

判別分析

分析の評価

村田 昇

講義の内容

- 第1日：判別分析の考え方
- 第2日：分析の評価

判別分析の復習

判別分析

- 個体の特徴量からその個体の属するクラスを予測する関係式を構成
- 事前確率： $\pi_k = P(Y = k)$ (prior probability)
 - $X = \mathbf{x}$ が与えられる前に予測されるクラス
- 事後確率： $p_k(\mathbf{x})$ (posterior probability)
 - $X = \mathbf{x}$ が与えられた後に予測されるクラス

$$p_k(\mathbf{x}) = P(Y = k | X = \mathbf{x})$$

- 所属する確率が最も高いクラスに個体を分類

判別関数

- 判別の手続き
 - 説明変数 $X = \mathbf{x}$ の取得
 - 事後確率 $p_k(\mathbf{x})$ の計算
 - 事後確率最大のクラスにデータを分類
- 判別関数： $\delta_k(\mathbf{x})$ ($k = 1, \dots, K$)

$$p_k(\mathbf{x}) < p_l(\mathbf{x}) \Leftrightarrow \delta_k(\mathbf{x}) < \delta_l(\mathbf{x})$$

事後確率の順序を保存する計算しやすい関数

- 判別関数 $\delta_k(\mathbf{x})$ を最大化するようなクラス k に分類

線形判別

- $f_k(\mathbf{x})$ の仮定
 - q 変量正規分布の密度関数
 - 平均ベクトル $\boldsymbol{\mu}_k$: クラスごとに異なる
 - 共分散行列 Σ : **すべてのクラスで共通**

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{q/2} \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

- 線形判別関数: \mathbf{x} の 1 次式

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k + \log \pi_k$$

2 次判別

- $f_k(\mathbf{x})$ の仮定
 - q 変量正規分布の密度関数
 - 平均ベクトル $\boldsymbol{\mu}_k$: クラスごとに異なる
 - 共分散行列 Σ_k : **クラスごとに異なる**

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{q/2} \sqrt{\det \Sigma_k}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

- 2 次判別関数: \mathbf{x} の 2 次式

$$\delta_k(\mathbf{x}) = -\frac{1}{2} \det \Sigma_k - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \log \pi_k$$

Fisher の線形判別

- 新しい特徴量 $Z = \boldsymbol{\alpha}^\top X$ を考える
- 良い Z の基準
 - クラス内では集まっているほど良い ($\boldsymbol{\alpha}^\top W \boldsymbol{\alpha}$ は小)
 - クラス間では離れているほど良い ($\boldsymbol{\alpha}^\top B \boldsymbol{\alpha}$ は大)
- Fisher の基準

$$\text{maximize } \boldsymbol{\alpha}^\top B \boldsymbol{\alpha} \quad \text{s.t. } \boldsymbol{\alpha}^\top W \boldsymbol{\alpha} = \text{const.}$$

- $\boldsymbol{\alpha}$ は $W^{-1}B$ の第 1 から第 $K-1$ 固有ベクトル
- 判別方法: 特徴量の距離を用いる
- $d_k = \sum_{l=1}^{K-1} (\alpha_l^\top \mathbf{x} - \alpha_l^\top \boldsymbol{\mu}_k)^2$ が最小のとなるクラス k に判別

2 値判別分析の評価

誤り率

- 単純な誤り

$$(\text{誤り率}) = \frac{(\text{誤って判別されたデータ数})}{(\text{全データ数})}$$

- 判別したいラベル: 陽性 (positive)
 - 真陽性: 正しく陽性と判定 (true positive; TP)
 - 偽陽性: 誤って陽性と判定 (false positive; FP) (第 I 種過誤)
 - 偽陰性: 誤って陰性と判定 (false negative; FN) (第 II 種過誤)
 - 真陰性: 正しく陰性と判定 (true negative; TN)

混同行列

	真値は陽性	真値は陰性
判別は陽性	真陽性 (True Positive)	偽陽性 (False Positive)
判別は陰性	偽陰性 (False Negative)	真陰性 (True Negative)

- **confusion matrix**
- 各条件にあてはまるデータ数を記載
- 転置で書く流儀もあるので注意 (次頁)

混同行列 (転置したもの)

