L2 MIASHS
Année 2021-22

Statistiques Descriptives 2

TD 3 de Statistiques Descriptives 2

On s'intéresse à la construction des arbres de décision pour la classification et la régression.

Exercice 1 On s'intéresse à un problème de diagnostique chez des patients atteints d'une tumeur afin de savoir si elle est maligne ou pas. On a 3 critères relevés pour chaque patient :

1. la tumeur est-elle grosse ?
2. la tumeur est-elle dure ?
3. la tumeur est-elle symétrique ?

Ces attributs sont respectivement abrégés L comme "Large", H comme "Hard" et S comme "Symmetric". Pour les données recueillies, les spécialistes savent également si la tumeur s'est avérée maligne ou non. Ce fait est indiqué dans la dernière colonne marquée d'un M, ou T signifie "True" et F signifie "False". On a les résultats suivants :

patient	feature X_1 =L	feature X_2 =H	feature X_3 =S	réponse Y=M
1	T	F	F	T
2	F	F	F	T
3	F	T	T	T
4	F	F	T	F
5	T	F	T	F

1. Faire un arbre de décision pour ce problème basé sur l'entropie conditionnelle.
   Solution : Il y a, indépendemment des facteurs $X_1 L$, $X_2 = H$, $X_3 = S$, une distribution empirique de 3 réponses T et 2 réponses F. L'entropie de la distribution empirique des réponses est donc

$$H(Y) = \frac{3}{5}\log\left(\frac{1}{\frac{3}{5}}\right) + \frac{2}{5}\log\left(\frac{1}{\frac{2}{5}}\right) = 0.292.$$

   On suppose d'abord que l'on choisit la variable (ou "le noeud" en terminologie théorie des graphes) $X_1 = L$ comme racine de l'arbre.

   — Si la réponse à L est T, on a 50% de réponse $Y = T$ et 50% de réponse $Y = F$.

   — L'entropie conditionnelle sachant $L = T$ est alors

$$H(Y \mid L = T) = 0.5\log\left(\frac{1}{0.5}\right) + 0.5\log\left(\frac{1}{0.5}\right) = \log(2) = 0.3$$

— Si la réponse à L est F, on a $2/3$ de réponse $Y = T$ et $1/3$ de réponse $Y = F$.

— L'entropie consitionnelle sachant $L = F$ est alors

$$H(Y \mid L = F) = \frac{2}{3}\log\left(\frac{1}{\frac{2}{3}}\right) + \frac{1}{3}\log\left(\frac{1}{\frac{1}{3}}\right) = 0.28$$

— L'entropie de Y conditionnelle à "L" est donc

$$H(Y \mid L) = \underbrace{\mathbb{P}(L = T)}_{\approx 2/5} H(Y \mid L = T) + \underbrace{\mathbb{P}(L = F)}_{\approx 3/5} H(Y \mid L = F) \tag{1}$$

$$= 2/5 \times 0.3 + 3/5 \times 0.28 = 0.288. \tag{2}$$

On suppose maintenant que l'on choisit le noeud H comme racine de l'arbre.

— Si la réponse à H est T, on a 100% de réponse $Y = T$ et 0% de réponse $Y = F$.

— L'entropie conditionnelle sachant $H = T$ est alors

$$H(Y \mid H = T) = 1\log\left(\frac{1}{1}\right) + 0\log\left(\frac{1}{0}\right) = 0$$

— Si la réponse à H est F, on a $2/4$ de réponse $Y = T$ et $2/4$ de réponse $Y = F$.

— L'entropie conditionnelle sachant $L = F$ est alors

$$H(Y \mid H = F) = \frac{2}{4}\log\left(\frac{1}{\frac{2}{4}}\right) + \frac{2}{4}\log\left(\frac{1}{\frac{2}{4}}\right) = \log(2) = 0.3$$

— L'entropie de Y conditionnelle à "H" est donc

$$H(Y \mid H) = \underbrace{\mathbb{P}(H = T)}_{\approx 1/5} H(Y \mid H = T) + \underbrace{\mathbb{P}(H = F)}_{\approx 4/5} H(Y \mid H = F) \tag{3}$$

$$= \frac{4}{5}0.3 = 0.24 \tag{4}$$

On remarque donc que la variable H a une plus petite entropie que la variable L

On suppose enfin que l'on choisit le noeud S comme racine de l'arbre.

— Si la réponse à S est T, on a $1/3$ de réponse $Y = T$ et $2/3$ de réponse $Y = F$.

