Machine learning

Laurent Rouvière

2021-02-10

Table des matières

Présentation **2**

I Algorithmes de référence **2**

1 Estimation du risque avec caret **2**
1.1 Notion de risque en apprentissage supervisé . . . 2
1.2 La validation croisée . . . 3
1.3 Le package caret . . . 5
1.4 La courbe ROC . . . 7
1.5 Compléments . . . 8

2 Analyse discriminante linéaire **10**
2.1 Prise en main : LDA et QDA sur les iris de Fisher . . . 11
2.2 Un cas avec beaucoup de classes . . . 12
2.3 Grande dimension : reconnaissance de phonèmes . . . 12
2.4 Exercices . . . 13

3 Arbres **14**
3.1 Coupures CART en fonction de la nature des variables . . . 15
3.2 Élagage . . . 18

II Algorithmes avancés **21**

4 Support Vector Machine (SVM) **21**
4.1 Cas séparable . . . 21
4.2 Cas non séparable . . . 23
4.3 L'astuce du noyau . . . 25
4.4 Support vector régression . . . 28
4.5 SVM sur les données spam . . . 29
4.6 Exercices . . . 30

5 Agrégation : forêts aléatoires et gradient boosting **31**
5.1 Forêts aléatoires . . . 32
5.2 Gradient boosting . . . 33

6 Réseaux de neurones avec Keras **35**

7 Données déséquilibrées **36**
7.1 Critères de performance pour données déséquilibrées . . . 36

7.2 Ré-équilibrage 37
7.3 Exercices supplémentaires 40

8 Comparaison d'algorithmes **40**

Présentation

Ce tutoriel présente une introduction au machine learning avec **R**. On pourra trouver :

— les supports de cours associés à ce tutoriel ainsi que les données utilisées à l'adresse suivante https://lrouviere.github.io/machine_learning/ ;
— le tutoriel sans les correction à l'url https://lrouviere.github.io/TUTO_ML/
— le tutoriel avec les corrigés (à certains moment) à l'url https://lrouviere.github.io/TUTO_ML/correction/.

Il est recommandé d'utiliser **mozilla firefox** pour lire le tutoriel.

Les thèmes suivants sont abordés :

— **Estimation du risque**, présentation du package **caret** ;
— **SVM**, cas séparable, non séparable et astuce du noyau ;
— **Arbres**, notamment l'algorithme CART ;
— **Agrégation d'arbres**, forêts aléatoires et gradient boosting ;
— **Réseaux de neurones et introduction au deep learning**, perceptron multicouches avec `keras`.

Il existe de nombreuses références sur le machine learning, la plus connue étant certainement Hastie et al. (2009), disponible en ligne à l'url https://web.stanford.edu/~hastie/ElemStatLearn/. On pourra également consulter Boehmke and Greenwell (2019) qui propose une présentation très claire des algorithmes machine learning avec **R**. Cet ouvrage est également disponible en ligne à l'url https://bradleyboehmke.github.io/HOML/.

Première partie

Algorithmes de référence

1 Estimation du risque avec caret

1.1 Notion de risque en apprentissage supervisé

L'apprentissage supervisé consiste à expliquer ou prédire une sortie $y \in \mathcal{Y}$ par des entrées $x \in \mathcal{X}$ (le plus souvent $\mathcal{X} = \mathbb{R}^p$). Cela revient à trouver un **algorithme** ou **machine** représenté par une fonction

$$f : \mathcal{X} \to \mathcal{Y}$$

qui à une nouvelle observation x associe la prévision $f(x)$. Bien entendu le problème consiste à chercher le **meilleur algorithme** pour le cas d'intérêt. Cette notion nécessite de meilleur algorithme la définition de **critères** que l'on va chercher à optimiser. Les critères sont le plus souvent définis à partir du fonction de perte

$$\begin{aligned} \ell : \mathcal{Y} \times \mathcal{Y} &\mapsto \mathbb{R}^+ \\ (y, y') &\to \ell(y, y') \end{aligned}$$

où $\ell(y, y')$ représentera l'erreur (ou la perte) pour la prévision y' par rapport à l'observation y. Si on représente le phénomène d'intérêt par un couple aléatoire (X, Y) à valeurs dans $\mathcal{X} \times \mathcal{Y}$, on mesurera la performance d'un algorithme f par son risque

$$\mathcal{R}(f) = \mathbf{E}[\ell(Y, f(X))].$$

Trouver le meilleur algorithme revient alors à trouver f qui minimise $\mathcal{R}(f)$. Bien entendu, ce cadre possède une utilité limitée en pratique puisqu'on ne connaît jamais la loi de (X, Y), on ne pourra donc jamais calculé le **vrai risque** d'un algorithme f. Tout le problème va donc être de trouver l'algorithme qui a le plus petit risque à partir de n observations $(x_1, y_1), \ldots, (x_n, y_n)$.

Nous verrons dans les chapitres suivants plusieurs façons de construire des algorithmes mais, dans tous les cas, un algorithme est représenté par une fonction

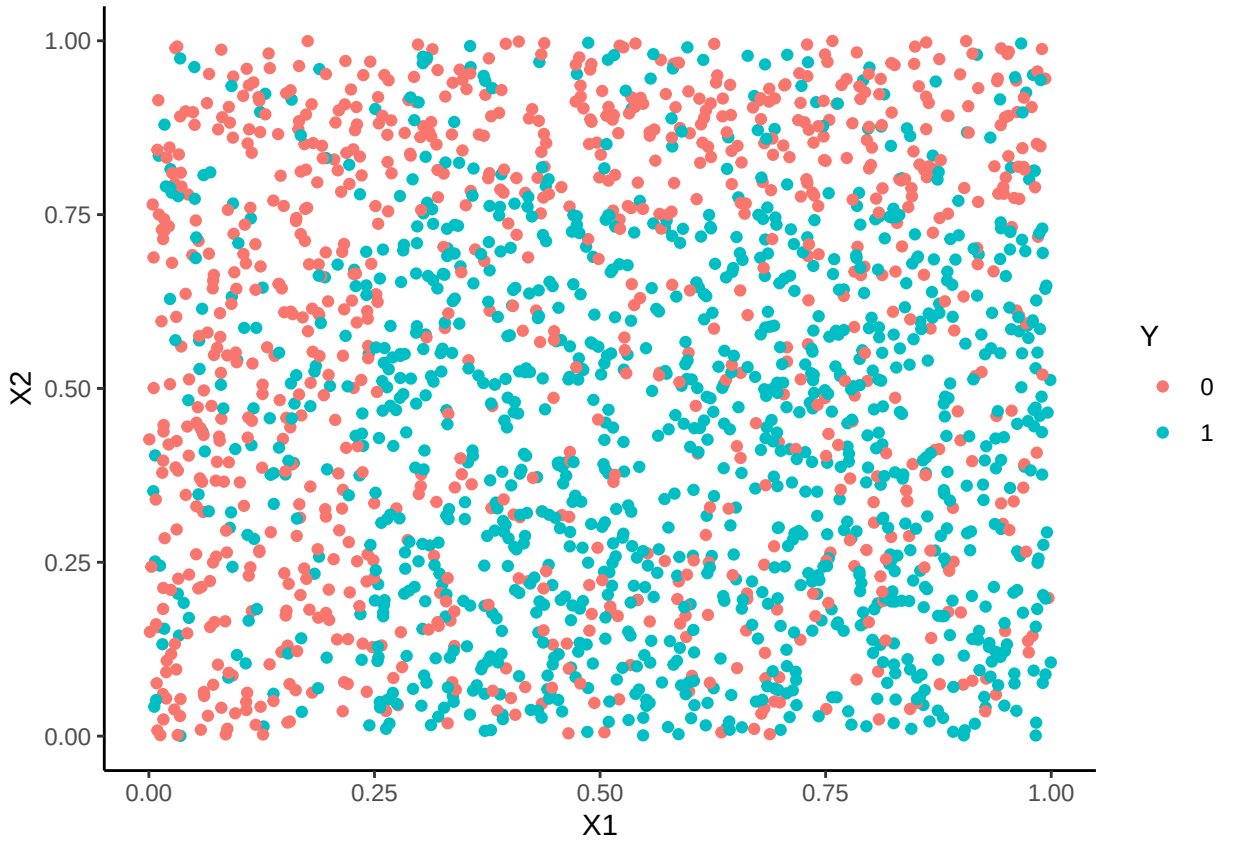
$$f_n : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}$$

qui, pour une nouvelle donnée x, renverra la prévision $f_n(x)$ calculée à partir de l'échantillon qui vit dans $(\mathcal{X} \times \mathcal{Y})^n$. Dès lors la question qui se pose est de calculer (ou plutôt d'estimer) le risque (inconnu) $\mathcal{R}(f_n)$ d'un algorithme f_n. Les techniques classiques reposent sur des algorithmes de type validation croisée. Nous les mettons en œuvre dans cette partie pour un algorithme simple : les k **plus proches voisins**. On commencera par programmer ces techniques "à la main" puis on utilisera le package **caret** qui permet de calculer des risques pour quasiment tous les algorithmes que l'on retrouver en apprentissage supervisé.

1.2 La validation croisée

On cherche à expliquer une variable binaire Y par deux variables quantitatives X_1 et X_2 à l'aide du jeu de données suivant

```
n <- 2000
set.seed(12345)
X1 <- runif(n)
X2 <- runif(n)
set.seed(9012)
R1 <- X1<=0.25
R2 <- (X1>0.25 & X2>=0.75)
R3 <- (X1>0.25 & X2<0.75)
Y <- rep(0,n)
Y[R1] <- rbinom(sum(R1),1,0.25)
Y[R2] <- rbinom(sum(R2),1,0.25)
Y[R3] <- rbinom(sum(R3),1,0.75)
donnees <- data.frame(X1,X2,Y)
donnees$Y <- as.factor(donnees$Y)
ggplot(donnees)+aes(x=X1,y=X2,color=Y)+geom_point()
```



On considère la perte indicatrice : $\ell(y, y') = \mathbf{1}_{y \neq y'}$, le risque d'un algorithme f est donc

$$\mathcal{R}(f) = \mathbf{E}[\mathbf{1}_{Y \neq f(X)}] = \mathbf{P}(Y \neq f(X)),$$

il est appelé **probabilité d'erreur** ou **erreur de classification**.

1. Séparer le jeu de données en un échantillon d'apprentissage **dapp** de taille 1500 et un échantillon test **dtest** de taille 500.
2. On considère la règle de classification des k plus proches voisins. Pour un entier k plus petit que n et un nouvel individu x, cette règle affecte à x le label majoritaire des k plus proches voisins de x. Sur **R** on utilise la fonction **knn** du package **class**. On peut par exemple obtenir les prévisions des individus de l'échantillon test de la règle des 3 plus proches voisins avec

```
library(class)
knn3 <- knn(dapp[,1:2],dtest[,1:2],cl=dapp$Y,k=3)
```

   Calculer l'erreur de classification de la règle des 3 plus proches voisins sur les données test (procédure **validation hold out**).
3. Expliquer la fonction **knn.cv**.
4. Calculer l'erreur de classification de la règle des 3 plus proches voisins par validation croisée **leave-one-out**.
5. On considère le vecteur de plus proches voisins suivant :

```
K_cand <- seq(1,500,by=20)
```

   Sélectionner une valeur de k dans ce vecteur à l'aide d'une **validation hold out** et d'un **leave-one-out** :

   — On calcule l'erreur de classification par **validation hold out** pour chaque valeur de k :

```
err.ho <- rep(0,length(K_cand))
for (i in 1:length(K_cand)){
  ...
  ...
}
```

— On de même chose avec la **validation croisée leave-one-out** :

```
err.loo <- rep(0,length(K_cand))
for (i in 1:length(K_cand)){
  ...
  ...
}
```

6. Faire la même chose à l'aide d'une validation croisée 10 blocs. On pourra construire les blocs avec

```
set.seed(2345)
blocs <- caret::createFolds(1:nrow(donnees),10,returnTrain = TRUE)
```

```
err.cv <- rep(0,length(K_cand))
prev <- donnees$Y
for (i in 1:length(K_cand)){
  for (j in 1:length(blocs)){
...
...
...
  }
  ...
}
K_cand[which.min(err.cv)]
```

1.3 Le package caret

Dans la partie précédente, nous avons utiliser des méthodes de validation croisée pour sélectionner le nombre de voisins dans l'algorithme des plus proches voisins. L'approche revenait à

— estimer un risque pour une grille de valeurs candidates de k
— choisir la valeur de k qui minimise le risque estimé.

Cette pratique est courante en machine learning : on la retrouve fréquemment pour calibrer les algorithmes. Le protocole est toujours le même, pour un méthode donnée il faut spécifier :

— une grille de valeurs pour les paramètres
— un risque
— un algorithme pour estimer le risque.

Le package **caret** permet d'appliquer ce protocole pour plus de 200 algorithmes machine learning. On pourra trouver une documentation complète à cette url http://topepo.github.io/caret/index.html. Deux fonctions sont à utiliser :

— **traincontrol** qui permettra notamment de spécifier l'algorithme pour estimer le risque ainsi que les paramètres de cet algorithme ;
— **train** dans laquelle on renseignera les données, la grille de candidats...

On reprend les données de la partie précédente.

1. Expliquer les sorties des commandes

```
library(caret)
set.seed(321)
ctrl1 <- trainControl(method="LGOCV",number=1)
KK <- data.frame(k=K_cand)
caret.ho <- train(Y~.,data=donnees,method="knn",trControl=ctrl1,tuneGrid=KK)
```

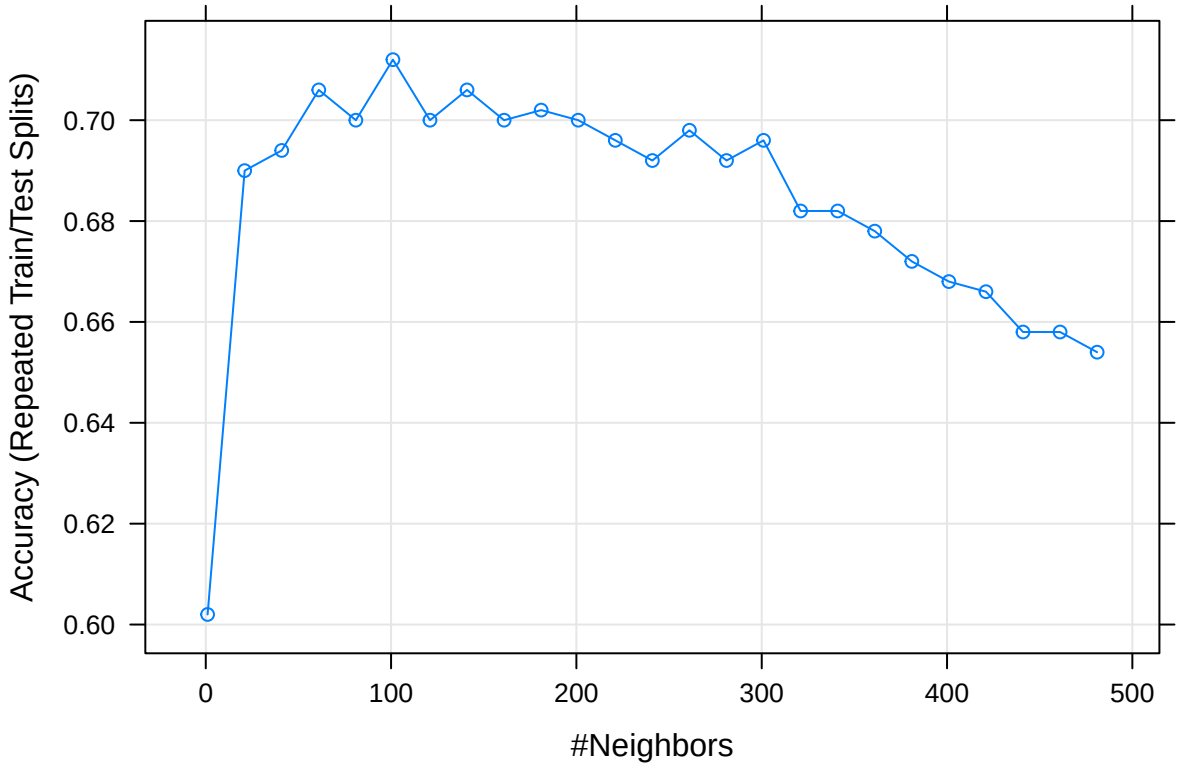
```
caret.ho
k-Nearest Neighbors

2000 samples
   2 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Repeated Train/Test Splits Estimated (1 reps, 75%)
Summary of sample sizes: 1500
Resampling results across tuning parameters:

  k    Accuracy  Kappa
    1  0.602     0.1956346
   21  0.690     0.3649415
   41  0.694     0.3736696
   61  0.706     0.3992546
   81  0.700     0.3867338
  101  0.712     0.4122641
  121  0.700     0.3882944
  141  0.706     0.4017971
  161  0.700     0.3903629
  181  0.702     0.3941710
  201  0.700     0.3898471
  221  0.696     0.3806637
  241  0.692     0.3714491
  261  0.698     0.3829078
  281  0.692     0.3693074
  301  0.696     0.3764358
  321  0.682     0.3474407
  341  0.682     0.3468831
  361  0.678     0.3352601
  381  0.672     0.3214167
  401  0.668     0.3113633
  421  0.666     0.3057172
  441  0.658     0.2853800
  461  0.658     0.2841354
  481  0.654     0.2732314

Accuracy was used to select the optimal model using
 the largest value.
The final value used for the model was k = 101.
plot(caret.ho)
```



2. En modifiant les paramètres du code précédent, retrouver les résultats de la validation hold out de la partie précédente. On pourra utiliser l'option `index` dans la fonction **trainControl**.
3. Utiliser **caret** pour sélectionner k par validation croisée leave-one-out.
4. Faire de même pour la validation croisée 10 blocs en gardant les mêmes blocs que dans la partie précédente.

