数理統計 補助資料
～線形判別分析～

2024年度2学期:　月曜1限, 水曜3限
担当教員:　石垣　司

Fisher の線形判別分析(Fisher 1936)

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

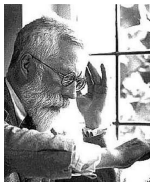
- 所属する群が既知のデータから群を区別する線形関数を構成
- 機械学習のデータ分類手法の元祖
- 応用例: 疾病の有無, 製品の不良発見, 優良顧客の判別など

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

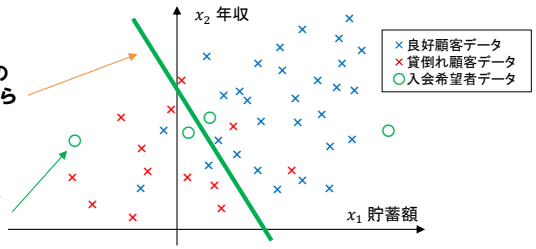
過去のデータ

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

既会員のデータから構成

新規入会希望者



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)による分類

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

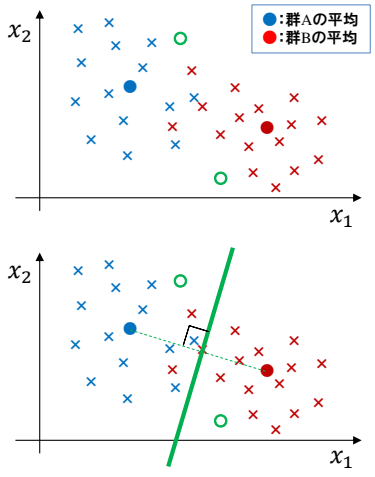
教師あり学習

- 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

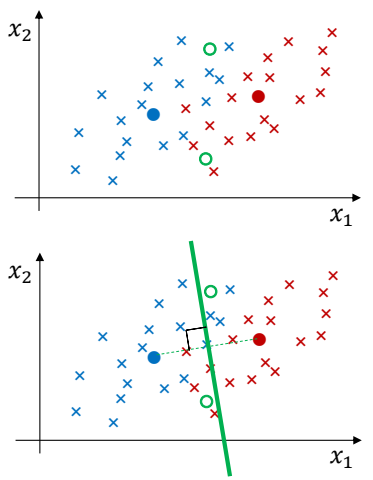
簡単じゃないですか？

2群間の平均のど真ん中を判別関数

【上手く判別できる例】



【上手く判別できない例】



Fisherの線形判別分析

群間分散と郡内分散のバランスが良い射影軸を構成

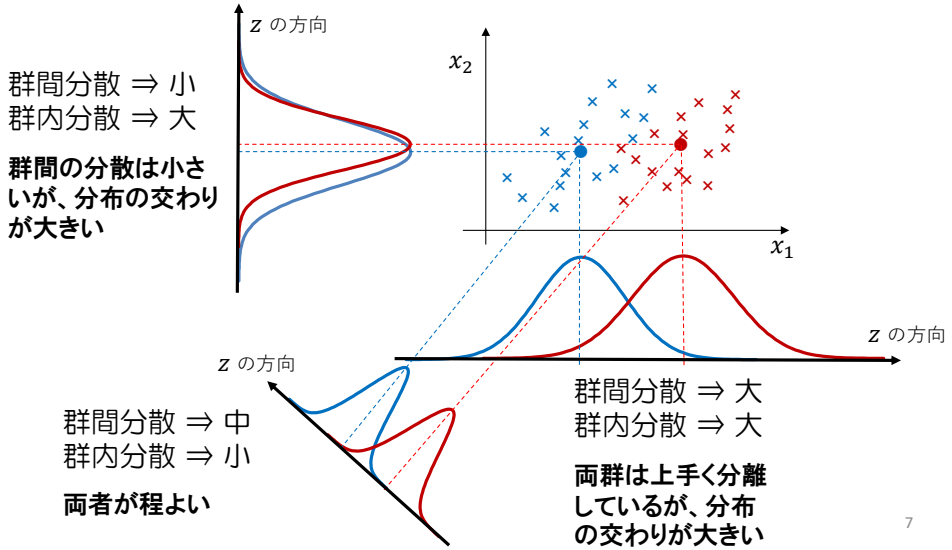
- **郡間分散：両群の分離度**
  群間分散を大きく ⇒ 2群間のデータを判別しやすくする
- **群内分散：射影後の各群のデータ分布の分散**
  群内分散を小さく ⇒ 2群間のデータの交わり部分を小さくする

