データ解析　課題　第５回

概要：　ベイズ判別

提出方法：レポート（紙）とファイル（メール添付）の両方

締め切り：６月３０日（月）の 15:00 **必着（紙とメールの両方）**

1. **レポート（紙）はいつもどおりレポートボックスへ提出**
2. **ファイルはメール添付にて，講義中に説明した「データ解析」のアドレスへ送信．添付するのは，以下で説明する**`"rep5-0412345.rda"`**と**`"rep5-0412345.R"`**です．メールの題名を「データ解析課題第５回** 0412345**」として，二つのファイルを添付します．ただし** 0412345 **を学籍番号におきかえます．**

5.1　正規混合分布

k 個の情報源があり，i 番目の情報源から出る信号 X は m 次元の多変量正規分布 $X \sim N_m(\boldsymbol{\mu}_i, \Sigma_i)$ とする．各 $\boldsymbol{\mu}_i$ は未知の m 次元ベクトル，Σ_i は未知の $m \times m$ 行列である．i 番目の情報源が選ばれる確率を π_i とする．このとき，パラメータ π_i, $\boldsymbol{\mu}_i$, Σ_i, $i = 1, \ldots, k$ を使って，観測信号 X の密度関数（周辺分布）を示せ．

5.2 パラメータ推定

学習用データでは，観測信号 x と情報源の番号 i がペアで観測される．サイズ n の学習用データ $(\boldsymbol{x}_t, i_t)$, $t = 1, \ldots, n$ を使って，モデルのパラメータ π_i, $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, $i = 1, \ldots, k$ を推定する式を示せ．推定したパラメータは $\hat{\pi}_i$, $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\Sigma}}_i$, $i = 1, \ldots, k$ と書くことにする．

5.3 ベイズ判別

新規に信号 x だけが観測されて，情報源の番号 i は観測されなかったとする．このとき，x および，推定したパラメータ $\hat{\pi}_i$, $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\Sigma}}_i$, $i = 1, \ldots, k$ を使って，i の事後確率 $\pi_i(\boldsymbol{x})$ を表す式を示せ．ベイズ判別では，この事後確率の値を最大にする i を予測値として出力する．

5.4 判別境界

上記のモデルでは，二つのクラスの境界が２次式になることを示せ（記号を簡単にするため，パラメータにハットをつけなくてよい）．とくに $\boldsymbol{\Sigma}_i$ が i によらず一定値 $\boldsymbol{\Sigma}$ のとき，その判別境界が１次式になるが，その式を示せ．なお次の数値例ではこの判別境界の式をつかわないほうが実装がラクなので注意．

5.5 ベイズ判別の最適性

情報源 i の信号が必ずしも正規分布ではなく，密度関数 $\boldsymbol{X} \sim f_i(\boldsymbol{x})$ とする．$\pi_i,\ f_i(\boldsymbol{x}),\ i = 1, \ldots, k$ は既知とする．あらゆる判別ルールのなかでベイズ法が正解率を最大にすることを示せ．

5.6 ベイズ判別の数値例

細胞における遺伝子 1 と遺伝子 2 の発現レベルを測定したデータから，その細胞の状態（１＝正常，２＝良性腫瘍あり，３＝悪性腫瘍あり）を分類する．あらかじめ３００個の細胞サンプルで発現レベルと状態を測定した学習用データがある．そこへ１００個の細胞サンプルで発現レベルのみを測定したテストデータを入手した．ベイズ判別をつかって，このテストデータにおける細胞の状態を予測せよ．
R のバイナリ形式ファイル `rep5-question.rda` **には，下記の３個のオブジェクトが含まれる．**

- `dat.train` **サイズ** 300 * 2 **の行列．** `dat.train[t,j]` **は** t **番目の細胞サンプルにおける遺伝子** j **の発現レベル．**
- `class.train` **長さ** 300 **のベクトル．** `class.train[t]` **は** t **番目の細胞サンプルにおける状態．**
- `dat.test` **サイズ** 100 * 2 **の行列．** `dat.test[t,j]` **はテストデータの** t **番目の細胞サンプルにおける遺伝子** j **の発現レベル．**

