project 1:手書き認識と精度評価
-手書き文字認識-

佐藤　好久
（九州工業大学大学院　情報工学研究院）

システム創成情報工学科　　システム創成プロジェクトⅠ

目標：2種類の手書き文字のデータを分析し，
その文字のパターンによる文字認識と
その精度評価を行う
（コンピュータで画像を認識させる問題）

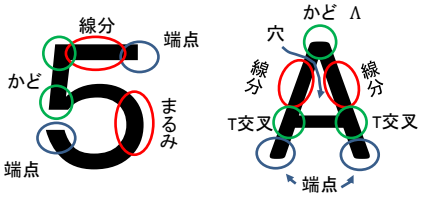
School of Computer Science and Systems Engineering　Kyushu Institute of Technology

1. 文字認識

パターン認識(pattern recognition) : 音声・音楽データ，文字や図形などの認識

- 文字や2次元図形のパターンの特徴例
  - 線分（直線の部分）
  - 円弧
  - 端点
  - かど
  - T交叉
  - 丸み
  - 穴
  - 文字の長さ，幅

  など



- パターン認識の手順
  1. 特徴抽出（「あいまいさ」があることを考慮して，より良い特徴を．）
  2. 特徴の分析 ・・・ 統計的方法（正規分布モデルによる特徴分類）
  3. パターンが，あらかじめ用意されたカテゴリーのどれに属するかを決定

手書き文字：あいまいさ，ばらつき，ゆらぎ，文字の変形，かすれ，書き癖，etc
これらを許容した上で特徴を取り出さなければならない．
文字の「大域的」特徴を捉えることが重要

1. 文字認識

手書き文字には，
多くの情報が含まれている

多くの「情報」

↓

特徴量の抽出

↓

パターン認識（分類）

データ数：各々88個（サンプル数をそろえた方が便利．正規化する手間が省ける）

「穴」：共通する特徴として「穴」の個数がある．「B」と「8」を区別する特徴として，
「穴」は使えない．
（一般に，似ている文字の認識には「位相幾何学」的な特徴は使えない．）

2. 統計的パターン認識の手法

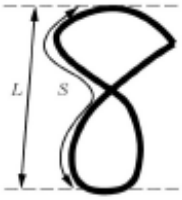
2. 1 Bと8の特徴(その1)

統計的パターン認識 (statistical pattern recognition)

特徴量(feature value) : 特徴を数量的に取り出したもの.
(特徴量は1種類だけとは限らない.)

① 左側が, B は直線的, 8 は曲線的.

1. 「直線的」「曲線的」を数値化
2. その値がどういう範囲にあるときに直線的であると判定するかを決定
([問題] : 「直線的」と「曲線的」の境界をどのように決めるか?)



L : 文字の高さ, S : 左側の長さ

特徴量① $x = \frac{L}{S}$

$x \approx 1$ → 直線的
? ?
$x \approx 0$ → 曲線的

2. 統計的パターン認識の手法

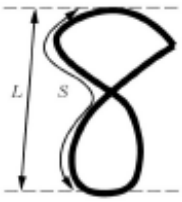
2. 1 Bと8の特徴(その1)

統計的パターン認識 (statistical pattern recognition)

特徴量(feature value) : 特徴を数量的に取り出したもの.
(特徴量は1種類だけとは限らない.)

① 左側が, B は直線的, 8 は曲線的.

1. 「直線的」「曲線的」を数値化
2. その値がどういう範囲にあるときに直線的であると判定するかを決定
([問題] : 「直線的」と「曲線的」の境界をどのように決めるか?)



L : 文字の高さ, S : 左側の長さ

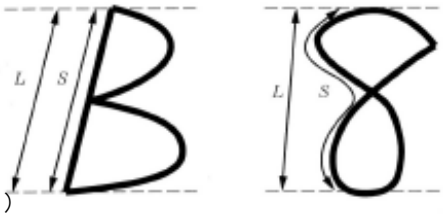
特徴量① $x = \frac{L}{S}$

$x \approx 1$ → 直線的
? ?
$x \approx 1$ → 曲線的

School of Computer Science and Systems Engineering Kyushu Institute of Technology

2. 統計的パターン認識の手法

2. 1 Bと8の特徴(その1)



