# データ解析　課題　第６回

**概要：　スパム判別**

**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：７月７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**"rep6-0412345.rda"**（復元画像）と**"rep6-0412345.R"**（スクリプト）です．メールの題名を「データ解析課題第６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．**

## 6.1　ロジスティック回帰の対数尤度関数

**データ** $(\boldsymbol{x}_t, y_t)$, $t = 1, \ldots, n$ **を観測したとき，ロジスティック回帰の対数尤度関数** $\log L(\boldsymbol{\beta})$ **を示せ．ただし，** $m$ **次元ベクトル** $\boldsymbol{x}_t$ **は説明変数，**0 **または** 1 **を値にとる** $y_t$ **は目的変数，回帰係数は** $\boldsymbol{\beta} = (\beta_0, \beta_1 \ldots, \beta_m)$ **と**

する．

## 6.2 対数尤度関数の微分

次式を示せ．

$$-\frac{\partial \log L}{\partial \beta_i} = \sum_{t=1}^{n}(p_t - y_t)x_{ti}, \quad -\frac{\partial^2 \log L}{\partial \beta_i \partial \beta_j} = \sum_{t=1}^{n} p_t(1-p_t)x_{ti}x_{tj}$$

## 6.3 ロジスティック回帰の数値例 1

スパムメール判別をロジスティック回帰を利用して行う．R のバイナリファイル `rep6-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 2500 * 57 の行列．`dat.train[t,j]` は $t$ 番目のメールにおける単語 $j$ の頻度．
- `spam.train` 長さ 2500 のベクトル．`spam.train[t]` は $t$ 番目のメールがスパム（$=1$）または非スパム $(=0)$ を表す．
- `dat.test` サイズ 1500 * 57 の行列．`dat.test[t,j]` はテストデータの $t$ 番目のメールにおける単語 $j$ の頻度．
- 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． `spam.test` 長さ 1500 の

ベクトル．spam.test[t] はテストデータの $t$ 番目のメールがスパム $(=1)$ または非スパム $(=0)$ を表す．

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 "rep6-0412345.R"を作成する．ただし，次の課題とまとめて一つのファイルにする．
- 結果は長さ 1500 のベクトル pred.test1　に格納し，バイナリファイル"rep6-0412345.rda"に保存すること．pred.test1[t] はテストデータの $t$ 番目のメールがスパム $(=1)$ または非スパム $(=0)$ を予測したもの．ただし，次の課題で計算する pred.test2 と一緒に同じファイルに保存すること．

  save(pred.test1,pred.test2,file="rep6-0412345.rda")
- spam.test と pred.test1 を比較して正解率を成績に反映させます．スパムを非スパムと判定してしまうエラーよりも，非スパムをスパムと判定してしまうエラーを１０倍の重みをつけてスコアを計算します．
- 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 "rep6-0412345.R"と"rep6-0412345.rda" はメール添付で提出する．
- 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）．

## 6.4　ロジスティック回帰の数値例 2

**まえの課題と同じ設定であるが，１０単語の変数だけを用いて予測せよ．結果は長さ 1500 のベクトル `pred.test2` に格納すること．**

- **どのようにして１０単語を選んだかをレポートで説明する．**
- **選んだ変数の頻度をそのまま使う必要はなく，自由に変換してよい．**

```
> load("rep6-0412345.rda") # 回答例の読み込み: pred.test1 と pred.test2
>
> ## スコア（大きいほどよい）を計算する関数の定義
> calcscore <- function(y,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
+   (p1-p0*10)*100 # ham を spam と間違えてしまうのは深刻なので１０倍する
+ }
>
> calcscore(pred.test1, spam.test) # 数値例１のスコア
[1] 60.72595
> calcscore(pred.test2, spam.test) # 数値例２のスコア（あれ，こっちのほうがよい！？）
[1] 62.94828
```