	判別は陽性	判別は陰性
真値は陽性	真陽性 (True Positive)	偽陰性 (False Negative)
真値は陰性	偽陽性 (False Positive)	真陰性 (True Negative)

- パターン認識や機械学習で多く見られた書き方
- 誤差行列 (error matrix) とも呼ばれる

基本的な評価基準

- 定義

$$(\text{真陽性率}) = \frac{TP}{TP + FN} \quad (\text{true positive rate})$$

$$(\text{真陰性率}) = \frac{TN}{FP + TN} \quad (\text{true negative rate})$$

$$(\text{適合率}) = \frac{TP}{TP + FP} \quad (\text{precision})$$

$$(\text{正答率}) = \frac{TP + TN}{TP + FP + TN + FN} \quad (\text{accuracy})$$

- 別名 (分野で異なるので注意)
 - 感度 (sensitivity) あるいは 再現率 (recall)

$$(\text{真陽性率}) = \frac{TP}{TP + FN}$$

- 特異度 (specificity)

$$(\text{真陰性率}) = \frac{TN}{FP + TN}$$

- 精度 (accuracy)

$$(\text{正答率}) = \frac{TP + TN}{TP + FP + TN + FN}$$

F-値

- 定義 (F-measure, F-score)

$$F_1 = \frac{2}{1/(\text{再現率}) + 1/(\text{適合率})}$$

$$F_\beta = \frac{\beta^2 + 1}{\beta^2/(\text{再現率}) + 1/(\text{適合率})}$$

- 再現率 (真陽性率) と適合率の (重み付き) 調和平均

Cohen の kappa 値

- 定義 (Cohen's kappa measure)

$$p_o = \frac{TP + TN}{TP + FP + TN + FN} \quad (\text{accuracy})$$

$$p_e = \frac{TP + FP}{TP + FP + TN + FN} \cdot \frac{TP + FN}{TP + FP + TN + FN}$$

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

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

- 観測された精度と偶然の精度の比較

受信者動作特性曲線

- ROC 曲線 (receiver operating characteristic curve)
- 2 値判別における判別関数を用いた判定方法の一般形

$$H(x; c) = \begin{cases} \text{陽性}, & \delta(x) > c \\ \text{陰性}, & \text{それ以外} \end{cases}$$

- 真陽性率と偽陽性率

$$TPR(c) = P(\text{陽性を正しく陽性と判別})$$

$$FPR(c) = P(\text{陰性を誤って陽性と判別})$$

$$= 1 - P(\text{陰性を正しく陰性と判別})$$

- **ROC 曲線** : $H(x; c)$ の c を自由に動かし x 軸に偽陽性率, y 軸に真陽性率を描画したもの
 - 定義から陽性率も偽陽性率もクラス事前分布によらない
 - 一般に ROC 曲線は $(0, 0)$ と $(1, 1)$ を結ぶ右肩上りの曲線
 - 曲線と x 軸で囲まれた面積が広い \Leftrightarrow 良い判別方法
- **AUC** : 上記の面積 (area under the ROC curve)
 - 2 値判別の難しさを測る基準の一つ

演習

問題

- 以下の問に答えなさい
 - F-値, 再現率, 適合率の大小関係はどのようになるか
 - 2 値判別 (陽性 = 1, 陰性 = 0 とする) において正解ラベル Y と予測ラベル \hat{Y} の相関係数を TP, FP, TN, FN およびデータ数 N を用いて表せ

解答例

- 最大最小と平均の関係から以下が成り立つ

$$\min(\text{再現率}, \text{適合率}) \leq F_1 \leq \max(\text{再現率}, \text{適合率})$$

さらに相加・相乗平均の関係から

$$F_1 \leq (\text{相乗平均}) \leq (\text{相加平均})$$

も成り立つ

- 相関係数の定義に従って計算すればよい

$$\rho = \frac{\text{Cov}(Y, \hat{Y})}{\sqrt{\text{Var}(Y)\text{Var}(\hat{Y})}}$$

