率(accuracy)
mean(pred$class==TestD[,"default"])
#適合率(precision)
sum(pYes & tYes)/sum(pYes)
#再現率(recall)、Hit rate
sum(pYes & tYes)/sum(tYes)library(MASS)
```

```
> summary(Default)
 default    student       balance           income
 No :9667   No :7056   Min.   :   0.0   Min.   :  772
 Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340
                       Median : 823.6   Median :34553
                       Mean   : 835.4   Mean   :33517
                       3rd Qu.:1166.3   3rd Qu.:43808
                       Max.   :2654.3   Max.   :73554
```

```
> #正答率(accuracy)
> mean(pred$class==TestD[,"default"])
[1] 0.9731429
> #精度、適合率(precision)
> sum(pYes & tYes)/sum(pYes)
[1] 0.75
> #再現率(recall)、Hit rate
> sum(pYes & tYes)/sum(tYes)
[1] 0.2844828
```


G. James, et al. "*An Introduction to Statistical Learning with Application in R*" Springer (2013)

演習問題

1. $\bar{x}_A = [1\ \ 1]^T, \bar{x}_B = [0\ \ 0]^T, S = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ **のとき,**
   新しく観測されたデータ $x = [0.6\ 0.4]^T$ **は群Aと群Bのどちらに判別されるか答えなさい**

2. **次の事例においてシステムの判別性能を評価するために適切な指標は何か答えなさい**
   - **癌検診の性能評価に適した指標**
     ヒント: 癌を発症している受診者の見逃しを避けたい
   - **迷惑メールフィルタの性能評価に適した指標**
     ヒント: 通常メールを迷惑メールとする誤分類は避けたい