① 左側が，Bは直線的，8は曲線的．

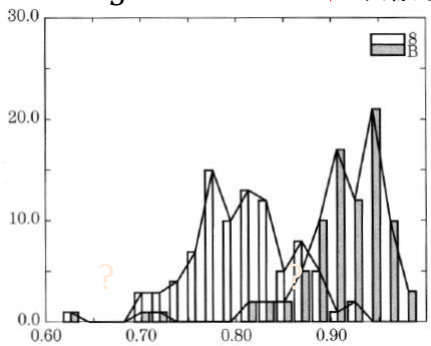
2. その値がどういう範囲にあるときに直線的であると判定するかを決定
([問題]：「直線的」と「曲線的」の境界をどのように決めるか？)

L：文字の高さ，　S：左側の長さ 特徴量① $x = \frac{L}{S}$　$x \approx 1$ → 直線的　$x \not\approx 1$ → 曲線的

- すべてのサンプルについて，特徴量 x の分布を調査する．
- 度数分布表を描く．

※ サンプル数を揃えているので，度数分布(表)は確率密度関数とと見なせる．(正規化する必要なし)

$p(x|B)$，　$p(x|8)$

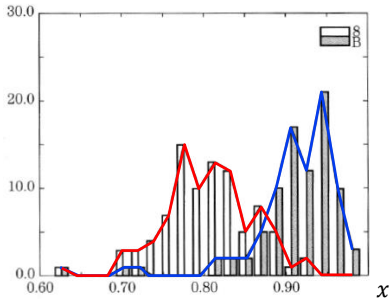


2. 統計的パターン認識の手法

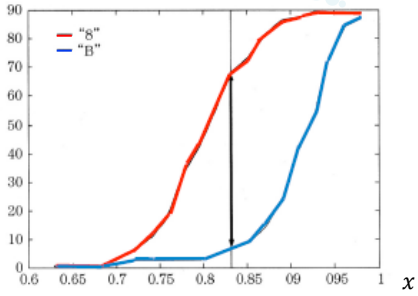
2. 1 Bと8の特徴(その1)

① 左側が，Bは直線的，8は曲線的．

一般には，ここを1に正規化する必要がある．サンプル数がそろっているので，このままでよい．



8およびBのそれぞれの88文字サンプルについての直線度 x の分布



8およびBの直線度 x の周辺分布

$\boldsymbol{p(x|8)}$：文字8の直線度 x の確率密度関数　$\boldsymbol{F(X|8)} = \int_{-\infty}^{X} p(x|8)\,dx$：文字8の直線度 x の周辺累積分布

文字8の直線度が $x \leq X$ である確率

$\boldsymbol{p(x|B)}$：文字Bの直線度 x の確率密度関数

$\boldsymbol{F(X|B)} = \int_{-\infty}^{X} p(x|B)dx$：文字Bの直線度 x の周辺累積分布

直線度の値が x である8(またはB)のサンプル数(度数分布)

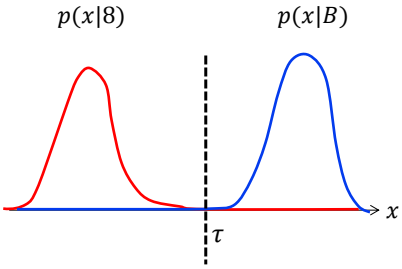
直線度 x の値が X 以下である8(またはB)のサンプル数

2. 統計的パターン認識の手法

① 左側が，B は直線的，8 は曲線的.

- 特徴量の分布に基づく識別閾値 τ の決め方

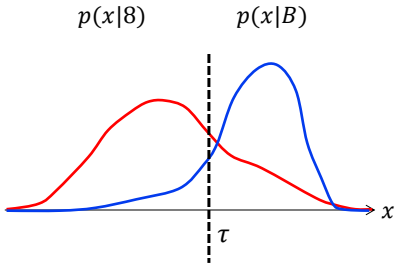
◆ 重なりがない



境界 $x = \tau$ で完全に識別可能

$x < \tau \implies$ 文字は8
$x > \tau \implies$ 文字はB

◆ 重なりがある



境界（閾値）$x = \tau$ をどこにとっても，
完全に識別することは不可能
（誤り判定がある）

2. 統計的パターン認識の手法

① 左側が，B は直線的，8 は曲線的.