- 例えば分子の共分散は以下のように計算される

$$\begin{aligned} \text{Cov}(Y, \hat{Y}) &= \mathbb{E}[(Y - \mathbb{E}[Y])(\hat{Y} - \mathbb{E}[\hat{Y}])] \\ &= \mathbb{E}[Y\hat{Y}] - \mathbb{E}[Y]\mathbb{E}[\hat{Y}] \\ &= \frac{TP}{N} - \frac{TP + FN}{N} \frac{TP + FP}{N} \\ &= \frac{TP(TP + FN + FP + TN)}{N^2} \\ &\quad - \frac{(TP + FN)(TP + FP)}{N^2} \\ &= \frac{TP \cdot TN - FP \cdot FN}{N^2} \end{aligned}$$

- 平均は標本平均で置き換えた

- 同様に分母の分散は以下ようになる

$$\begin{aligned}\text{Var}(Y) &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\ &= \frac{(TP + FN)(TN + FP)}{N^2} \\ \text{Var}(\hat{Y}) &= \mathbb{E}[\hat{Y}^2] - \mathbb{E}[\hat{Y}]^2 \\ &= \frac{(TP + FP)(TN + FN)}{N^2}\end{aligned}$$

- したがって以下のようにまとめられる

$$\rho = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- これは Matthews correlation coefficient (MCC) と呼ばれる評価指標の一つである

解析の事例

データについて

- 気象庁より取得した東京の気候データ
 - 気象庁 <https://www.data.jma.go.jp/gmd/risk/obsdl/index.php>
 - データ https://noboru-murata.github.io/multivariate-analysis/data/tokyo_weather.csv

気温と湿度による月の判別

- 温度と湿度による 8,9 月の線形判別
- 温度と湿度による 8,9 月の 2 次判別

混同行列の比較

さまざまな評価指標の比較

Table 1: 線形判別

指標	値
accuracy	0.721
kap	0.442
sens	0.742
spec	0.700
ppv	0.719
npv	0.724
mcc	0.442
j_index	0.442
bal_accuracy	0.721
detection_prevalence	0.525
precision	0.719
recall	0.742
f_meas	0.730

