Project Ⅰ
識別精度について

講義担当:
本田あおい

プロジェクトⅠ　日程(1)

- 1週目 画像認識(1)
  - 3限目　講義(特徴量を使った識別)(佐藤)
  - 4限目　講義(演習説明)(齊藤)
  - 5限目　演習
- 2週目 画像認識(2)
  - 3限目　講義(演習説明)(宮野)
  - 4～5限目　演習
- 3週目 画像認識(3)(徳永)
  - 3～5限目　演習
- 計画書提出(3週目16:10まで)

プロジェクトⅠ　日程(2)

- プログラム提出(4週目の前日まで)
- 課題画像収集作業を4週目3限目までに終える。スキャン作業は4週目5限目までに終える。
- 4週目 識別精度
  - 3限目　講義(識別精度)(本田)
  - 4限目　講義(演習説明)(尾下or宮野or齊藤or徳永)
  - 5限目　演習
- 5週目 自由演習
  - 3～5限目　演習
- 6週目 プレゼンテーション
  - 3～4限目　プレゼンテーション

判別器

入力

数理モデル

出力

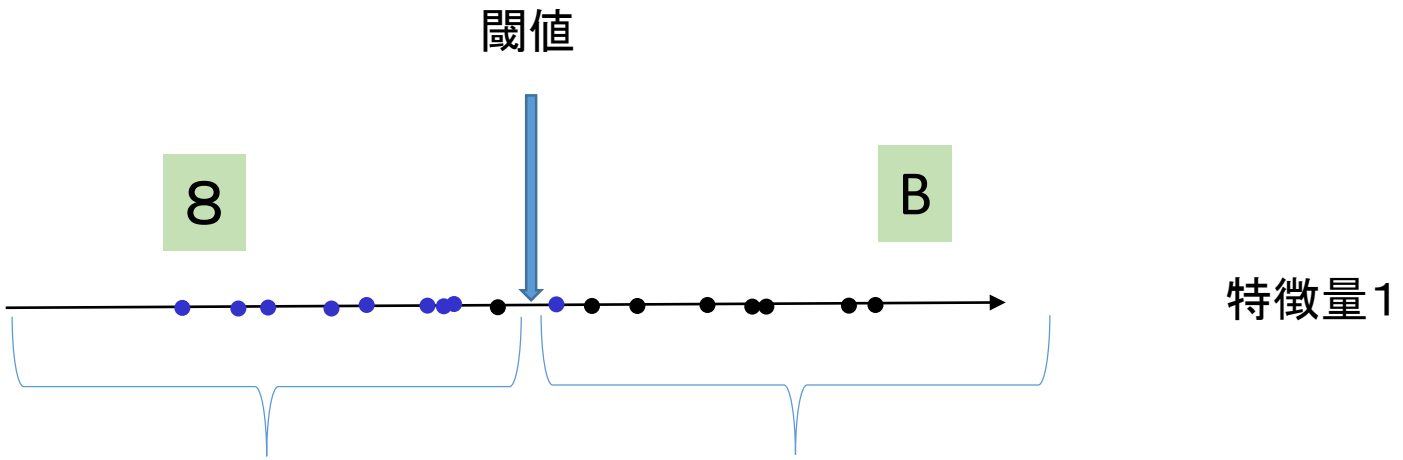
識別精度

Bと8の識別結果

予測 / 元データ	B	8	計
B	n_{11}	n_{12}	$n_{11} + n_{12}$
8	n_{21}	n_{22}	$n_{21} + n_{22}$
計	$n_{11} + n_{21}$	$n_{12} + n_{22}$	N

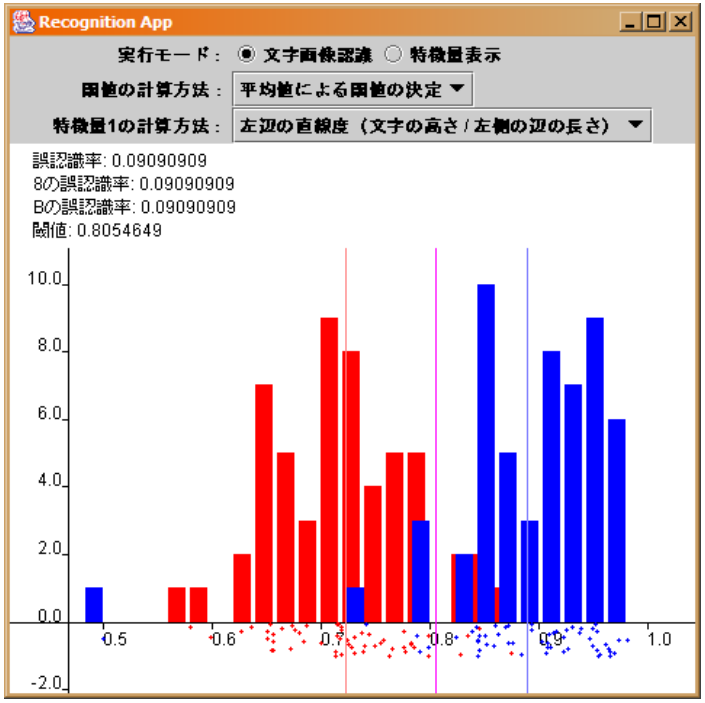
$$\text{誤認識率} = \frac{n_{12} + n_{21}}{N}$$

判別器とは



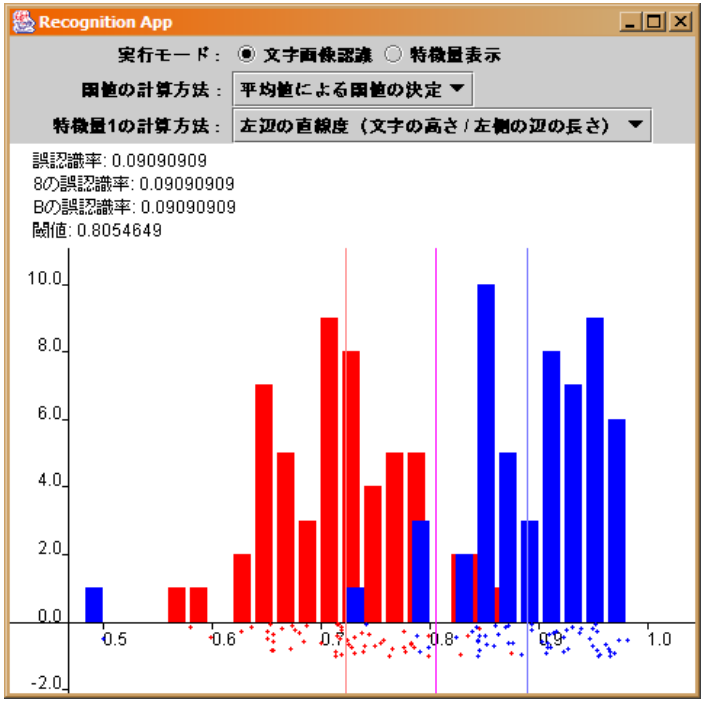
ヒストグラムの計算と描画

- 特徴量の値をいくつかの区分に分ける
- 各区分に含まれるデータの数をカウント
- 区分を自動的に決定
  - 区分の幅を指定する方法
    makeHistogramsBySize()
  - 区分の数を指定する方法
    makeHistogramsByWidth()