1.4 La courbe ROC

C'est un critère fréquemment utilisé pour mesurer la performance d'un **score**. Etant donné (X, Y) un couple aléatoire à valeurs dans $\mathcal{X} \times \{-1, 1\}$, on rappelle qu'un score est une fonction $S : \mathcal{X} \to \mathbb{R}$. Dans la plupart des cas, un score s'obtient en estimant la probabilité $\mathbf{P}(Y = 1|X = x)$. Pour un seuil $s \in \mathbb{R}$ fixé, un score possède deux types d'erreur

$$\alpha(s) = \mathbf{P}(S(X) \geq s|Y = -1) \quad \text{et} \quad \beta(s) = \mathbf{P}(S(X) < s|Y = 1).$$

La courbe ROC est la **courbe paramétrée** définie par :

$$\begin{cases} x(s) = \alpha(s) = \mathbf{P}(S(X) > s|Y = -1) \\ y(s) = 1 - \beta(s) = \mathbf{P}(S(X) \geq s|Y = 1) \end{cases}$$

Elle permet donc de visualiser sur une seul graphe 2D ces deux erreurs pour toutes les valeurs de seuil s.

Exercice 1.1 (Étude de la courbe ROC)**.**

On considère dans cet exercice une fonction de score S que l'on suppose absolument continue.

1. Montrer que la courbe ROC vit dans le carré $[0, 1]^2$.
2. On suppose dans cette question que S est **parfait**, ce qui revient à dire qu'il sépare parfaitement les 2 groupes. Mathématiquement on traduit cela par l'existence d'un seuil $s^\star \in \mathbb{R}$ tel que

   $$\mathbf{P}(Y = 1|S(X) \geq s^\star) = 1 \quad \text{et} \quad \mathbf{P}(Y = -1|S(X) < s^\star) = 1.$$

   Analyser la courbe ROC du score parfait.

3. On suppose dans cette question que S est **aléatoire** dans le sens où $S(X)$ est indépendante de Y (cela revient à dire que les notes $S(X)$ n'ont aucun lien avec le groupe). Analyser la courbe ROC d'un tel score.

Exercice 1.2 (Courbe ROC avec R)**.**

On dispose de 4 fonctions de score $S_j(x)$ dont on souhaite visualiser les courbes ROC à partir des valeurs de score calculés sur un échantillon. On trouvera dans le tableau `df` les scores $S_j(X_i), i = 1, \ldots, n$ ainsi que les observations des groupes Y_i

```
set.seed(12345)
n <- 200
Y <- rbinom(n,1,0.5)
S1 <- runif(n)
S2 <- S1
S2[Y==1] <- runif(sum(Y==1),0.6,1)
S2[Y==0] <- runif(sum(Y==0),0,0.6)
S3 <- S2
S3[Y==1][1:10] <- runif(10,0,0.6)
S3[Y==0][1:10] <- runif(10,0.6,1)
df <- data.frame(S1,S2,S3,Y=Y)
head(df)
         S1        S2          S3 Y
1 0.5885923 0.6301922 0.419557838 1
2 0.8925918 0.7897537 0.006244182 1
3 0.1237949 0.7058358 0.302344248 1
4 0.5133090 0.6922984 0.438635526 1
5 0.6636402 0.2034380 0.806095919 0
6 0.7655420 0.4036820 0.665537030 0
```

1. Visualiser, pour chaque score, les valeurs de score en fonction de Y. Commenter
2. Visualiser sur un même graphe les trois courbes ROC. On pourra utiliser d'abord utiliser la fonction **roc** du package **pROC** puis la fonction **geom_roc** du **plotROC**.
3. Calculer les `AUC` à l'aide de la fonction **auc**.
4. On rappelle que l'AUC vérifier la propriété suivante : si (X_1, Y_1) et (X_2, Y_2) sont indépendantes et de même loi que (X, Y), on a

$$AUC(S) = \mathbf{P}(S(X_1) \geq S(X_2) | (Y_1, Y_2) = (1, -1)).$$

Utiliser cette propriété pour retrouver l'AUC de S3.

1.5 Compléments

1.5.1 Calcul parallèle

Les validations croisées peuvent se révéler coûteuses en temps de calcul. On utilise souvent des techniques de parallélisation pour améliorer les performances computationnelles. Ces techniques sont relativement facile à mettre en œuvre avec **caret**, on peut par exemple utiliser la librairie **doParallel** pour utiliser plusieurs cœurs de la machine. On compare les temps de calculs pour une même validation croisée 10 blocs exécutée avec 1 cœur et 4 cœurs :

```
library(doParallel)
ctrl4 <- trainControl(method="cv",index=blocs)
cl <- makePSOCKcluster(1)
registerDoParallel(cl)
```

```
temps1 <- system.time(ee3 <- train(Y~.,data=donnees,method="knn",trControl=ctrl4,tuneGrid=KK))
stopCluster(cl)
cl <- makePSOCKcluster(4)
registerDoParallel(cl)
temps4 <- system.time(ee3 <- train(Y~.,data=donnees,method="knn",trControl=ctrl4,tuneGrid=KK))
stopCluster(cl)
```

On compare ces deux temps de calcul

```
temps1
   user  system elapsed
 12.985   0.062  13.124
temps4
   user  system elapsed
  0.625   0.018   5.935
```

Sans surprise, l'exécution est beaucoup plus rapide avec 4 cœurs.

1.5.2 Répéter les méthodes de rééchantillonnage

Les méthodes d'estimation du risque présentées dans cette partie (hold out, validation croisée) sont basées sur du **rééchantillonnage**. Elles peuvent se révéler sensible à la manière de couper l'échantillon. C'est pourquoi il est recommandé de les **répéter plusieurs fois** et de moyenner les erreurs sur les répétitions. Ces répétitions sont très faciles à mettre en œuvre avec **caret**, par exemple pour

— la validation hold out on utilise l'option `number` :

```
ctrl <- trainControl(method="LGOCV",number=5)
caret.ho.rep <- train(Y~.,data=donnees,method="knn",trControl=ctrl,tuneGrid=KK)
```

— la validation croisée on utilise les options `repeatedcv` et `repeats` :

```
ctrl <- trainControl(method="repeatedcv",repeats=5)
caret.ho.rep <- train(Y~.,data=donnees,method="knn",trControl=ctrl,tuneGrid=KK)
```

1.5.3 Modifier le risque

Enfin nous avons uniquement considéré l'erreur de classification. Il est bien entendu possible d'utiliser d'autres **risques** pour évaluer les performances. C'est l'option `metric` de la fonction **train** qui permet généralement de spécifier le risque, si on est par exemple intéressé par l'**aire sur la courbe ROC (AUC)** on fera :

```
donnees1 <- donnees
names(donnees1)[3] <- "Class"
levels(donnees1$Class) <- c("G0","G1")
ctrl <- trainControl(method="LGOCV",number=1,classProbs=TRUE,summary=twoClassSummary)
caret.auc <- train(Class~.,data=donnees1,method="knn",trControl=ctrl,tuneGrid=KK,metric="ROC")
```

On obtient ici pour chaque valeur de k, l'AUC ainsi que les sensibilité et spécificité :

```
caret.auc
k-Nearest Neighbors

2000 samples
   2 predictor
```

```
   2 classes: 'G0', 'G1'

No pre-processing
Resampling: Repeated Train/Test Splits Estimated (1 reps, 75%)
Summary of sample sizes: 1500
Resampling results across tuning parameters:

  k    ROC        Sens       Spec
    1  0.6338315  0.5937500  0.6739130
   21  0.7488031  0.6473214  0.8152174
   41  0.7599072  0.6696429  0.8115942
   61  0.7554590  0.6785714  0.8188406
   81  0.7586698  0.6875000  0.8224638
  101  0.7628025  0.6785714  0.8333333
  121  0.7603196  0.6830357  0.8333333
  141  0.7605703  0.6830357  0.8297101
  161  0.7625194  0.6875000  0.8224638
  181  0.7616945  0.6875000  0.8188406
  201  0.7609990  0.6830357  0.8188406
  221  0.7582411  0.6696429  0.8188406
  241  0.7567854  0.6607143  0.8224638
  261  0.7563406  0.6473214  0.8188406
  281  0.7546260  0.6383929  0.8297101
  301  0.7530328  0.6294643  0.8333333
  321  0.7554914  0.6205357  0.8297101
  341  0.7530166  0.6205357  0.8369565
  361  0.7518925  0.5848214  0.8405797
  381  0.7500970  0.5357143  0.8550725
  401  0.7472179  0.5133929  0.8586957
  421  0.7472907  0.4866071  0.8695652
  441  0.7432550  0.4732143  0.8768116
  461  0.7429720  0.4598214  0.8840580
  481  0.7404487  0.4241071  0.8876812

ROC was used to select the optimal model using the
 largest value.
The final value used for the model was k = 101.
```

Et on choisira la valeur de k qui maximise l'AUC :

```
caret.auc$bestTune
    k
6 101
```

2 Analyse discriminante linéaire

L'analyse discriminante linéaire est un algorithme de référence en classification supervisée. Il peut être appréhendé de deux façons complémentaires :

— une approche **géométrique** qui revient à chercher des hyperplans qui séparent au mieux les groupes ;
— une approche **modèle** qui fait l'hypothèse que les lois des covariables sont des vecteurs gaussiens avec des valeurs de paramètres différentes pour chaque groupe.

On considère $(x_1, y_1), \ldots, (x_n, y_n)$ un échantillon où x_i est à valeurs dans $\mathbb{R}^d$ et y_i dans $\{0, 1\}$. L'approche **géométrique** revient à chercher une droite de $\mathbb{R}^d$ d'équation $a_1 x_1 + \cdots + a_d x_d = 0$ telle que :

— les centres de gravité de chaque groupe projeté sur cette droite soit au mieux séparé $\Longrightarrow$ maximiser la distance inter-classe.
— les observations projetés soient proches de leur centre de gravité projeté $\Longrightarrow$ minimiser la distance intra-classe.

Le compromis entre ces deux distances s'obtient en maximisant le **coefficient de Rayleigh** qui est le quotient entre ces deux distance :

$$J(a) = \frac{B(a)}{W(a)} = \frac{a^t B a}{a^t W a}$$

où B et W sont les matrices inter et intra classes définies pas

$$B = \frac{1}{n}\sum_{k=1}^{K} n_k (g_k - g)(g_k - g)^t \quad \text{et} \quad W = \frac{1}{n}\sum_{k=1}^{K} n_k V_k \quad \text{avec} \quad V_k = \frac{1}{n_k}\sum_{i:Y_i=k} (X_i - g_k)(X_i - g_k)^t.$$

Ici g désigne le centre de gravité du nuage $x_i, i = 1, \ldots, n$ et $g_k, k = 0, 1$ les centres de gravité des deux groupes. La solution est donnée par un vecteur propre associé à la plus grande valeur propre de $W^{-1}B$.

L'approche **modèle** fait l'hypothèse que les vecteurs $X|Y = k, k = 0, 1$ sont des vecteurs gaussiens d'espérance $\mu_k \in \mathbb{R}^d$ et de matrice de variance covariance Σ. Ces paramètres sont estimés par maximum de vraisemblance et on déduit les probablités a posteriori par la **formule de Bayes** :

$$\mathbf{P}(Y = k|X = x) = \frac{\pi_k f_{X|Y=k}(x)}{f(x)}$$

Le lien entre ces deux approches est établi dans l'exercice 2.4. Nous proposons dans cette partie quelques exercices pour mettre en œuvre et analyser des analyses discriminantes avec **R**.

2.1 Prise en main : LDA et QDA sur les iris de Fisher

On considère les données sur les iris de Fisher.

```
data(iris)
```

1. A l'aide de la fonction **PCA** du package **FactoMineR**, réaliser une ACP en utilisant comme variables actives les 4 variables quantitatives du jeu de données. On mettra la variable `Species` comme variable qualitative supplémentaire (option `quali.sup`).
2. Représenter le nuage des individus sur les 2 premiers axes de l'ACP en utilisant une couleur différente pour chaque espèce d'iris (option `habillage`).
3. A l'aide de la fonction **lda** du package **MASS**, effectuer une analyse discriminante linéaire permettant d'expliquer l'espèce par les 4 autres variables explicatives.
4. Représenter le nuage des individus sur les deux premiers axes de l'analyse discriminante linéaire (en utilisant une couleur différente pour chaque espèce d'iris).
5. Rappeler comment sont obtenues les coordonnées des individus sur chaque axe. En déduire une interprétation de la position des individus.
6. Comparer les représentations des questions 2 et 4.
7. Expliquer les sorties des commandes suivantes (`mod.lda` est l'objet construit avec la fonction **lda**).

```
score <- predict(mod.lda)$x
ldahist(score[,1],iris[,5])
ldahist(score[,2],iris[,5])
```

“ “

8. Exécuter et analyser les sorties de la commande

```
mod.lda2 <- lda(Species~.,data=iris,CV=TRUE)
```

9. Comparer, en terme d'erreur de prévision, les performances de LDA et QDA.

2.2 Un cas avec beaucoup de classes

On considère les jeux de données **Vowel** (training et test) qui se trouvent à cet url. On peut les importer avec

```
dapp <- read_csv("https://web.stanford.edu/~hastie/ElemStatLearn/datasets/vowel.train")[,-1]
dtest <- read_csv("https://web.stanford.edu/~hastie/ElemStatLearn/datasets/vowel.test")[,-1]
```

1. Expliquer le problème.
2. Effectuer une analyse discriminante linéaire (uniquement avec les données d'apprentissage) et visualiser les individus sur les 2 premiers axes de l'analyse discriminante. On pourra utiliser **predict**.
3. La fonction suivante permet de choisir les axes à visualiser, ainsi que les centres de gravité projetés des groupes.

```
repres_axes <- function(prev,cdg,axe1=1,axe2=2){
  cdg <- prev %>% group_by(y) %>% summarise_all(mean)
  nom1 <- paste("LD",as.character(axe1),sep="")
  nom2 <- paste("LD",as.character(axe2),sep="")
  ggplot(prev)+aes_string(x=as.name(nom1),y=as.name(nom2))+geom_point(aes(color=y))+geom_point(data
}
```

   Étudier la pertinence des axes.
4. Représenter les individus sur le premier plan factoriel de l'ACP, on utilisera une couleur différente pour chaque groupe. On pourra utiliser le package **FactoMineR**.
5. Comparer cette projection avec celle obtenue par l'analyse discriminante linéaire.
6. Évaluer la performance de la **lda** sur les données test. Comparer avec l'analyse discriminante quadratique.
7. Expliquer comment on peut faire de la prévision en réduisant la dimension de l'espace des X.
8. Proposer une méthode permettant de choisir le meilleur nombre d'axes. On pourra notamment utiliser l'option `dimen` de la fonction **predict.lda**.

2.3 Grande dimension : reconnaissance de phonèmes

On considère le jeu de données **phoneme** téléchargeable à l'url https://github.com/cran/ElemStatLearn/blob/master/data/phoneme.RData.

```
load("data/phoneme.RData")
data(phoneme)
donnees <- phoneme[,-258]
```

1. Expliquer le problème et représenter pour chaque groupe la courbe moyenne.
2. Séparer les données en un échantillon d'apprentissage de taille 3000 et un échantillon test de taille 1509.
3. Effectuer une analyse discriminante linéaire et une analyse discriminante quadratique sur les données d'apprentissage uniquement. Évaluer les performances de ces deux approches sur les données test.
4. Quels peuvent être les intérêts d'effectuer une analyse discriminante régularisée dans ce contexte ? Effectuer une telle analyse à l'aide de la fonction **rda** du package **klaR**.
5. Sélectionner les paramètres de régularisation à l'aide du package **caret**. Comparer le nouveau modèle aux précédents.

2.4 Exercices

Exercice 2.1 (Optimalité de la règle de Bayes)**.**

On dispose de n observations $(x_1, y_1), \ldots, (x_n, y_n)$ telles que $x_i \in \mathbb{R}^p$ et $y_i \in \{0, 1\}$ pour $i = 1, \ldots, n$. On souhaite expliquer les sorties y_i par les entrées x_i.

