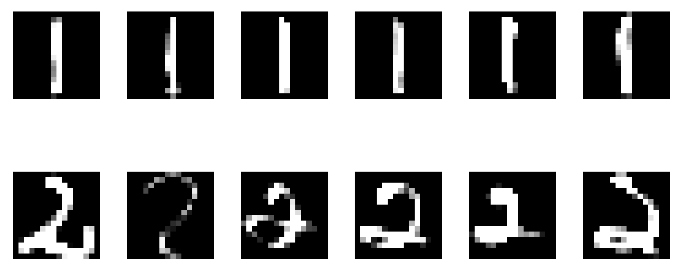
統計的機械学習
第2回　識別関数のよさを測る規準

2013/06/04

表現工学科　尾形研究室　ゼミ
野田　邦昭

手書き文字認識の例

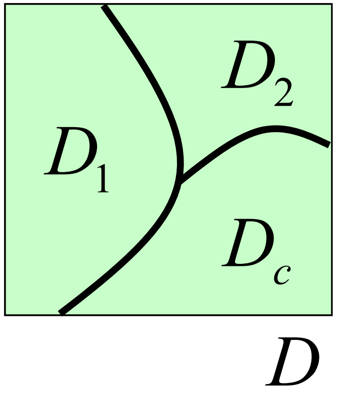
- スキャナで取り込んだ文字画像が16×16画素のとき，パターン x は各画素の濃度を縦に並べた256次元のベクトル．
- 厳密には画素値は実数ではない（例えば8ビット，即ち256階調の離散値）が，$[0,1]$ に正規化した実数値として扱う．
- このとき，パターン空間は $D=[0,1]^{256}$ ．
- カテゴリは各文字に対応．



識別関数･決定領域･決定境界

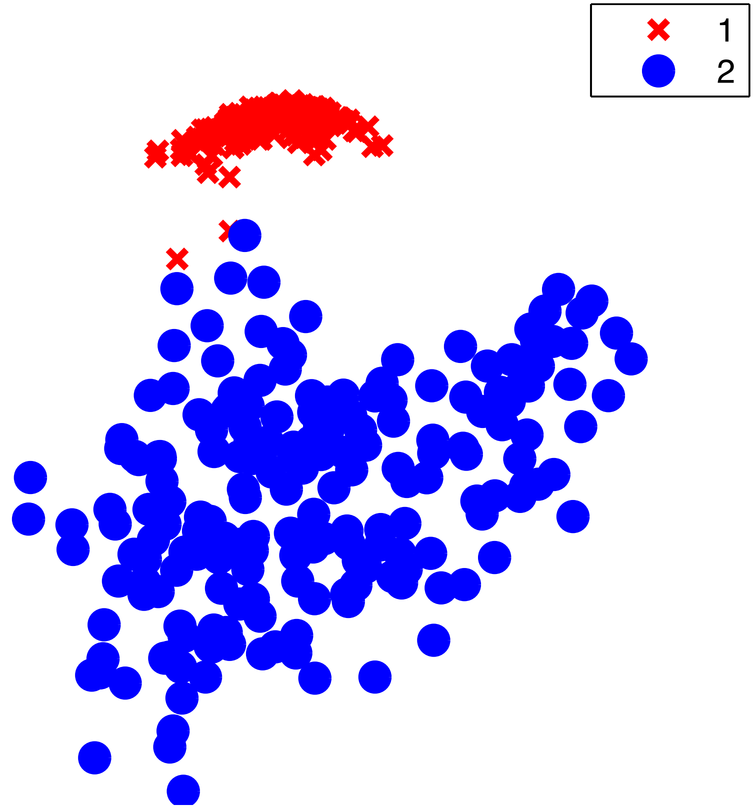
- 識別関数(discrimination function) $f(x)$: パターン xをそれが属するカテゴリ y に対応づける関数
- 決定領域(decision region) D_y : カテゴリ y のパターンが属する領域
- 決定境界(decision boundary): いくつかの決定領域どうしの境界