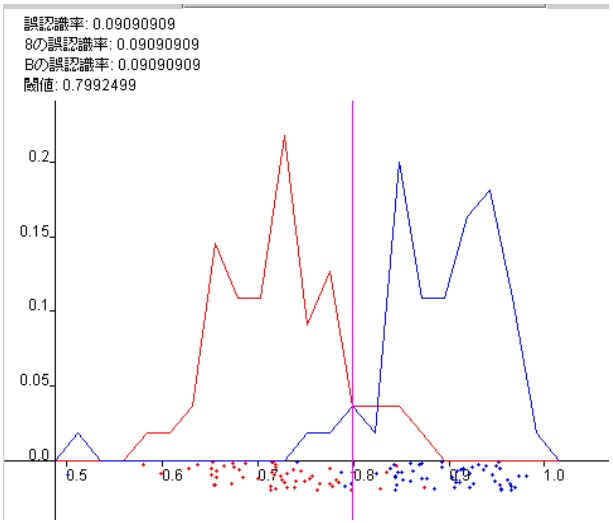
ヒストグラムの計算と描画

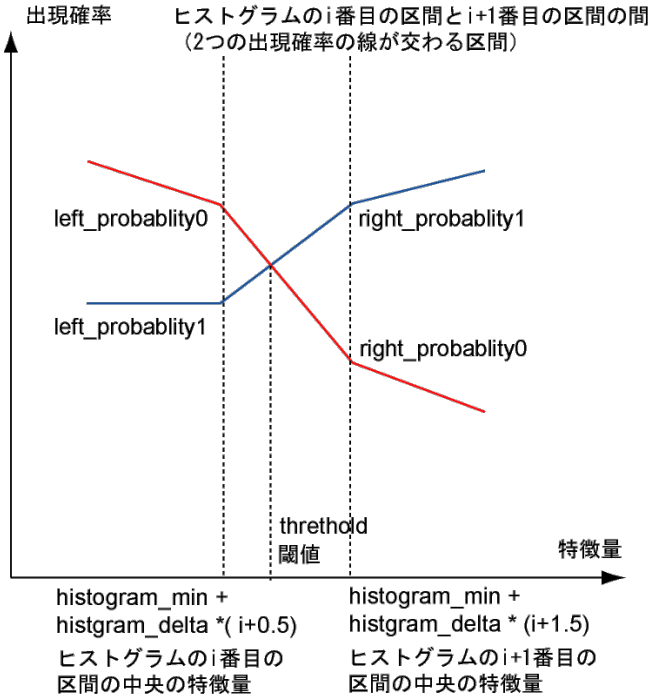
- 特徴量の値をいくつかの区分に分ける
- 各区分に含まれるデータの数をカウント
- 区分を自動的に決定
  - 区分の幅を指定する方法
    makeHistogramsBySize()
  - 区分の数を指定する方法
    makeHistogramsByWidth()



出現率が等しくなる閾値

- 出現率
  - 各区間でデータが出現する確率
    - 区間でのデータ数／全データ数
  - 各隣接区間での出現率
    から、線分の交点を計算

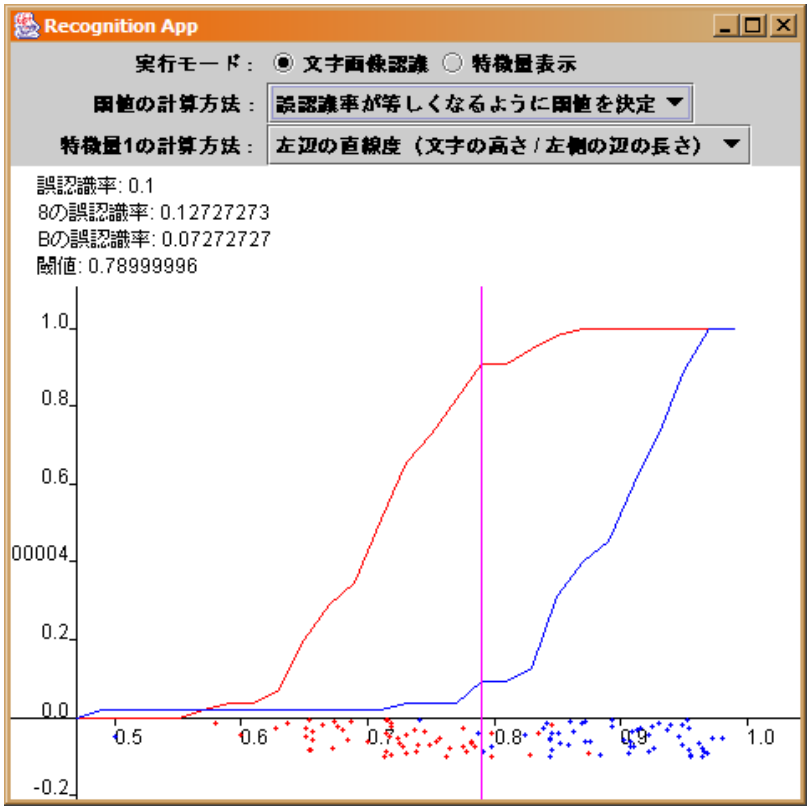




誤認識率が等しくなる閾値

- 誤認識率
  - その値を閾値にしたとき、どの程度の割合のデータを誤認識するか
- 誤認識率が等しくなる
  - 左右にある誤りデータの出現率が等しい
  - 出現確率の累計の和が 1 になる値