1. Rappeler la définition d'une règle de prévision.
2. Rappeler la définition de la règle de Bayes $g^\star$ et de l'erreur de Bayes $L^\star$.
3. Soit g une règle de décision. Montrer que
$$\mathbf{P}(g(X) \neq Y | X = x) = 1 - (\mathbf{1}_{g(x)=1}\eta(x) + \mathbf{1}_{g(x)=0}(1 - \eta(x)))$$
où $\eta(x) = \mathbf{P}(Y = 1|X = x)$.
4. En déduire que pour tout $x \in \mathcal{X}$ et pour toute règle g
$$\mathbf{P}(g(X) \neq Y | X = x) - \mathbf{P}(g^\star(X) \neq Y | X = x) \geq 0.$$
Conclure.
5. On considère (X, Y) un couple aléatoire à valeurs dans $\mathbb{R} \times \{0, 1\}$ tel que
$$X \sim \mathcal{U}[-2, 2] \quad \text{et} \quad (Y|X = x) \sim \begin{cases} \mathcal{B}(1/5) & \text{si } x \leq 0 \\ \mathcal{B}(9/10) & \text{si } x > 0 \end{cases}$$
où $\mathcal{U}[a, b]$ désigne la loi uniforme sur $[a, b]$ et $\mathcal{B}(p)$ la loi de Bernoulli de paramètre p. Calculer la règle de Bayes et l'erreur de Bayes.

Exercice 2.2 (MV pour LDA)**.**

On cherche à expliquer une variable aléatoire Y à valeurs dans $\{0, 1\}$ par une variable aléatoire X à valeurs dans $\mathbb{R}$.

1. Quels sont les paramètres à estimer dans le modèle d'analyse discriminante linéaire.
2. Calculer la vraisemblance conditionnelle à Y et en déduire les estimateurs des paramètres des lois gaussiennes.
3. Comparer les estimateurs obtenus avec ceux du cours.

Exercice 2.3 (Fonctions linéaires discriminantes)**.**

On cherche à expliquer une variable aléatoire Y à valeurs dans $\{0, 1\}$ par une variable aléatoire X à valeurs dans $\mathbb{R}^p$.

1. Rappeler le modèle d'analyse discriminante linéaire.
2. Soit $x \in \mathbb{R}^p$ un nouvel individu. Montrer que la règle qui consiste à affecter x dans le groupe qui maximise $\mathbf{P}(Y = k|X = x)$ est équivalente à la règle qui consiste à affecter x dans le groupe qui maximise les fonctions linéaires discriminantes (on prendra soin de rappeler la définition des fonctions linéaires discriminantes).

Exercice 2.4 (Approche géométrique de la LDA)**.**

On considère un n-échantillon i.i.d. $(x_1, y_1), \ldots, (x_n, y_n)$ où x_i est à valeurs dans $\mathbb{R}^2$ et y_i dans $\{0, 1\}$. On cherche une droite vectorielle a telle que les projections de chaque groupe sur a soient séparées "au mieux". Dit autrement, on cherche a telle que

— la distance entre les centres de gravité
$$g_0 = \frac{1}{\text{card}\{i : y_i = 0\}} \sum_{i:y_i=0} x_i \quad \text{et} \quad g_1 = \frac{1}{\text{card}\{i : y_i = 1\}} \sum_{i:y_i=1} x_i$$
projetés sur a soit maximale (cette distance est appelée distance interclasse) ;

— la distance entre les projections des individus et leur centre de gravité soit minimale (distance interclasse).

Pour un vecteur u de $\mathbb{R}^2$, on désigne par $\pi_a(u)$ son projeté sur la droite engendrée par a. Sans perte de généralité on supposera dans un premier temps que a est de norme 1.

1. Rappeler les définitions des variances totale V, intra W et inter B des observations $(x_1, y_1), \ldots, (x_n, y_n)$.
2. Pour u fixé dans $\mathbb{R}^2$, exprimer $\pi_a(u)$ en fonction de u et a et en déduire que $\|\pi_a(u)\|^2 = a^t u u^t a$.
3. Exprimer les variances totale $V(a)$, intra $W(a)$ et inter $B(a)$ projetées sur a en fonction des variances calculées à la question 1.
4. On cherche maintenant à maximiser
$$J(a) = \frac{B(a)}{W(a)}$$
ou encore à
$$\text{maximiser } B(a) \quad \text{sous la contrainte} \quad W(a) = 1. \tag{1}$$
La méthode des multiplicateurs de Lagrange permet de résoudre un tel problème. La solution du problème de maximisation d'une fonction $f(x)$ sujette à $h(x) = 0$ s'obtient en résolvant l'équation
$$\frac{\partial L(x, \lambda)}{\partial x} = 0, \quad \text{où} \quad L(x, \lambda) = f(x) + \lambda h(x).$$
   a. Montrer que la solution du problème (1) est un vecteur propre de $W^{-1}B$ associé à la plus grande valeur propre de $W^{-1}B$. On note $a^\star$ cette solution.
   b. Montrer que $a^\star$ est colinéaire à $W^{-1}(g_1 - g_0)$. On pourra admettre que, dans le cas de 2 groupes, on a
$$B = \frac{n_0 n_1}{n^2}(g_1 - g_0)(g_1 - g_0)^t.$$
   c. On considère la règle géométrique d'affectation qui consiste à classer un nouvel individu $x \in \mathbb{R}^p$ au groupe 1 si son projeté sur $a^\star$ est plus proche de $\pi_{a^\star}(g_1)$ que de $\pi_{a^\star}(g_0)$. Montrer que x sera affecté au groupe 1 si
$$S(x) = x^t W^{-1}(g_1 - g_0) > s$$
   où on exprimera s en fonction de g_0, g_1 et W.
   d. Montrer que cette règle est équivalente à choisir le groupe qui minimise la distance de Mahalanobis
$$d(x, g_k) = (x - g_k)^t W^{-1}(x - g_k), \quad k = 0, 1.$$
   e. On revient maintenant à l'approche probabiliste de l'analyse discriminante linéaire vue en cours et on considère la règle d'affectation qui consiste à décider "groupe 1" si $\mathbf{P}(Y = 1 | X = x) \geq 0.5$. Montrer que dans ce cas, un nouvel individu x est affecter au groupe 1 si :
$$S(x) = x^t \Sigma^{-1}(\mu_1 - \mu_0) > \frac{1}{2}(\mu_1 + \mu_0)^t \Sigma^{-1}(\mu_1 - \mu_0) - \log\left(\frac{\pi_1}{\pi_0}\right).$$
   Conclure.

3 Arbres

Les méthodes par arbres sont des algorithmes où la prévision s'effectue à partir de **moyennes locales**. Plus précisément, étant donné un échantillon $(x_1, y_1) \ldots, (x_n, y_n)$, l'approche consiste à :

— construire une partition de l'espace de variables explicatives $(\mathbb{R}^p)$;
— prédire la sortie d'une nouvelle observation x en faisant :
   — la moyenne des y_i pour les x_i qui sont dans la même classe que x si on est en régression ;

— un vote à la majorité parmi les y_i tels que les x_i qui sont dans la même classe que x si on est en classification.

Bien entendu toute la difficulté est de trouver la "bonne partition" pour le problème d'intérêt. Il existe un grand nombre d'algorithmes qui permettent de trouver une partition. Le plus connu est l'algorithme **CART** (Breiman et al., 1984) où la partition est construite par **divisions successives** au moyen d'hyperplan orthogonaux aux axes de $\mathbb{R}^p$. L'algorithme est récursif : il va à chaque étape séparer un groupe d'observations (**nœuds**) en deux groupes (**nœuds fils**) en cherchant la meilleure variable et le meilleur seuil de coupure. Ce choix s'effectue à partir d'un critère **d'impureté** : la meilleure coupure est celle pour laquelle l'impureté des 2 nœuds fils sera minimale. Nous étudions cet algorithme dans cette partie.

3.1 Coupures CART en fonction de la nature des variables

Une partition CART s'obtient en séparant les observations en 2 selon une coupure parallèle aux axes puis en itérant ce procédé de séparation binaire sur les deux groupes... Par conséquent la première question à se poser est : pour un ensemble de données $(x_1, y_1), \ldots, (x_n, y_n)$ fixé, comment obtenir la meilleure coupure ?

Comme souvent ce sont les données qui vont répondre à cette question. La sélection de la meilleur coupure s'effectue en introduisant une **fonction d'impureté** $\mathcal{I}$ qui va mesurer le degrés d'hétérogénéité d'un nœud $\mathcal{N}$. Cette fonction prendra de

— grandes valeurs pour les nœuds hétérogènes (les valeurs de Y diffèrent à l'intérieur du nœud) ;
— faibles valeurs pour les nœuds homogènes (les valeurs de Y sont proches à l'intérieur du nœud).

On utilise souvent comme fonction d'impureté :

— la **variance** en régression
$$\mathcal{I}(\mathcal{N}) = \frac{1}{|\mathcal{N}|} \sum_{i:x_i \in \mathcal{N}} (y_i - \overline{y}_{\mathcal{N}})^2,$$
où $\overline{y}_{\mathcal{N}}$ désigne la moyenne des y_i dans $\mathcal{N}$.
— l'impureté de **Gini** en classification binaire
$$\mathcal{I}(\mathcal{N}) = 2p(\mathcal{N})(1 - p(\mathcal{N}))$$
où $p(\mathcal{N})$ représente la proportion de 1 dans $\mathcal{N}$.

Les coupures considérées par l'algorithme CART sont des hyperplans orthogonaux aux axes de $\mathbb{R}^p$, choisir une coupure revient donc à choisir une variable j parmi les p variables explicatives et un seuil s dans $\mathbb{R}$. On peut donc représenter une coupure par un couple (j, s). Une fois l'impureté définie, on choisira la coupure (j, s) qui **maximise le gain d'impureté** entre le noeud père et ses deux noeuds fils :

$$\Delta(\mathcal{I}) = \mathbf{P}(\mathcal{N})\mathcal{I}(\mathcal{N}) - (\mathbf{P}(\mathcal{N}_1(j,s))\mathcal{I}(\mathcal{N}_1(j,s)) + \mathbf{P}(\mathcal{N}_2(j,s))\mathcal{I}(\mathcal{N}_2(j,s))$$

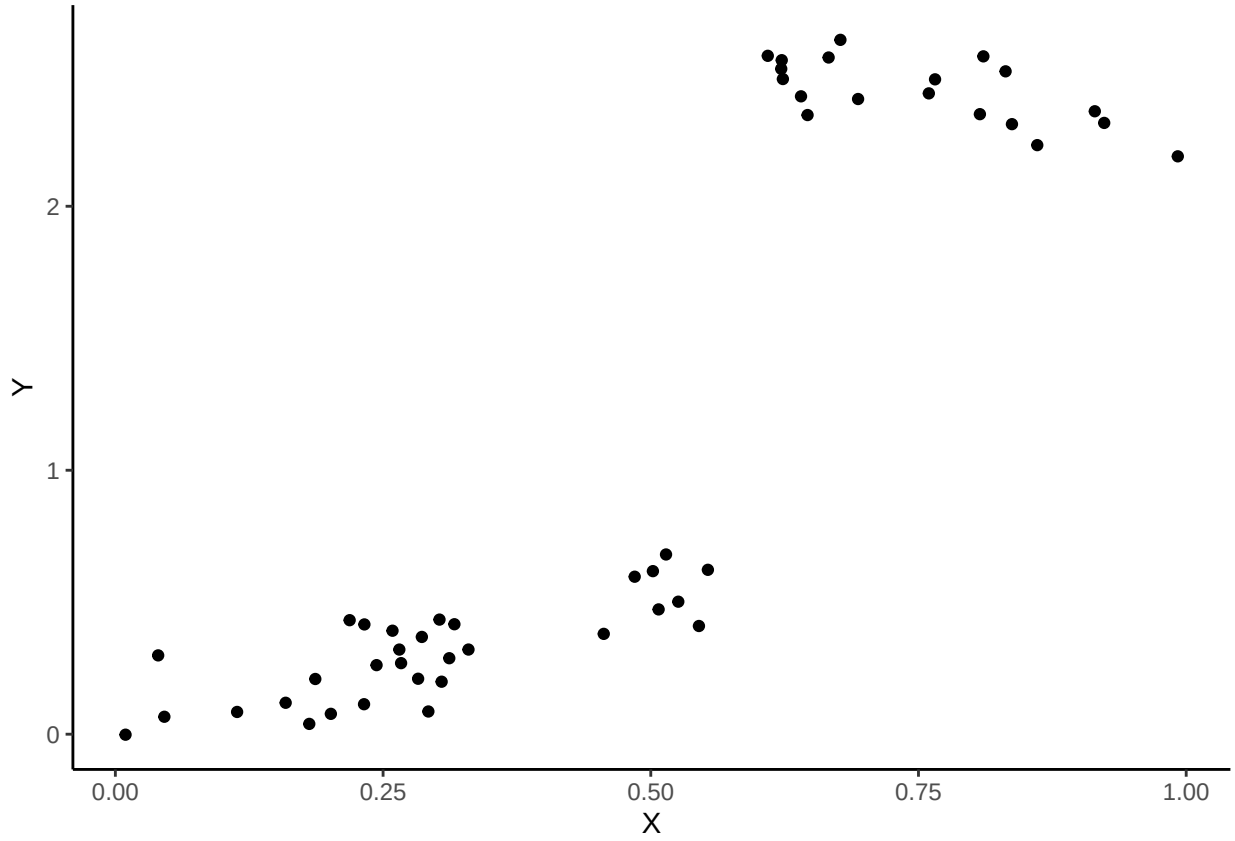
où * $\mathcal{N}_1(j, s)$ et $\mathcal{N}_2(j, s)$ sont les 2 nœuds fils de $\mathcal{N}$ engendrés par la coupure (j, s) ; * $\mathbf{P}(\mathcal{N})$ représente la proportion d'observations dans le nœud $\mathcal{N}$.

3.1.1 Arbres de régression

On considère le jeu de données suivant où le problème est d'expliquer la variable quantitative Y par la variable quantitative X.

```
n <- 50
set.seed(1234)
X <- runif(n)
set.seed(5678)
```

```
Y <- 1*X*(X<=0.6)+(-1*X+3.2)*(X>0.6)+rnorm(n,sd=0.1)
data1 <- data.frame(X,Y)
ggplot(data1)+aes(x=X,y=Y)+geom_point()
```



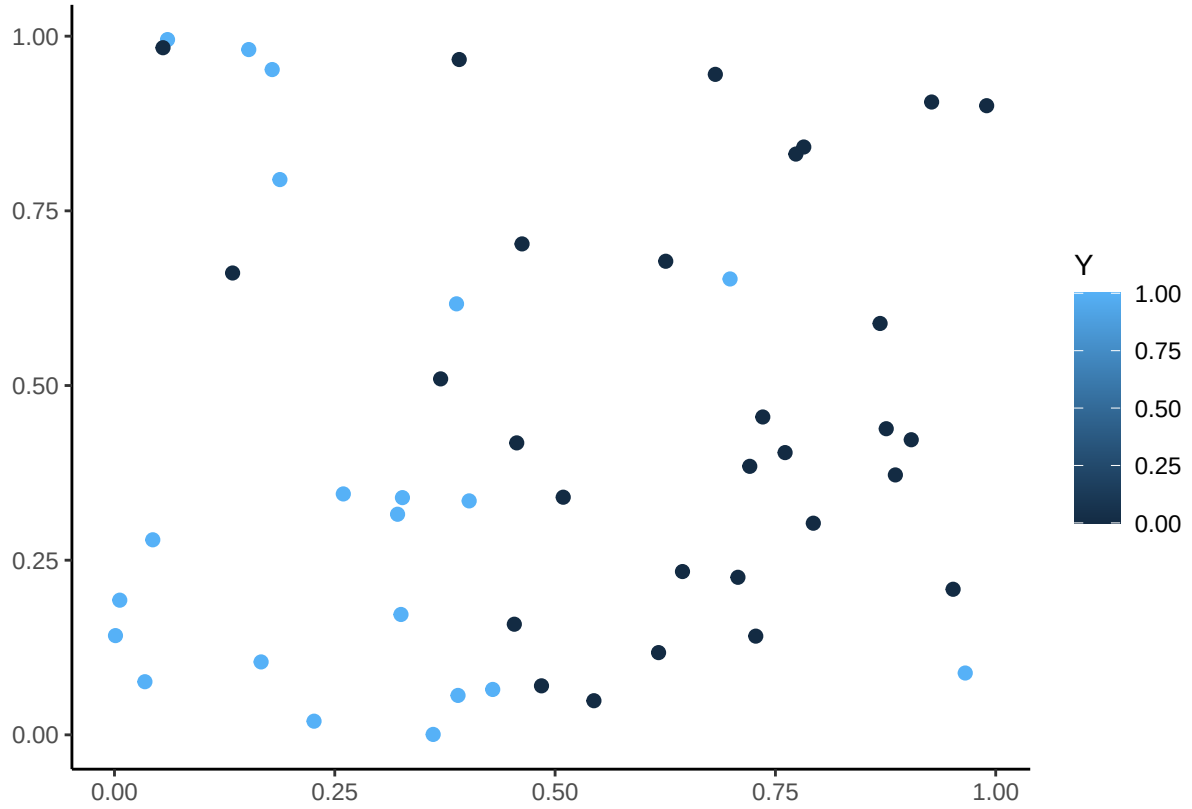
1. A l'aide de la fonction **rpart** du package **rpart**, construire un arbre permettant d'expliquer Y par X.
```
library(rpart)
```
2. Visualiser l'arbre à l'aide des fonctions **prp** et **rpart.plot** du package **rpart.plot**.
3. Écrire l'estimateur associé à l'arbre.
4. Ajouter sur le graphe de la question 1 la partition définie par l'arbre ainsi que les valeurs prédites.

