

第9回講義「統計学2」

線形判別分析による

手書き数字の認識(1)

[1] 手書き数字データの読み込み

と表示

[2] 線形判別分析の実装

手書き数字データの読み込み(1)

講義ノートのある下記URL:

<http://cis.k.hosei.ac.jp/~wakahara>

から手書き数字のデータファイル

digit.mat

をダウンロードする。続いて、データを読み込み

```
> load digit.mat
```

whosコマンドを用いて、変数の一覧を確認する。

```
> whos
```

手書き数字データの読み込み(2)

手書き数字データは、3階のテンソル X と T に格納されている。1つの手書き文字データは 16×16 画素のサイズで、各画素の濃淡値はdouble型8bytesで表現され、黒画素は値-1、白画素は値1である。

変数 X には'0'から'9'までの各数字が500文字ずつ含まれており、訓練用データとする。

変数 T には'0'から'9'までの各数字が200文字ずつ含まれており、テスト用データとする。

手書き数字データの表示

訓練用の手書き数字'3'の128番目のデータを変数 x に取り出すときは

```
> x = X(:, 128, 3);
```

とすればよい。ただ、手書き数字'0'のデータを指定するには3番目の引数を10とする。

取り出した手書き文字のデータの画像は、以下のようにして表示することができる。

```
> imagesc(reshape(x, [16, 16]))
```

線形判別分析の実装(1)

手書き数字の‘1’と‘2’を分類する文字識別器を作ってみよう。

ここで、‘1’と‘2’の各カテゴリに属するパターンが、期待値は異なるが、分散共分散行列は等しい、正規分布に従うと仮定する。すなわち、

$$\{\mathbf{x}_i\}_{i:y_i=1} \sim N(\boldsymbol{\mu}_1, \Sigma), \quad \{\mathbf{x}_i\}_{i:y_i=2} \sim N(\boldsymbol{\mu}_2, \Sigma)$$

この条件のもとで、テストパターンが与えられたときの、各カテゴリに属する事後確率を計算する。

線形判別分析の実装(2)

訓練データによる最尤推定量 $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$ を代入すると、それぞれのカテゴリの対数事後確率は次式となる。ただし、事前確率は等しいとした。

$$\log p(y = 1 | \mathbf{x}) = \mathbf{x}^T \hat{\Sigma}^{-1} \hat{\mu}_1 - \frac{1}{2} \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1$$

$$\log p(y = 2 | \mathbf{x}) = \mathbf{x}^T \hat{\Sigma}^{-1} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_2^T \hat{\Sigma}^{-1} \hat{\mu}_2$$

この対数事後確率が大きい方のカテゴリに分類するのがフィッシャーの線形判別分析である。

線形判別分析の実装(3)

期待値の推定:

```
> mu1 = mean(X(:, :, 1), 2); mu2 = mean(X(:, :, 2), 2);
```

分散共分散行列の推定:

```
> S = (cov(X(:, :, 1)') + cov(X(:, :, 2)'))/2; invS = inv(S);
```

対数事後確率の計算:

```
> t = T(:, 1, 2);
```

```
> p1 = t'*invS*mu1-mu1'*invS*mu1/2;
```

```
> p2 = t'*invS*mu2-mu2'*invS*mu2/2;
```

線形判別分析の実装(4)

対数事後確率の大小による2カテゴリ分類:

```
> sign(p1-p2)
```

```
ans = -1
```

この例では、 $p1 < p2$ となり、テストパターンは正しくカテゴリ2(数字'2')に分類される。

上記値が1のときは、 $p1 > p2$ となり、カテゴリ1(数字'1')に分類されたことになる。

線形判別分析の実装(5)

カテゴリ2(数字'2')に属すると分かっているテストパターン全ての識別結果は次のように計算する。

```
> t = T(:, :, 2);
```

```
> p1 = mu1'*invS*t-mu1'*invS*mu1/2;
```

```
> p2 = mu2'*invS*t-mu2'*invS*mu2/2;
```

```
> result = sign(p1-p2);
```

ここで、p1とp2が横ベクトルであることに注意する。

線形判別分析の実装(6)

識別結果をまとめると、次のようになる。

```
> sum(result == -1)
```

```
ans = 198
```

```
> sum(result == 1)
```

```
ans = 2
```

すなわち、200個の手書きの‘2’のテストパターンのうち198個は正しく識別され、正解率は99%である。

練習問題14

上記の設定のもと、カテゴリ1(数字'1')に属すると分かっているテストパターン全ての識別結果を調べなさい。また、誤識別されたパターンを表示して確認しなさい。

ヒント: カテゴリ2に誤識別されたパターンの番号は次のようにして分かる。

```
> find(result ~= 1)
```