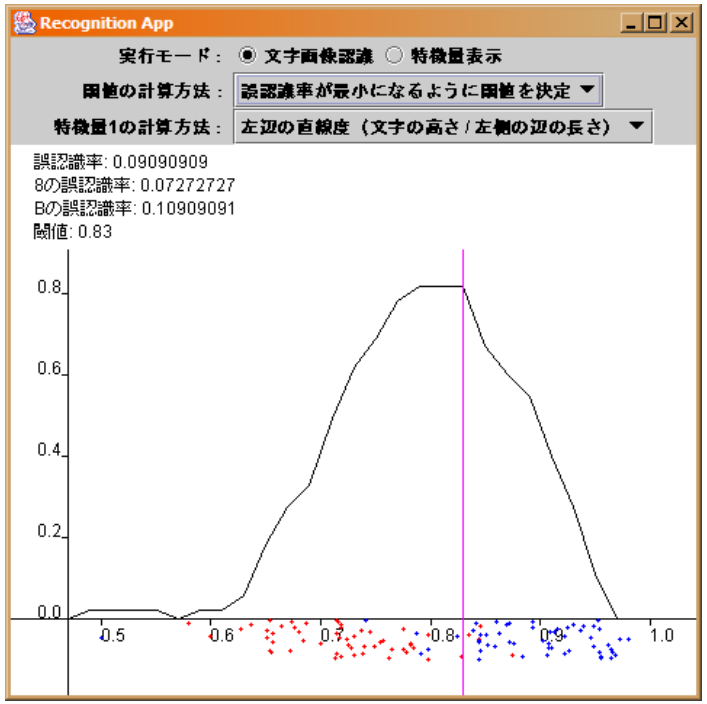
出現確率の累計のグラフ



誤認識率が最小になる閾値

- 2つの特徴量の出現確率の累積の差（正しく認識されるデータの割合）が最大になる値

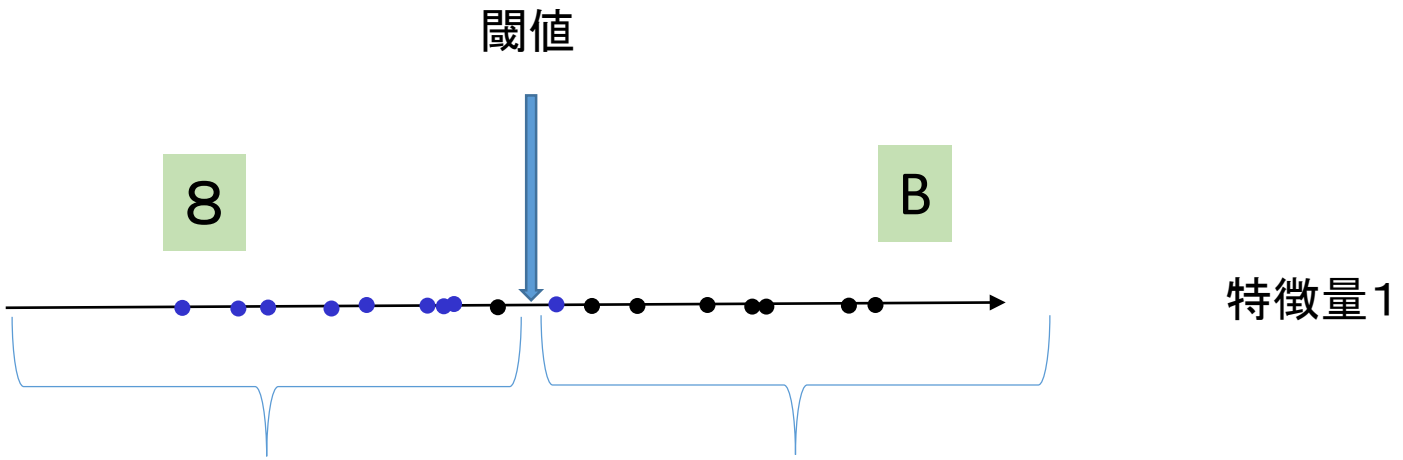
出現確率の累計の差のグラフ



2次元への拡張

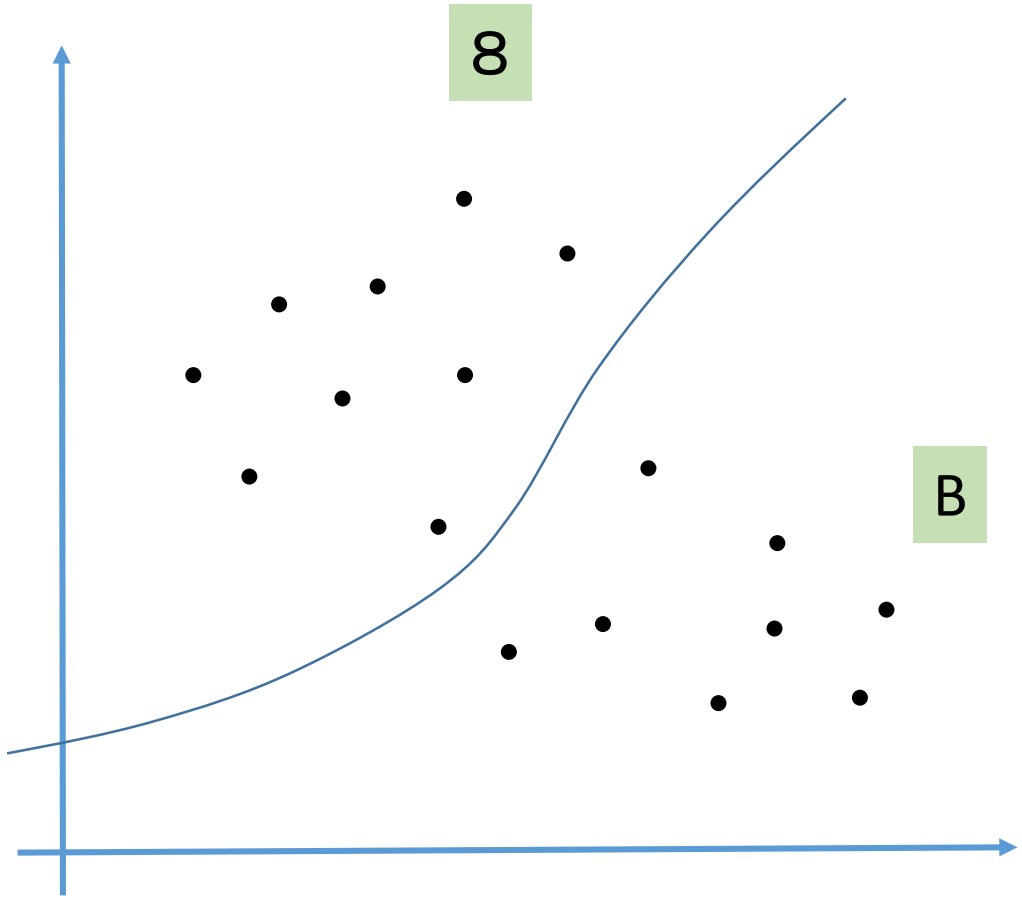
- 2つの特徴量を使った画像認識に拡張
  - 特徴量の計算はそのままで問題ない
    - 2つ目の特徴量の計算を追加する
  - 閾値の計算に、2つの特徴量を入力できるように変更する必要がある
    - 資料28ページのインターフェース定義を参照
- 具体的な内容は資料を参照