3.1.2 Arbres de classification

On considère les données suivantes où le problème est d'expliquer la variable binaire Y par deux variables quantitatives X_1 et X_2.

```
n <- 50
set.seed(12345)
X1 <- runif(n)
set.seed(5678)
X2 <- runif(n)
Y <- rep(0,n)
set.seed(54321)
Y[X1<=0.45] <- rbinom(sum(X1<=0.45),1,0.85)
set.seed(52432)
```

```
Y[X1>0.45] <- rbinom(sum(X1>0.45),1,0.15)
data2 <- data.frame(X1,X2,Y)
ggplot(data2)+aes(x=X1,y=X2,color=Y)+geom_point(size=2)+scale_x_continuous(name="")+
  scale_y_continuous(name="")+theme_classic()
```



1. Construire un arbre permettant d'expliquer Y par X_1 et X_2. Représenter l'arbre et identifier l'éventuel problème.
2. Écrire la règle de classification ainsi que la fonction de score définies par l'arbre.
3. Ajouter sur le graphe de la question 1 la partition définie par l'arbre.

3.1.3 Entrée qualitative

On considère les données

```
n <- 100
X <- factor(rep(c("A","B","C","D"),n))
set.seed(1234)
Y[X=="A"] <- rbinom(sum(X=="A"),1,0.9)
Y[X=="B"] <- rbinom(sum(X=="B"),1,0.25)
Y[X=="C"] <- rbinom(sum(X=="C"),1,0.8)
Y[X=="D"] <- rbinom(sum(X=="D"),1,0.2)
Y <- as.factor(Y)
data3 <- data.frame(X,Y)
```

1. Construire un arbre permettant d'expliquer Y par X.
2. Expliquer la manière dont l'arbre est construit dans ce cadre là.

3.2 Élagage

Le procédé de coupe présenté précédemment permet de définir un très grand nombre d'arbres à partir d'un jeu de données (arbre sans coupure, avec une coupure, deux coupures...). Se pose alors la question de trouver le **meilleur arbre** parmi tous les arbres possibles. Une première idée serait de choisir parmi tous les arbres possibles celui qui optimise un critère de performance. Cette approche, bien que cohérente, n'est généralement pas possible à mettre en œuvre en pratique car le nombre d'arbres à considérer est souvent trop important.

La méthode CART propose une procédure permettant de choisir automatiquement un arbre en 3 étapes :

— On construit un **arbre maximal** (très profond) $\mathcal{T}_{max}$;
— On sélectionne une **suite d'arbres emboités** :

$$\mathcal{T}_{max} = \mathcal{T}_0 \supset \mathcal{T}_1 \supset \cdots \supset \mathcal{T}_K.$$

La sélection s'effectue en optimisant un critère **Cout/complexité** qui permet de réguler le compromis entre **ajustement** et **complexité** de l'arbre.
— On **sélectionne un arbre** dans cette sous-suite en optimisant un critère de performance.

Cette approche revient à choisir un sous-arbre de l'arbre $\mathcal{T}_{\max}$, c'est-à-dire à enlever des branches à $T_{\max}$, c'est pourquoi on parle **d'élagage**.

3.2.1 Élagage pour un problème de régression

On considère les données **Carseats** du package **ISLR**.

```
library(ISLR)
data(Carseats)
summary(Carseats)
     Sales          CompPrice       Income
 Min.   : 0.000   Min.   : 77   Min.   : 21.00
 1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75
 Median : 7.490   Median :125   Median : 69.00
 Mean   : 7.496   Mean   :125   Mean   : 68.66
 3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00
 Max.   :16.270   Max.   :175   Max.   :120.00
  Advertising       Population        Price
 Min.   : 0.000   Min.   : 10.0   Min.   : 24.0
 1st Qu.: 0.000   1st Qu.:139.0   1st Qu.:100.0
 Median : 5.000   Median :272.0   Median :117.0
 Mean   : 6.635   Mean   :264.8   Mean   :115.8
 3rd Qu.:12.000   3rd Qu.:398.5   3rd Qu.:131.0
 Max.   :29.000   Max.   :509.0   Max.   :191.0
  ShelveLoc        Age          Education    Urban
 Bad    : 96   Min.   :25.00   Min.   :10.0   No :118
 Good   : 85   1st Qu.:39.75   1st Qu.:12.0   Yes:282
 Medium:219   Median :54.50   Median :14.0
              Mean   :53.32   Mean   :13.9
              3rd Qu.:66.00   3rd Qu.:16.0
              Max.   :80.00   Max.   :18.0
   US
 No :142
 Yes:258
```

On cherche ici à expliquer la variable quantitative **Sales** par les autres variables.

1. Construire un arbre permettant de répondre au problème.
2. Expliquer les sorties de la fonction **printcp** appliquée à l'arbre de la question précédente et calculer le dernier terme de la colonne **rel error**.
3. Construire une suite d'arbres plus grandes en jouant sur les paramètres `cp` et `minsplit` de la fonction **rpart**.
4. Expliquer la sortie de la fonction **plotcp** appliquée à l'arbre de la question précédente.
5. Sélectionner le "meilleur" arbre dans la suite construite.
6. Visualiser l'arbre choisi (utiliser la fonction **prune**).
7. On souhaite prédire les valeurs de Y pour de nouveaux individus à partir de l'arbre sélectionné. Pour simplifier on considèrera ces 4 individus :

```
new_ind <- Carseats %>% slice(3,58,185,218) %>% dplyr::select(-Sales)
new_ind
    CompPrice Income Advertising Population Price ShelveLoc
3         113     35          10        269    80    Medium
58         93     91           0         22   117       Bad
185       132     33           7         35    97    Medium
218       106     44           0        481   111    Medium
    Age Education Urban  US
3    59        12   Yes Yes
58   75        11   Yes  No
185  60        11    No Yes
218  70        14    No  No
```

   Calculer les valeurs prédites.
8. Séparer les données en un échantillon d'apprentissage de taille 250 et un échantillon test de taille 150.
9. On considère la suite d'arbres définie par

```
set.seed(4321)
tree <- rpart(Sales~.,data=train,cp=0.000001,minsplit=2)
```

   Dans cette suite, sélectionner

   — un arbre très simple (avec 2 ou 3 coupures)
   — un arbre très grand
   — l'arbre optimal (avec la procédure d'élagage classique).
10. Calculer l'erreur quadratique de ces 3 arbres en utilisant l'échantillon test.
11. Refaire la comparaison avec une validation croisée 10 blocs.

3.2.2 Élagage en classification binaire et matrice de coût

On considère ici les mêmes données que précédemment mais on cherche à expliquer une version binaire de la variable **Sales**. Cette nouvelle variable, appelée `High` prend pour valeurs `No` si `Sales` est inférieur ou égal à 8, `Yes` sinon. On travaillera donc avec le jeu `data1` défini ci-dessous.

```
High <- ifelse(Carseats$Sales<=8,"No","Yes")
data1 <- Carseats %>% dplyr::select(-Sales) %>% mutate(High)
```

1. Construire un arbre permettant d'expliquer `High` par les autres variables (sans `Sales` évidemment !) et expliquer les principales différences par rapport à la partie précédente précédente.
2. Expliquer l'option `parms` dans la commande :

```
tree1 <- rpart(High~.,data=data1,parms=list(split="information"))
tree1$parms
$prior
   1    2
0.59 0.41

$loss
     [,1] [,2]
[1,]    0    1
[2,]    1    0

$split
[1] 2
```

3. Expliquer les sorties de la fonction **printcp** sur le premier arbre construit et retrouver la valeur du dernier terme de la colonne **rel error**.
4. Sélectionner un arbre optimal dans la suite.
5. On considère la suite d'arbres

```
tree2 <- rpart(High~.,data=data1,parms=list(loss=matrix(c(0,5,1,0),ncol=2)),
           cp=0.01,minsplit=2)
```

Expliquer les sorties des commandes suivantes. On pourra notamment calculer le dernier terme de la colonne **rel error** de la table **cptable**.

```
tree2$parms
$prior
   1    2
0.59 0.41

$loss
     [,1] [,2]
[1,]    0    1
[2,]    5    0

$split
[1] 1
printcp(tree2)

Classification tree:
rpart(formula = High ~ ., data = data1, parms = list(loss = matrix(c(0,
    5, 1, 0), ncol = 2)), cp = 0.01, minsplit = 2)

Variables actually used in tree construction:
[1] Advertising Age         CompPrice   Education
[5] Income      Population  Price       ShelveLoc

Root node error: 236/400 = 0.59

n= 400
```

```
         CP nsplit rel error xerror    xstd
1  0.101695      0   1.00000 5.0000 0.20840
2  0.050847      2   0.79661 3.8136 0.20909
3  0.036017      3   0.74576 3.2034 0.20176
4  0.035311      5   0.67373 3.1271 0.20038
5  0.025424      9   0.50847 2.6144 0.19069
6  0.016949     11   0.45763 2.3475 0.18307
7  0.015537     16   0.37288 2.1992 0.17905
8  0.014831     21   0.28814 2.1992 0.17905
9  0.010593     23   0.25847 2.0466 0.17367
10 0.010000     25   0.23729 2.0297 0.17292
```

6. Comparer les valeurs ajustées par les deux arbres considérés.

Deuxième partie

Algorithmes avancés

4 Support Vector Machine (SVM)

Etant donnée un échantillon $(x_1, y_1), \ldots, (x_n, y_n)$ où les x_i sont à valeurs dans $\mathbb{R}^p$ et les y_i sont binaires à valeurs dans $\{-1, 1\}$, l'approche **SVM** cherche le **meilleur hyperplan** en terme de séparation des données. Globalement on veut que les `1` se trouvent d'un coté de l'hyperplan et les `-1` de l'autre. Dans cette partie on propose d'étudier la mise en œuvre de cet algorithme tout d'abord dans le cas idéal où les données sont séparables puis dans le cas plus réel où elles ne le sont pas. Nous verrons ensuite comment introduire de la non linéarité ne utilisant l'**astuce du noyau**.

4.1 Cas séparable

Le cas séparable est le cas facile : il correspond à la situation où il existe effectivement un (même plusieurs) hyperplan(s) qui sépare(nt) parfaitement les 1 des -1. Il ne se produit quasiment jamais en pratique mais il convient de l'étudier pour comprendre comment est construit l'algorithme. Dans ce cas on cherche l'hyperplan d'équation $\langle w, x\rangle + b = w^t x + b = 0$ tel que la **marge** (qui peut être vue comme la distance entre les observations les plus proches de l'hyperplan et l'hyperplan) soit maximale. Mathématiquement le problème se réécrit comme un problème d'optimisation sous contraintes :

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{2}$$
$$\text{sous les contraintes } y_i(w^t x_i + b) \geq 1,\ i = 1, \ldots, n.$$

La solution s'obtient de façon classique en résolvant le problème dual et elle s'écrit comme une combinaison linéaire des x_i

$$w^\star = \sum_{i=1}^{n} \alpha_i^\star y_i x_i.$$

De plus, les conditions **KKT** impliquent que pour tout $i = 1, \ldots, n$:

— $\alpha_i^\star = 0$

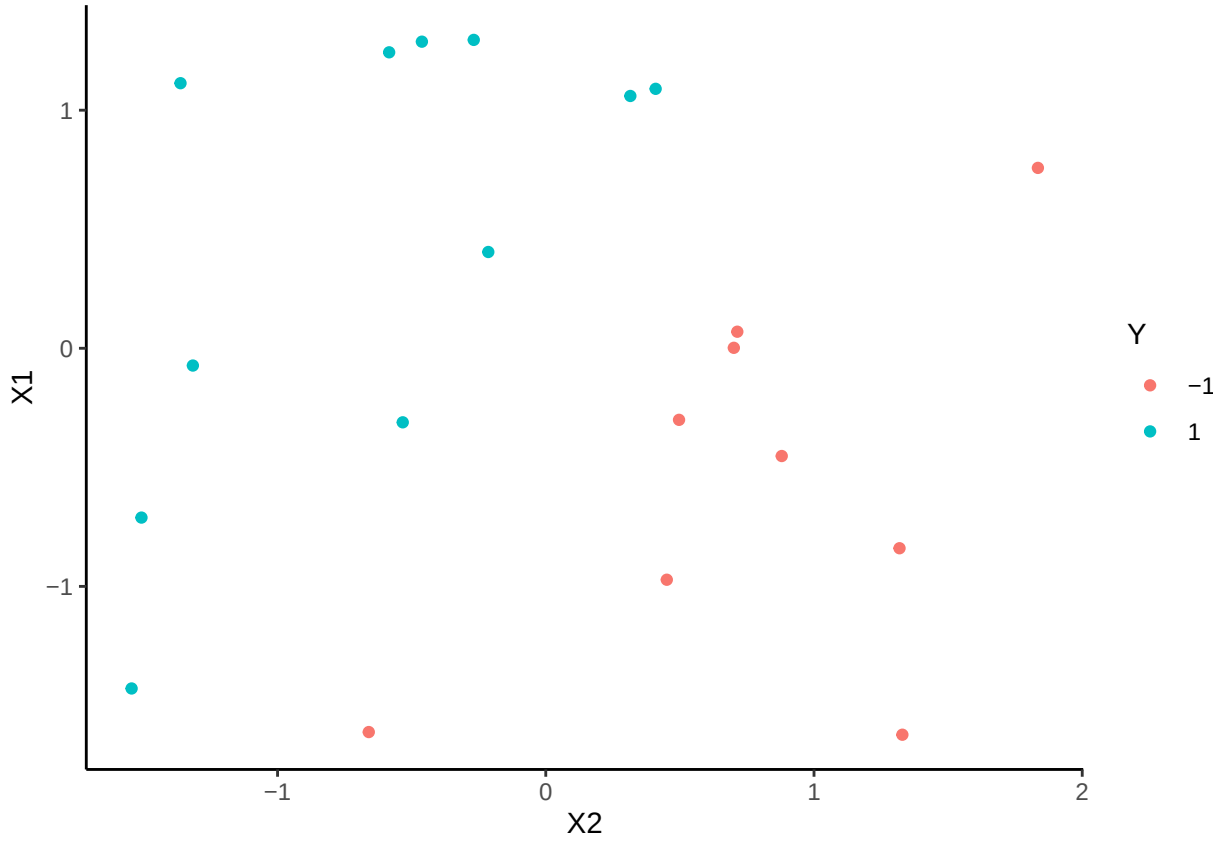
ou

— $y_i(x_i^t w + b) - 1 = 0$.

Ces conditions impliquent que $w^\star$ s'écrit comme une combinaison linéaire de quelques points, appelés **vecteurs supports** qui se trouvent **sur la marge**. Nous proposons maintenant de retrouver ces points et de tracer la marge sur un exemple simple.

On considère le nuage de points suivant :

```
n <- 20
set.seed(123)
X1 <- scale(runif(n))
set.seed(567)
X2 <- scale(runif(n))
Y <- rep(-1,n)
Y[X1>X2] <- 1
Y <- as.factor(Y)
donnees <- data.frame(X1=X1,X2=X2,Y=Y)
p <- ggplot(donnees)+aes(x=X2,y=X1,color=Y)+geom_point()
p
```



La fonction **svm** du package **e1071** permet d'ajuster une SVM :

```
library(e1071)
mod.svm <- svm(Y~.,data=donnees,kernel="linear",cost=10000000000)
```

1. Récupérer les vecteurs supports et visualiser les sur le graphe (en utilisant une autre couleur par exemple). On les affectera à un **data.frame** dont les 2 premières colonnes représenteront les valeurs de X_1 et X_2 des vecteurs supports.

```
ind.svm <- mod.svm$index
sv <- donnees %>% slice(ind.svm)
...
```

2. Retrouver ce graphe à l'aide de la fonction **plot**.
3. Rappeler la règle de décision associée à la méthode SVM. Donner les estimations des paramètres de la règle de décision sur cet exemple. On pourra notamment regarder la sortie `coef` de la fonction **svm**.
4. On dispose d'un nouvel individu $x = (-0.5, 0.5)$. Expliquer comment on peut prédire son groupe.
5. Retrouver les résultats de la question précédente à l'aide de la fonction **predict**. On pourra utiliser l'option `decision.values = TRUE`.
6. Obtenir les probabilités prédites à l'aide de la fonction **predict**. On pourra utiliser `probability=TRUE` dans la fonction **svm**.