群間分散と群内分散の比 r の最大化

- **目的関数 $r = \dfrac{\text{群間分散}}{\text{群内分散}}$ の制約付き最大化問題を考える**
- **ラグランジュ未定乗数法と固有値問題へ帰結**

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 z 軸へ射影した値：$z_{Ai} = \boldsymbol{w}^T \boldsymbol{x}_{Ai}$**
- **群Bのデータ $\boldsymbol{x}_{Bi}$ を z 軸へ射影した値：$z_{Bi} = \boldsymbol{w}^T \boldsymbol{x}_{Bi}$**

- **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

- z_{Ai} と z_{Bi} の平均

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

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

#メモ：導出計算は前回の講義と同様

- 群間分散の定義

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

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 z 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

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

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

#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 z 軸上での標本分散の重み付け和

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

ここで, $S = \frac{1}{N_A + N_B - 2}\{(N_A - 1)S_A + (N_B - 1)S_B\}$

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：

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

check!

#メモ：マハラノビス距離を用いる定式化も有名。その解は w^* に一致する

線形判別関数の求め方 #5

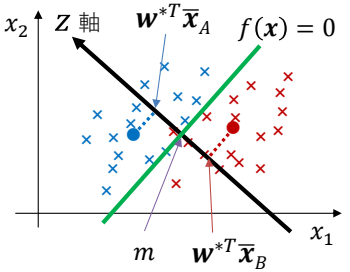
Fisher の線形判別関数

- 両群データの平均の z 軸上への正射影の中点で定義

z 軸上での群Aの平均：$w^{*T}\bar{x}_A$

z 軸上での群Bの平均：$w^{*T}\bar{x}_B$

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



- 判別関数 $f(\boldsymbol{x})$

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

- 判別線ー：$f(\boldsymbol{x}) = 0$

$f(\boldsymbol{x}) \geq 0$ ならば未知の $\boldsymbol{x}$ を群Aに分類

$f(\boldsymbol{x}) < 0$ ならば未知の $\boldsymbol{x}$ を群Bに分類

分類の良さの評価 #1

分類の間違い方(誤分類)の種類

これ以降は線形判別分析のみに限らないより一般的な分類の話

		分類結果	
		群A(陽性)	群B(陰性)
正答	群A（陽性）	True Positive(TP)	False Negative(FN)
	群B（陰性）	False Positive(FP)	True Negative(TN)

