Master MAS
Classification supervisé

TP4 : Règle de Bayes avec coûts

Exercice 1. Les objectifs :

- Découvrir le problème des données déséquilibrées.
- Appliquer de manière empirique la règle de Bayes avec matrice de coûts.

On considère dans cet exercice un jeu de données simulées où $n = 1000$ observations (par exemple des entreprises) sont décrites par deux variables quantitatives X^1 et X^2 (par exemple deux ratios financiers) et une variable qualitatative Y à deux modalités (par exemple "saine" et "faillite").

1. Charger les données.

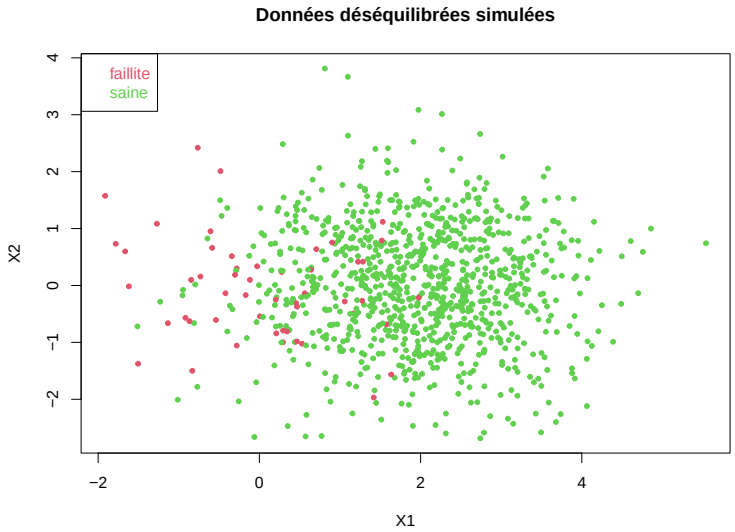
```
load("../data/simu_gauss1.rda")
head(don)
```

```
##       Y    X1      X2
## 1 saine 2.869 -0.6800
## 2 saine 2.173 -0.1594
## 3 saine 2.793  1.6944
## 4 saine 3.240  0.0294
## 5 saine 3.657  1.1312
## 6 saine 0.598  0.1568
```

```
class(don)
```

```
## [1] "data.frame"
```

2. Visualiser ce jeu de données.



3. Quelle est la répartition des classes dans ce jeu de données ?

```
##
## faillite    saine
##    0.051    0.949
```

4. Découper les données en 80% de données d'apprentissage et 20% de données test.

```
# Tirage aléatoire des indices des 800 observations
# de l'échantillon d'apprentissage
set.seed(10) # pour avoir des résultats reproductibles
n <-length(don$Y)
tr <- sample(1:n,800)
Xtrain <- don[tr, -1]
Ytrain <- don$Y[tr]
Xtest <- don[-tr, -1]
Ytest <- don$Y[-tr]
```

   Vérifier que les proportions des deux classes sont bien respectées dans les échantillons d'apprentissage et test.

5. On veut d'abord prédire les classes (saine ou en faillite) des entreprises de l'échantillon test avec la règle (un peu bête) du maximum à priori qui consiste à prédire la classe la plus probable à priori.
   (a) Estimer les **probabilités à priori** par les proportions des classes dans l'échantillon d'apprentissage.
   (b) Comment seront prédites avec la règle du maximum à priori les classes des entreprises l'échantillon test ?
   (c) Quel est le taux d'erreur test de cette règle ?

6. On veut maintenant prédire les classes des entreprises de l'échantillon test avec la règle du maximum à prosteriori (**règle de Bayes avec matrice de coûts 0-1**), qui consiste à prédire la classe la plus probable à priori.
   (a) Utiliser la fonction `lda` pour estimer les **probabilités à posteriori** des 200 entreprises de l'échantillon test.
   (b) Prédire la classe des 200 entreprises de l'échantillon test à partir du vecteur des probabilités à posteriori d'être en faillite et vérifier qu'il s'agit bien prédictions de la méthode `lda`.

   Remarque : La méthode `lda` met en oeuvre la règle de Bayes avec matrice de coût 0-1. En d'autres termes, la méthode `lda` met en oeuvre la règle du maximum à posteriori ou plutôt sa version empirique à partir des probabilités à posteriori estimées par `lda`.