— L'entropie consitionnelle sachant $S = T$ est alors

$$H(Y \mid S = T) = \frac{1}{3}\log\left(\frac{1}{\frac{1}{3}}\right) + \frac{2}{3}\log\left(\frac{1}{\frac{2}{3}}\right) = 0.276.$$

— Si la réponse à S est F, on a $2/2$ de réponse T et $0/2$ réponse F.
  — L'entropie consitionnelle sachant $S = F$ est alors

$$H(Y \mid S = F) = 1 \log\left(\frac{1}{1}\right) + 0 \log\left(\frac{1}{0}\right) = 0$$

— L'entropie de Y conditionnelle à "S" est donc

$$H(Y \mid S) = \underbrace{\mathbb{P}(S = T)}_{\approx 3/5} H(Y \mid S = T) + \underbrace{\mathbb{P}(S = F)}_{\approx 2/5} H(Y \mid S = F) = 0.276 \times \frac{3}{5} = 0.167. \tag{5}$$

Et donc c'est la variable S qui a la plus petite entropie vis à vis de la malignité.
On choisir S comme premier noeud de l'arbre de décision et on veut maintenant choisir le deuxième noeud de l'arbre : L ou H ou pas d'autre question ?

— On suppose que $S = T$, et on calcule l'entropie de Y sachant X et $S = T$ pour $X = L$ et $X = H$.
  — On commence par calculer l'entropie de Y sachant L

$$\begin{aligned} H(Y \mid L, S = T) = &\underbrace{\mathbb{P}(L = T \mid S = T)}_{\approx 1/3} H(Y \mid L = T, S = T) \\ &+ \underbrace{\mathbb{P}(L = F \mid S = T)}_{\approx 2/3} H(Y \mid L = F, S = T) \end{aligned}$$

Or

$$H(Y \mid L = T, S = T) = 1 \log(\frac{1}{1}) + 0 \log(\frac{0}{0}) = 0$$

et

$$H(Y \mid L = F, S = T) = \frac{1}{2} \log(\frac{1}{\frac{1}{2}}) + \frac{1}{2} \log(\frac{1}{\frac{1}{2}}) = \log(2) = 0.3.$$

Donc

$$H(Y \mid L, S = T) = \frac{1}{3} \times 0 + \frac{2}{3} \times 0.3 = 0.2. \tag{6}$$

  — On fait le même calcul pour l'entropie de Y sachant H

$$\begin{aligned} H(Y \mid H, S = T) = &\underbrace{\mathbb{P}(H = T \mid S = T)}_{\approx 1/3} H(Y \mid H = T, S = T) \\ &+ \underbrace{\mathbb{P}(H = F \mid S = T)}_{\approx 2/3} H(Y \mid H = F, S = T) \end{aligned}$$

Or

$$H(Y \mid H = T, S = T) = 1 \log(\frac{1}{1}) + 0 \log(\frac{0}{0}) = 0$$

et

$$H(Y \mid H = F, S = T) = \frac{2}{2} \log(\frac{1}{\frac{2}{2}}) + \frac{0}{2} \log(\frac{1}{\frac{0}{2}}) = 0.$$

Donc

$$H(Y \mid H, S = T) = \frac{1}{3} \times 0 + \frac{2}{3} \times 0 = 0. \quad (7)$$

— La variable H a donc la plus petite entropie sachant $S = T$.
— On suppose maintenant que $S = F$ alors $H(Y \mid S = F) = 0$ et donc on n'a pas plus de question à poser.

La construction de l'arbre de décision est la suivante :

— Si $S = F$ alors $Y = T$
— Si $S = T$ alors
  — si $H = T$ alors $Y = T$
  — si $H = F$ alors $Y = F$

Le fait que les conclusions soient sans appel et que aucune donnée ne contredise les décisions est tout de même très suspect. Il est impossible qu'avec des données réelles, on n'ait pas de fluctuations contradictoires. Ceci n'est pas surprenant si on réalise que l'on n'a pris en considération que 5 données.

2. Comment prédire la malignité pour un patient $L = F$, $H = F$ et $S = T$? et $L = T$, $H = T$ et $S = T$?

   Solution : On a une prédiction de classification égale à F dans le premier cas et T dans le second cas.

Exercice 2 Un site de vente en ligne de musique voudrait analyser le comportement de ses clients. Les facteurs sont les suivants :

— Il y a trois types de musique : Heavy Metal, Pop, RnB.
— Le prix est catégorisé en "E" (expencive) et "C" (cheap).

