HW2

0513250 邱郁雯

1. 清洗 Accident.csv 的資料

(1) A1/A2 資料合併成 A2

將 A1/A2 資料合併成一類(A2)，A3 另成一類分析。

(2) 資料不列入

不將編號、天候名稱、光線名稱、路面狀況名稱、當事者性別名稱、車種名稱、保護裝備名稱及飲酒情形名稱列入資料分析內容，因為我認為編號對於分析結果應沒有幫助，只是普通的數字，其他天候、光線、路面狀況等名稱，因為他們有各自對應的代碼，所以不將這些說明列入分析中。

(3) 資料內容改動

因 accident 內的天候代碼與天候名稱對不上，例如天候代碼 8 對應的名稱就有「晴」與「暴雨」，無法做惡劣天氣的分類，因此我將每個代碼都只有對應一個天氣名稱。

天氣代碼	1	2	4	6	7	8
天氣名稱	暴雨	強風	霧或煙	雨	陰	晴

(4) 遺失資料處理

- 一開始匯入遺失的資料筆數：

```
> data<-accident[,c(3,4,5,7,9,11,13,14,16,18,20)]
> sapply(data,num_na)
          GPS經度           GPS緯度         天候代碼         光線代碼
                1                1              129             5797
路面狀況.路面狀態代碼       當事者性別代碼       當事者年齡         車種代碼
             5828                1                1                1
      保護裝備代碼       飲酒情形代碼       事故類別名稱
             5697             5698                0
```

- 將空白的資料以比數最多的資料取代：

找出比數最多者、將資料取代後，確認沒有遺漏的資料。

```
> table(data$天候代碼)#1 最多

   1    2    4    6    7    8
5614    4    1  255  171 3469
> table(data$光線代碼)#1 最多

   1    2    3    4
2778   54  937   77
> table(data$路面狀況.路面狀態代碼)#5 最多

   1    2    3    4    5
   4    4    5  313 3489
> table(data$保護裝備代碼)#3 最多

   1    2    3    4
 690    1 3118  137
> table(data$飲酒情形代碼)#2 最多

   1    2    3    4    5    6    7    8    9   10   11
317 3176    9    7    4    9   11   18   53  279   62
```

```
> data$天候代碼[is.na(data$天候代碼)]<-1
> data$光線代碼[is.na(data$光線代碼)]<-1
> data$路面狀況.路面狀態代碼[is.na(data$路面狀況.路面狀態代碼)]<-5
> data$保護裝備代碼[is.na(data$保護裝備代碼)]<-3
> data$飲酒情形代碼[is.na(data$飲酒情形代碼)]<-2
> sapply(data,num_na)
          GPS經度           GPS緯度         天候代碼         光線代碼
                0                0                0                0
路面狀況.路面狀態代碼       當事者性別代碼       當事者年齡         車種代碼
                0                0                0                0
      保護裝備代碼       飲酒情形代碼       事故類別名稱
                0                0                0
```

- 車種代碼為 factor，所以使用 dummy variable 分析：

```
車種代碼 = factor(data$車種代碼)#車種代碼變dummy
dummies = model.matrix(~車種代碼+0)
dummies<-dummies[,-1]
data1<-cbind(data,dummies)
```

2. 使用 SVM 方法及 Logistics 方法進行 10 次預測，預測目標為「事故類別」。

- 進行 10 次的 Logistics 方法與 SVM 方法預測事故類別：
  在 logistics 中，我認為事故類別裡 A3 的發生次數會較多，所以機率應該較 A1, A2 高，所以當機率大於 0.5 則分類為 A3。

```
for(i in 1:10){
  n<-0.1*nrow(data1)
  index_d<-sample(1:nrow(data1),n)
  train_d<-data1[-index_d,]
  test_d<-data1[index_d,]
  #logit
  model_d<-glm(事故類別名稱~.,
               family = binomial(link = 'logit'),data=train_d)
  table(train_d$事故類別名稱)
  p1<-predict(model_d,test_d,type = "response")
  p1_result<-ifelse(p1>0.5,"A3","A2")
  x<-mean(p1_result==test_d$事故類別名稱)
  accuracy.logit[i]<-x
  #svm
  model.svm<-svm(事故類別名稱~.,data=train_d)
  result02<-predict(model.svm,test_d)
  true_value.svm<-test_d$事故類別名稱
  #table(true_value.svm,result02)
  compare02<-
    ifelse(result02==true_value.svm,1,0)
  y<-sum(compare02)/length(compare02)
  accuracy.svm[i]<-y
}
```

- Logistics 和 SVM 的準確度：

這兩個方法預測事故種類的準確度都高達 95%以上，在圖中基本上 logit 的準確度都比 SVM 高一些，而在第 3,4,7,9 筆資料中，SVM 和 logit 的準確度一樣，所以在圖片中只有一個點。

```
> accuracy.logit
 [1] 0.9730290 0.9636929 0.9512448 0.9678423 0.9678423 0.9616183 0.9564315 0.9564315 0.9636929
[10] 0.9688797
> accuracy.svm
 [1] 0.9719917 0.9626556 0.9512448 0.9678423 0.9668050 0.9626556 0.9564315 0.9543568 0.9636929
[10] 0.9668050
```