判別器とは



データを多次元にすると

判別対象をベクトルで表す



線形判別法

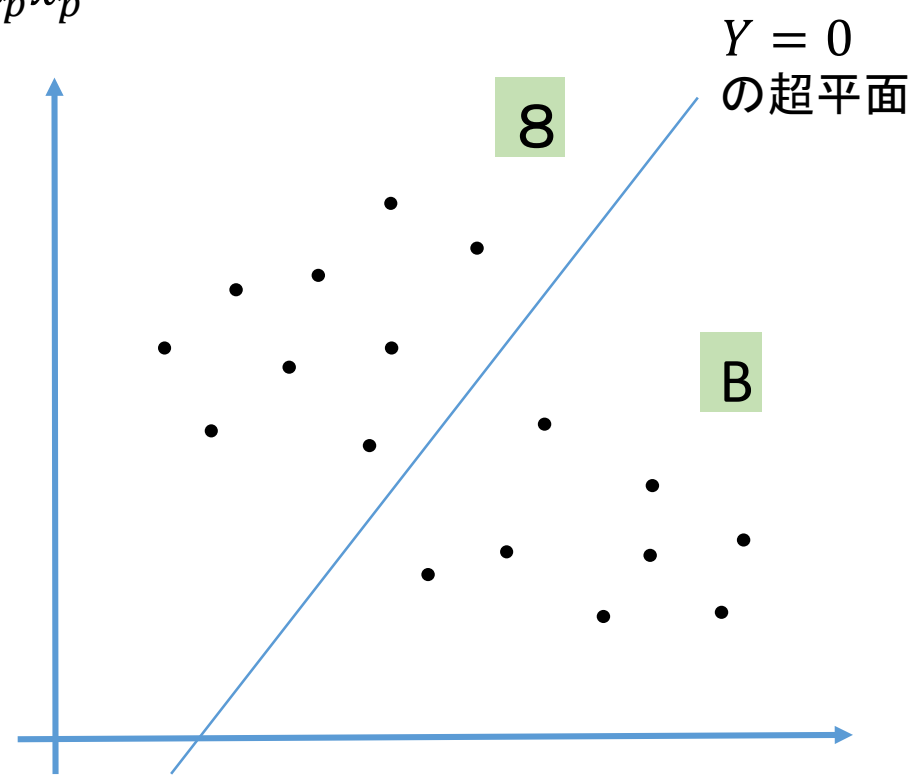
既知のデータ（学習データ）では正しく0, 1を出力する判別器を作る

最も基本的な判別法　線形判別関数

$$Y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$$

$Y \geq 0$　のとき 1
$Y < 0$　のとき 0

とする。

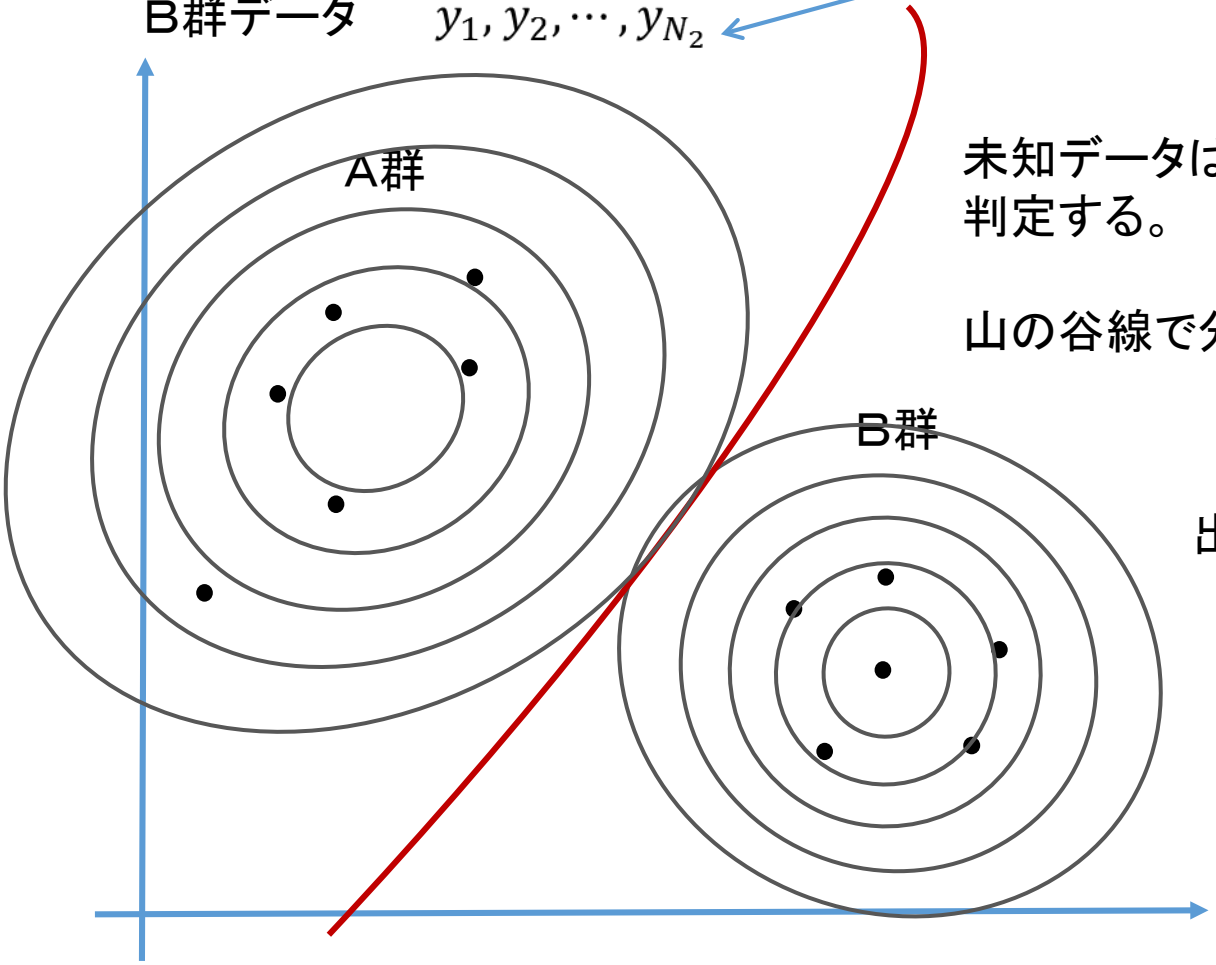


正規分布モデルによる分類

A群データ　$x_1, x_2, \cdots, x_{N_1}$

B群データ　$y_1, y_2, \cdots, y_{N_2}$

正規分布に従い出現すると仮定する



未知データはA群、B群のうち、近い方と判定する。

山の谷線で分離する

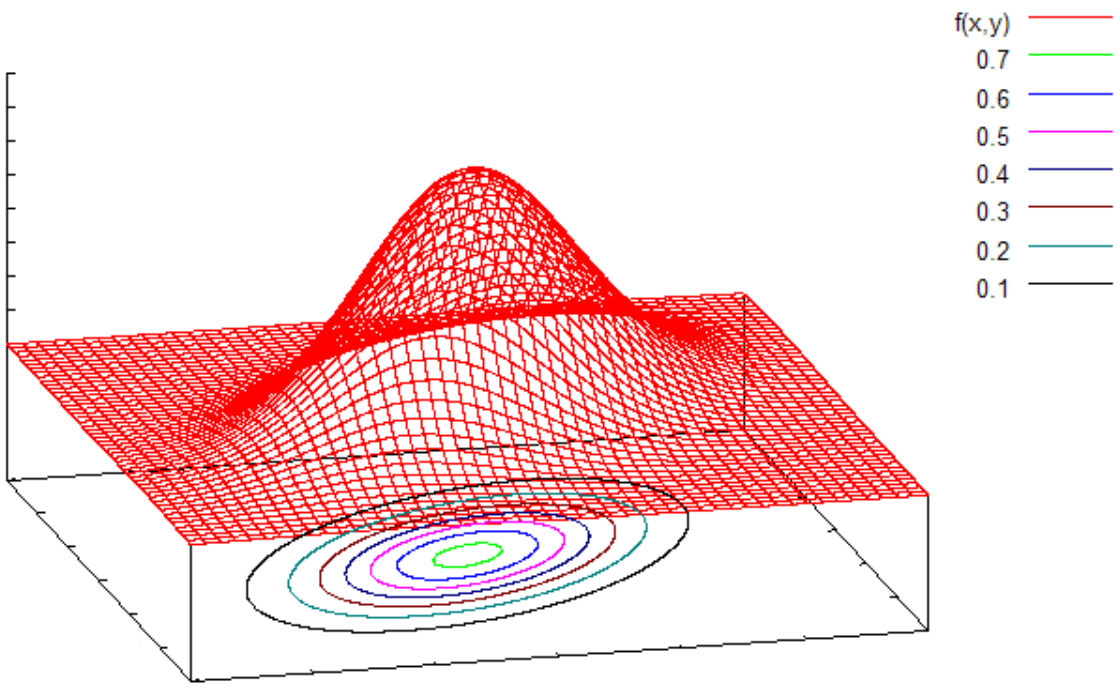
出現割合＝密度関数と考える

密度関数のパラメータは、既知データから計算する

正規分布モデルによる分類

2次元正規分布 $N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ の同時確率密度関数

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(x-\mu_y)^2}{\sigma_y^2}\right\}\right]$$