   (c) Quel est le taux d'erreur test de cette règle ? Est-il bien meilleur que celui de la règle du maximum à priori ?
   (d) En considérant que positif="prédire la faillite" trouver **sur les données test** le nombre de vrais positifs (VP), vrais négatifs (VN), faux positifs (FP), faux négatifs (FN).
   (e) En déduire :
   — le taux de vrais positifs : $TVP = \dfrac{VP}{VP+FN}$,
   — le taux de vrais negatifs : $TVN = \dfrac{VN}{FP+VN}$.

Il y a donc uniquement 22% des entreprises en faillites de l'échantillon test qui sont bien classées. En revanche, il y a 100% des entreprises saines de l'échantillon test qui sont bien classées. La classe "saine" est donc mieux prédite que la classe "faillite". En effet, lorsque les classes sont très déséquilibrées (ce qui est le cas ici), la classe la mieux prédite est par construction la plus fréquente. On retrouve couramment ce phénomène pour la recherche d'anomalies par exemple où la classe d'interêt est rarement observée et c'est pourtant elle que l'on veut prédire sans se tromper. Un autre exemple où l'on peut vouloir favoriser le taux de vrais positifs (au détriment du taux d'erreur) est celui de la prédiction de spam. Il est en effet plus grave de mettre un ham (non spam) en spam qu'un spam en ham. C'est également le cas pour les méthodes de diagnostique médical où il est plus grave de ne pas prédire une maladie que l'inverse.

Nous allons voir comment modifier la règle de décision en introduisant des coûts afin de favoriser la prediction d'une classe d'intérêt. Ici, nous allons chercher à mieux prédire les faillites (et forcément moins bien la classe "saine").

7. Afin de mieux pédire la classe "faillite", on applique maintenant la règle du risque minimum à posteriori (**règle de Bayes avec matrice de coûts quelconques**) qui consiste à prédire la classe la moins risquées à posteriori. Pour cela, on considère que le coût de mauvaise classification d'une entreprise en faillite est 5 fois plus important le coût de mauvaise classification d'une entreprise saine.

   (a) En considérant que "1=faillite" et "2=saine", quelle sera alors la matrice de coût

   $$\begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

   où $C_{k\ell}$ est le coût de mauvaise classification d'une donnée de la classe k dans la classe ℓ ?

   (b) Utiliser la fonction `lda` et cette matrice de coûts pour estimer les risques à posteriori des 200 entreprises de l'échantillon test. On rappelle que :

   $$R(1|x) = C_{21}\mathbb{P}(Y = 2|X = x)$$
   $$R(2|x) = C_{12}\Pr(Y = 1|X = x)$$

   (c) Prédire la classe des 200 entreprises de l'échantillon test à partir de ces deux vecteurs de risques à posteriori.

   Quel est le taux d'erreur test de cette règle ? Pourquoi est-ce normal qu'il soit plus grand que celui de la règle du maximum à posteriori ?

   (d) Quelle est la nouvelle matrice de confusion ? Commenter.

   (e) En déduire le taux de vrais positifs et le taux de vrais negatifs ? Comparer aux résultats trouvés avec la règle du maximum à posteriori. Pourquoi ces résultats étaient attendus ?

Exercice 2. On s'intéresse aux jeux de données `infarctus` avec 101 observations et 8 variables :

- FRCAR : Frequence Cardiaque
- INCAR : Index Cardique
- INSYS : Index Systolique
- PRDIA : Pression Diastolique
- PAPUL : Pression Arterielle Pulmonaire
- PVENT : Pression Ventriculaire
- REPUL : Resistance Pulmonaire

— PRONO : Pronostic : une variable qualitative avec deux modalités (DECES, SURVIE)

1. Charger le jeu de données.

```
load("../data/infarctus.rda")
```

2. Quelles sont les variables d'entrées et quelle est la variable de sortie ?
3. Découper les données en 70% de données d'apprentissage (70 données) et 20% de données test (31 données) en fixant la graine à 30 (pour obtenir des résultats reproductibles).
4. Quel est le taux d'erreur test de la règle (un peu bête) du max à priori ?

5. Quel est le taux d'erreur test de la méthode `qda`, le taux de vrais positifs et le taux de vrais négatifs (en considérant que positif=prédire DECES).

6. Introduire maintenant une matrice de coûts pour mieux prédire les décès sans faire trop de faux positifs (cas de SURVIE prédit en DECES).
7. Prédire les données test en utilisant la fonction `qda` et en cette matrice de coût. Calculer le taux d'erreur test, le taux de vrais positifs (test) et le taux de vrais négatifs (test). Vérifier que l'introduction de cette matrice de coût a eu l'effet escompté.