4.2 Cas non séparable

Dans la vraie vie, les groupes ne sont généralement pas séparables et il n'existe donc pas de solution au problème (2). On va donc autoriser certains points à être :

— mal classés

et/ou

— bien classés mais à l'intérieur de la marge.

Mathématiquement, cela revient à introduire des **variables ressorts** (**slacks variables**) $\xi_1, \dots, \xi_n$ positives telles que :

— $\xi_i \in [0, 1] \Longrightarrow i$ bien classé mais **dans** la région définie par la **marge** ;
— $\xi_i > 1 \Longrightarrow i$ **mal classé**.

Le problème d'optimisation est alors de minimiser en (w, b, ξ)

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{sous les contraintes } \begin{cases} y_i(w^t x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, \dots, n. \end{cases}$$

Le paramètre $C > 0$ est à **calibrer** et on remarque que le cas séparable correspond à $C \to +\infty$. Les solutions de ce nouveau problème d'optimisation s'obtiennent de la même façon que dans le cas séparable, en particulier $w^\star$ s'écrite toujours comme une combinaison linéaire

$$w^\star = \sum_{i=1}^{n}\alpha_i^\star y_i x_i.$$

de **vecteurs supports** sauf qu'on distingue deux types de vecteurs supports ($\alpha_i^\star > 0$) :

— ceux **sur la frontière** définie par la marge : $\xi_i^\star = 0$;
— ceux **en dehors** : $\xi_i^\star > 0$ et $\alpha_i^\star = C$.

Le choix de C est crucial : ce paramètre régule le **compromis biais/variance** de la svm :

— $C \searrow$: la marge est privilégiée et les $\xi_i \nearrow \Longrightarrow$ beaucoup d'observations dans la marge ou **mal classées** (et donc **beaucoup de vecteurs supports**).

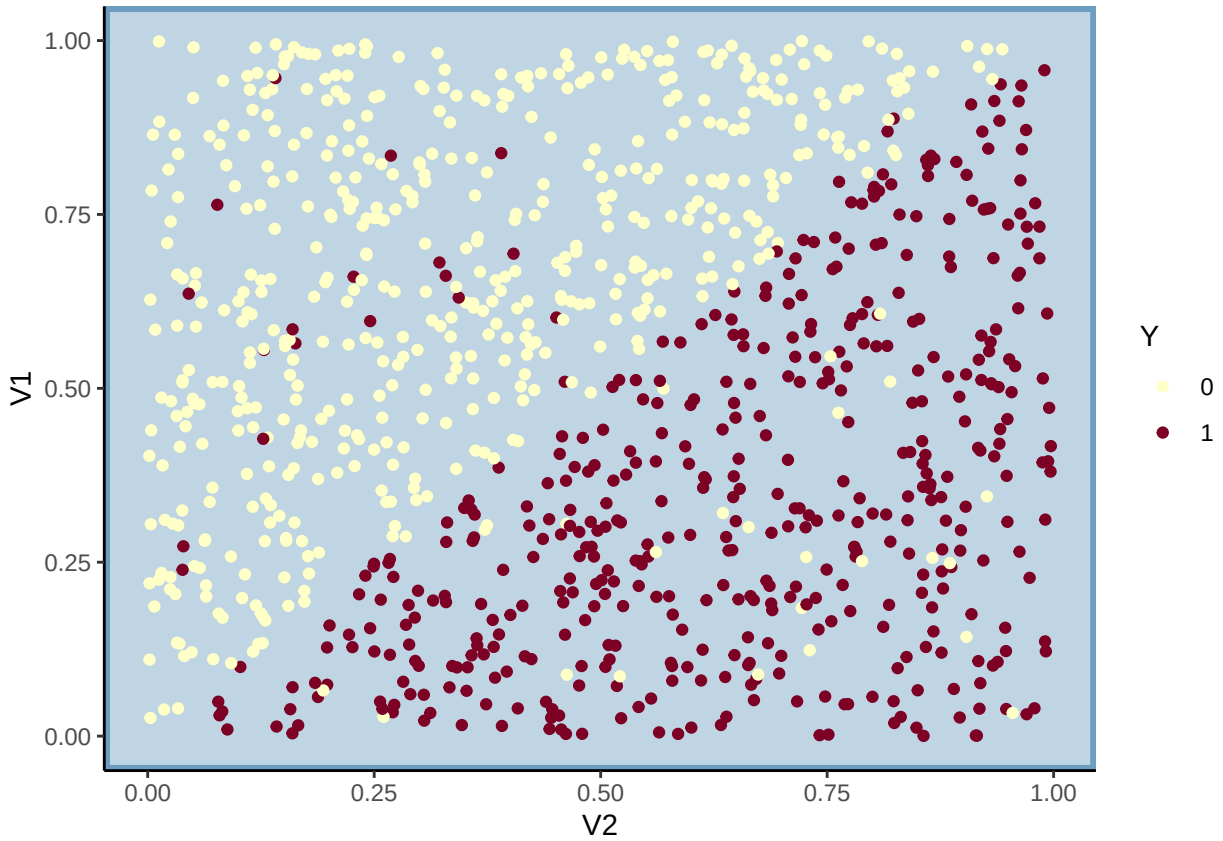
— $C \nearrow \Longrightarrow \xi_i \searrow$ donc moins d'observations mal classées $\Longrightarrow$ **meilleur ajustement** mais petite marge $\Longrightarrow$ risque de **surajustement**.

On choisit généralement ce paramètre à l'aide des techniques présentées dans le chapitre 1 :

— choix d'une grille de valeurs de C et d'un critère ;
— choix d'une méthode de ré-échantillonnage pour estimer le critère ;
— choix de la valeur de C qui minimise le critère estimé.

On considère le jeu de données `df3` définie ci-dessous.

```
n <- 1000
set.seed(1234)
df <- as.data.frame(matrix(runif(2*n),ncol=2))
df1 <- df %>% filter(V1<=V2)%>% mutate(Y=rbinom(nrow(.),1,0.95))
df2 <- df %>% filter(V1>V2)%>% mutate(Y=rbinom(nrow(.),1,0.05))
df3 <- bind_rows(df1,df2) %>% mutate(Y=as.factor(Y))
ggplot(df3)+aes(x=V2,y=V1,color=Y)+geom_point()+
  scale_color_manual(values=c("#FFFFC8", "#7D0025"))+
  theme(panel.background = element_rect(fill = "#BFD5E3", colour = "#6D9EC1",size = 2, linetype = "soli
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())
```



1. Ajuster 3 svm en considérant comme valeur de C : 0.000001, 0.1 et 5. On pourra utiliser l'option `cost`.

```
mod.svm1 <- svm(Y~.,data=df3,kernel="linear",...)
mod.svm2 <- svm(Y~.,data=df3,kernel="linear",...)
mod.svm3 <- svm(Y~.,data=df3,kernel="linear",...)
```

2. Calculer les nombres de vecteurs supports pour chaque valeur de C.
3. Visualiser les 3 svm obtenues. Interpréter.

4.3 L'astuce du noyau

Les SVM présentées précédemment font l'hypothèse que les groupes sont **linéairement séparables**, ce qui n'est bien entendu pas toujours le cas en pratique. L'**astuce du noyau** permet de mettre de la non linéarité, elle consiste à :

— plonger les données dans un nouvel espace appelé **espace de représentation** ou **feature space** ;
— appliquer une **svm** linéaire dans ce nouvel espace.

Le terme **astuce** vient du fait que ce procédé ne nécessite pas de connaître explicitement ce nouvel espace : pour résoudre le problème d'optimisation dans le **feature space** on a juste besoin de connaître le **noyau** associé au feature space. D'un point de vu formel un noyau est une fonction

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

dont les propriétés sont proches d'un produit scalaire. Il existe donc tout un tas de noyau avec lesquels on peut faire des SVM, par exemple

— **Linéaire** (sur $\mathbb{R}^d$) : $K(x, x') = x^t x'$.
— **Polynomial** (sur $\mathbb{R}^d$) : $K(x, x') = (x^t x' + 1)^d$.
— **Gaussien** (Gaussian radial basis function ou RBF) (sur $\mathbb{R}^d$)

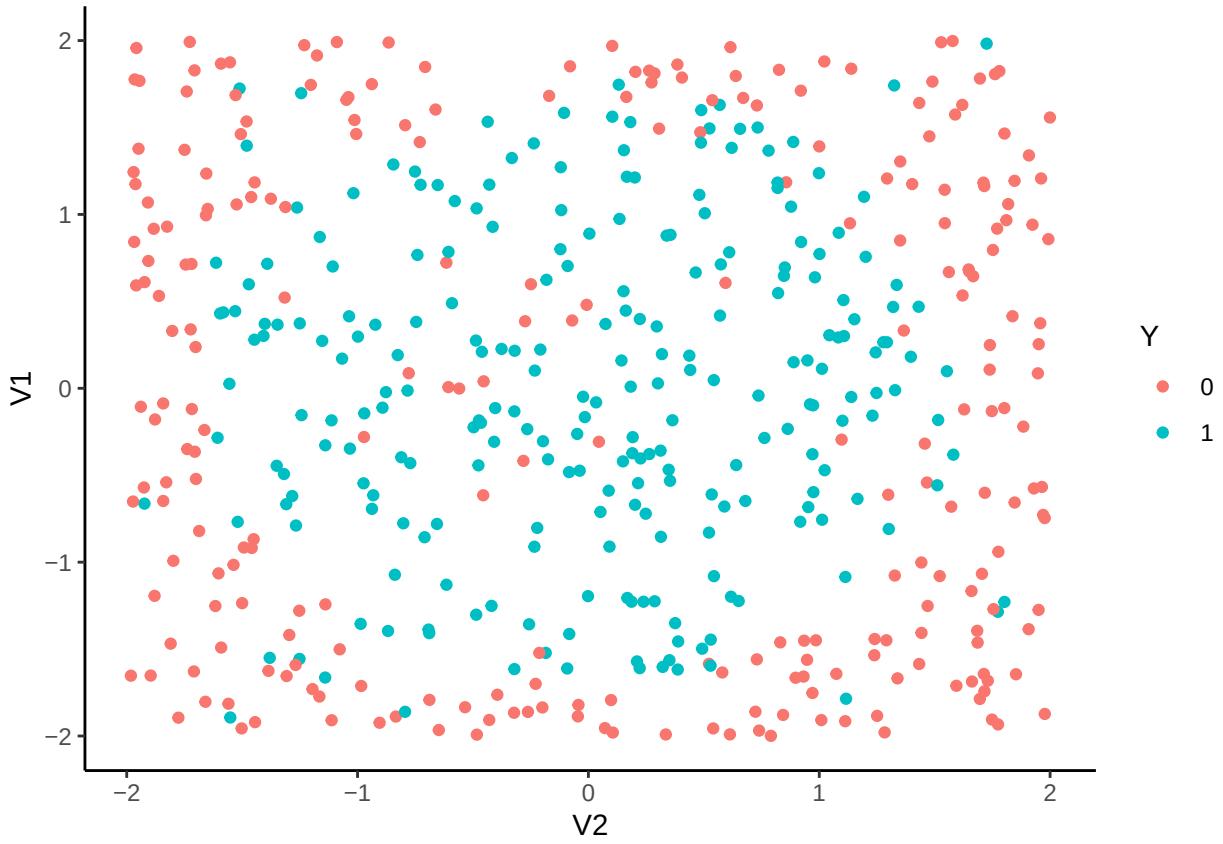
$$K(x, x') = \exp\left(-\frac{\|x - x'\|}{2\sigma^2}\right).$$

— **Laplace** (sur $\mathbb{R}$) : $K(x, x') = \exp(-\gamma|x - x'|)$.
— **Noyau min** (sur $\mathbb{R}^+$) : $K(x, x') = \min(x, x')$.
— ...

Bien entendu, en pratique tout le problème va consister à **trouver le bon noyau** !

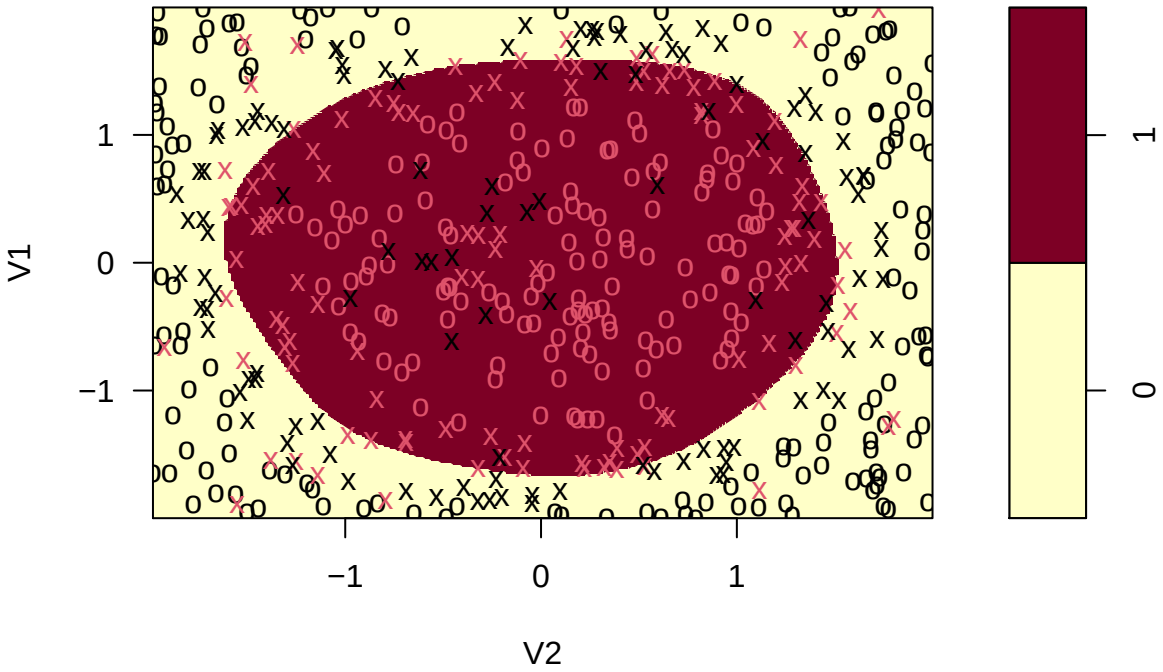
On considère le jeu de données suivant où le problème est d'expliquer Y par $V1$ et $V2$.

```
n <- 500
set.seed(13)
X <- matrix(runif(n*2,-2,2),ncol=2) %>% as.data.frame()
Y <- rep(0,n)
cond <- (X$V1^2+X$V2^2)<=2.8
Y[cond] <- rbinom(sum(cond),1,0.9)
Y[!cond] <- rbinom(sum(!cond),1,0.1)
df <- X %>% mutate(Y=as.factor(Y))
ggplot(df)+aes(x=V2,y=V1,color=Y)+geom_point()+theme_classic()
```



1. Ajuster une svm linéaire et visualiser l'hyperplan séparateur. Que remarquez-vous ?
2. Exécuter la commande suivante et commenter la sortie.

```
mod.svm1 <- svm(Y~.,data=df,kernel="radial",gamma=1,cost=1)
plot(mod.svm1,df,grid=250)
```



3. Faire varier les paramètres **gamma** et **cost**. Interpréter (on pourra notamment étudier l'évolution du nombre de vecteurs supports en fonction du paramètre **cost**).

```
mod.svm2 <- svm(Y~.,data=df,kernel="radial",gamma=...,cost=...)
mod.svm3 <- svm(Y~.,data=df,kernel="radial",gamma=...,cost=...)
mod.svm4 <- svm(Y~.,data=df,kernel="radial",gamma=...,cost=...)

plot(mod.svm2,df,grid=250)
plot(mod.svm3,df,grid=250)
plot(mod.svm4,df,grid=250)

mod.svm2$nSV
mod.svm3$nSV
mod.svm4$nSV
```

4. Sélectionner automatiquement ces paramètres. On pourra utiliser la fonction **tune** en faisant varier **C** dans **c(0.1,1,10,100,1000)** et **gamma** dans **c(0.5,1,2,3,4)**.

```
tune.out <- tune(svm,Y~.,data=...,kernel="...",
             ranges=list(cost=...,gamma=...))
```

5. Faire de même avec **caret**, on utilisera **method="svmRadial"** et **prob.model=TRUE**.

```
C <- c(0.001,0.01,1,10,100,1000)
sigma <- c(0.5,1,2,3,4)
gr <- expand.grid(C=C,sigma=sigma)
ctrl <- trainControl(...)
res.caret1 <- train(...,prob.model=TRUE)
res.caret1
```