- 特徴量の分布に基づく識別閾値 τ の決め方

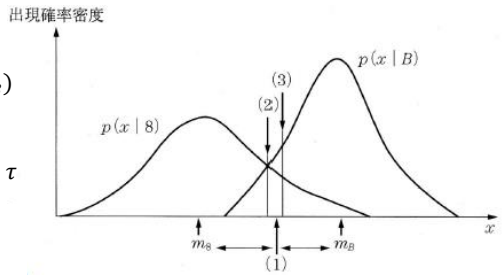
識別閾値 τ の設定方法

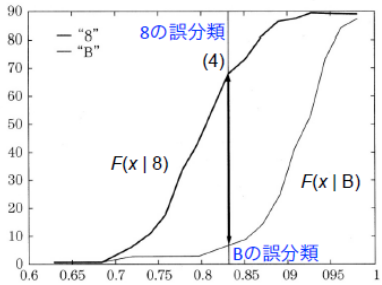
(1) 平均 m_8, m_B の平均 $\tau = \frac{1}{2}(m_8 + m_B)$

(2) $p(\tau|8) = p(\tau|B)$ となる τ

(3) $F(\tau|8) + F(\tau|B) = 1$ (100%)となる τ

◆ 重なりがある





識別閾値 τ の設定方法

(4) $F(\tau|8) - F(\tau|B)$ の差が最大となるτ

2. 統計的パターン認識の手法

① 左側が，B は直線的，8 は曲線的．

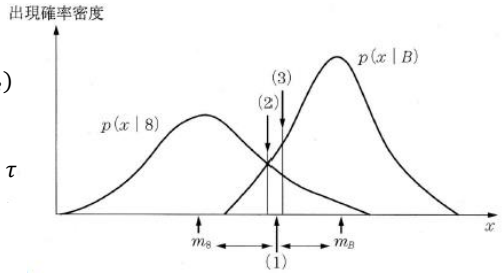
- 特徴量の分布に基づく識別閾値 τ の決め方　　◆ 重なりがある

識別閾値 τ の設定方法

(1) 平均 m_8, m_B の平均 $\tau = \frac{1}{2}(m_8 + m_B)$

(2) $p(\tau|8) = p(\tau|B)$ となる τ

(3) $F(\tau|8) + F(\tau|B) = 1$　(100%)となる τ



(1) 平均値はデータの代表値として典型的なもの．この2つの典型的な値から最も離れたところを境界とする考え方

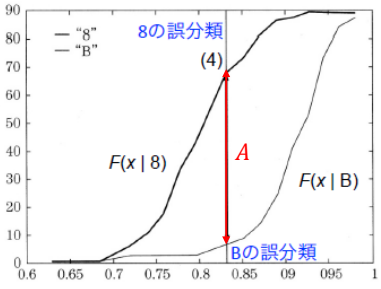
(2) $p(x|8) \lessgtr p(x|B)$ により判定しようとする考え方

(3) このような τ を選ぶと，$1 - F(\tau|8)$ は「8」であるにも関わらず「B」と判定される割合，$F(\tau|B)$ は「B」であるにも関わらず「8」と判定される割合．
8 と B とのお互いで間違って識別される割合を等しくする意図のもの