多次元に一般化される

正規分布モデルによる分類

二つの正規分布の密度関数の谷線は二次曲線になる

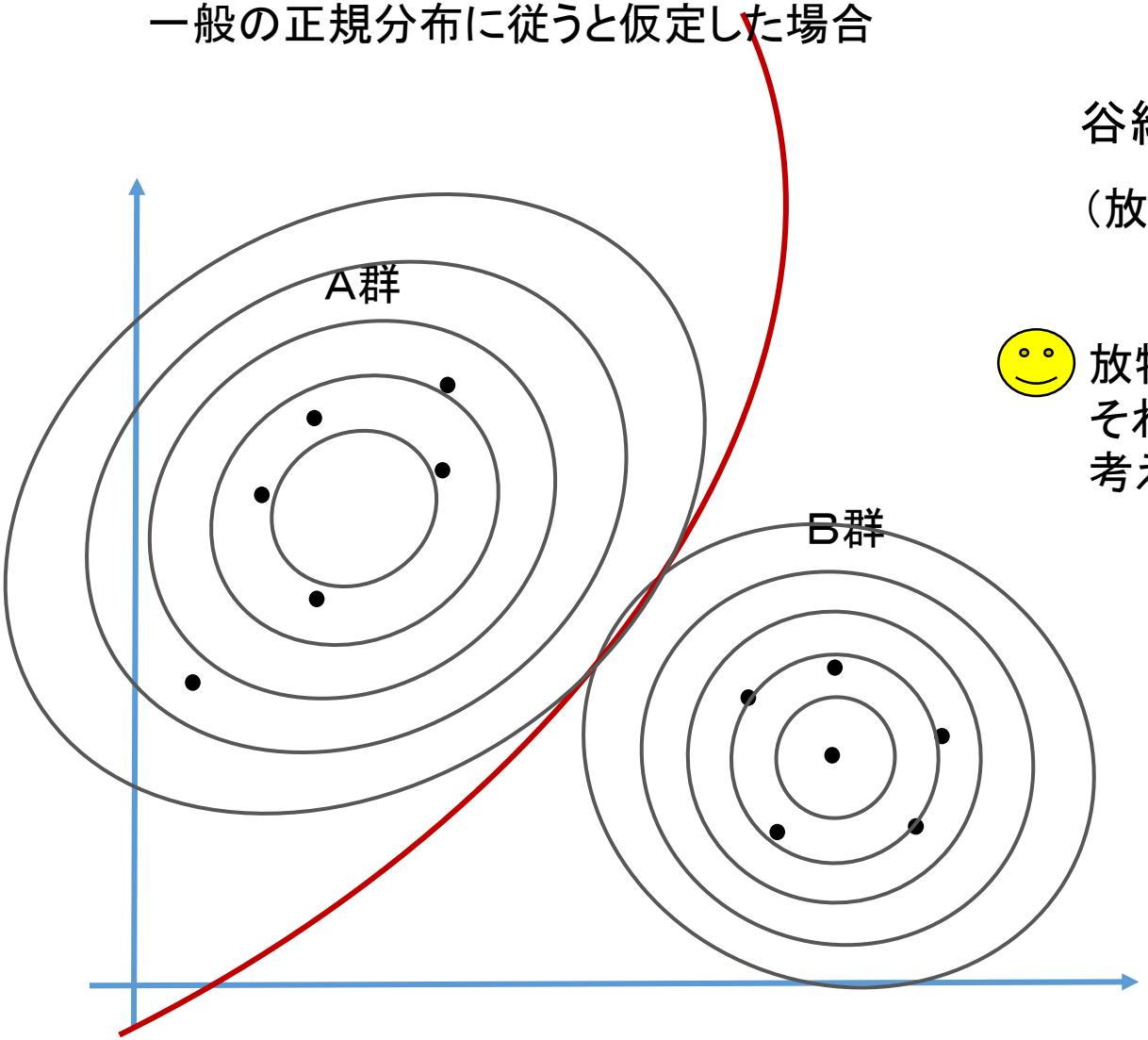
谷線は、確率密度関数が等しいところ

$$\frac{1}{2\pi\sqrt{1-\rho_1^2}\sigma_{x_1}\sigma_{y_1}}\exp\left[-\frac{1}{2(1-\rho_1^2)}\left\{\frac{(x-\mu_{x_1})^2}{\sigma_{x_1}^2}-2\rho\frac{(x-\mu_{x_1})(y-\mu_{y_1})}{\sigma_{x_1}\sigma_{y_1}}+\frac{(x-\mu_{y_1})^2}{\sigma_{y1}^2}\right\}\right]$$

$$=\frac{1}{2\pi\sqrt{1-\rho_2^2}\sigma_{x_2}\sigma_{y_2}}\exp\left[-\frac{1}{2(1-\rho_2^2)}\left\{\frac{(x-\mu_{x_2})^2}{\sigma_{x_2}^2}-2\rho\frac{(x-\mu_{x_2})(y-\mu_{y_2})}{\sigma_{x_2}\sigma_{y_2}}+\frac{(x-\mu_{y_2})^2}{\sigma_{y2}^2}\right\}\right]$$