6. Visualiser la règle sélectionnée.

4.4 Support vector régression

Dans un contexte de régression (lorsque $y_i \in \mathbb{R}$), on ne recherche plus la l'hyperplan qui va séparer au mieux. On va dans ce cas là cherche à approcher au mieux les valeurs de y_i. Cela revient à chercher $w \in \mathbb{R}^p$ et $b \in \mathbb{R}$ tels que

$$|\langle w, x_i \rangle + b - y_i| \leq \varepsilon$$

avec $\varepsilon > 0$ petit à choisir par l'utilisateur. Par analogie avec la **SVM** binaire, on va ainsi chercher (w, b) qui minimisent

$$\frac{1}{2}\|w\|^2$$

$$\text{sous les contraintes } |y_i - \langle w, x_i \rangle - b| \leq \varepsilon, \ i = 1, \dots, n,$$

Les contraintes impliquent que toute les observations doivent se définir dans une **marge** ou **bande** de taille 2ε. Cette hypothèse peut amener l'utilisateur à utiliser des valeurs de ε très grandes et empêcher la solution de bien ajuster le nuage de points. Pour pallier à cela, on introduit, comme dans le cas de la SVM binaire, des **variables ressorts** qui vont autoriser certaines observations à se situer en dehors de la marge. Le problème revient alors à trouver $(w, b, \xi, \xi^\star)$ qui minimise

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^\star)$$

$$\text{sous les contraintes } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i, \ i = 1, \dots, n, \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^\star, \ i = 1, \dots, n \\ \xi_i \geq 0, \xi_i^\star \geq 0, \ i = 1, \dots, n \end{cases}$$

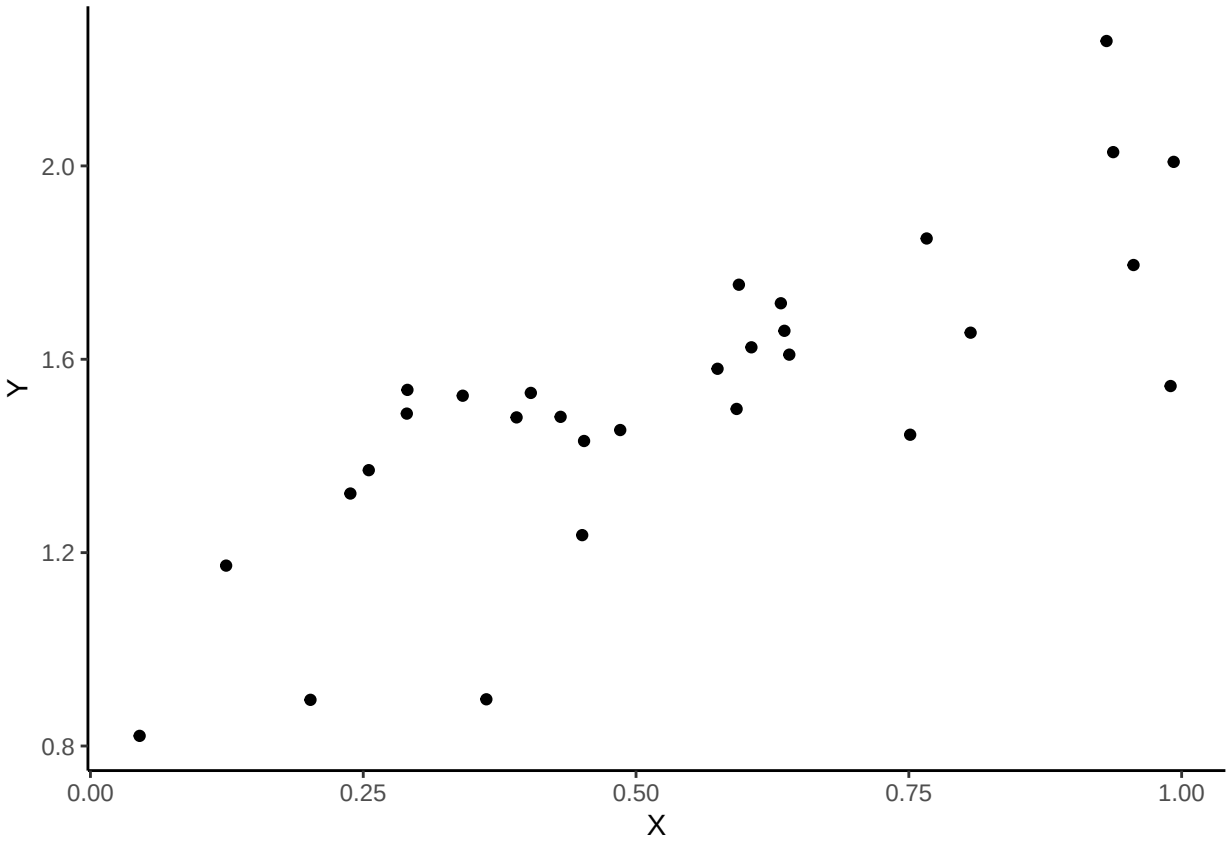
Les solutions s'obtiennent exactement de la même façon que dans le cas binaire. On montre notamment que $w^\star$ s'écrit comme une combinaison linéaire de vecteurs supports :

$$w^\star = \sum_{i=1}^{n}(\alpha_i^\star - \alpha_i)x_i.$$

Les vecteurs supports sont les observations vérifiant $\alpha_i^\star - \alpha_i \neq 0$. Ici encore il faudra calibrer le paramètre C et on pourra utiliser l'astuce du noyau.

On considère le nuage de points $(x_i, y_i), i = 1, \dots, n$ définie ci-dessous :

```
set.seed(321)
n <- 30
X <- runif(n)
eps <- rnorm(n,0,0.2)
Y <- 1+X+eps
df <- data.frame(X,Y)
p1 <- ggplot(df)+aes(x=X,y=Y)+geom_point()
p1
```



On souhaite faire une **SVR** permettant de prédire Y par X. On peut l'obtenir sur **R** toujours avec la fonction **svm** de **e1071** :

```
svr1 <- svm(Y~.,data=df,kernel="linear",epsilon=0.5,cost=100,scale=FALSE)
```

On choisit ici exceptionnellement de ne pas réduire les X.

1. Écrire une fonction **R** qui, à partir d'un objet `svm`, calcule l'équation de la droite de la SVR. Cette fonction pourra également tracer cette doite ainsi que la marge.
2. Comparer la **SVR** précédente avec celle utilisant `epsilon=0.7`.
3. On ajoute le point de coordonnées $(0.05, 3)$ aux données. Discuter de la **SVR** pour ce nouveau jeu de données en utilisant plusieurs valeurs pour `C` et `epsilon`.

```
df1 <- df %>% bind_rows(data.frame(X=0.05,Y=3))
```

4.5 SVM sur les données spam

On considère le jeu de données `spam` où le problème est d'expliquer la variable `type` par les autres.

```
data(spam)
summary(spam$type)
nonspam    spam
   2788    1813
```

On veut comparer plusieurs **svm** en utilisant le package `kernlab`. On pourra trouver un descriptif du package à cette adresse https://www.jstatsoft.org/article/view/v011i09.

1. Utiliser la fonction **ksvm** pour faire une svm linéaire et une svm à noyau gaussien. On prendra comme paramètre 1 pour `C` et pour le paramètre du noyau gaussien.
2. Évaluer la performance des 2 svm précédentes en calculant l'erreur de classification par validation croisée 5 blocs. Comparer ces deux algorithmes.
3. Refaire la svm à noyau gaussien avec l'option `kpar='automatic'`. Expliquer.
4. On s'intéresse maintenant à l'AUC. À partir de validation croisée, sélectionner un noyau (linéaire ou gaussien) ainsi que des valeurs de paramètres associés au noyau, sans oublier le paramètre C. On pourra utiliser le package **caret** et comparer le résultat obtenu à celui d'une forêt aléatoire.

4.6 Exercices

Exercice 4.1 (Résolution du problème d'optimisation dans le cas séparable)**.**

On considère n observations $(x_1, y_1), \ldots, (x_n, y_n)$ telles que $(x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}$. On cherche à expliquer la variable Y par X. On considère l'algorithme SVM et on se place dans le cas où les données sont séparables.

1. Soit $\mathcal{H}$ un hyperplan séparateur d'équation $\langle w, x\rangle + b = 0$ où $w \in \mathbb{R}^p, b \in \mathbb{R}$. Exprimer la distance entre $x_i, i = 1, \ldots, n$ et $\mathcal{H}$ en fonction de w et b.
2. Expliquer la logique du problème d'optimisation

$$\max_{w,b,\|w\|=1} M$$

$$\text{sous les contraintes } y_i(\langle w, x_i\rangle + b) \geq M,\ i = 1, \ldots, n.$$

3. Montrer que ce problème peut se réécrire

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$\text{sous les contraintes } y_i(\langle w, x_i\rangle + b) \geq 1,\ i = 1, \ldots, n.$$

4. On rappelle que pour la minimisation d'une fonction $h : \mathbb{R}^p \to \mathbb{R}$ sous contraintes affines $g_i(u) \geq 0, i = 1, \ldots, n$, le Lagrangien s'écrit

$$L(u, \alpha) = h(u) - \sum_{i=1}^{n} \alpha_i g_i(u).$$

Si on désigne par $u_\alpha = \text{argmin}_u\, L(u, \alpha)$, la fonction duale est alors donnée par

$$\theta(\alpha) = L(u_\alpha, \alpha) = \min_{u \in \mathbb{R}^p} L(u, \alpha),$$

et le problème dual consiste à maximiser $\theta(\alpha)$ sous les contraintes $\alpha_i \geq 0$. En désignant par $\alpha^\star$ la solution de ce problème, on déduit la solution du problème primal $u^\star = u_{\alpha^\star}$. Les conditions de Karush-Kuhn-Tucker sont données par

— $\alpha_i^\star \geq 0$.
— $g_i(u_{\alpha^\star}) \geq 0$.
— $\alpha_i^\star g_i(u_{\alpha^\star}) = 0$.

a. Écrire le Lagrangien du problème considéré et en déduire une expression de w en fonction des α_i et des observations.
b. Écrire la fonction duale.
c. Écrire les conditions KKT et en déduire les solutions $w^\star$ et $b^\star$.
d. Interpréter les conditions KKT.

Exercice 4.2 (Règle svm à partir de sorties R)**.**

On considère n observations $(x_1, y_1), \ldots, (x_n, y_n)$ telles que $(x_i, y_i) \in \mathbb{R}^3 \times \{-1, 1\}$. On cherche à expliquer la variable Y par $X = (X_1, X_2, X_3)$. On considère l'algorithme SVM et on se place dans le cas où les données sont séparables. On rappelle que cet algorithme consiste à chercher une droite d'équation $w^t x + b = 0$ où $(w, b) \in \mathbb{R}^3 \times \mathbb{R}$ sont solutions du problème d'optimisation (problème primal)

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$\text{sous les contraintes } y_i(w^t x_i + b) \geq 1, \ i = 1, \ldots, n.$$

On désigne par $\alpha_i^\star, i = 1, \ldots, n$, les solutions du problème dual et par $(w^\star, b^\star)$ les solutions du problème ci-dessus.

1. Donner la formule permettant de calculer $w^\star$ en fonction des $\alpha_i^\star$.
2. Expliquer comment on classe un nouveau point $x \in \mathbb{R}^3$ par la méthode **svm**.
3. Les données se trouvent dans un dataframe `df`. On exécute

```
set.seed(1234)
n <- 100
X <- data.frame(X1=runif(n),X2=runif(n),X3=runif(n))
X <- data.frame(X1=scale(runif(n)),X2=scale(runif(n)),X3=scale(runif(n)))
Y <- rep(-1,100)
Y[X[,1]<X[,2]] <- 1
#Y <- (apply(X,1,sum)<=0) %>% as.numeric() %>% as.factor()
df <- data.frame(X,Y=as.factor(Y))
```

```
mod.svm <- svm(Y~.,data=df,kernel="linear",cost=10000000000)
```

et on obtient

```
df[mod.svm$index,]
     X1   X2   X3  Y
51 -1.1 -1.0 -1.0  1
92  0.7  0.8  1.1  1
31  0.7  0.5 -1.0 -1
37 -0.5 -0.6  0.3 -1
mod.svm$coefs
     [,1]
[1,]   59
[2,]   49
[3,]  -30
[4,]  -79
mod.svm$rho
[1] -0.5
```

Calculer les valeurs de $w^\star$ et $b^\star$. En déduire la règle de classification.

4. On dispose d'une nouvelle observation $x = (1, -0.5, -1)$. Dans quel groupe (`-1` ou `1`) l'algorithme affecte cette nouvelle donnée ?

5 Agrégation : forêts aléatoires et gradient boosting

Les méthodes par arbres présentées précédemment sont des algorithmes qui possèdent tout un tas de qualités (facile à mettre en œuvre, interprétable...). Ce sont néanmoins rarement les algorithmes qui se révèlent les plus performants. Les méthodes d'agrégation d'arbres présentées dans cette partie sont souvent beaucoup

plus pertinentes, notamment en terme de qualité de prédiction. Elles consistent à construire un très grand nombre d'arbres "simples" : $g_1, \ldots, g_B$ et à les agréger en faisant la moyenne :

$$\frac{1}{B}\sum_{k=1}^{B} g_k(x).$$

Les forêts aléatoires (Breiman, 2001) et le gradient boosting (Friedman, 2001) utilisent ce procédé d'agrégation.

5.1 Forêts aléatoires

L'algorithme des forêts aléatoires consiste à construire des arbres sur des échantillons bootstrap et à les agréger. Il peut s'écrire de la façon suivante :

Entrées :

- $x \in \mathbb{R}^d$ l'observation à prévoir, $\mathcal{D}_n$ l'échantillon ;
- B nombre d'arbres ; n_{max} nombre max d'observations par nœud
- $m \in \{1, \ldots, d\}$ le nombre de variables candidates pour découper un nœud.

Algorithme : pour $k = 1, \ldots, B$:

1. Tirer un échantillon *bootstrap* dans $\mathcal{D}_n$
2. Construire un *arbre CART sur cet échantillon bootstrap*, chaque coupure est sélectionnée en minimisant la fonction de coût de CART sur un ensemble de m *variables choisies au hasard* parmi les d. On note $T(., \theta_k, \mathcal{D}_n)$ l'arbre construit.

Sortie : l'estimateur $T_B(x) = \frac{1}{B}\sum_{k=1}^{B} T(x, \theta_k, \mathcal{D}_n)$.

Cet algorithme peut être utilisé sur **R** avec la fonction **randomForest** du package **randomForest**. Nous la présentons à travers l'exemple du jeu de données **spam** du package **kernlab**.

```
library(kernlab)
data(spam)
set.seed(1234)
spam <- spam[sample(nrow(spam)),]
```

Le problème est d'expliquer la variable binaire **type** par les autres.

1. A l'aide de la fonction **randomForest** du package **randomForest**, ajuster une forêt aléatoire pour répondre au problème posé.
2. Appliquer la fonction **plot** à l'objet construit avec **randomForest** et expliquer le graphe obtenu. A quoi peut servir ce graphe en pratique ?
3. Construire la forêt avec **mtry=1** et comparer ses performances avec celle construite précédemment.
4. Utiliser la fonction **train** du package **caret** pour choisir le paramètre **mtry** dans la grille **seq(1,30,by=5)**.
5. Construire la forêt avec le paramètre **mtry** sélectionné. Calculer l'importance des variables et représenter ces importance à l'aide d'un diagramme en barres.
6. La fonction **ranger** du package **ranger** permet également de calculer des forêts aléatoires. Comparer les temps de calcul de cette fonction avec **randomForest**

5.2 Gradient boosting

Les algorithmes de gradient boosting permettent de minimiser des pertes empiriques de la forme

$$\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f(x_i)).$$

où $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ est une fonction de coût convexe en son second argument. Il existe plusieurs type d'algorithmes boosting. Un des plus connus et utilisés a été proposé par Friedman (2001), c'est la version que nous étudions dans cette partie.

Cette approche propose de chercher la meilleure combinaison linéaire d'arbres binaires, c'est-à-dire que l'on recherche $g(x) = \sum_{m=1}^{M}\alpha_m h_m(x)$ qui minimise

$$\mathcal{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, g(x_i)).$$

Optimiser sur toutes les combinaisons d'arbres binaires se révélant souvent trop compliqué, Friedman (2001) utilise une descente de gradient pour construire la combinaison d'abres de façon récursive. L'algorithme est le suivant :

Entrées :

— $d_n = (x_1, y_1), \ldots, (x_n, y_n)$ l'échantillon, λ un paramètre de régularisation tel que $0 < \lambda \leq 1$.
— $M \in \mathbb{N}$ le nombre d'itérations.
— paramètres de l'arbre (nombre de coupures...)