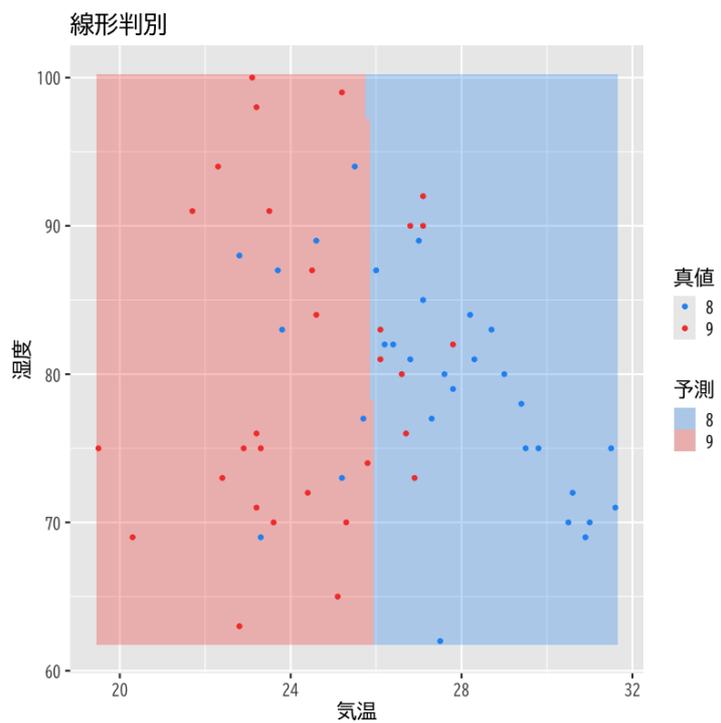


Figure 1: 線形判別

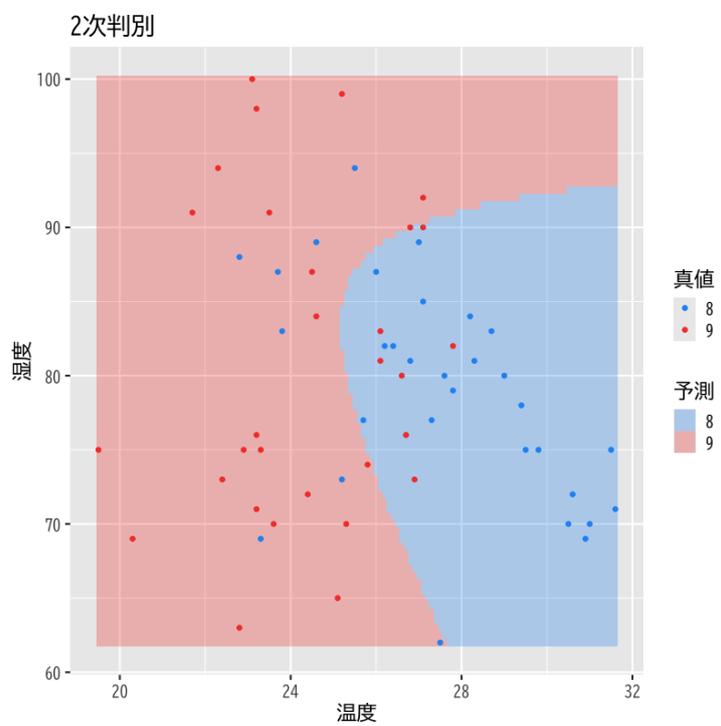


Figure 2: 2次判別

Linear Discriminant

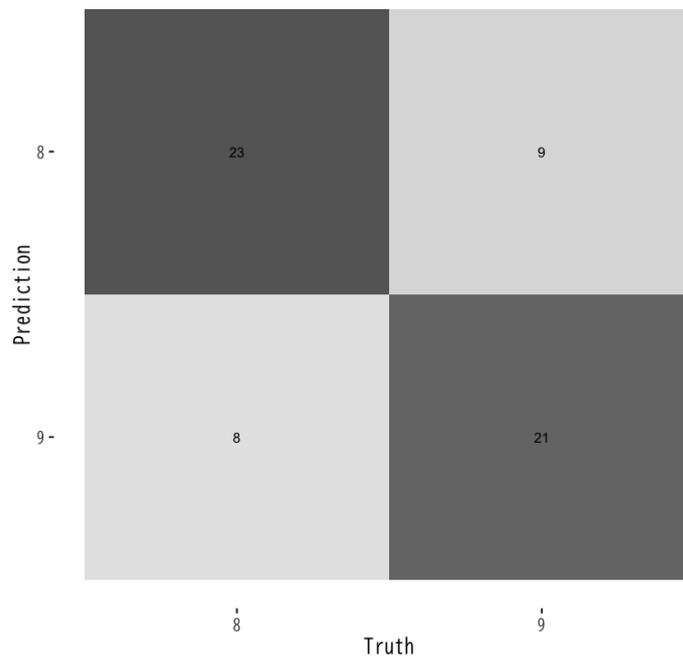


Figure 3: 線形判別の混同行列

Quadratic Discriminant

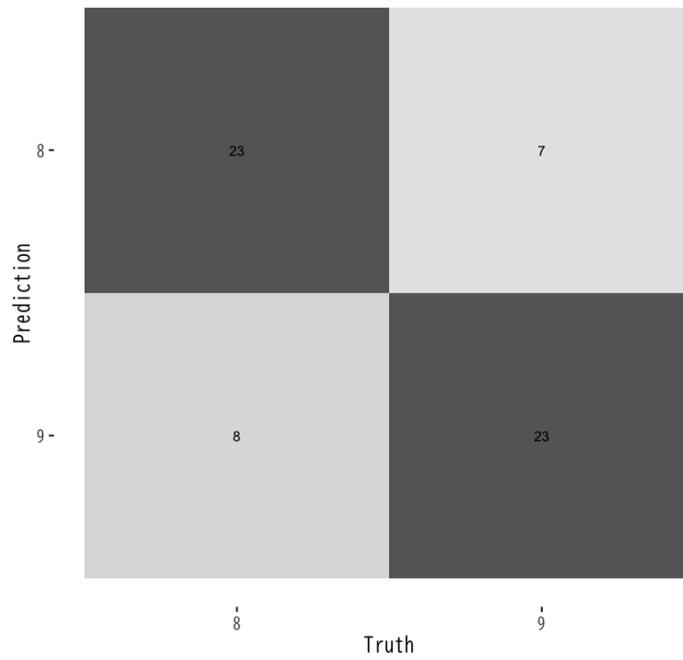


Figure 4: 2次判別の混同行列

Table 2: 2次判別

指標	値
accuracy	0.754
kap	0.508
sens	0.742
spec	0.767
ppv	0.767
npv	0.742
mcc	0.509
j_index	0.509
bal_accuracy	0.754
detection_prevalence	0.492
precision	0.767
recall	0.742
f_meas	0.754