識別関数を求めること
＝決定領域を求めること
＝決定境界を求めること

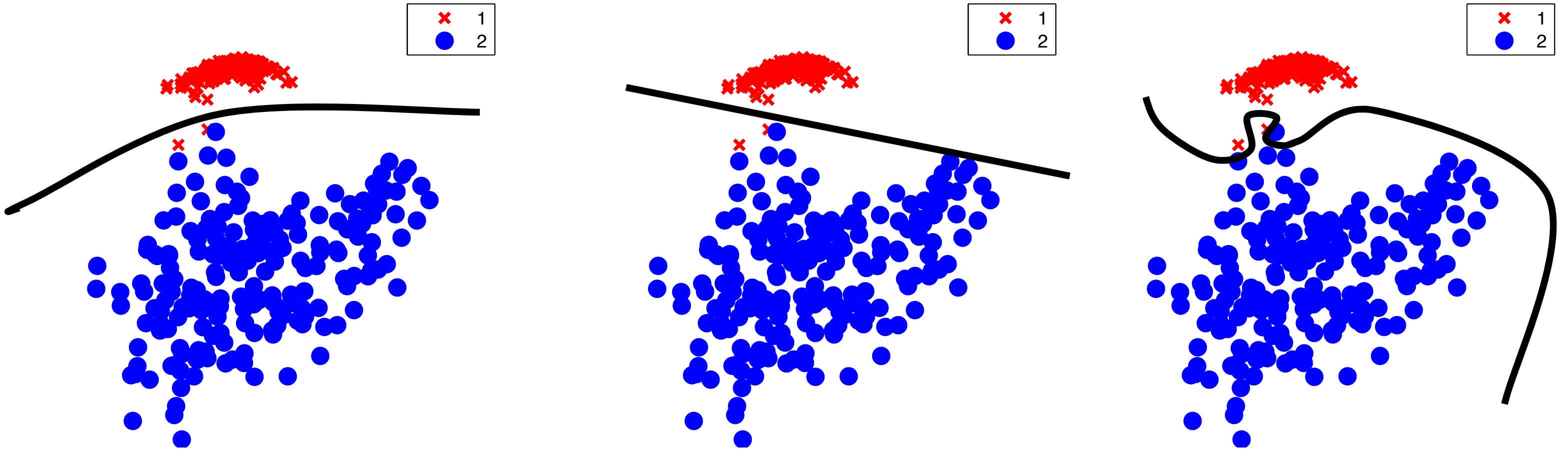


パターンの分布のイメージ

- 256次元空間内に分布しているパターンを適当な2次元部分空間に射影すると



どのような決定境界がよいか？



- 手持ちのパターンだけでなく，未知のパターンも正しく分類できるように，決定境界を定めたい．

識別関数のよさを測る規準

- よい識別関数を構成するためには，まず識別関数の「よさ」を測る規準が必要
    - 最大事後確率則
    - 最小誤識別率則
    - ベイズ決定則

最大事後確率則(1)

- 最大事後確率則(maximum a posteriori probability rule): 入力パターンが属する可能性が最も高いカテゴリを選ぶ

- これは, xを事後確率が最大になるカテゴリに分類することに対応:

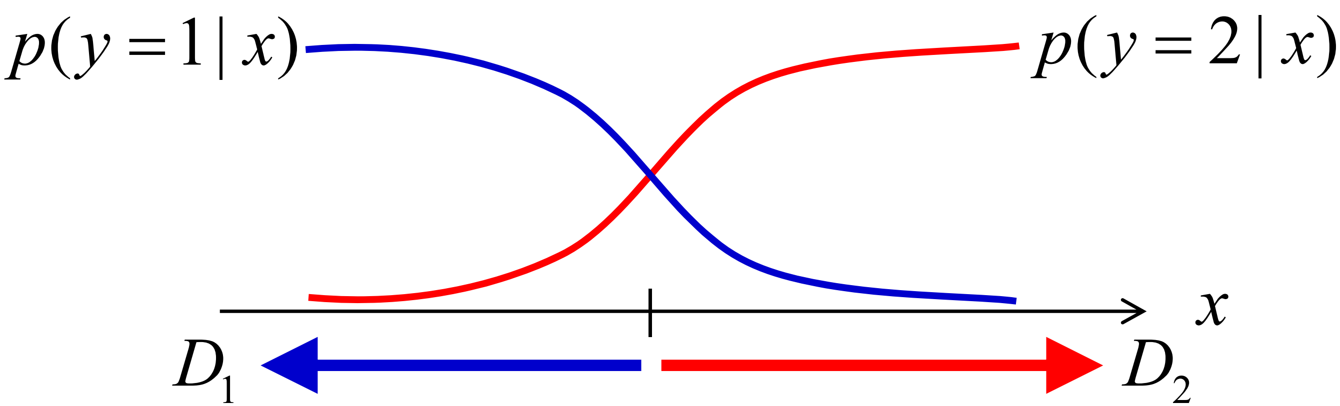
$$\operatorname*{arg\,max}_{y}\ p(y \mid x)$$

最大事後確率則(2)

$$\arg\max_{y} p(y \mid x)$$

■ 決定領域を次のように設定することとも等価:

$$D_y = \{x \mid p(y \mid x) \geq p(y' \mid x) \text{ for all } y' \neq y\}$$

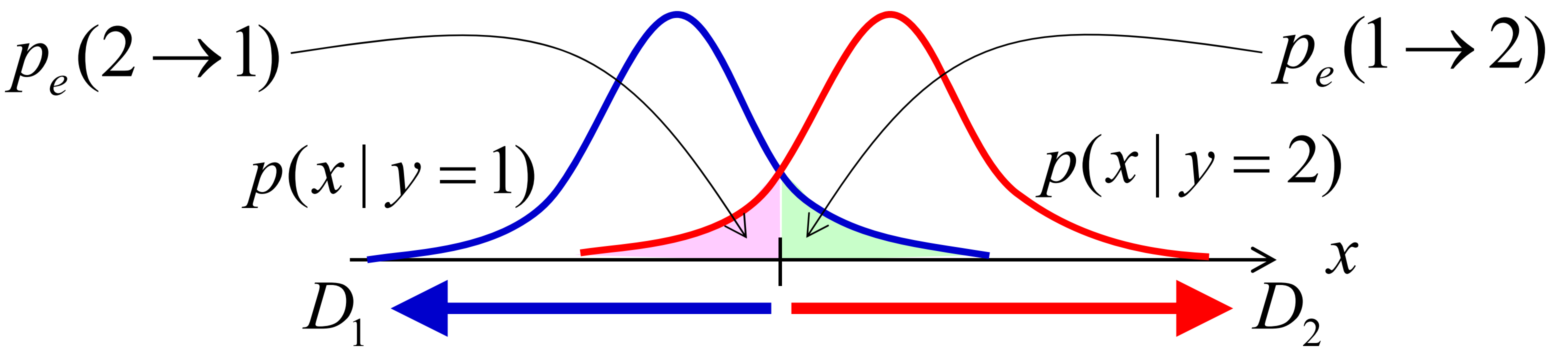


$p(y=1 \mid x) + p(y=2 \mid x) = 1$ (カテゴリ数 $c=2$ と仮定)

最小誤識別率則(1)

- 最小誤識別率則(minimum misclassification rate rule): パターンが誤って分類される確率を最小にするように識別関数を決定
- $p_e(y \to y')$:カテゴリ y に属するパターンが誤ってカテゴリ y' に分類される確率