2. 統計的パターン認識の手法

① 左側が，B は直線的，8 は曲線的．

- 特徴量の分布に基づく識別閾値 τ の決め方　　◆ 重なりがある



識別閾値 τ の設定方法

(4) $F(\tau|8) - F(\tau|B)$ の差が最大となるτ

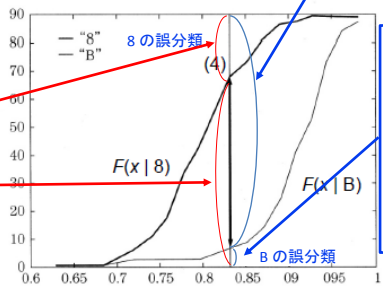
A を最大= 誤判定の割合の合計を最小

Bのデータでの $x \geq \tau$ の割合（Bを正しくBと判定する割合）

32%

8のデータでの$x \geq \tau$ の割合（8をBと誤判定）

8のデータでの$x \leq \tau$ の割合（8を正しく8と判定する割合）



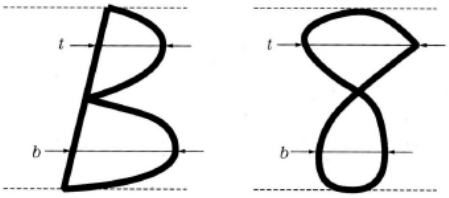
Bのデータで $x \leq \tau$ の割合（Bを8と誤判定）

7%

サンプルでの誤判定の合計= 32 + 7 = 39 %
両方の誤判定を20%以下にすることは困難

2. 統計的パターン認識の手法

2. 2 B と 8 の特徴（その2）



1つの特徴量では，誤差の和が一定以下にできない．認識率向上のためには，別の特徴量を考える必要あり．

② 上下の和の幅の比
B は下が大きい，8 は上が大きい．

特徴量② $x = \frac{b}{t}$

特徴量②についても特徴量①と同様に識別閾値の分析をすることができるが，特徴量①と同様に，単独では識別向上には限界がある．

2. 統計的パターン認識の手法

複数個の特徴量を識別に用いて，識別向上を図る．

- n個の特徴量 $x_1, x_2, \cdots, x_n$ から作られるベクトル（**特徴量ベクトル**feature vector）
$$\mathbb{x} = (x_1, x_2, \cdots, x_n)$$
- **特徴空間**(feature space) ：特徴量ベクトルを座標とする空間（n次元空間）

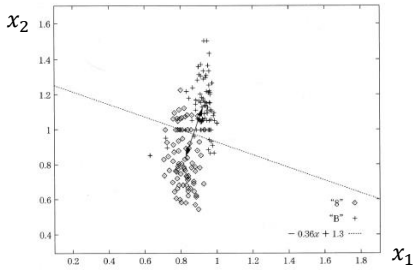
➢ 「8」と「B」

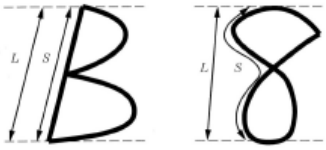
特徴量①：$x_1 = \frac{L}{S}$

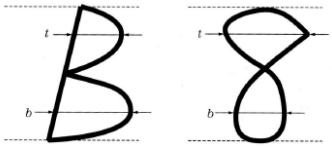
特徴量②：$x_2 = \frac{b}{t}$

特徴量ベクトル (x_1, x_2)

特徴空間：2次元空間（平面）







2. 統計的パターン認識の手法

2. 3　2次元特徴空間における識別境界の設定方法

1次元特徴空間では，識別のために，識別閾値 τ を設定する方法を考察した．
2次元特徴空間では，識別するための境界線（**識別境界** decision boundary）を設定しなければならない．以後，特徴量が正規分布に従うものと仮定．

➢ **平均値は異なるが，分散が同一の独立した正規分布に従う場合**

x_i の確率密度関数：

$$p(x_i|8) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-m_i^8)^2}{2\sigma^2}\right]$$
$$p(x_i|B) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-m_i^B)^2}{2\sigma^2}\right]$$

m_i^8 :「8」の特徴量 x_i の平均値
m_i^B :「B」の特徴量 x_i の平均値
σ^2 : 両正規分布で同一の分散

特徴ベクトル $\mathbb{x} = (x_1, x_2)$ の同次密度関数は，

$$p(\mathbb{x}|8) = p(x_1|8)p(x_2|8) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2\sigma^2}\|\mathbb{x}-\mathbb{m}_8\|^2\right]$$
$$p(\mathbb{x}|\mathrm{B}) = p(x_1|\mathrm{B})p(x_2|\mathrm{B}) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2\sigma^2}\|\mathbb{x}-\mathbb{m}_B\|^2\right]$$

$\mathbb{m}_8 = (m_1^8, m_2^8)$
$\mathbb{m}_B = (m_1^B, m_2^B)$

ただし，$\mathbb{x} = (x_1, x_2)$，$\mathbb{y} = (y_1, y_2)$ に対して，$\|\mathbb{x}-\mathbb{y}\| = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$