La réponse à analyser est la CATEGORY qui lorsqu'elle est "+" indique que le client est allé au bout de la transaction et "-" s'il a abandonné la transaction. Les données sont ci-dessous (évidemment, dans le vrai fichier, il y a des centaines de milliers de données !).

client	Type de musique	Prix	Transaction
1	TYPE = H	PRICE = E	CATEGORY = +
2	TYPE = R	PRICE = C	CATEGORY = +
3	TYPE = R	PRICE = E	CATEGORY = +
4	TYPE = H	PRICE = C	CATEGORY = +
5	TYPE = P	PRICE = C	CATEGORY = +
6	TYPE = R	PRICE = E	CATEGORY = -
7	TYPE = P	PRICE = E	CATEGORY = -
8	TYPE = P	PRICE = C	CATEGORY = -
9	TYPE = H	PRICE = E	CATEGORY = +
10	TYPE = P	PRICE = E	CATEGORY = -
11	TYPE = R	PRICE = E	CATEGORY = -
12	TYPE = P	PRICE = C	CATEGORY = +
13	TYPE = R	PRICE = E	CATEGORY = -

1. Faire un arbre de décision pour ces données.

Solution : On calcule l'entropie de la réponse $Y = Transaction$

$$H(Y) = \frac{7}{13}\ \log\left(\frac{1}{\frac{7}{13}}\right) + \frac{6}{13}\ \log\left(\frac{1}{\frac{6}{13}}\right) = 0.2998.$$

On prend la première variable X_1=Type de musique, et on calcule l'entropie conditionnelle de Y sachant X_1. Pour cela, il faut d'abord calculer l'entropie conditionnelle de Y sachant $X_1 = H$, puis celle de Y sachant $X_1 = R$, puis celle de Y sachant $X_1 = P$. On commence par

$$\begin{aligned} H(Y \mid X_1 = H) &= \underbrace{\mathbb{P}(Y = + \mid X_1 = H)}_{\approx 1} \log\left(\frac{1}{\mathbb{P}(Y = + \mid X_1 = H)}\right) \\ &\quad + \underbrace{\mathbb{P}(Y = - \mid X_1 = H)}_{\approx 0} \log\left(\frac{1}{\mathbb{P}(Y = - \mid X_1 = H)}\right) \\ &= 0. \end{aligned}$$

On calcule maintenant

$$\begin{aligned} H(Y \mid X_1 = R) &= \underbrace{\mathbb{P}(Y = + \mid X_1 = R)}_{\approx 2/5} \underbrace{\log\left(\frac{1}{\mathbb{P}(Y = + \mid X_1 = R)}\right)}_{\approx \log(5/2)} \\ &\quad + \underbrace{\mathbb{P}(Y = - \mid X_1 = R)}_{\approx 3/5} \underbrace{\log\left(\frac{1}{\mathbb{P}(Y = - \mid X_1 = R)}\right)}_{\approx \log(5/3)} \\ &= 0.292. \end{aligned}$$

On calcule enfin

$$\begin{aligned} H(Y \mid X_1 = P) &= \underbrace{\mathbb{P}(Y = + \mid X_1 = P)}_{\approx 2/5} \underbrace{\log\left(\frac{1}{\mathbb{P}(Y = + \mid X_1 = P)}\right)}_{\approx \log(5/2)} \\ &\quad + \underbrace{\mathbb{P}(Y = - \mid X_1 = P)}_{\approx 3/5} \underbrace{\log\left(\frac{1}{\mathbb{P}(Y = - \mid X_1 = P)}\right)}_{\approx \log(5/3)} \\ &= 0.292. \end{aligned}$$

L'entropie de Y conditionnellement à X_1 (on dit encore "sachant X_1") est donc donnée par

$$\begin{aligned} H(Y \mid X_1) &= \underbrace{\mathbb{P}(X_1 = H)}_{\approx 3/13} \underbrace{H(Y \mid X_1 = H)}_{\approx 0} + \underbrace{\mathbb{P}(X_1 = R)}_{\approx 5/13} \underbrace{H(Y \mid X_1 = R)}_{\approx 0.292} \\ &\quad + \underbrace{\mathbb{P}(X_1 = P)}_{\approx 5/13} \underbrace{H(Y \mid X_1 = P)}_{\approx 0.292} \\ &= 0.225. \end{aligned}$$

