データ解析 応用課題 第２回

「スパムメール判別」を行う．学習用データを利用して判別アルゴリズムのパラメータを調整し，テスト用データで予測を行う．予測結果および作成したプログラムをメールで提出する．

- 提出物は，以下で説明する"rep2-0412345.rda"（予測結果）と"rep2-0412345.R"（プログラム）です．メールの題名を「データ解析 応用課題 第２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．
- 課題のデータファイルは次のようにして読み込めます．

```
> load("rep2-question.rda") # データファイルの読み込み
> dim(dat.train) # 学習用データの次元
[1] 2000   57
> ## 変数は 1:54=単語か文字頻度，55:57 大文字に関する情報 （連の長さの平均，最大値，和）
> length(spam.train) # 学習用データにおける spam=1,ham=0
[1] 2000
> spam.train[1:10] # 最初の 10 個
 [1] 0 1 0 0 1 0 0 1 1 1
> dim(dat.test) # テスト用データの次元
[1] 1000   57
```

- プログラムは次の形式で作成してから実行してください．seed の値は，各自適当に変えてください．

```
## データ解析 応用課題 第２回
```

```
## 東工太郎 0412345
set.seed(123) # 擬似乱数を利用する場合は，かならず seed を明示的に指定する（再現性のため）
load("rep2-question.rda") # データファイルの読み込み
... 中略 ...
... 結果はかならず y という名前のオブジェクトにする ...
save(y,file="rep2-0412345.rda") # 出力ファイル名はかならずこの形式で！
```

- プログラムファイルの名前は rep2-0412345.R の形式にしてください．Windows なら R の「ファイル」-「新しいスクリプト」を選択して R のなかでも作成できます．「編集」-「すべて実行」でプログラムを実行できます．Unix 系ならばプログラムは適当な編集ソフトで作成して，R CMD BATCH rep2-0412345.R などとして実行できます．いずれのプラットフォームでも R の中から source("rep2-0412345.R") で実行できます．
- 「誤判別」を次のように計算して成績に反映させます．y0 が dat.test に対する「真の判別結果」(dat.train に対する spam.train に相当) です．各自，提出前に以下の y0 を y に置き換えてエラーにならないかを確認することを薦めます．なお，y の各要素は spam=1,ham=0 または spam=TRUE,ham=FALSE とします (TRUE=1,FALSE=0 として扱われます)．

```
> load("rep2-0412345.rda")
> if(exists("y") && is.vector(y) && length(y)==length(y0)
+    && sum(y==1)+sum(y==0)==length(y0)) {
+   p0 <- mean(y[y0==0]) # ham メールを spam と判別する確率
+   p1 <- mean(y[y0==1]) # spam メールを spam と判別する確率
+ } else {
+   p0 <- 1; p1 <- 1
+ }
```

```
> p <- p0*50 + (1-p1) # 「誤判別」の点数．この数値が小さいほどよい．
>
> p
[1] 1.192143
```

- **データの出典は** Spambase **データセット**（UCI Repository of machine learning databases http://www.ics.uci.edu/~mlearn/MLRepository.html）**です．これを一部変更したものを出題しています．**
- **提出されたプログラムも成績に反映させます．読みやすいように適宜コメントを入れてください．また，氏名と学籍番号，アルゴリズムの概要や工夫した点などもコメントとして記入してください．講義で説明した方法にとらわれず，創意工夫のあるものを評価します．**