正規分布モデルによる分類

一般の正規分布に従うと仮定した場合



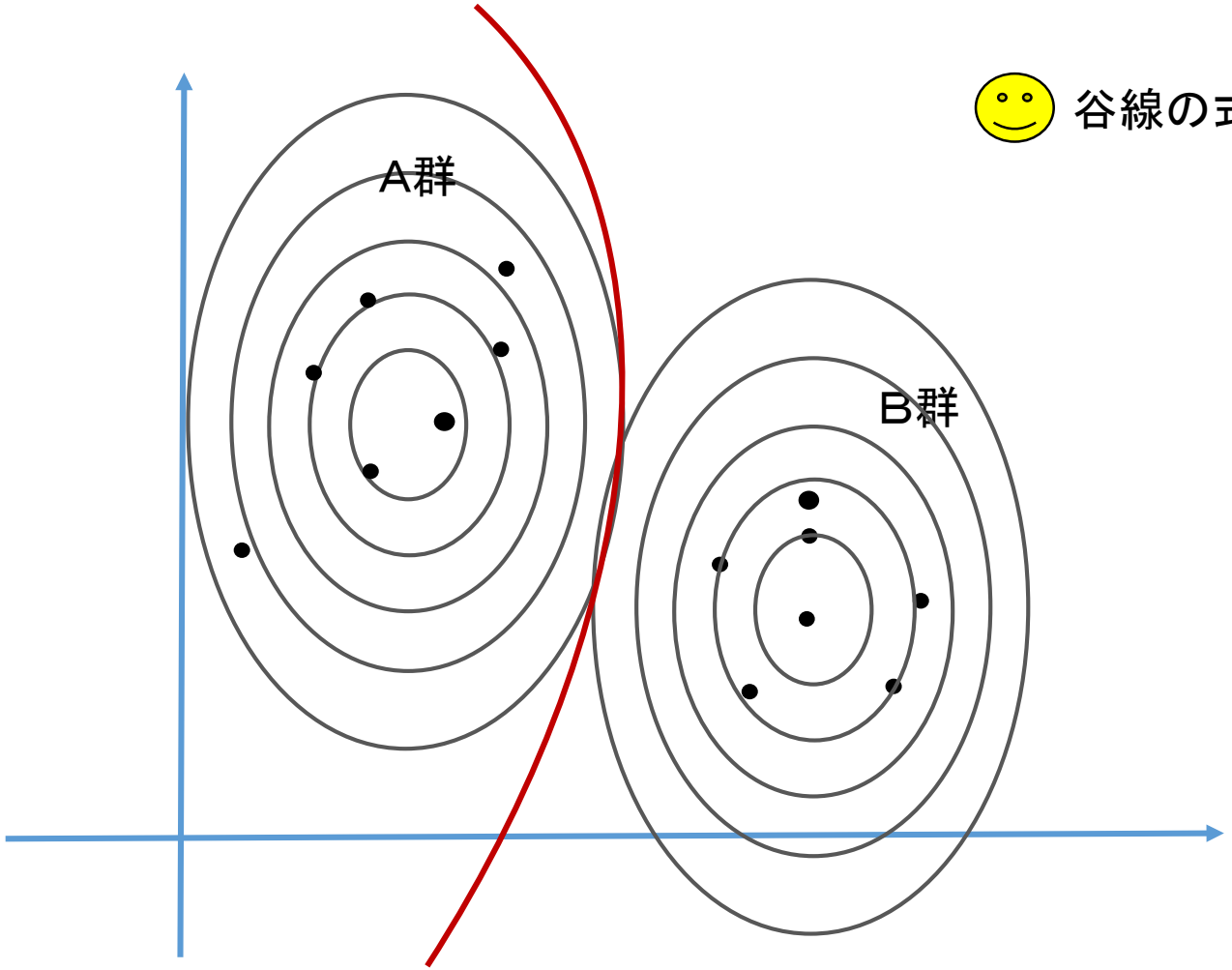
谷線は2次曲線となる

（放物線、楕円、双曲線）

放物線、楕円、双曲線のとき
それぞれどのような状況か
考えてみよう

正規分布モデルによる分類

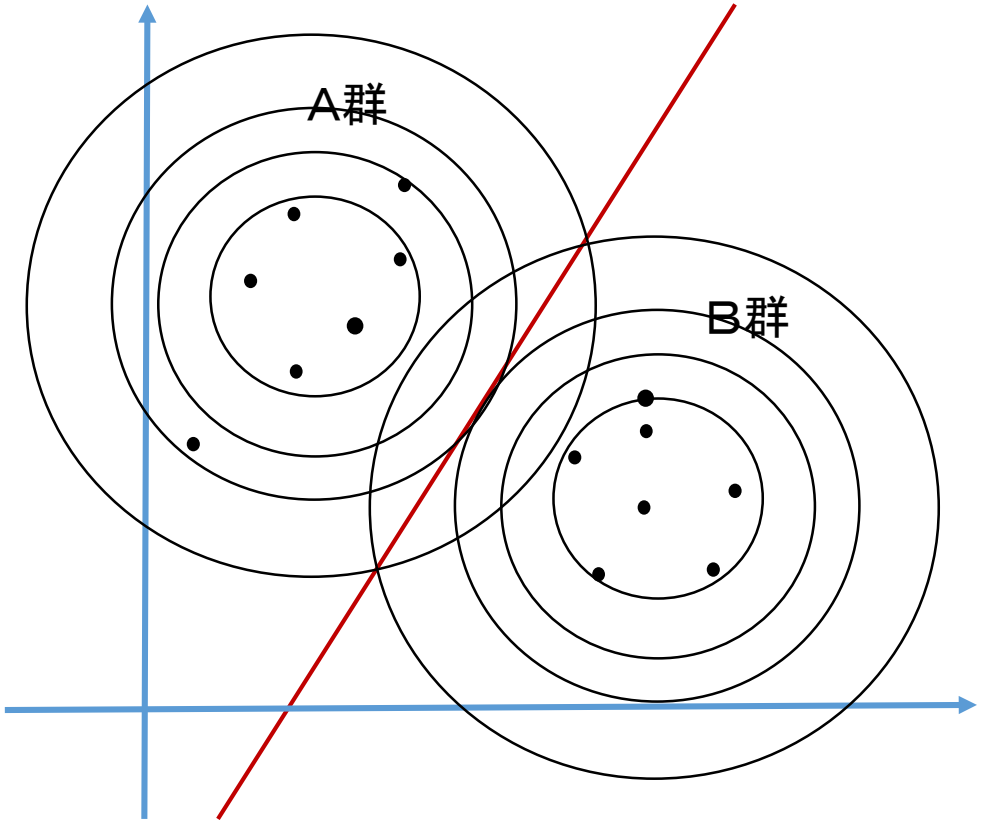
二つの特徴量が独立で、特徴量ごとに別の分散をもつと仮定した場合は
式が少し簡単になる



谷線の式を計算してみよう

正規分布モデルによる分類

仮定を強くすると、直線になる
二つの特徴量が独立で、かつ分散も等しいと仮定した場合



正規分布モデルによる分類

マハラノビスの距離とは

データの分散・共分散を考慮に入れた、グループの重心からの距離
インドの数理統計学者が考案

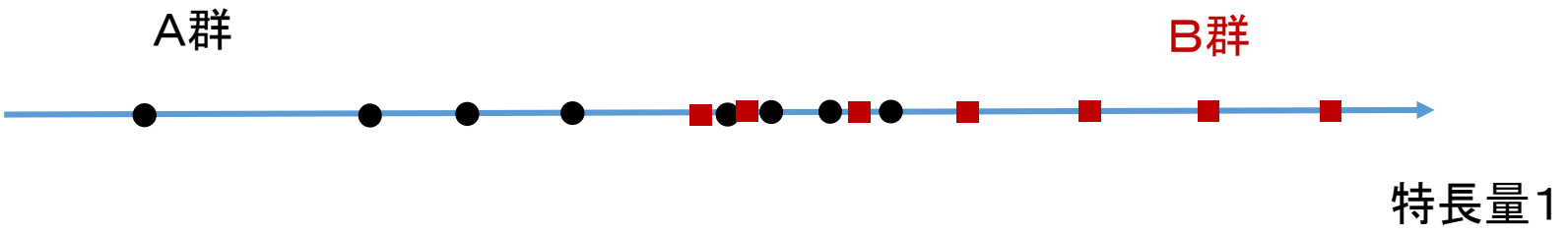
正規分布モデルによる分類は、未知データをマハラノビス距離の近い群に
入れることに相当している。

二次元データ(x, y)と二次元正規分布 $N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ のマハラノビス距離

$$D((x,y), N_2) := \sqrt{(x-\mu_x, y-\mu_y) \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}}$$

$$= \sqrt{\frac{\sigma_x^2(x-\mu_x)^2 + \sigma_y^2(x-\mu_y)^2 - 2\rho\sigma_x\sigma_y(x-\mu_x)(y-\mu_y)}{\sigma_x^2\sigma_y^2(1-\rho)^2}}$$