2. 統計的パターン認識の手法

➢ **平均値は異なるが，分散が同一の独立した正規分布に従う場合**

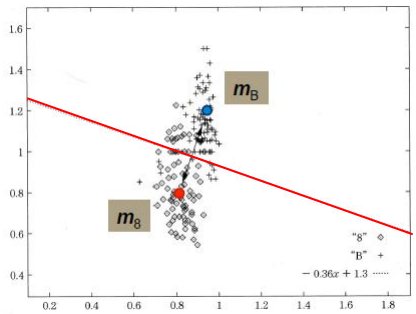
識別境界の設定：$p(\mathbb{x}|8) = p(\mathbb{x}|B)$ となる $\mathbb{x}$ の描く曲線

データ $\mathbb{x}$ について，

$$\begin{cases} p(\mathbb{x}|8) \geq p(\mathbb{x}|B) & \Rightarrow \quad 8 \text{ と判定} \\ p(\mathbb{x}|8) \leq p(\mathbb{x}|B) & \Rightarrow \quad B \text{ と判定} \end{cases}$$

$$※\ p(\mathbb{x}|8) = p(\mathbb{x}|B) \Leftrightarrow \frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2\sigma^2}\|\mathbb{x}-\mathbb{m}_8\|^2\right] = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2\sigma^2}\|\mathbb{x}-\mathbb{m}_B\|^2\right]$$
$$\Leftrightarrow \|\mathbb{x}-\mathbb{m}_8\|^2 = \|\mathbb{x}-\mathbb{m}_B\|^2$$

識別境界線：$\|\mathbb{x}-\mathbb{m}_8\| = \|\mathbb{x}-\mathbb{m}_B\|$
（$\mathbb{m}_8$ と $\mathbb{m}_B$ を結ぶ線分の垂直2等分線）



データ $\mathbb{x}$ について，

$$\begin{cases} \|\mathbb{x}-\mathbb{m}_8\| \geq \|\mathbb{x}-\mathbb{m}_B\| & \Rightarrow \quad 8 \text{ と判定} \\ \|\mathbb{x}-\mathbb{m}_8\| \leq \|\mathbb{x}-\mathbb{m}_B\| & \Rightarrow \quad B \text{ と判定} \end{cases}$$

2. 統計的パターン認識の手法

➢ **に従う場合**

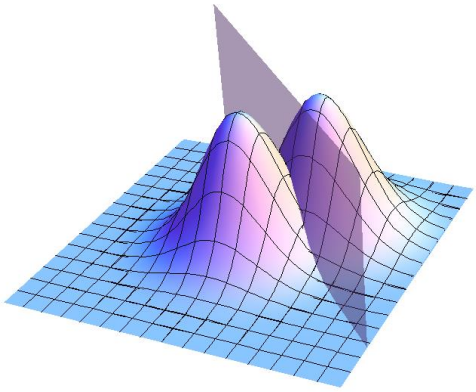
識

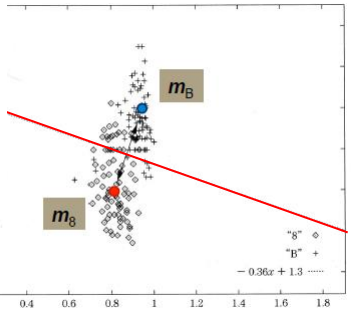
と判定

と判定

※ $= \frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2\sigma^2}\|\mathbb{x}-\mathbb{m}_B\|^2\right]$

識





テ

$\begin{cases}\|\mathbb{x}-\mathbb{m}_8\| \le \|\mathbb{x}-\mathbb{m}_B\| & \Rightarrow \quad B \text{ と判定} \end{cases}$

2. 統計的パターン認識の手法

➢ **各特徴量毎に独立で同一の正規分布に従う場合**
（分散が文字に依らない．特徴量①の分散 σ_1^2，特徴量②の分散 σ_2^2）

$\mathbb{x}=(x_1,x_2)$ の共分散行列 $\Sigma=\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}$

$$\begin{cases} p(\mathbb{x}|8)=\dfrac{1}{2\pi|\Sigma|^{1/2}}\exp\left[-\dfrac{1}{2}(\mathbb{x}-\mathbb{m}_8)\Sigma^{-1}{}^t(\mathbb{x}-\mathbb{m}_8)\right] \\ p(\mathbb{x}|B)=\dfrac{1}{2\pi|\Sigma|^{1/2}}\exp\left[-\dfrac{1}{2}(\mathbb{x}-\mathbb{m}_B)\Sigma^{-1}{}^t(\mathbb{x}-\mathbb{m}_B)\right] \end{cases}$$