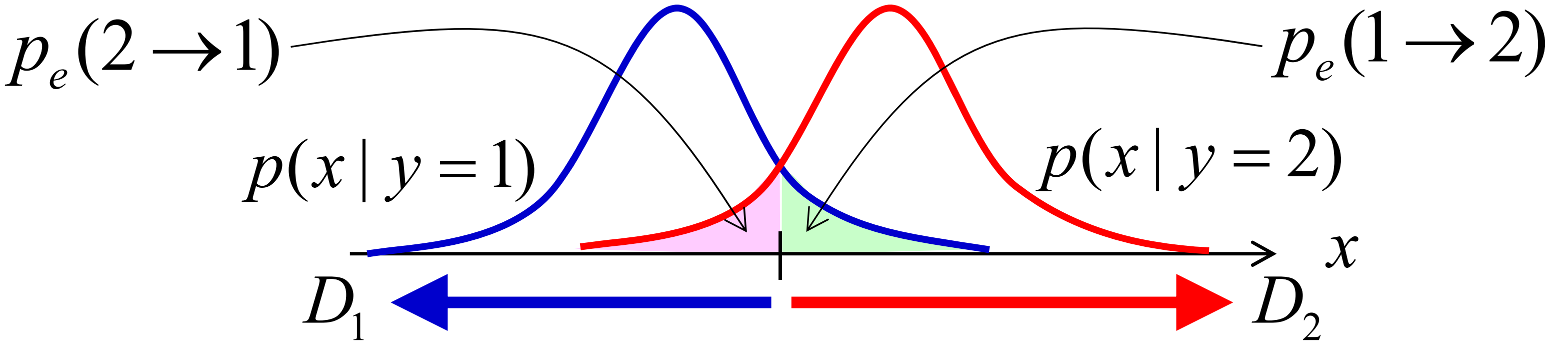
$$p_e(y \to y') = \int_{x \in D_{y'}} p(x \mid y) dx$$



最小誤識別率則(2)

$$p_e(y \to y') = \int_{x \in D_{y'}} p(x \mid y)dx$$

- これは，カテゴリ y に属するパターンが決定領域 $D_{y'}$ に入る確率と等価



最小誤識別率則(3)

- $p_e(y)$:カテゴリ y に属するパターンが誤って他のカテゴリに分類される確率

$$p_e(y) = \sum_{y' \neq y} p_e(y \to y')$$

- これは，以下のように分解できる：

$$\begin{aligned} p_e(y) &= \sum_{y' \neq y} \int_{x \in D_{y'}} p(x \mid y)dx \\ &\qquad + \int_{x \in D_y} p(x \mid y)dx - \int_{x \in D_y} p(x \mid y)dx \\ &= 1 - \underbrace{\int_{x \in D_y} p(x \mid y)dx}_{\text{正解率}} \end{aligned}$$

最小誤識別率則(4)

- 全体の誤識別率 p_e :

  $p_e(y)$ を全カテゴリーに対して平均したもの

$$p_e = \sum_{y=1}^{c} p_e(y) p(y)$$

- 最小誤識別率則では, p_e が最小になるように識別関数を決定する.
- 実は, 最小誤識別率則は最大事後確率則と等価である(証明は宿題).

誤識別と損失

- 最小誤識別率則に従えば，降水確率40%の時は雨が降らないと識別する.
- 雨が降らないならば傘を持っていく必要はないが，多くの人は降水確率40%ならば傘を持っていくであろう.
- それは，傘を持っていかなくて雨が降ったときの損失（雨にぬれて風邪をひく）が，傘を持っていって雨が降らなかったときの損失（かばんが少し重くなる）よりもずっと大きいからである.
- 宿題：他のおもしろい例を考えよ

ベイズ決定則(1)

- ベイズ決定則(Bayes decision rule): 誤って識別した時の損失を最小にするように識別
- $l_{y,y'}$:カテゴリ y に属するパターンを誤ってカテゴリ y' に分類したときの損失(loss)
- 条件付きリスク(conditional risk) $R(y' \mid x)$: パターン x をカテゴリ y' に分類したときの損失の期待値

$$R(y' \mid x) = \sum_{y=1}^{c} l_{y,y'} p(y \mid x)$$

ベイズ決定則（2）

$$R(y' \mid x) = \sum_{y=1}^{c} l_{y,y'} p(y \mid x)$$

- ベイズ決定則では，条件付きリスクが最小になるカテゴリにパターンを分類する

$$\arg\min_{y} R(y \mid x)$$

- これは，決定領域を次のように設定することと等価である.

$$D_y = \{x \mid R(y \mid x) \le R(y' \mid x) \text{ for all } y' \ne y\}$$

ベイズ決定則（3）

- 全リスク(total risk) R : 条件付きリスクの全ての x に関する期待値

$$R = \int_D R(\hat{y} \mid x) p(x) dx$$

但し, $\hat{y}$ は識別機の出力を表す.