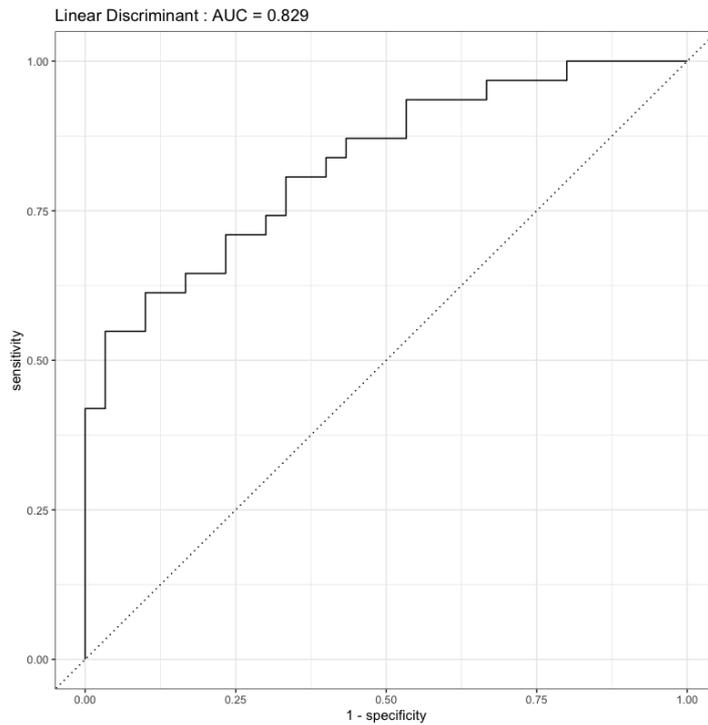


Figure 5: 線形判別の ROC 曲線

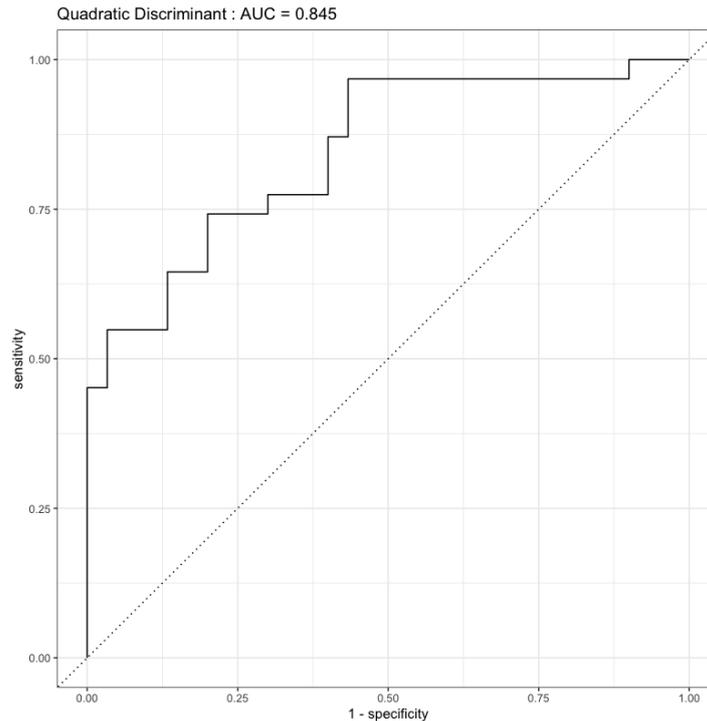


Figure 6: 2次判別の混同行列

ROC 曲線の比較

予測誤差

訓練誤差と予測誤差

- 訓練誤差：既知データに対する誤り (training error)
- 予測誤差：未知データに対する誤り (predictive error)
- 訓練誤差は予測誤差より良くなることが多い
- 既知データの判別に特化している可能性がある
 - 過適応 (over-fitting)
 - 過学習 (over-training)
- 予測誤差が小さい \Leftrightarrow 良い判別方法

交叉検証

- データを訓練データと試験データに分割して用いる
 - 訓練データ：判別関数を構成する (training data)
 - 試験データ：予測精度を評価する (test data)
- データの分割に依存して予測誤差の評価が偏る
- 偏りを避けるために複数回分割を行ない評価する
- “交差” と書く場合もある