（注意）(1) $|\Sigma|$ は共分散行列 Σ の行列式，${}^t(\mathbb{x}-\mathbb{m}_8)$ は $\mathbb{x}-\mathbb{m}_8$ の転置

(2) 8 と B について，別々に σ_1^2，σ_2^2 を計算して，大きく異なっている場合は，この方法は使えない．

(3) 分散は文字に依らず，特徴量毎に共通であると仮定するので，
σ_1^2：8，B の全てのデータを合わせて計算

σ_2^2：8，B の全てのデータを合わせて計算

識別境界の設定：$p(\mathbb{x}|8)=p(\mathbb{x}|B)$ となる $\mathbb{x}$ の描く曲線

識別：データ $\mathbb{x}$ について，$\begin{cases} p(\mathbb{x}|8)\ge p(\mathbb{x}|B) & \Rightarrow \quad 8 \text{ と判定} \\ p(\mathbb{x}|8)\le p(\mathbb{x}|B) & \Rightarrow \quad B \text{ と判定} \end{cases}$

2. 統計的パターン認識の手法

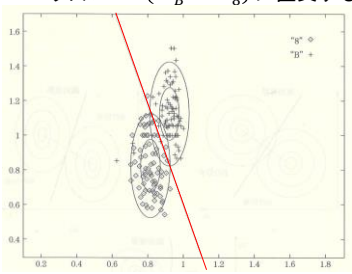
➢ **各特徴量毎に独立で同一の正規分布に従う場合**
（分散が文字に依らない．特徴量①の分散 σ_1^2，特徴量②の分散 σ_2^2）

識別境界の設定：$p(\mathbb{x}|8) = p(\mathbb{x}|B)$ となる$\mathbb{x}$ の描く曲線

$$※\ p(\mathbb{x}|8) = p(\mathbb{x}|B) \Leftrightarrow \exp\left[-\frac{1}{2}(\mathbb{x}-\mathbb{m}_8)\Sigma^{-1}{}^t(\mathbb{x}-\mathbb{m}_8)\right] = \exp\left[-\frac{1}{2}(\mathbb{x}-\mathbb{m}_B)\Sigma^{-1}{}^t(\mathbb{x}-\mathbb{m}_B)\right]$$
$$\Leftrightarrow (\mathbb{x}-\mathbb{m}_8)\Sigma^{-1}{}^t(\mathbb{x}-\mathbb{m}_8) = (\mathbb{x}-\mathbb{m}_B)\Sigma^{-1}{}^t(\mathbb{x}-\mathbb{m}_B)$$
$$\Leftrightarrow\ \mathbb{x}\Sigma^{-1}{}^t(\mathbb{m}_B-\mathbb{m}_8) = \frac{1}{2}(\mathbb{m}_B+\mathbb{m}_8)\Sigma^{-1}{}^t(\mathbb{m}_B-\mathbb{m}_8)$$

よって，

識別境界：点 $\frac{1}{2}(\mathbb{m}_B+\mathbb{m}_8)$ を通り，
ベクトル$\Sigma^{-1}{}^t(\mathbb{m}_B-\mathbb{m}_8)$ に直交する直線



$\mathbb{x}\cdot\mathbb{a} = \boldsymbol{b}$ の形
1次方程式なので，n 次元空間内の「超平面」
2次元空間（平面）の場合は直線．
3次元空間の場合は平面．
$\mathbb{a}$ は法線ベクトル

2. 統計的パターン認識の手法

➢ **一般の正規分布に従う場合**

Σ_8 :「8」のパターン $\mathbb{x} = (x_1, x_2)$ の共分散行列，　$\mathbb{m}_8$:平均値

Σ_B :「B」のパターン $\mathbb{x} = (x_1, x_2)$ の共分散行列，　$\mathbb{m}_B$:平均値

$$\begin{cases} p(\mathbb{x}|8) = \dfrac{1}{2\pi|\Sigma_8|^{1/2}}\exp\left[-\dfrac{1}{2}(\mathbb{x}-\mathbb{m}_8)\Sigma_8^{-1}{}^t(\mathbb{x}-\mathbb{m}_8)\right] \\ p(\mathbb{x}|B) = \dfrac{1}{2\pi|\Sigma_B|^{1/2}}\exp\left[-\dfrac{1}{2}(\mathbb{x}-\mathbb{m}_B)\Sigma_B^{-1}{}^t(\mathbb{x}-\mathbb{m}_B)\right] \end{cases}$$