- ベイズリスク(Bayes risk): ベイズ決定則に対する全リスクの値

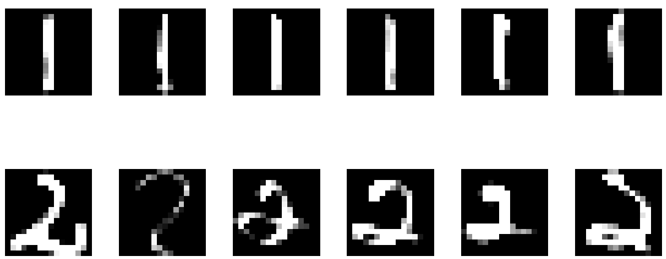
付録

主成分分析に関する補足

主成分分析 (PCA) 1

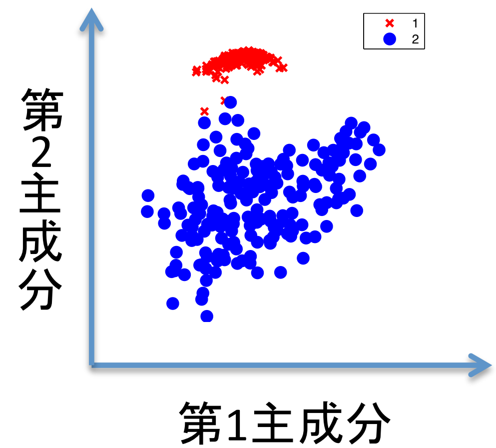
- 多次元のデータを低次元化するときに用いる

256次元



圧縮

2次元

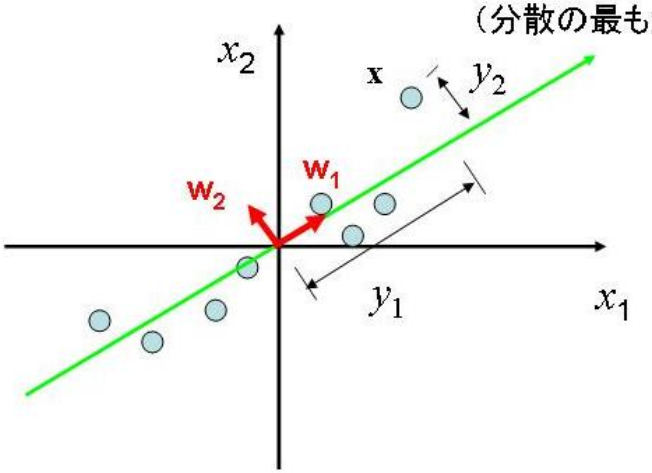


データの大まかな傾向を知ることができる

例 : 1より2の方がパターンのばらつきが大きい

主成分分析 (PCA) 2

- 統計データから互いに無関係（無相関）の成分を取り出して、観測値をそれらの成分の線形結合で説明する



$$\begin{cases} y_1 = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1n}x_n = \mathbf{w}_1^T\mathbf{x} \\ y_2 = w_{21}x_1 + w_{22}x_2 + \cdots + w_{2n}x_n = \mathbf{w}_2^T\mathbf{x} \\ \vdots \\ y_n = w_{n1}x_1 + w_{n2}x_2 + \cdots + w_{nn}x_n = \mathbf{w}_n^T\mathbf{x} \end{cases}$$

ただし，$\mathbf{w}_i^T\mathbf{w}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$　主成分同士は互いに直交し，大きさは1

$$\mathbf{y} = \mathbf{W}^T\mathbf{x}$$

W=[$\mathbf{w}_1$, $\mathbf{w}_2$,…, $\mathbf{w}_n$]