On prend maintenant la variable X_2 et on calcule l'entropie de Y sachant X_2. Pour cela ; il faut calculer l'entropie de Y sachant $X_2 = E$ et l'entropie de Y sachant $X_2 = C$. On a

$$H(Y \mid X_2 = E) = 0.287 \tag{8}$$

et

$$H(Y \mid X_2 = C) = 0.217 \tag{9}$$

et donc

$$H(Y \mid X_2) = \underbrace{\mathbb{P}(X_2 = E)}_{\approx 8/13} H(Y \mid X_2 = E) + \underbrace{\mathbb{P}(X_2 = C)}_{\approx 5/13} H(Y \mid X_2 = C) \tag{10}$$

$$= 0.26 \tag{11}$$

On en conclut que X_1 est la question la plus informative, c'est à dire qui donne l'entropie la plus petite pour la réponse. Et pour construire l'arbre de décision, on pose donc d'abord la question X_1 du type de musique, puis la question X_2 du prix, et on peut décider de la réponse.

On peut construire l'arbre de décision qui dans notre cas avec deux questions correspond à poser la question X_1 puis la question X_2. Qu'est-ce que cela donne en pratique ?

- Supposons que $X_1 = H \Rightarrow Y = +$
- Supposons $X_1 = R$,
  - supposons $X_2 = E$ alors la majorité des cas donne $Y = -$ donc on prendra $Y = -$ dans notre arbre

- supposons $X_2 = C$ alors $Y = +$ dans la majorité des cas (1 seul ici !) et donc on prendra $Y = +$ dans notre arbre
- Supposons $X_1 = P$
  - si $X_2 = E$ la majorité donne $Y = -$
  - si $X_2 = C$ la majorité donne $Y = +$.

2. Peut-on prédire si la transaction va avoir lieu pour un client avec les valeurs $TYPE = H$ et $PRICE = E$

   L'arbre de décision va prédire comme réponse pour chaque nouveau client, selon la majorité des réponses pour les clients de notre base de donnée ayant répondu à X_1 et X_2 de la même façon. Vote par majorité !

   Solution : Comme 100% des clients de la base de donnée avec les mêmes caractéristiques ont une réponse $Y = +$, la majorité est obtenue pour la réponse $+$ (très très largement car aucune réponse - n'a été observée.

Exercice 3 On va constuire un arbre de régression sur des features (c'est à dire des co-variables) catégorielles (c'est à dire qualitatives) concernant les ventes d'un site web d'une marque de vêtements selon les conditions météo et le fait de savoir si on est en période de soldes ou non. La variable réponse est le nombre de ventes. Les résultats sont donnés dans le tableau de la page suivante.

Day	Météo	Temp.	Humidité	Soldes	Nombre d'achats
1	Ensoleillé	Chaud	Forte	OuiOui !	25
2	Ensoleillé	Chaud	Forte	NonNon	30
3	Nuageux	Chaud	Forte	OuiOui !	46
4	Pluvieux	Moyenne	Forte	OuiOui !	45
5	Pluvieux	Frais	Normale	OuiOui !	52
6	Pluvieux	Frais	Normale	NonNon	23
7	Nuageux	Frais	Normale	NonNon	43
8	Ensoleillé	Moyenne	Forte	OuiOui !	35
9	Ensoleillé	Frais	Normale	OuiOui !	38
10	Pluvieux	Moyenne	Normale	OuiOui !	46
11	Ensoleillé	Moyenne	Normale	NonNon	48
12	Nuageux	Moyenne	Forte	NonNon	52
13	Nuageux	Chaud	Normale	OuiOui !	44
14	Pluvieux	Moyenne	Forte	NonNon	30

1. Donner la moyenne et l'écart-type des réponses (nombre de ventes)

Solution :

ventes $(y_i,\ i = 1, \ldots, 14\ = \{25, 30, 46, 45, 52, 23, 43, 35, 38, 46, 48, 52, 44, 30\}$

$$\bar{y} = (25 + 30 + 46 + 45 + 52 + 23 + 43 + 35 + 38 + 46 + 48 + 52 + 44 + 30)/14 = 39.78$$

$$s_y = \sqrt{((25 - 39.78)^2 + (30 - 39.78)^2 + (46 - 39.78)^2 + \ldots + (30 - 39.78)^2)\ /14} = 9.32.$$

2. On prend la variable Météo et on regarde comment cette variable prédit la réponse :

   (a) Calculer la moyenne et l'écart-type pour la modalité "Ensoleillé".