– **もちろん**, FP**と**FN**の割合が小さい分類が望ましい**

① 正答率(精度, Accuracy): $\frac{TP+TN}{TP+FP+FN+TN}$

– **未知データを正しく分類できた割合**

全て陰性と分類した意味のない分類結果

		分類結果	
		陽性	陰性
正答	陽性	45	5
	陰性	5	45

Accuracy = 0.9

		分類結果	
		陽性	陰性
正答	陽性	25	25
	陰性	25	25

Accuracy = 0.5

		分類結果	
		陽性	陰性
正答	陽性	0	5
	陰性	0	95

Accuracy = 0.95

分類の良さの評価 #2

正答率は自然で直感的な分類性能の評価指標であるが, 正答率による評価が適さない場合もあるので注意

– **例: 健康診断や定期がん検診の評価**

多くの受診者の正答は陰性(H28年肺がん発見率 0.03%**)**

日本医師会HP(最終閲覧2024,12,16)
https://www.med.or.jp/forest/gankenshin/data/detection/

		分類結果	
		陽性	陰性
正答	陽性	0	3
	陰性	0	9997

テキトーな検査で全員を「問題なし」と分類したとき, 正答率(Accuracy) = 0.9997

– **例: 迷惑メールフィルタの分類性能の評価**

陽性: 迷惑メール, 陰性: 仕事のメール

		分類結果	
		陽性	陰性
正答	陽性	45	0
	陰性	10	45

迷惑メールを仕事メールと分類

仕事メールを迷惑メールと分類

正答率(Accuracy) = 0.9 だが, 仕事のメールを10%も取り逃している

分類の良さの評価 #3

		分類結果	
		群A(陽性)	群B(陰性)
正答	群A	True Positive(TP)	False Negative(FN)
	群B	False Positive(FP)	True Negative(TN)