- 以上の３個がファイルに含まれている．次の１個のファイルは含まれない． class.test 長さ 100 のベクトル．class.test[t] はテストデータの t 番目の細胞サンプルにおける状態．

```
> load("rep5-question.rda") # データファイルの読み込み
> plot(dat.train,log="xy",pch=as.character(class.train),col=class.train+1)
> points(dat.test)
> load("rep5-0412345.rda") # 回答例の読込 (pred.test)
> mean(pred.test == class.test)*100 # 正解率 (%)
[1] 85
```

解答はつぎの手順にしたがうこと．ただし 0412345 を学籍番号におきかえます．

- データをよみこみ，予測を行うスクリプト "rep5-0412345.R"を作成する．
- 結果は長さ 100 のベクトル pred.test　に格納し，バイナリファイル"rep5-0412345.rda"に保存すること．pred.test[t] はテストデータの t 番目の細胞サンプルにおける状態の予測値 (1,2,3 のどれかの値をとる)．
- class.test と pred.test を比較して正解率を成績に反映させます．
- スクリプトの定義と実行の様子がわかるようにコンソール出力を印刷してレポートに含める．
- 二つのファイル "rep5-0412345.R"と"rep5-0412345.rda" はメール添付で提出する．
- もし正規混合分布をつかったベイズ判別を行うのなら，発現レベルをそのまま用いるのではなく，なんらかの変換を前処理として行うと良い．どうしてその変換を選んだのか，根拠を図などで示して説明

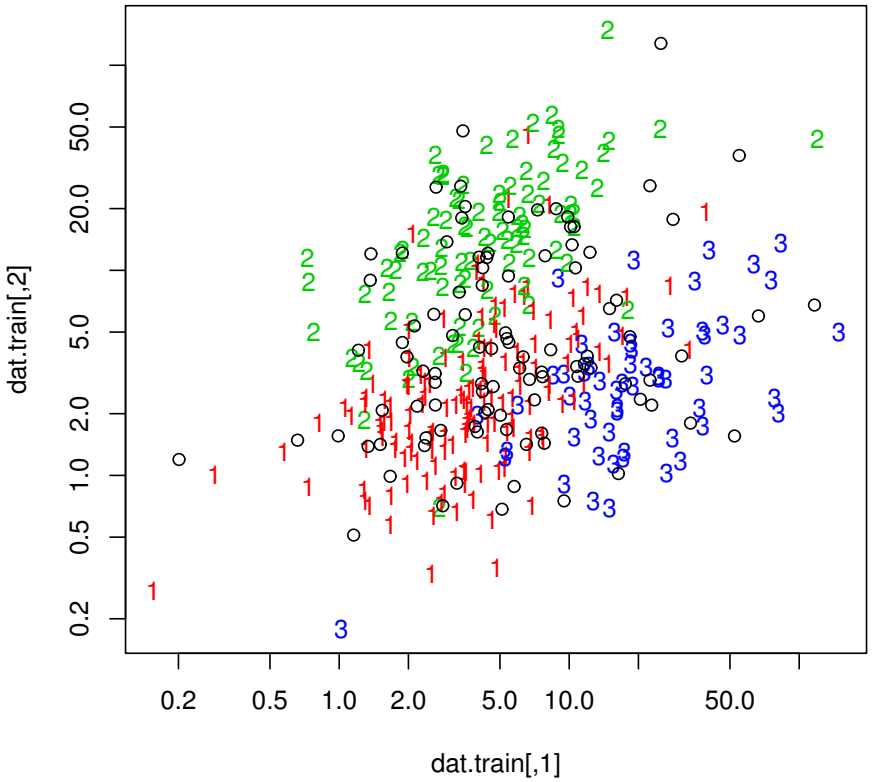

図 1　発現レベルのデータ

せよ．

- **実装に関しては，R に標準で含まれる関数は自由に用いて良い（アドオンパッケージはのぞく）．**