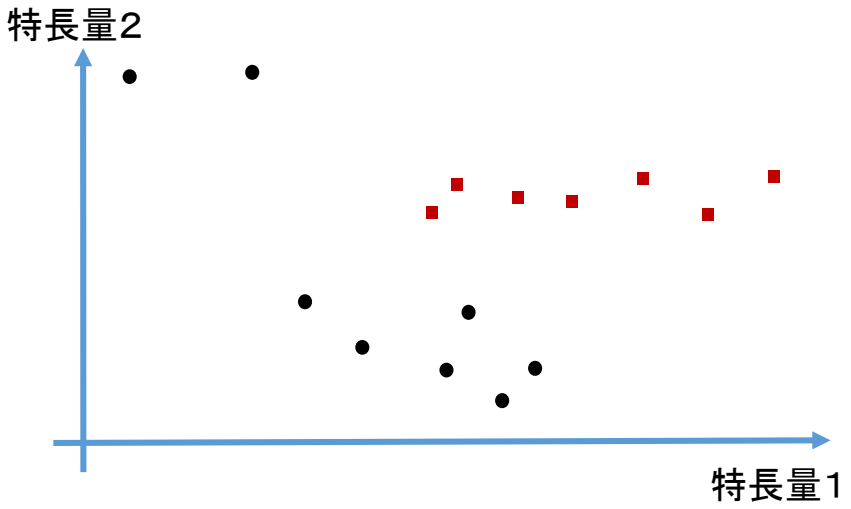
特長量の組み合わせ

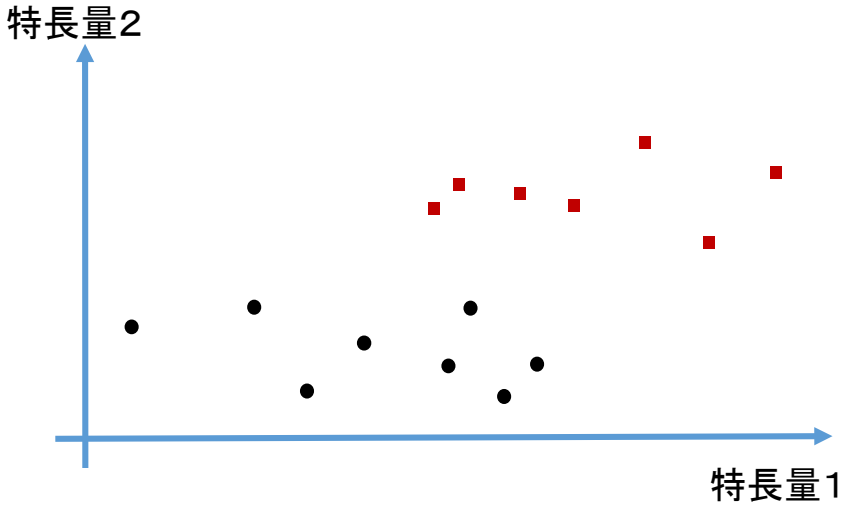


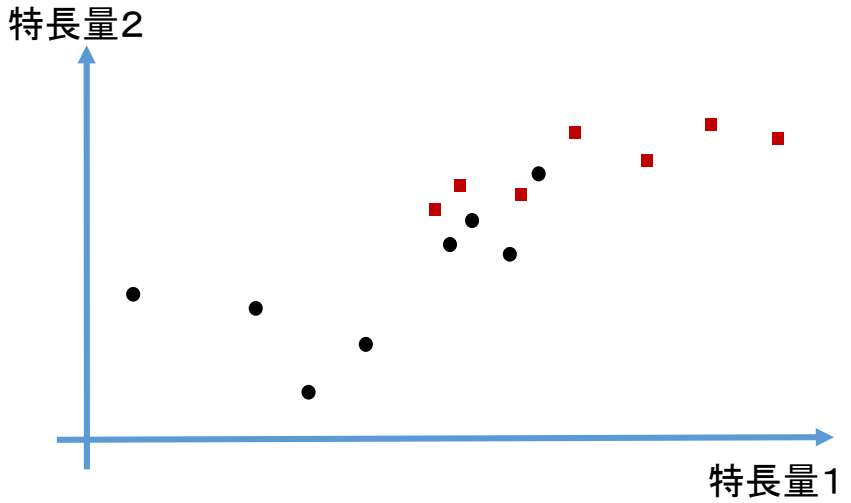
特長量2はどんなものがよいか？

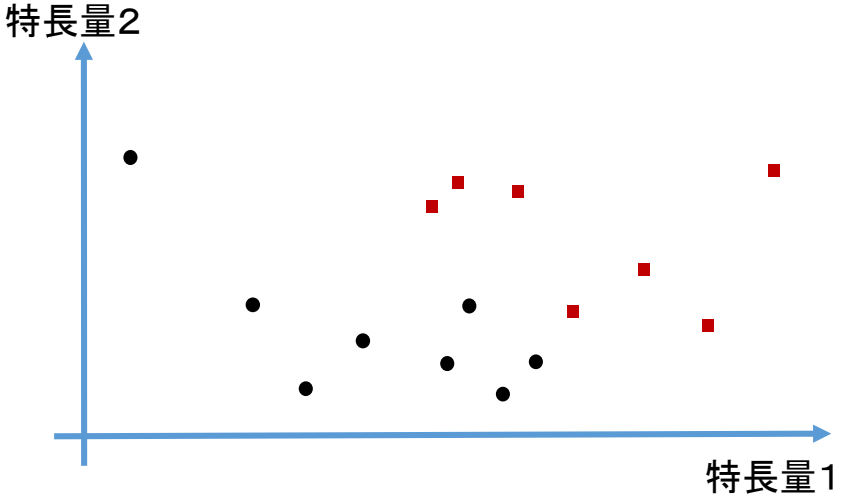
特長量の組み合わせ

どれが一番よい組み合わせでしょう？









判別器の性能測定

作成した判定器の性能は？

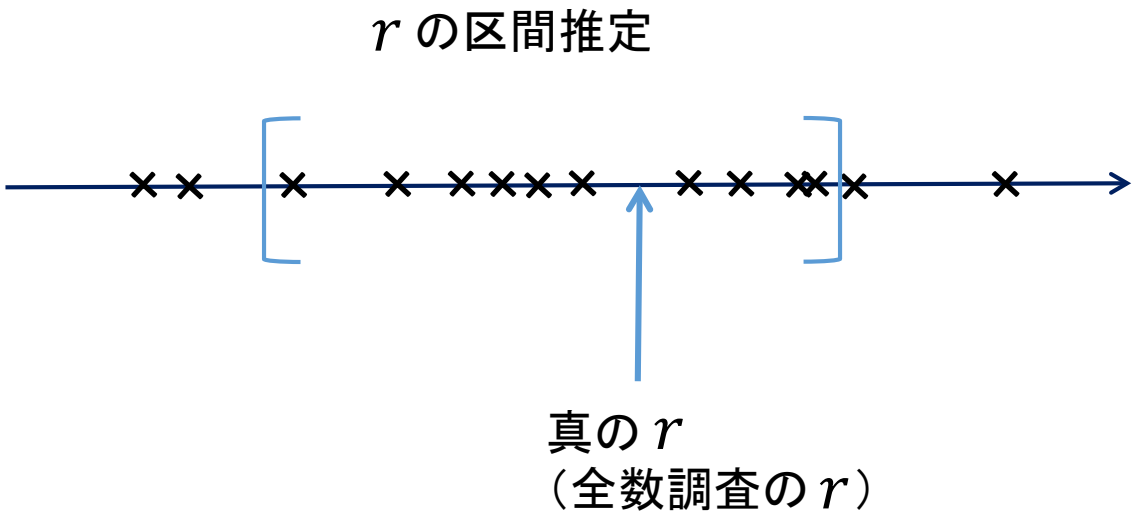
全数調査で得られる認識率が真の性能であるが・・・

世の中の全ての
手書き「B」と「8」を
集める？

それは
無理！

真の認識率に近い認識率を得るために

何度もテストを繰り返し得られた多くの認識率より、真の認識率を推定する



精度のよい推定値を得るためには、適切な量のデータ数と
サンプリング回数が必要

Test sample 法

- 検証データは学習データと異なるものが入っているのが望ましい

データを学習データと検証データに分割して性能を判定する

データ数が十分でないとき、test sample 法では

良い識別をするためには ⇒ 学習データを多く

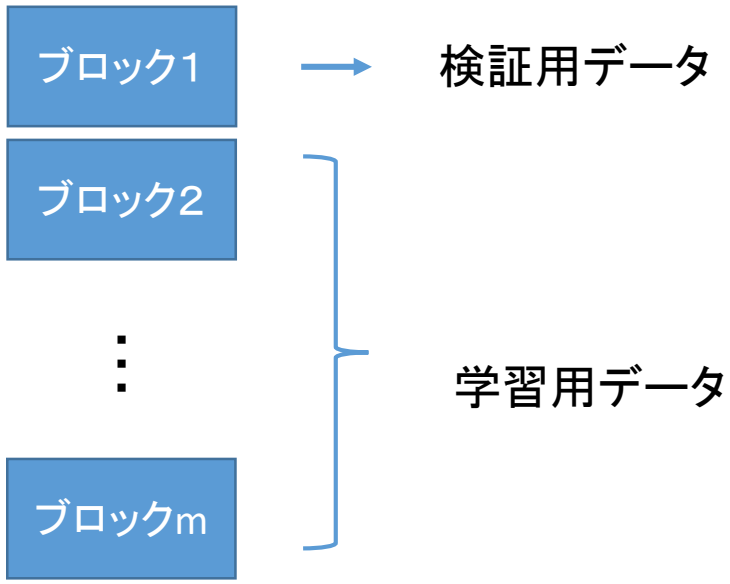
↓

判別性能の誤差が大きくなる ⇐ 検証データが少なくなり

しかも、そもそも1回しかテストできない

Cross-Validation 法

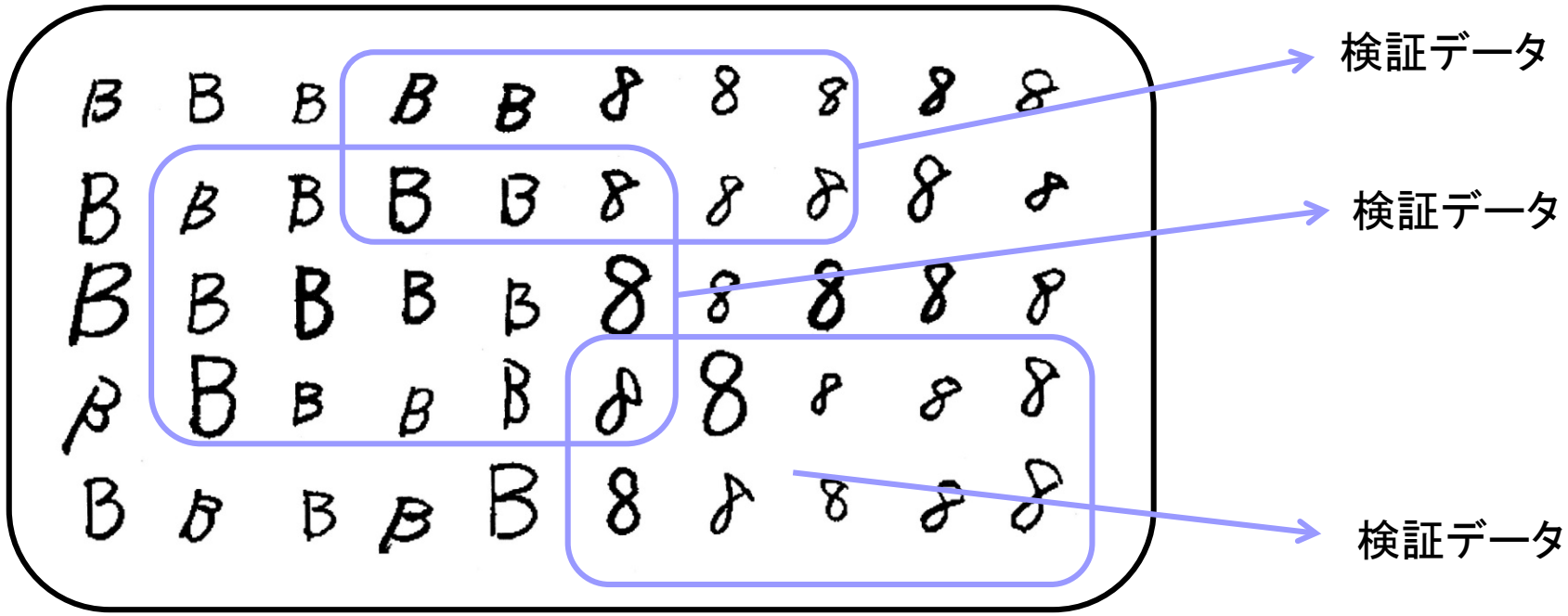
- データをいくつかのブロックに分割
- ブロックの１つを検証データに、残りを学習データに用いる



利点　全てのデータが検証に使われる
ブロックの数だけ検証ができる

Bootstrap法 （ブーツのつまみ革、自動の）

Efron(1979, 1982) により提案された
標本化によって目的とするものの真の値を推定する統計的手法
今回は、検証データの抽出に利用（残りを学習データとする）