識別境界の設定：$p(\mathbb{x}|8) = p(\mathbb{x}|B)$ となる$\mathbb{x}$ の描く曲線

$$※\ p(\mathbb{x}|8) = p(\mathbb{x}|B)$$
$$\Leftrightarrow \frac{1}{|\Sigma_8|^{1/2}}\exp\left[-\frac{1}{2}(\mathbb{x}-\mathbb{m}_8)\Sigma_8^{-1}{}^t(\mathbb{x}-\mathbb{m}_8)\right] = \frac{1}{|\Sigma_B|^{1/2}}\exp\left[-\frac{1}{2}(\mathbb{x}-\mathbb{m}_B)\Sigma_B^{-1}{}^t(\mathbb{x}-\mathbb{m}_B)\right]$$
$$\Leftrightarrow \log|\Sigma_8| + (\mathbb{x}-\mathbb{m}_8)\Sigma_8^{-1}{}^t(\mathbb{x}-\mathbb{m}_8) = \log|\Sigma_B| + (\mathbb{x}-\mathbb{m}_B)\Sigma_B^{-1}{}^t(\mathbb{x}-\mathbb{m}_B)$$

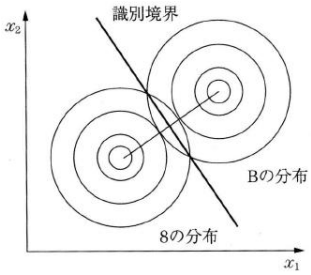
よって，これは2次方程式であるので，

識別境界：放物線，楕円，双曲線などの2次曲線

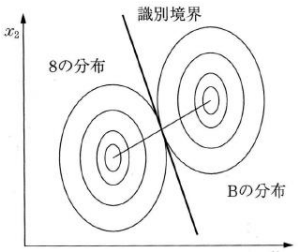
$\mathbb{x}\boldsymbol{A}^t\mathbb{x} = \boldsymbol{b}$ の形
A :対称行列

2. 統計的パターン認識の手法

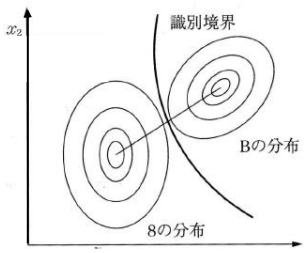
(1) 特徴量毎に独立で同一の正規分布を仮定



(2) 特徴量毎に分散のみ等しい独立し正規分布を仮定



(3) 一般的な正規分布を仮定



※ 図中の楕円は「8」と「B」のそれぞれの分布に対応した，平均からのマハラノビス距離が標準偏差の1倍，2倍，．．．となる点を模式的に表している．

マハラノビス距離(Mahalanobis distance) : $\left\|(\mathbb{x}-\mathbb{m}_8)\Sigma_8^{-1t}(\mathbb{x}-\mathbb{m}_8)\right\|$

$\left\|(\mathbb{x}-\mathbb{m}_8)\Sigma_8^{-1t}(\mathbb{x}-\mathbb{m}_8)\right\| = r$: 分布の中心 $\mathbb{m}_8$ から計って等出現確率となる $\mathbb{x}$ までの距離が等距離 r となる点 $\mathbb{x}$ の全体

まとめ

- 文字を識別するために，特徴量を考え，1次元特徴空間・2次元特徴空間を用いて判別する．
- 判別するために必要な，1次元特徴空間の「閾値」や2次元特徴空間の「識別境界線」の設定方法を考察した．

演習:

自分たちが集めたデータの「標本集団」を基に，識別のための「閾値」や「識別境界線」を設定 ⟹ 誤認識率

母集団 : 日本国民や全世界の人々 vs 標本集団

「標本集団」から得られた結果（「閾値」，「識別境界線」 特に，誤認識率）は信頼できるか？ （母集団に対しても当てはまるか？）

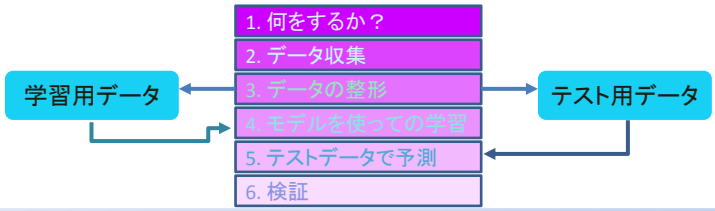
結果に対しての精度評価を行う必要がある．
（10-fold Cross-Validation, Bootstrap法による精度評価：本田先生の講義）

まとめ

機械学習：
大量のデータに対して人間が行うような判断をコンピュータに自動的に実行させる手段
その判断の内容は，大きく「分類」（識別タスク）と「予測」（予測タスク）の2つからなる．

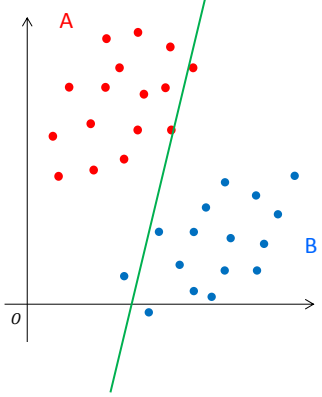
教師あり学習	(1-1) 識別（クラス分類）
	(1-2) 予測
教師なし学習	(2-1) クラスタリング
	(2-2) 次元圧縮
	(2-3) 密度推定，相関ルール抽出

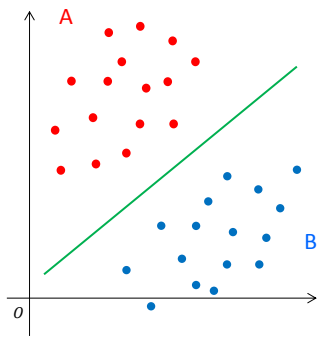
今回の創作プロジェクトのテーマ：文字認識
「教師あり学習」（ラベル「B」，「8」付データに対する「識別」）



今回の創成プロジェクトのテーマは機械学習の中の **判別分析** という手法

判別分析：**線形判別分析**と非線形判別分析

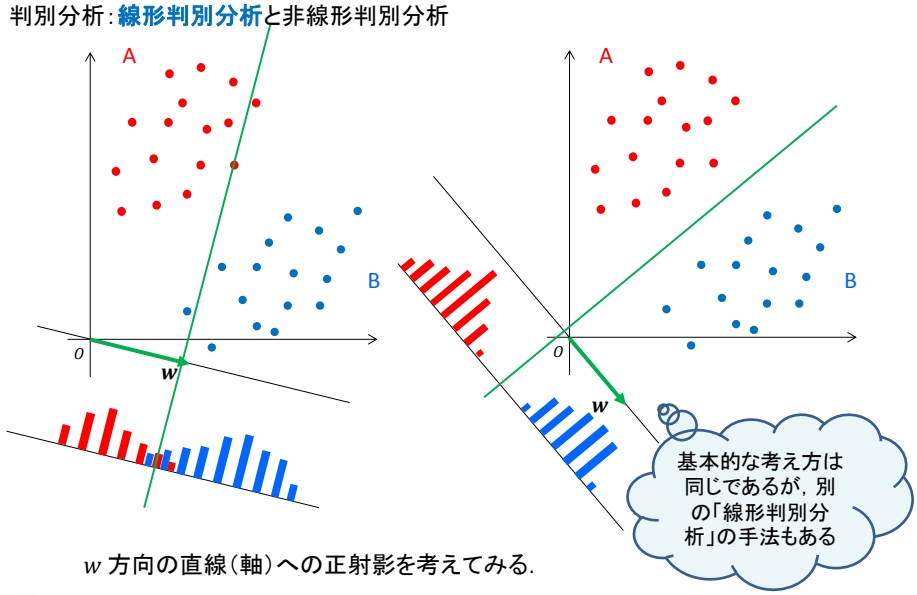




2つのグループを分ける「直線」を引きたい．
どのような引き方をすればよいか．

今回の創成プロジェクトのテーマは機械学習の中の **判別分析** という手法

判別分析：**線形判別分析**と非線形判別分析



w 方向の直線（軸）への正射影を考えてみる．