Itérations :

1. Initialisation : $g_0(.) = \text{argmin}_c \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, c)$
2. Pour $m = 1$ à M :
   a. Calculer l'opposé du gradient $-\frac{\partial}{\partial g(x_i)}\ell(y_i, g(x_i))$ et l'évaluer aux points $g_{m-1}(x_i)$:

$$U_i = -\frac{\partial}{\partial g(x_i)}\ell(y_i, g(x_i))\Big|_{g(x_i)=g_{m-1}(x_i)}, \quad i = 1, \ldots, n.$$

   b. Ajuster un arbre sur l'échantillon $(x_1, U_1), \ldots, (x_n, U_n)$, on le note h_m.
   c. Mise à jour : $g_m(x) = g_{m-1}(x) + \lambda h_m(x)$.

Sortie : la suite $(g_m(x))_m$.

Sur **R** On peut utiliser différents packages pour faire du gradient boosting. Nous utilisons ici le package **gbm** (Ridgeway, 2006).

5.2.1 Un exemple simple en régression

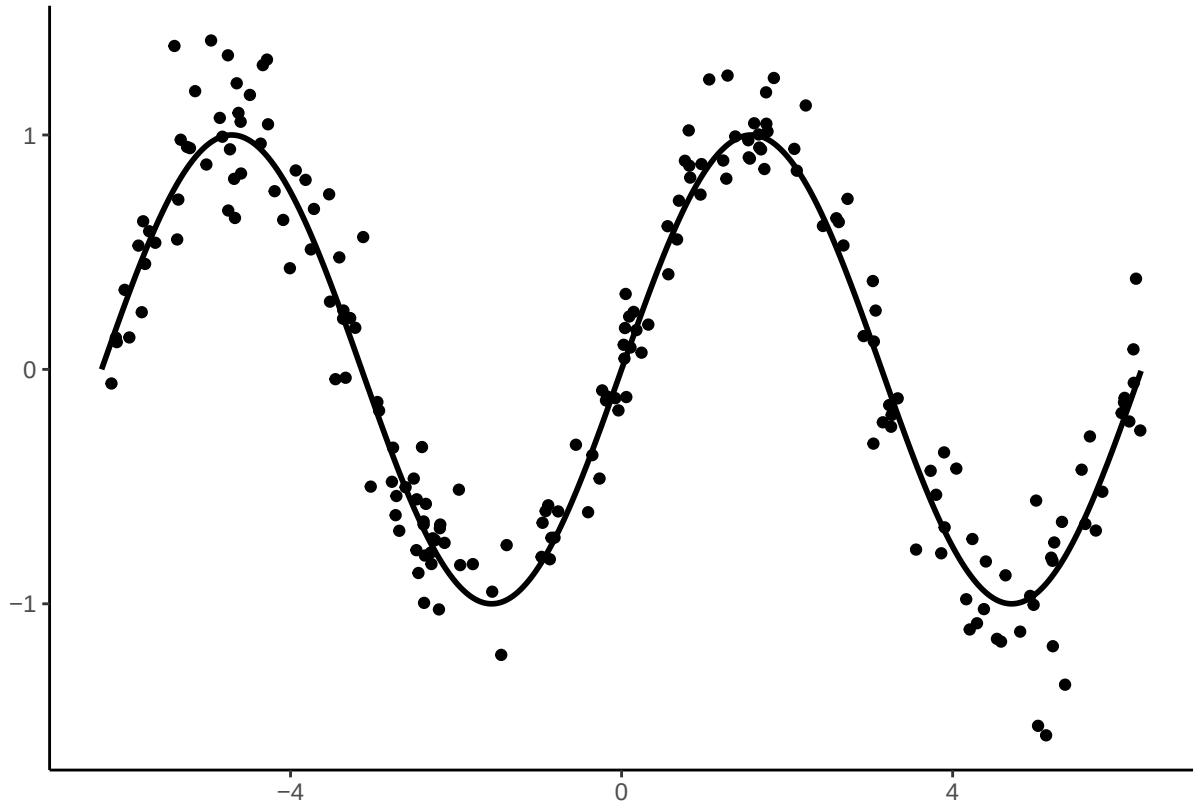
On considère un jeu de données $(x_i, y_i), i = 1, \ldots, 200$ issu d'un modèle de régression

$$y_i = m(x_i) + \varepsilon_i$$

où la vraie fonction de régression est la fonction **sinus** (mais on va faire comme si on ne le savait pas).

```
x <- seq(-2*pi,2*pi,by=0.01)
y <- sin(x)
set.seed(1234)
```

```
X <- runif(200,-2*pi,2*pi)
Y <- sin(X)+rnorm(200,sd=0.2)
df1 <- data.frame(X,Y)
df2 <- data.frame(X=x,Y=y)
p1 <- ggplot(df1)+aes(x=X,y=Y)+geom_point()+geom_line(data=df2,size=1)+xlab("")+ylab("")
p1
```



1. Rappeler ce que siginifie le L_2-boosting.
2. A l'aide de la fonction **gbm** du package **gbm** construire un algorithme de L_2-boosting. On utilisera 500000 itérations et gardera les autres valeurs par défaut de paramètres.
3. Visualiser l'estimateur à la première itération. On pourra faire un **predict** avec l'option `n.trees`.
4. Faire de même pour les itérations 1000 et 500000.
5. Sélectionner le nombre d'itérations par la procédure de votre choix.

5.2.2 Adaboost et logitboost pour la classification binaire.

On considère le jeu de données **spam** du package **kernlab**.

```
library(kernlab)
data(spam)
set.seed(1234)
spam <- spam[sample(nrow(spam)),]
```

1. Exécuter la commande

```
model_ada1 <- gbm(type~.,data=spam,distribution="adaboost",interaction.depth=2,
              shrinkage=0.05,n.trees=500)
```

2. Proposer une correction permettant de faire fonctionner l'algorithme.
3. Expliciter le modèle ajusté par la commande précédente.
4. Effectuer un **summary** du modèle ajusté. Expliquer la sortie.
5. Utiliser la fonction **vip** du package **vip** pour retrouver ce sorties.
6. Sélectionner le nombre d'itérations pour l'algorithme adaboost en faisant de la validation croisée 5 blocs.
7. Faire la même procédure en changeant la valeur du paramètre **shrinkage**. Interpréter.
8. Expliquer la différence entre **adaboost** et **logitboost** et précisez comment on peut mettre en œuvre ce dernier algorithme.

5.2.3 Exercices

1. Rappeler la fonction de risque adaboost.
2. Montrer que le risque est minimum en

$$f^\star(x) = \frac{1}{2}\log\frac{\mathbf{P}(Y=1|X=x)}{\mathbf{P}(Y=-1|X=x)}.$$

3. Mêmes questions pour le risque logitboost.

6 Réseaux de neurones avec Keras

Nous présentons ici une introduction au réseau de neurones à l'aide du package **keras**. On pourra trouver une documentation complète ainsi qu'un très bon tutoriel aux adresses suivantes https://keras.rstudio.com et https://tensorflow.rstudio.com/tutorials/beginners/basic-ml/. On commence par charger la librairie

```
library(keras)
#install_keras() 1 seule fois sur la machine
```

On va utiliser des réseaux de neurones pour le jeu de données **spam** où le problème est d'expliquer la variable binaires `type`par les 57 autres variables du jeu de données :

```
library(kernlab)
data(spam)
spamX <- as.matrix(spam[,-58])
#spamY <- to_categorical(as.numeric(spam$type)-1, 2)
spamY <- as.numeric(spam$type)-1
```

On sépare les données en un échantillon d'apprentissage et un échantillon test

```
set.seed(5678)
perm <- sample(4601,3000)
appX <- spamX[perm,]
appY <- spamY[perm]
validX <- spamX[-perm,]
validY <- spamY[-perm]
```

1. A l'aide des données d'apprentissage, entrainer un perceptron simple avec une fonction d'activation **sigmoïde**. On utilisera 30 epochs et des batchs de taille 5.

```
#Définition du modèle
percep.sig <- keras_model_sequential()
percep.sig %>% layer_dense(units=...,input_shape = ...,activation="...")
summary(percep.sig)
percep.sig %>% compile(
  loss="binary_crossentropy",
  optimizer="adam",
  metrics="accuracy"
)
#Entrainement
p.sig <- percep.sig %>% fit(
  x=...,
  y=...,
  epochs=...,
  batch_size=...,
  validation_split=...,
  verbose=0
)
```

2. Faire de même avec la fonction d'activation **softmax**. On utilisera pour cela 2 neurones avec une sortie Y possédant la forme suivante.

```
spamY1 <- to_categorical(as.numeric(spam$type)-1, 2)
appY1 <- spamY1[perm,]
validY1 <- spamY1[-perm,]
```

3. Comparer les performances des deux perceptrons sur les données de validation à l'aide de la fonction **evaluate**.
4. Construire un ou deux réseaux avec deux couches cachées. On pourra faire varier les nombre de neurones dans ces couches. Comparer les performances des réseaux construits.

7 Données déséquilibrées

On parle de **données déséquilibrées** lorsque les deux modalités de la variable cible Y ne sont pas représentées de façon égale dans l'échantillon, ou plus précisément lorsqu'une des deux modalités est fortement majoritaire. Ce contexte est fréquemment rencontré en pratique, on peut citer les cas de détection de fraudes (peu de fraudeurs), de la présence d'une maladie rare (peu de patients atteints), du risque de crédit (peu de mauvais payeurs)... Les algorithmes standards peuvent être mis en difficultés et de nouvelles stratégies doivent être élaborées. Les stratégies classiques permettant de répondre à ce problème consistent à

— utiliser des critères de performance adaptés au déséquilibre ;
— ré-échantillonner les données pour se rapprocher d'une situation d'équilibre.

Nous présentons ces stratégies à travers quelques exercices.

7.1 Critères de performance pour données déséquilibrées

La notion de **risque** en machine learning est capitale puisque c'est à partir de l'estimation de ces risques que l'on **calibre des algorithmes** et que l'on **choisit un algorithme de prévision**. En présence de données déséquilibré, il convient de choisir un risque adapté. En effet, il est le plus souvent important de parvenir à bien identifier des individus de la classe minoritaire. Des critères tels que l'accuracy ou l'erreur de classification ne sont pas pertinents pour ce cadre. On va privilégier des critères comme

— le **balanced accuracy**

$$\text{Bal Acc} = \frac{1}{2}\mathbf{P}(g(X)=1|Y=1) + \frac{1}{2}\mathbf{P}(g(X)=-1|Y=-1) = \frac{\text{TPR}+\text{TNR}}{2}.$$

— le F_1**-score**

$$F_1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

avec

$$\text{Precision} = \mathbf{P}(Y=1|g(X)=1) \quad \text{et} \quad \text{Recall} = \mathbf{P}(g(X)=1|Y=1).$$

— le **kappa de Cohen**

$$\kappa = \frac{\mathbf{P}(a) - \mathbf{P}(e)}{1 - \mathbf{P}(e)}$$

où $\mathbf{P}(a)$ représente l'accuracy et $\mathbf{P}(e)$ l'accuracy sous une hypothèse d'indépendance.

— la courbe ROC et l'AUC...

Comme d'habitude, ces critères sont inconnus et doivent être estimés par des méthodes de ré-échantillonnage de type validation croisée.

Exercice 7.1 (Calculer des critères)**.**

1. Générer un vecteur d'observations $\mathbf{Y}$ de taille 500 selon une loi de Bernoulli de paramètre 0.05.
2. Générer un vecteur de prévisions **P1** de taille 500 selon une loi de Bernoulli de paramètre 0.01.
3. Générer un vecteur de prévision **P2** de taille 500 tel que

$$\mathcal{L}(P2|Y=0) = \mathcal{B}(0.10) \quad \text{et} \quad \mathcal{L}(P2|Y=1) = \mathcal{B}(0.85).$$

4. Dresser les tables de contingence de **P1** et **P2** à l'aide de **table**. Commenter.
5. Pour **P2**, calculer, avec les fonctions usuelles de `R`, l'**accuracy**, le **recall** et la **précision**.
6. En déduire le F1-score.
7. Même question pour le κ de Cohen.
8. Retrouver ces indicateurs à l'aide de la fonction **confusionMatrix** de **caret** puis comparer les prévisions **P1** et **P2**.

7.2 Ré-équilibrage

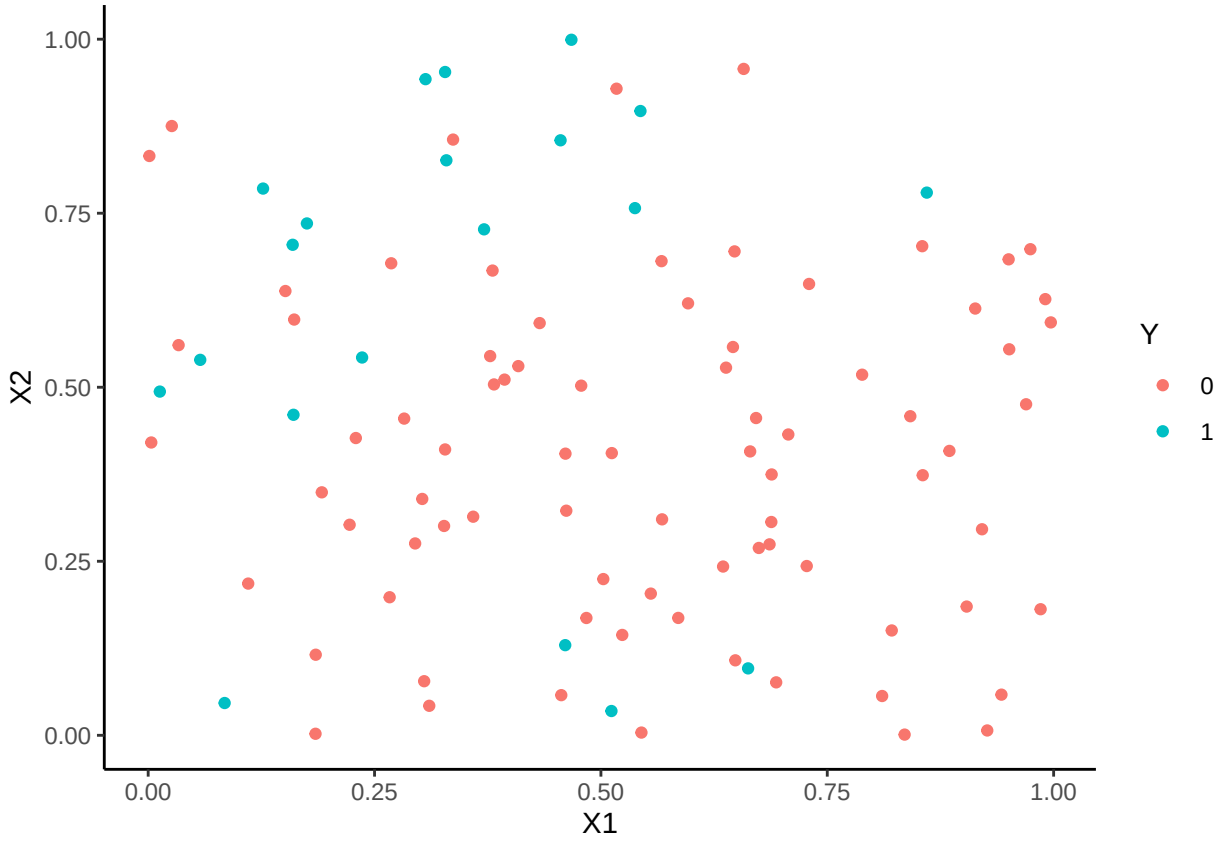
En complément du choix d'un **critère pertinent**, il peut être intéressant de tenter de **ré-équilibrer** l'échantillon pour aider les algorithmes à mieux **détecter les individus de la classe minoritaire**. Les méthodes classiques consistent à créer de nouvelles observations de la classe minoritaire (**oversampling**) et/ou supprimer des individus de la classe minoritaire (**undersampling**).

Exercice 7.2 (Quelques algorithmes de ré-équilibrage)**.**

On considère le jeu de données `df` ci-dessous où on cherche à prédire `Y` par `X1` et `X2`.

```
n <- 2000
set.seed(1234)
X1 <- runif(n)
set.seed(5678)
X2 <- runif(n)
set.seed(9012)
R1 <- X1<=0.25
R2 <- (X1>0.25 & X2>=0.75)
R3 <- (X1>0.25 & X2<0.75)
Y <- rep(0,n)
```

```
Y[R1] <- rbinom(sum(R1),1,0.75)
Y[R2] <- rbinom(sum(R2),1,0.75)
Y[R3] <- rbinom(sum(R3),1,0.25)
df1 <- data.frame(X1,X2,Y)
df1$Y <- factor(df1$Y)
indDY1 <- which(df1$Y==1)
df1.1 <- df1[-indDY1[1:650],]
df1.2 <- df1.1[sample(nrow(df1.1),1000),]
df <- df1.2[sample(nrow(df1.2),100),]
rownames(df) <- NULL
p1 <- ggplot(df)+aes(x=X1,y=X2,color=Y)+geom_point()
p1
```



On a ici 4 fois plus d'observations dans le groupe 0.

```
summary(df$Y)
 0  1
80 20
```

1. On commence par faire du **oversampling** avec la fonction `RandOverClassif`.
   a. Effectuer le ré-échantillonnage et expliquer.
   b. Corriger les paramètres de la fonction de manière à avoir 80 observations dans le groupe 0 et 60 dans le groupe 1.
2. On s'intéresse maintenant à l'algorithme **SMOTE**

a. Exécuter la fonction **SmoteClassif** avec `k=3` et les les paramètres par défaut
b. Visualiser les **observations smote**.
c. Corriger les paramètres de la fonction de manière à avoir 80 observations dans le groupe 0 et 60 dans le groupe 1.