② 適合率(Precision): $\frac{TP}{TP+FP}$

– **陽性と分類したデータが本当に陽性である割合**

		分類結果	
		陽性	陰性
正答	陽性	45	5
	陰性	5	45

Precision = 0.9

		分類結果	
		陽性	陰性
正答	陽性	25	25
	陰性	25	25

Precision = 0.5

		分類結果	
		陽性	陰性
正答	陽性	0	5
	陰性	0	95

Precision 算出不可

③ 再現率(網羅率, Recall, Hit rate): $\frac{TP}{TP+FN}$

– **正答が陽性であるデータが正しく陽性に分類されている割合**

		分類結果	
		陽性	陰性
正答	陽性	45	5
	陰性	5	45

Recall = 0.9

		分類結果	
		陽性	陰性
正答	陽性	25	25
	陰性	25	25

Recall = 0.5

		分類結果	
		陽性	陰性
正答	陽性	0	5
	陰性	0	95

Recall = 0

分類の良さの評価 #4

不均衡データ(imbalanced data)に注意

– **不均衡データ: 一方の群のサンプルサイズが極端に小さいデータセット**

– **健康診断の検査, クレジットカードの貸倒れ, 不良品の検査など, 実応用の多くの場面で現れる**

– **妥当ではない分類でも各指標が良好な値を示す場合あり**

妥当な分類結果

		分類結果	
		陽性	陰性
正答	陽性	45	5
	陰性	5	45

Accuracy = 0.9
Precision = 0.9
Recall = 0.9

		分類結果	
		陽性	陰性
正答	陽性	25	25
	陰性	25	25

Accuracy = 0.5
Precision = 0.5
Recall = 0.5

妥当ではない分類結果

		分類結果	
		陽性	陰性
正答	陽性	95	0
	陰性	5	0

Accuracy = 0.95
Precision = 0.95
Recall = 1.0

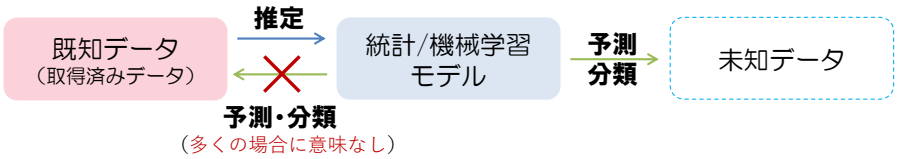
		分類結果	
		陽性	陰性
正答	陽性	0	5
	陰性	0	95

Accuracy = 0.95
Precision = 算出不可
Recall = 0

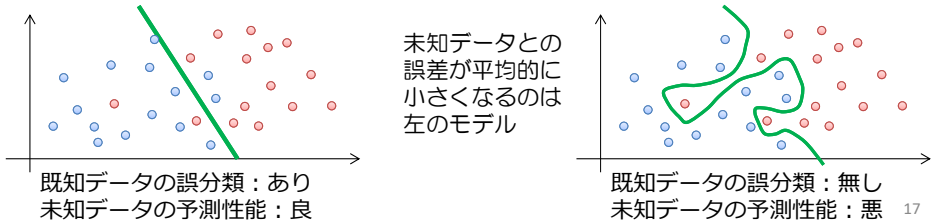
未知データの分類性能の評価 #1

汎化性能(復習)：未知データをうまく分類できる性能

- **実問題への応用の多くで必要なのは未知データの予測**



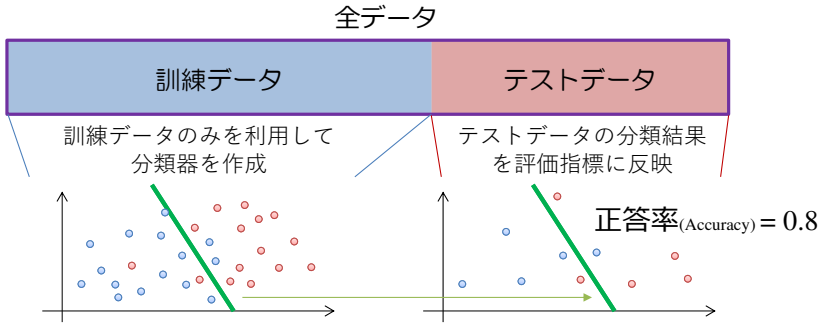
- **既知データへの過剰適合(過学習)は未知データの分類性能を低下させる**



未知データの分類性能の評価 #2

ホールドアウトサンプルを用いた分類性能評価

- **ホールドアウトサンプル：分類器の作成に用いないテストデータ**
- **訓練データ(学習データ)とテストデータを分けて分類性能を検証**



- **分類器は訓練に利用したデータの正答を知っているので，訓練データを正確に分類できるのは当たりまえ**

未知データの分類性能の評価 #3

再掲：交差検証法(Cross-validation)

- **例：** 3-fold **クロスバリデーション**　　**全観測データ** 1 2 3
- **データを3分割し，その3つの分類性能の平均で性能を評価**

評価： この 分類モデル の分類性能は正答率 0.86

Leave one out **法**

- **1つのサンプルのみをテストデータ，それ以外のデータを訓練データとして，交差検証を行う**
  - **比較的サンプルサイズが少数の時に利用**

線形判別分析の実行 #1

クレジットカードの貸倒れデータ

- ISLR**パッケージ内の**10,000**人分の疑似データ**
- **分類したい変数：** Default (Yes or No)
- **分類に利用する変数：** Student, Balance, Income

	Default	Student	Balance	Income
Yes	333	2944		
No	9667	7056		
最小値			0	772
最大値			2654	73554
平均値			835	33517

不均衡データ

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

線形判別分析の実行 #2

10-fold **交差検証法による分類性能評価：**

Accuracy = 0.97, Precision = 0.80, Recall = 0.24

分類の結果と解釈

- 分類の正答率は高い
  （ただし，不均衡データ）
- 陽性と分類された人の80%は実際に貸倒れを起こした人
- 実際に貸倒れを起こした人の24%を正しく分類

貸倒れを起こす人の76%を見逃してしまうが，陽性と判断された人に対しては80%の確率で正しく結果を予測できる手法である

複数の検証内でこのような判別線

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. **次の分類性能を評価するための適切な指標を正答率**(Accuracy), **適合率**(Precision), **再現率**(Recall)**の中から選び，その理由も答えなさい**
   - **健康診断や定期がん検診の評価**
     ヒント：がんや病気が発生している受診者の見逃しは許されない
   - **迷惑メールフィルタの分類性能の評価**
     ヒント：迷惑メールを仕事メールとする誤分類は許せるが，仕事メールを迷惑メールとする誤分類は許されない