交叉検証法

- **cross-validation (CV)**
- k -重交叉検証法 (k -fold cross-validation; k -fold CV)
 - n 個のデータを k ブロックにランダムに分割
 - 第 i ブロックを除いた $k-1$ ブロックで判別関数を推定
 - 除いておいた第 i ブロックで予測誤差を評価
 - $i = 1, \dots, k$ で繰り返し k 個の予測誤差で評価 (平均や分散)
- leave-one-out 法 (leave-one-out CV; LOO-CV)
 - $k = n$ として上記を実行

解析の事例

データについて

- UC Irvine Machine Learning Repository の公開データ
 - <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>
Wine Quality Data Set
P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.
* 以下では winequality-red.csv を利用
- データ概要
 - データ数 1599
 - 説明変数 (based on physicochemical tests)
 - 1 - fixed acidity
 - 2 - volatile acidity
 - 3 - citric acid
 - 4 - residual sugar
 - 5 - chlorides
 - 6 - free sulfur dioxide
 - 7 - total sulfur dioxide
 - 8 - density
 - 9 - pH
 - 10 - sulphates
 - 11 - alcohol
 - 目的変数 (based on sensory data)
 - 12 - quality (score between 0 and 10)
 - * ただし解析では A,B,C,D の 4 値に集計
- 実際のデータの一部

線形判別の訓練誤差と予測誤差

2次判別の訓練誤差と予測誤差

LOO 交叉検証による予測誤差の評価

次回の予定

- 第 1 日: クラスタ分析の考え方と階層的方法

Table 3

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	grade
7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5	C
7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.800	5	C
7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650	9.800	5	C
11.200	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580	9.800	6	B
7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5	C
7.400	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560	9.400	5	C
7.900	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460	9.400	5	C
7.300	0.650	0	1.200	0.065	15	21	0.995	3.390	0.470	10	7	A
7.800	0.580	0.020	2	0.073	9	18	0.997	3.360	0.570	9.500	7	A
7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	5	C
6.700	0.580	0.080	1.800	0.097	15	65	0.996	3.280	0.540	9.200	5	C
7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	5	C
5.600	0.615	0	1.600	0.089	16	59	0.994	3.580	0.520	9.900	5	C
7.800	0.610	0.290	1.600	0.114	9	29	0.997	3.260	1.560	9.100	5	C
8.900	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880	9.200	5	C
8.900	0.620	0.190	3.900	0.170	51	148	0.999	3.170	0.930	9.200	5	C
8.500	0.280	0.560	1.800	0.092	35	103	0.997	3.300	0.750	10.500	7	A
8.100	0.560	0.280	1.700	0.368	16	56	0.997	3.110	1.280	9.300	5	C
7.400	0.590	0.080	4.400	0.086	6	29	0.997	3.380	0.500	9	4	D
7.900	0.320	0.510	1.800	0.341	17	56	0.997	3.040	1.080	9.200	6	B