射影先（低次元）　射影元（高次元）

分散が最大となる，互いに独立な方向ベクトル，w1, w2, ..., wn を求めるのがPCA
（理屈上，結果的にそうなる）

主成分分析 (PCA) 3

xj を部分空間 w1, w2, ..., wm (m<=n) へ射影した後，元の空間に復元して得た x^j

$$\hat{\mathbf{x}}_j = y_1\mathbf{w}_1 + y_2\mathbf{w}_2 + \cdots + y_m\mathbf{w}_m = \sum_{i=1}^{m}\mathbf{w}_i^T\mathbf{x}_j\mathbf{w}_i$$

$y_i = w_i^T x_j$

評価関数（サンプル点とその復元との距離） → 最小化

$$E(\mathbf{w}_i) = \sum_{j=1}^{N}\left\|\mathbf{x}_j - \hat{\mathbf{x}}_j\right\|^2 = \sum_{j=1}^{N}\left\|\mathbf{x}_j - \sum_{i=1}^{m}\mathbf{w}_i^T\mathbf{x}_j\mathbf{w}_i\right\|^2$$

$$= \sum_{j=1}^{N}\left\|\mathbf{x}_j\right\|^2 - \sum_{i=1}^{m}\mathbf{w}_i^T\left(\sum_{j=1}^{N}\mathbf{x}_j\mathbf{x}_j^T\right)\mathbf{w}_i = \sum_{j=1}^{N}\left\|\mathbf{x}_j\right\|^2 - \sum_{i=1}^{m}\mathbf{w}_i^T\mathbf{S}\mathbf{w}_i$$

なお，$\mathbf{S} = \sum_{j=1}^{n}\mathbf{x}_j\mathbf{x}_j^T$

サンプルの共分散行列

↓

wiの関数ではない

↓

E(wi)の最小化は，$\sum_{i=1}^{m}\mathbf{w}_i^T\mathbf{S}\mathbf{w}_i$ の最大化と同等

主成分分析 (PCA) 4

拘束条件付き最適化問題

$$\max \sum_{i=1}^{m} \mathbf{w}_i^T \mathbf{S} \mathbf{w}_i \quad subject \ \|\mathbf{w}_i\|^2 = 1 \quad (i = 1, 2, \cdots, m)$$

ラグランジュ関数の定義（ラグランジュの未定乗数法）

$$L(\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m) = \sum_{i=1}^{m} \mathbf{w}_i^T \mathbf{S} \mathbf{w}_i - \sum_{i=1}^{m} \lambda_i (\mathbf{w}_i^T \mathbf{w}_i - 1)$$

wiについて偏微分

$$\frac{1}{2}\frac{\partial L}{\partial \mathbf{w}_1} = \mathbf{S}\mathbf{w}_1 - \lambda_1 \mathbf{w}_1 = 0 \quad \Rightarrow \mathbf{S}\mathbf{w}_1 = \lambda_1 \mathbf{w}_1$$

$$\frac{1}{2}\frac{\partial L}{\partial \mathbf{w}_2} = \mathbf{S}\mathbf{w}_2 - \lambda_2 \mathbf{w}_2 = 0 \quad \Rightarrow \mathbf{S}\mathbf{w}_2 = \lambda_2 \mathbf{w}_2$$

$$\vdots$$

$$\frac{1}{2}\frac{\partial L}{\partial \mathbf{w}_m} = \mathbf{S}\mathbf{w}_m - \lambda_m \mathbf{w}_m = 0 \quad \Rightarrow \mathbf{S}\mathbf{w}_m = \lambda_m \mathbf{w}_m$$

主成分 w1, w2, ..., wm は，サンプルの分散共分散行列 S の固有ベクトルになることがわかる

主成分分析 (PCA) 5

w1, w2, ..., wm は，サンプルの分散共分散行列 S の固有ベクトルになる

主成分は，分散共分散行列に対する固有値分解によって求めることができる

分散が最大となる，互いに独立な方向ベクトルを求めることに等しい