   Solution : Les données concernées sont

Day	Météo	Temp.	Humidité	Soldes	Nombre de ventes
1	Ensoleillé	Chaud	Forte	OuiOui !	25
2	Ensoleillé	Chaud	Forte	NonNon	30
8	Ensoleillé	Moyen	Forte	OuiOui !	35
9	Ensoleillé	Frais	Normale	OuiOui !	38
11	Ensoleillé	Moyen	Normale	NonNon	48

$$\bar{y} \text{ sachant "Météo= Ensoleillé"} = (25 + 30 + 35 + 38 + 48)/5 = 35.2$$

$$RSS \text{ sachant "Météo= Ensoleillé"} = (25 - 35.2)^2 + (30 - 35.2)^2 + \ldots = 302.8$$

   (b) Calculer la moyenne et l'écart-type pour la modalité Nuageux

Day	Météo	Temp.	Humidité	Soldes	Nombre de ventes
3	Nuageux	Chaud	Forte	OuiOui !	46
7	Nuageux	Frais	Normale	NonNon	43
12	Nuageux	Moyen	Forte	NonNon	52
13	Nuageux	Chaud	Normale	OuiOui !	44

$$\bar{y} \text{ sachant "Météo= Nuageux"} = (46 + 43 + 52 + 44)/4 = 46.25$$
$$RSS \text{ sachant "Météo= Nuageux"} = (46 - 46.25)^2 + (43 - 46.25)^2 + \ldots = 48.75$$

   (c) Calculer la moyenne et l'écart-type pour la modalité Pluvieux.

   Solution :

Day	Météo	Temp.	Humidité	Soldes	Nombre de ventes
4	Pluvieux	Moyen	Forte	OuiOui !	45
5	Pluvieux	Frais	Normale	OuiOui !	52
6	Pluvieux	Frais	Normale	NonNon	23
10	Pluvieux	Moyen	Normale	OuiOui !	46
14	Pluvieux	Moyen	Forte	NonNon	30

$$\bar{y} \text{ sachant "Météo= Pluvieux"} = (45 + 52 + 23 + 46 + 30)/5 = 39.2$$
$$RSS \text{ sachant "Météo= Pluvieux"} = (45 - 39.2)^2 + (52 - 39.2)^2 + \ldots = 590.8$$

(d) Donner la RSS sachant la Météo

Météo	Std dev des ventes	Effectifs
Nuageux	48.75	4
Pluvieux	590.8	5
Ensoleillé	302.8	5

$$\text{RSS sachant "Météo"} = 48.75 + 590.8 + 302.8 = 942.35.$$

3. On prend la variable Température et on regarde comment cette variable prèdit la réponse en fonction du RSS.

(a) Calculer la RSS pour la modalité Chaud
Solution :

$$\bar{y} \text{ sachant "Température=Chaud"} = (25 + 30 + 46 + 44)/4 = 36.25$$
$$RSS \text{ sachant "Température=Chaud"} = (25 - 36.25)^2 + (30 - 36.25)^2 + (46 - 36.25)^2 + (44 - 36.25)^2 = 320.75.$$

(b) Calculer la RSS pour la modalité Moyen
Solution :

$$\bar{y} \text{ sachant "Température=Chaud"} = (45 + 35 + 46 + 48 + 52 + 30)/6 = 42.67$$
$$RSS \text{ sachant "Température=Chaud"} = (45 - 42.67)^2 + (35 - 42.67)^2 + (46 - 42.67)^2 + (48 - 42.67)^2 + (52 - 42.67)^2 + (30 - 42.67)^2 = 351.34.$$

(c) Calculer la RSS pour la modalité Frais
Solution :

$$\bar{y} \text{ sachant "Température=Frais"} = (52 + 23 + 43 + 38)/4 = 39$$
$$RSS \text{ sachant "Température=Chaud"} = (52 - 39)^2 + (23 - 39)^2 + (43 - 39)^2 + (38 - 39)^2 = 442.$$

(d) Donner la RSS sachant la Température
Solution :

$$RSS \text{ sachant "Température"} = 442 + 351.34 + 320.75 = 1114.09.$$

4. On prend la variable Humidité et on regarde comment cette variable prédit la réponse en fonction du RSS.

(a) Calculer la variance pour la modalité Forte
Solution :

$$\bar{y} \text{ sachant "Humidité=Forte"} = (25 + 30 + 46 + 45 + 35 + 52 + 30)/7 = 37.57$$
$$RSS \text{ sachant "Température=Chaud"} = (25 - 37.57)^2 + (30 - 37.57)^2 + (46 - 37.57)^2 + (45 - 37.57)^2 + (35 - 37.5$$