LDA (training data) : acc = 0.615

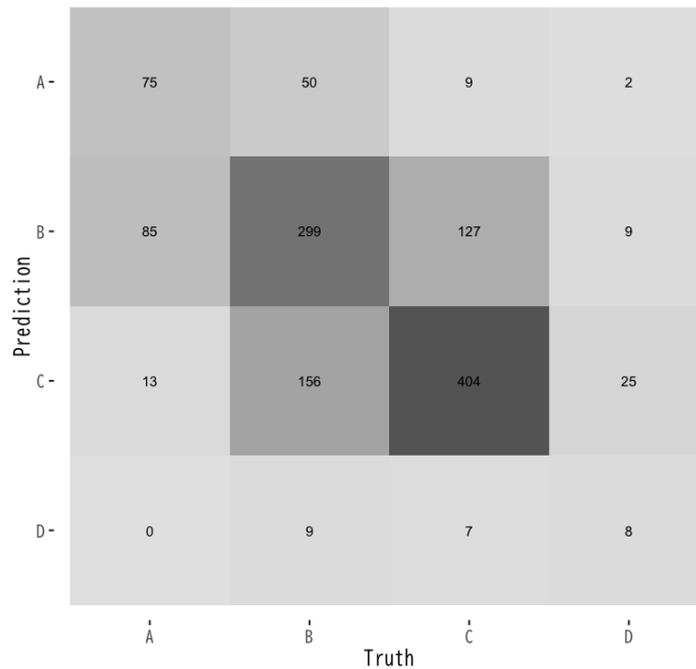


Figure 7: 訓練誤差

LDA (test data) : acc = 0.607

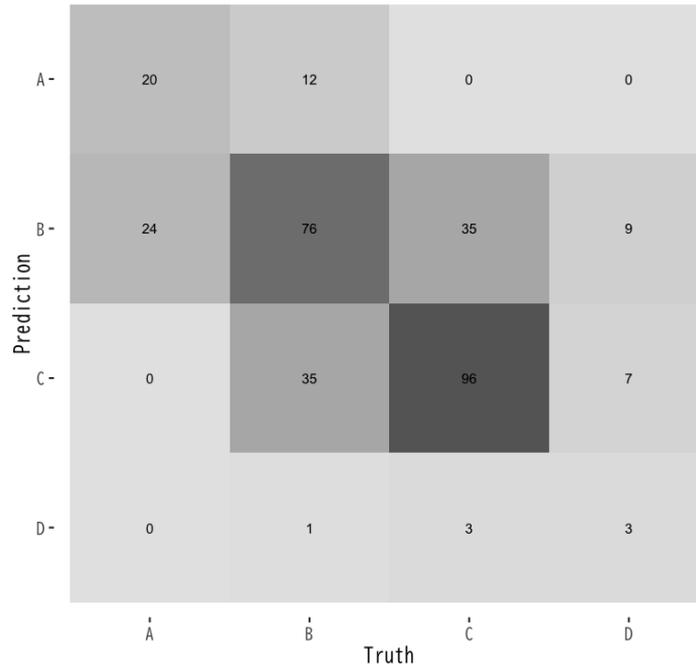


Figure 8: 予測誤差

QDA (training data) : acc = 0.626

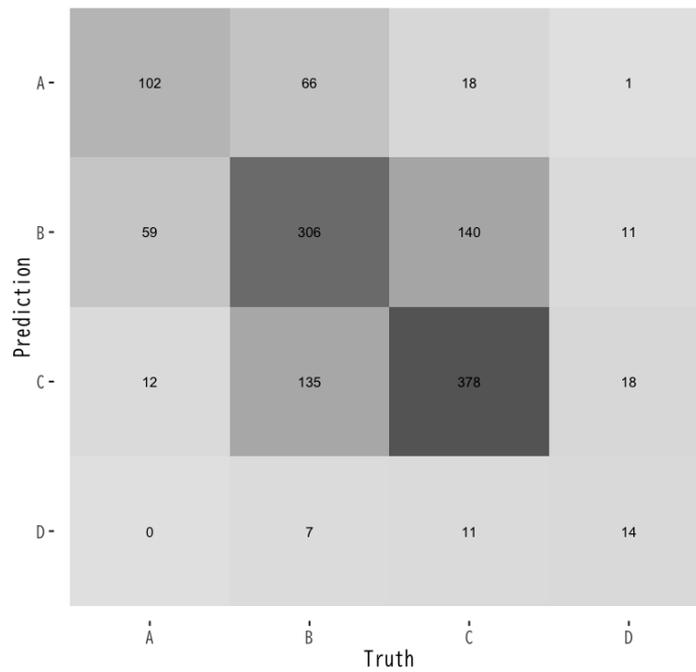


Figure 9: 訓練誤差