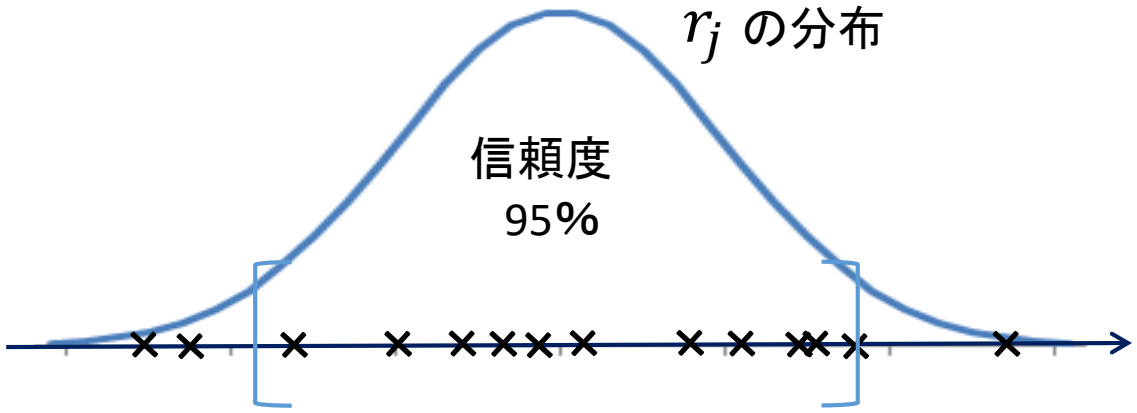
何度でも、学習＆検証が可能　⇒　たくさんのデータ（認識率）を得ることができる

認識率の区間推定　正規分布近似法

正規分布近似法

得られたたくさんの r_j から平均、分散を定め、真の r を区間推定する

r_j の出現は正規分布に従うと仮定している

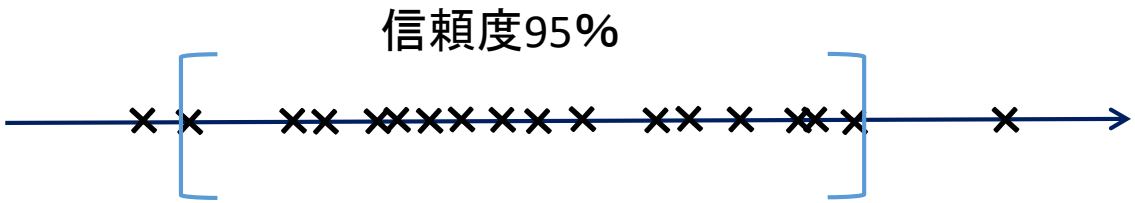


認識率の区間推定　パーセンタイル法

パーセンタイル法

得られたたくさんの r_j のうち両側から何パーセント（信頼度に合わせて）
か外した区間とする

分布を仮定しない（ノンパラメトリックと呼ばれる）



実際には信頼度を高めるために、多数（数千～数万）のデータを取得する。
分布を仮定しなくても、データから分布関数（ヒストグラム）を得ることができる。

区間推定　練習

テキスト「スッキリわかる確率統計」7.4.2節（p.218）を参考に、母平均の区間推定をしてみよう。

使うのはこの定理

定理　7.3 母平均 μ と母分散 σ^2 がともに未知の正規母集団からの無作為標本を $X_1, X_2, \ldots, X_n$ とし、不偏分散を U^2 とすれば、母平均 μ の信頼度 100(1-α)の信頼区間は

$$\overline{X} - t_{n-1}\left(\frac{\alpha}{2}\right)\frac{U}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1}\left(\frac{\alpha}{2}\right)\frac{U}{\sqrt{n}}$$

となる。ただし、$\overline{X}$ は標本平均で、$t_{n-1}(\alpha/2)$は自由度 n-1 の t 分布の上側 α/2 点である。

例題　2018年に生まれた新生児10人の体重（g）は、3470, 2550, 2920, 2530, 3280, 2840, 2520, 3350, 3610, 3430 であった。
2018年に生まれた新生児の平均体重を信頼度95%で区間推定せよ。