(b) Calculer la variance pour la modalité Normale
Solution :

$$\bar{y} \text{ sachant "Humidité=Forte"} = (52 + 23 + 43 + 38 + 46 + 48 + 44)/7 = 42$$
$$RSS \text{ sachant "Température=Chaud"} = (52 - 42)^2 + (23 - 42)^2 + (43 - 42)^2 + (38 - 42)^2 + (46 - 42)^2 + (48 - 42)$$

(c) Donner la RSS sachant l'Humidité
Solution :

$$RSS \text{ sachant "Humidité"} = 534 + 613.71 = 1147.71.$$

5. On prend la variable Soldes et on regarde comment cette variable prèdit la réponse en fonction du RSS.

(a) Calculer la RSS pour les Soldes
Solution :

$$\bar{y} \text{ sachant "Soldes=OuiOui"} = (25 + 46 + 45 + 52 + 35 + 38 + 46 + 44)/8 = 41.375$$
$$RSS \text{ sachant "Température=Chaud"} = (25 - 41.375)^2 + (46 - 41.375)^2 + (45 - 41.375)^2 + (52 - 41.375)^2 + (35 -$$

(b) Calculer la RSS pour la période de non Soldes
Solution :

$$\bar{y} \text{ sachant "Soldes=NonNon"} = (30+23+43+48+52+30)/6 = 37.66$$
$$RSS \text{ sachant "Soldes=NonNon"} = (30-37.67)^2+(23-37.67)^2+(43-37.67)^2+(48-37.67)^2+(52-37.67)^2+$$

(c) Donner la RSS sachant les Soldes ou pas
Solution :

$$RSS \text{ sachant "Soldes"} = 495.875 + 673.33 = 1169.205.$$

6. Quelle est la meilleure feature à la première étape ?
Solution : la meilleure feature est celle qui a la plus petite valeur de RSS, c'est à dire ... ma Météo !
7. On s'intéresse maintenant au cas où la météo est Ensoleillée et on veut savoir quelle est alors la meilleure Feature à choisir.

(a) Calculer la RSS sachant la Température et que la météo est "Ensoleillée"
Solution :
Pour la modalité "Chaud", on a

$$\bar{y} \text{ sachant "Temp=Chaud" et "Ensoleillé} = (25+30)/2 = 27.5$$
$$RSS \text{ sachant "Temp=Chaud" et "Ensoleillé} = (25-27.5)^2+(30-27.5)^2 = 2\times 2.5^2 = 12.5.$$

Pour la modalité "Moyenne", on a

$$\bar{y} \text{ sachant "Temp=Moyenne" et "Ensoleillé} = (35+48)/2 = 41.5$$
$$RSS \text{ sachant "Temp=Moyenne" et "Ensoleillé} = (35-41.5)^2+(48-41.5)^2 = 84.5.$$

Pour la modalité "Frais", on a une seule donnée : 38 donc la moyenne conditionnelle à Ensoleillé et Frais est 38 et la RSS est 0.
Ainsi, la somme des résidus au carré (RSS) pour la Température sachant Ensoleillé est 97.

(b) Calculer la RSS sachant l'Humidité et que la météo est "Ensoleillée"
Pour la modalité "Forte", on a

$$\bar{y} \text{ sachant "Hum=Forte" et "Ensoleillé} = (25+30+35)/3 = 30$$
$$RSS \text{ sachant "Hum=Forte" et "Ensoleillé} = (25-30)^2+(30-30)^2+(35-30)^2 = 50.$$

Pour la modalité "Moyenne", on a

$$\bar{y} \text{ sachant "Hum=Normale" et "Ensoleillé} = (38+48)/2 = 43$$
$$RSS \text{ sachant "Hum=Normale" et "Ensoleillé} = (38-43)^2 + (48-43)^2 = 50.$$

Ainsi, la somme des résidus au carré (RSS) pour l'humidité sachant Ensoleillé est 100.

(c) Calculer la RSS sachant les Soldes et que la météo est "Ensoleillée"
Pour la modalité "Soldes OuiOui", on a

$$\bar{y} \text{ sachant "Soldes=OuiOui" et "Ensoleillé} = (25+38+35)/3 = 32.67$$
$$RSS \text{ sachant "Soldes=OuiOui" et "Ensoleillé} = (25-32.67)^2 + (38-32.67)^2 + (35-32.67)^2 = 92.67.$$

Pour la modalité "Soldes=NonNon", on a

$$\bar{y} \text{ sachant "Soldes=NonNon" et "Ensoleillé} = (30+48)/2 = 39$$
$$RSS \text{ sachant "Soldes=NonNon" et "Ensoleillé} = (30-39)^2 + (48-39)^2 = 162.$$

Ainsi, la somme des résidus au carré (RSS) pour les soldes sachant Ensoleillé est 254.66.