QDA (test data) : acc = 0.604

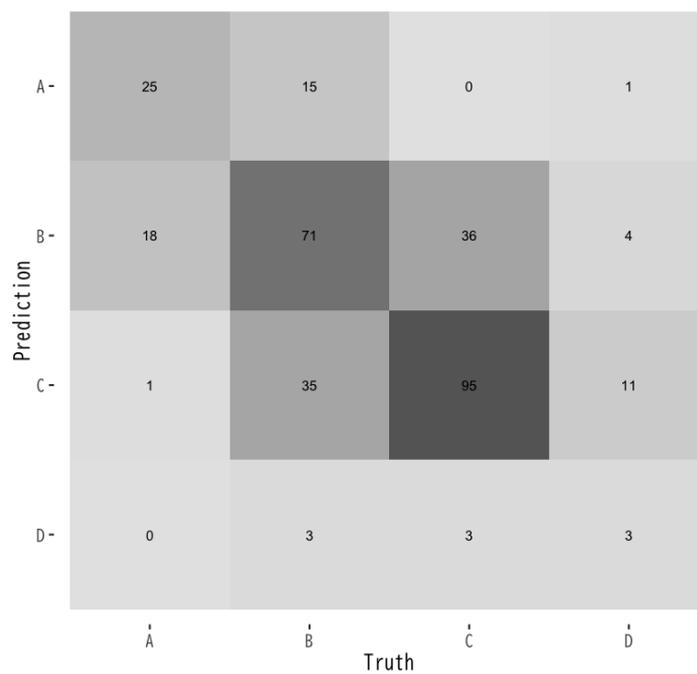


Figure 10: 予測誤差

LDA (loo cv error) : acc = 0.6

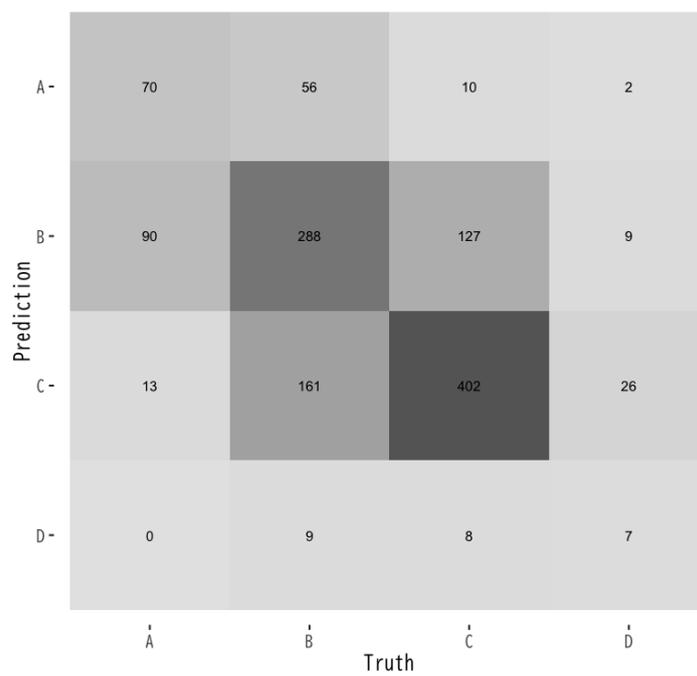


Figure 11: 線形判別

QDA (loo cv error) : acc = 0.576

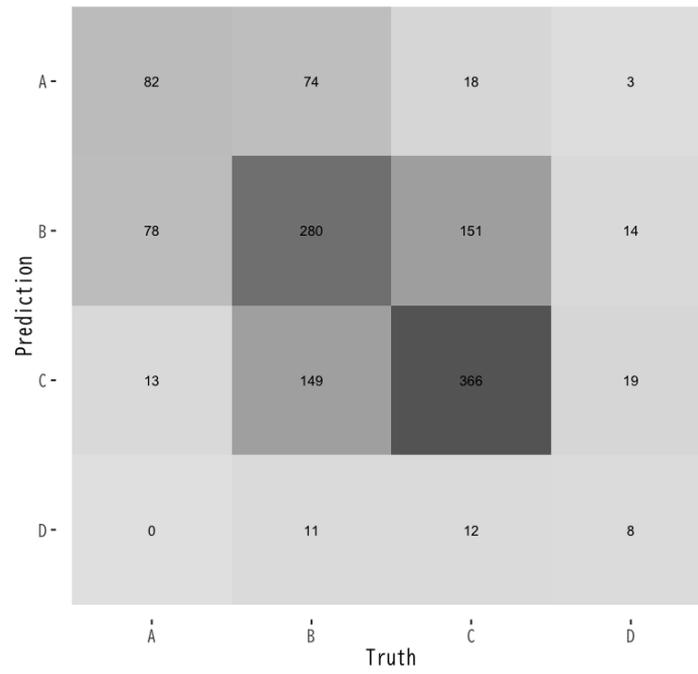


Figure 12: 2次判別

- 第2日：非階層的方法と分析の評価