3. On souhaite maintenant ré-équilibrer par **random undersampling**. Utiliser la fonction **RandUnderClassif** pour effectuer un tel ré-équilibrage. Ici encore on pourra faire varier les paramètres.
4. On passe maintenant à l'algorithme **Tomek**.
   a. Sans utiliser la fonction **TomekClassif** identifier les paires d'observations qui ont un **lien de Tomek**. On pourra utiliser la fonction **nng** du package `cccd`.
   b. Retrouver ces paires à l'aide de la fonction Tomek LinK.
   c. Visualiser les observations supprimées. On prendra soin d'expliquer l'option `rem` de **TomekClassif**.

Exercice 7.3 (Comparaison de méthodes de ré-équilibrage)**.**

On considère 3 jeux de données `df1`, `df2` et `df3`.

```
n <- 2000
set.seed(12345)
X1 <- runif(n)
set.seed(5678)
X2 <- runif(n)
set.seed(9012)
R1 <- X1<=0.25
R2 <- (X1>0.25 & X2>=0.75)
R3 <- (X1>0.25 & X2<0.75)
Y <- rep(0,n)
Y[R1] <- rbinom(sum(R1),1,0.75)
Y[R2] <- rbinom(sum(R2),1,0.75)
Y[R3] <- rbinom(sum(R3),1,0.25)
df1 <- data.frame(X1,X2,Y)
df1$Y <- factor(df1$Y)
indDY1 <- which(df1$Y==1)
df2 <- df1[-indDY1[1:400],]
df3 <- df1[-indDY1[1:700],]
df1 <- df1[sample(nrow(df1),1000),]
df2 <- df2[sample(nrow(df2),1000),]
df3 <- df3[sample(nrow(df3),1000),]
```

1. Comparer la distribution de **Y** pour ces trois jeux de données et visualiser les observations.
2. On sépare ces 3 échantillons en un échantillon d'apprentissage et un échantillon test.

```
set.seed(123)
a1 <- createDataPartition(1:nrow(df1),p=2/3)
a2 <- createDataPartition(1:nrow(df2),p=2/3)
a3 <- createDataPartition(1:nrow(df3),p=2/3)
train1 <- df1[a1$Resample1,]
train2 <- df2[a2$Resample1,]
train3 <- df3[a3$Resample1,]
test1 <- df1[-a1$Resample1,]
test2 <- df2[-a2$Resample1,]
test3 <- df3[-a3$Resample1,]
```

Ajuster une forêt aléatoire sur les 3 échantillon d'apprentissage, calculer les labels prédits sur les échantillons tests et estimer les différents indicateurs vus en cours à l'aide de **confusionMatrix**.

3. On considère uniquement l'échantillon **df3**. Refaire l'analyse précédente en utilisant des techniques de ré-échantillonnage.

7.3 Exercices supplémentaires

Exercice 7.4 (Echantillonnage rétrospectif)**.**

Dans le cadre de l'échantillonnage rétrospectif pour le modèle logistique vu en cours, démontrer la propriété qui lie le modèle logistique initial au modèle ré-équilibré.

Exercice 7.5 (Echantillonnage rétrospectif)**.**

Une étude cas/témoins est réalisée pour mesurer l'effet du tabac sur une pathologie. Pour ce faire, on choisit $n_1 = 250$ patients atteints de la pathologie (cas) et $n_0 = 250$ patients sains (témoins). Les résultats de l'étude sont présentés ci-dessous

	Fumeur	Non fumeur
Non malade	48	202
Malade	208	42

1. A partir des données obtenues, estimer à l'aide d'un modèle logistique la probabilité d'être atteint pour un fumeur, puis pour un non fumeur.
2. Comment interpréter ces deux probabilités ? Est-ce qu'elles estiment la probabilité d'être atteint pour un individu quelconque dans la population ?
3. Des études précédentes ont montré que cinq individus sur mille sont atteints par la pathologie dans la population entière. En utilisant la propriété de l'exercice précédent, en déduire les probabilités d'être atteint pour un fumeur et un non fumeur dans la population.

8 Comparaison d'algorithmes

Les chapitres précédents ont présenté plusieurs algorithmes permettant de répondre à un problème posé, le plus souvent de classification supervisée. Se pose bien entendu la question de choisir un unique algorithme. Etant donné un échantillon $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_y)\}$ on rappelle qu'un algorithme de prévision est une fonction

$$g : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}$$

qui, à une nouvelle observation $x \in \mathcal{X}$ renverra la prévision $g(x, \mathcal{D}_n)$ calculée à partir de l'échantillon $\mathcal{D}_n$. Cette fonction g peut contenir tout un tas d'étapes comme :

— la gestion des données manquantes
— une procédure de choix de variables
— une méthode pour ré-équilibrer les données
— des procédures pour calibrer des paramètres (qui peuvent éventuellement inclure des validations croisées)
— ...

Le machine learning se focalisant sur la capacité d'un algorithme à bien prédire, les stratégies classiques pour choisir un algorithme vont (une fois de plus) consister à évaluer le pouvoir prédictif de chaque algorithme. Il n'y a rien de bien nouveau puisque cela va reposer sur les techniques présentées aux chapitres 1 :

— choisir un ou plusieurs critères (erreur de classification, AUC, F_1-score...)
— choisir une procédure de ré-échantillonnage pour estimer ce critère (validation hold-out, validation croisée, OOB...).

Nous proposons de développer une stratégie pour choisir un algorithme sur le jeu de données **Internet Advertisements Data Set** disponible sur cette page https://archive.ics.uci.edu/ml/datasets/internet+advertisements. Le problème est d'identifier la présence d'une image publicitaire sur des pages webs. Il comporte

```
ad.data <- read.table("data/ad_data.txt",header=FALSE,sep=",",dec=".",na.strings = "?",strip.white = TRU
dim(ad.data)
[1] 3279 1559
```

Ce jeu de données contient 1558 variables explicatives, ces variables contiennent différentes caractériques de la page web (voir le site où sont présentées les données pour plus d'information). La dernière variable est la variable à expliquer, elle vaut `ad.` si présence d'une publicité, `nonad.` sinon.

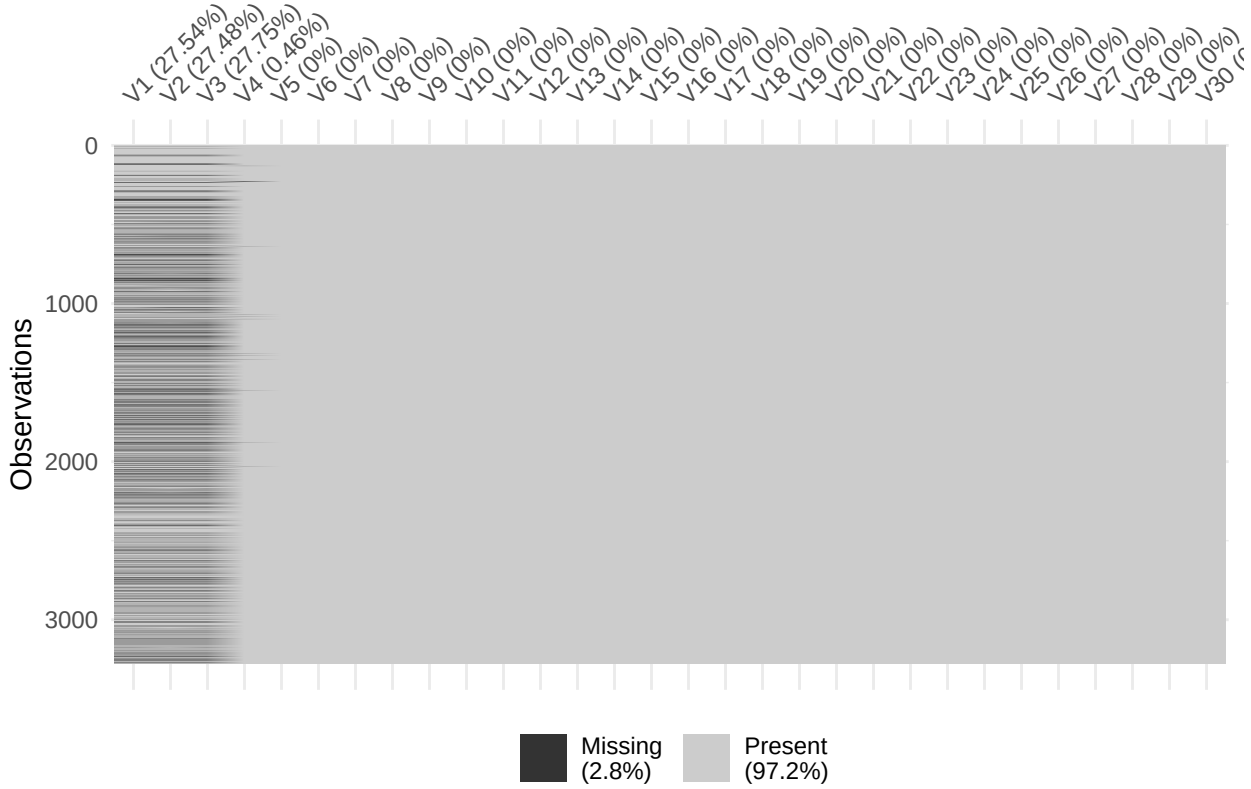
```
names(ad.data)[ncol(ad.data)] <- "Y"
ad.data$Y <- as.factor(ad.data$Y)
summary(ad.data$Y)
   ad. nonad.
   459   2820
```

Ce jeu de données contient des données manquantes.

```
sum(is.na(ad.data))
[1] 2729
```

On peut les visualiser avec

```
library(visdat)
vis_miss(ad.data[,1:30])
```



On remarque que :

— 920 lignes
— 4 colonnes (les 4 premières)

ont au moins une valeur manquante.

```
apply(is.na(ad.data),1,any) %>% sum()
[1] 920
var.na <- apply(is.na(ad.data),2,any)
names(ad.data)[var.na]
[1] "V1" "V2" "V3" "V4"
```

On choisit de retirer ces 4 variables de l'analyse (il faudrait peut-être réfléchir un peu plus...).

```
ad.data1 <- ad.data[,var.na==FALSE]
dim(ad.data1)
[1] 3279 1555
sum(is.na(ad.data1))
[1] 0
```

On se retrouve donc en présence de 3279 individus et 1554 variables explicatives. On construit la matrice des `X` et le vecteur des `Y` qui sont nécessaires pour certaines fonctions comme `glmnet` :

```
X.ad <- model.matrix(Y~.,data=ad.data1)[,-1]
Y.ad <- ad.data1$Y
```

et on transforme la variable cible en 0-1 pour utiliser `gbm` :

```
ad.data2 <- ad.data1 %>% mutate(Y=recode(Y,"ad."=0,"nonad."=1))
```

On souhaite comparer les algorithmes présentés précédemment. Ils nécessitent les packages suivants

```
library(e1071)
library(caret)
library(rpart)
library(glmnet)
library(ranger)
library(gbm)
```

On commence tout d'abord par représenter un algorithme par une fonction **R** qui admettra en entrée un jeu de données et renverra une unique prévision pour de nouveaux individus. On illustre ces fonctions pour prédire ce nouvel individu.

```
newX <- ad.data1[1000,]
newX.X <- matrix(X.ad[1000,],nrow=1)
```

On stockera les prévisions dans l'objet suivant

```
prev <- tibble(algo=c("SVM","arbre","ridge","lasso","foret","ada","logit"),prev=0)
```

— **SVM** à noyau gaussien où le choix des paramètres du noyau se fait par validation croisée 4 blocs :

```
prev.svm <- function(df,newX){
  C <- c(0.01,1,10)
  sigma <- c(0.1,1,3)
  gr <- expand.grid(C=C,sigma=sigma)
  ctrl <- trainControl(method="cv",number=4)
```

```
  cl <- makePSOCKcluster(3)
  registerDoParallel(cl)
  res.svm <- train(Y~.,data=df,method="svmRadial",trControl=ctrl,
                tuneGrid=gr,prob.model=TRUE)
  stopCluster(cl)
  predict(res.svm,newX,type="prob")[2]
}
prev[1,2] <- prev.svm(ad.data1,newX)
```

— **Arbre de classification** où l'élagage est fait selon la procédure **CART** présentée dans le chapitre 3.

```
prev.arbre <- function(df,newX){
  arbre <- rpart(Y~.,data=df,cp=1e-8,minsplit=2)
  cp_opt <- arbre$cptable %>% as.data.frame() %>% filter(xerror==min(xerror)) %>%
dplyr::select(CP) %>% slice(1) %>% as.numeric()
  arbre.opt <- prune(arbre,cp=cp_opt)
  predict(arbre,newdata=newX,type="prob")[,2]
}
prev[2,2] <- prev.arbre(ad.data1,newX)
```

— **Lasso et Ridge** où le paramètre de régularisation est choisi par validation croisée 10 blocs en minimisant la déviance binomiale :

```
prev.ridge <- function(df.X,df.Y,newX){
  ridge <- cv.glmnet(df.X,df.Y,family="binomial",alpha=0)
  as.vector(predict(ridge,newx = newX,type="response"))
}
prev.lasso <- function(df.X,df.Y,newX){
  lasso <- cv.glmnet(df.X,df.Y,family="binomial",alpha=1)
  as.vector(predict(lasso,newx = newX,type="response"))
}
prev[3,2] <- prev.ridge(X.ad,Y.ad,newX.X)
prev[4,2] <- prev.lasso(X.ad,Y.ad,newX.X)
```

— **Forêt aléatoire** avec les paramètres par défaut :

```
prev.foret <- function(df,newX){
  foret <- ranger(Y~.,data=df,probability=TRUE)
  predict(foret,data=newX,type="response")$predictions[,2]
}
prev[5,2] <- prev.foret(ad.data1,newX)
```

— **Adaboost et logitboost** avec le nombre d'itérations choisi par validation croisée 5 blocs :

```
prev.ada <- function(df,newX){
  ada <- gbm(Y~.,data=df,distribution="adaboost",interaction.depth=2,
         bag.fraction=1,cv.folds = 5,n.trees=500)
  nb.it <- gbm.perf(ada,plot.it=FALSE)
  predict(ada,newdata=newX,n.trees=nb.it,type="response")
}

prev.logit <- function(df,newX){
  logit <- gbm(Y~.,data=df,distribution="bernoulli",interaction.depth=2,
           bag.fraction=1,cv.folds = 5,n.trees=500)
```

```r
  nb.it <- gbm.perf(logit,plot.it=FALSE)
  predict(logit,newdata=newX,n.trees=nb.it,type="response")
}

prev[6,2] <- prev.ada(ad.data2,newX)
prev[7,2] <- prev.logit(ad.data2,newX)
```

On peut visualiser la prévision de chaque algorithme

```r
prev
# A tibble: 7 x 2
  algo   prev
  <chr> <dbl>
1 SVM   0.950
2 arbre 0.990
3 ridge 0.984
4 lasso 0.980
5 foret 0.979
6 ada   0.974
7 logit 0.983
```

Exercice 8.1 (Choix d'un algorithme par validation croisée)**.**

Choisir un algorithme parmi les précédents en utilisant comme critère l'erreur de classification ainsi que la courbe ROC et l'AUC. On pourra faire une validation croisée 10 blocs (même si ça peut être un peu long...).

Exercice 8.2 (Choix d'un algorithme de ré-équilibrage par validation croisée)**.**

On considère le même jeu de données que précédemment. Choisir un algorithme de ré-équilibrage par validation croisée. Il s'agira de combiner des méthodes de ré-équilibrage (random over/under sampling, smote, tomek...) avec des algorithmes de prévision de machine learning. On pourra se restreindre au modèle logistique avec calcul des estimateurs par maximum de vraisemblance, ridge, lasso...

Références

Boehmke, B. and Greenwell, B. (2019). *Hands-On Machine Learning with R*. CRC Press.

Breiman, L. (2001). Random forests. *Machine learning*, 45 :5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Wadsworth & Brooks.

Friedman, J. H. (2001). Greedy function approximation : A gradient boosting machine. *Annals of Statistics*, 29 :1189–1232.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, second edition.

Ridgeway, G. (2006). Generalized boosted models : A guide to the gbm package.