(d) Quelle est la meilleure feature sachant que la météo est Ensoleillée ?
Solution : La température a le meilleur RSS sachant que la météo est ensoleillée.

8. On s'intéresse maintenant au cas où la météo est Nuageuse et on veut savoir quelle est alors la meilleure Feature à choisir.
   (a) Calculer la RSS sachant la Température et que la météo est "Nuageuse"
   (b) Calculer la RSS sachant l'Humidité et que la météo est "Nuageuse"
   (c) Calculer la RSS sachant les Soldes et que la météo est "Nuageuse"
   (d) Quelle est la meilleure feature sachant que la météo est Nuageuse ?
9. On s'intéresse maintenant au cas où la météo est Pluvieuse et on veut savoir quelle est alors la meilleure Feature à choisir.
   (a) Calculer la RSS sachant la Température et que la météo est "Pluvieuse"
   (b) Calculer la RSS sachant l'Humidité et que la météo est "Pluvieuse"
   (c) Calculer la RSS sachant les Soldes et que la météo est "Pluvieuse"
   (d) Quelle est la meilleure feature sachant que la météo est Ensoleillée ?
10. Construire l'arbre de régression sous R
    Solution :
    day = range(1,14)
    Outcast = c('Sun','Sun','Cloud','Rain','Rain','Rain','Cloud','Sun','Sun','Rain'

```
,'Sun','Cloud','Cloud','Rain')
Temp = c('H','H','H','M','C','C','C','M','C','M','M','M','H','M')
Hum = c('H','H','H','H','N','N','N','H','N','N','N','H','N','H')
Soldes = c('Y','N','Y','Y','Y','N','N','Y','Y','Y','N','N','Y','N')
Resp = c(25,30,46,45,52,23,43,35,38,46,48,52,44,30)
data = data.frame(day,Temp,Hum,Soldes,Resp)
```

Exercice 4 On considère les régions décrites dans la figure 1. Donnez l'arbre de décision ou de régression auquel ce découpage correspond.

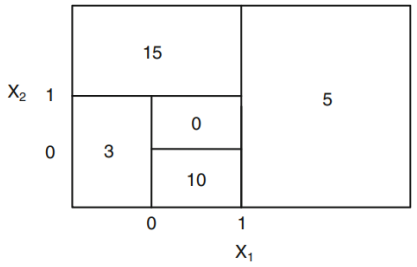

FIGURE 1 – Régions à traduire en arbre

Solution : voir la Figure 2

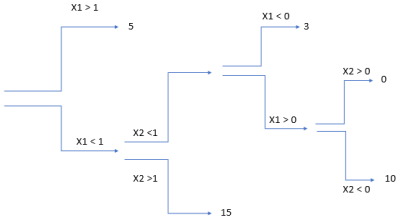

FIGURE 2 – Arbre à partir des régions

Exercice 5 On considère l'arbre décrit dans la figure 3. Donnez le découpage en régions auquel cet arbre de décision ou de régression correspond.

Solution : voir la Figure 4

Exercice 6 On va constuire un arbre de régression sur des features (c'est à dire des co-variables) catégorielles (c'est à dire qualitatives) concernant les ventes d'un site web d'une marque de vêtements selon les conditions météo et le fait de savoir si on est en période de soldes ou non. La variable réponse est le nombre de ventes. Les résultats sont donnés dans le tableau de la page suivante.

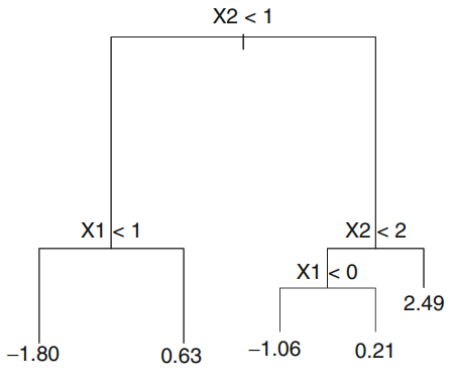

FIGURE 3 – Arbre à traduire en régions

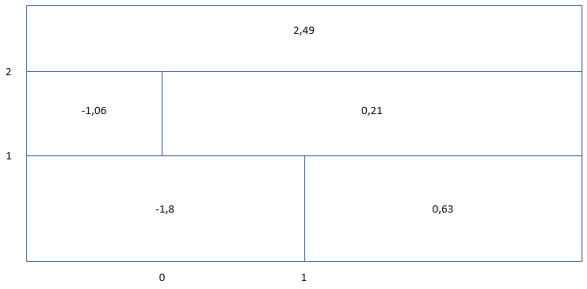

FIGURE 4 – Régions à partir de l'arbre

Day	Météo	Temp.	Humidité	Soldes	Nombre d'achats
1	Ensoleillé	28	Forte	OuiOui !	25
2	Ensoleillé	30	Forte	NonNon	30
3	Nuageux	35	Forte	OuiOui !	46
4	Pluvieux	16	Forte	OuiOui !	45
5	Pluvieux	5	Normale	OuiOui !	52
6	Pluvieux	7	Normale	NonNon	23
7	Nuageux	6	Normale	NonNon	43
8	Ensoleillé	12	Forte	OuiOui !	35
9	Ensoleillé	4	Normale	OuiOui !	38
10	Pluvieux	18	Normale	OuiOui !	46
11	Ensoleillé	14	Normale	NonNon	48
12	Nuageux	15	Forte	NonNon	52
13	Nuageux	31	Normale	OuiOui !	44
14	Pluvieux	12	Forte	NonNon	30

1. Faire un arbre de régression sur ce problème.

Solution : On commence par la variable Température car on a déjà analysé les autres variables en termes de RSS dans l'Exercice 3.

On découpe la variable température en classes. Dans l'esprit de la méthode CART de Breiman,

on découpe en 2 classes. On peut aussi être un peu plus aventureux et découper en 3 classes, 4 classes, etc.

Dans CART, on prend les valeurs t de temp. pour t de 4 à 35 une par une et on calcule la RSS pour chaque splitting entre inférieur ou égal à t et supérieur t : on prend la moyenne des ventes pour les cas de température inférieur à t puis on calcule la somme des carrés des erreurs pour ce groupe si on prédit par la moyenne sur ce groupe. Même chose pour les valeurs de températures supérieures à t. Comme on travaille à la main, on ne va pas tester tous les splits possibles. On peut essayer de splitter à 10. Cela forme deux groupes :

— $\{52, 23, 43, 38\}$ sur lequel la moyenne va être de $(52 + 23 + 43 + 38)/4 = 39$. La somme des carrés des erreurs de prédiction (RSS) est donc

$$RSS = (52 - 39)^2 + (23 - 39)^2 + (43 - 39)^2 + (38 - 39)^2 = 442$$

— $(25 + 30 + 46 + 45 + 35 + 46 + 48 + 52 + 44 + 30)/10 = 40.1$ La somme des carrés des erreurs de prédiction (RSS) est donc

$$RSS = (25 - 40.1)^2 + 30 - 40.1)^2 + 46 - 40.1)^2 + 45 - 40.1)^2 + 35 - 40.1)^2 + 46 - 40.1)^2 + (48 - 40.1)^2 + (52 - 40.1)^2 + (44 - 40.1)^2 + (30 - 40.1)^2) = 770.9$$

— La somme des deux RSS est donc 1212.9. C'est donc moins bien que la température en terme de RSS.

— On continue par étapes à choisir des variables les unes après les autres.

— Construire l'arbre de régression sous R

```
# Load CART packages
# install rpart package
# install.packages("rpart.plot")
# library(rpart.plot)
day = seq(1,14)
Outcast = c('Sun','Sun','Cloud','Rain','Rain','Rain','Cloud',
'Sun','Sun','Rain','Sun','Cloud','Cloud','Rain')
Temp = c('H','H','H','M','C','C','C','M','C','M','M','M','H','M')
Temp2 = c(28,30,35,16,5,7,6,12,4,18,14,15,31,12)
Hum = c('H','H','H','H','N','N','N','H','N','N','N','H','N','H')
Soldes = c('Y','N','Y','Y','Y','N','N','Y','Y','Y','N','N','Y','N')
Resp = c(25,30,46,45,52,23,43,35,38,46,48,52,44,30)
input.dat = data.frame(Temp2,Hum,Soldes,Resp)
output.tree <- rpart(Resp   Outcast+ Temp + Hum+Soldes,
data=input.dat,control =rpart.control(minsplit =2,minbucket=2, cp=0.15))
plot(output.tree)
output.tree
```