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

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.

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 !

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)
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".
   (b) Calculer la moyenne et l'écart-type pour la modalité Nuageux
   (c) Calculer la moyenne et l'écart-type pour la modalité Pluvieux.
   (d) Donner la RSS sachant la Météo
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
   (b) Calculer la RSS pour la modalité Frais
   (c) Donner la RSS sachant la Température
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
   (b) Calculer la variance pour la modalité Normale
   (c) Donner la RSS sachant l'Humidité
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
   (b) Calculer la RSS pour la période de non Soldes
   (c) Donner la RSS sachant les Soldes ou pas
6. Quelle est la meilleure feature à la première étape ?
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"
   (b) Calculer la RSS sachant l'Humidité et que la météo est "Ensoleillée"
   (c) Calculer la RSS sachant les Soldes et que la météo est "Ensoleillée"
   (d) Quelle est la meilleure feature 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

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

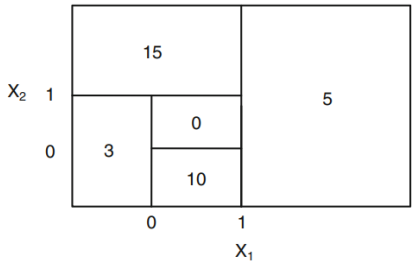

FIGURE 1 – Régions à traduire en arbre

Exercice 5 On considère l'arbre décrit dans la figure 2. Donnez le découpage en régions auquel cet arbre de décision ou de régression correspond.

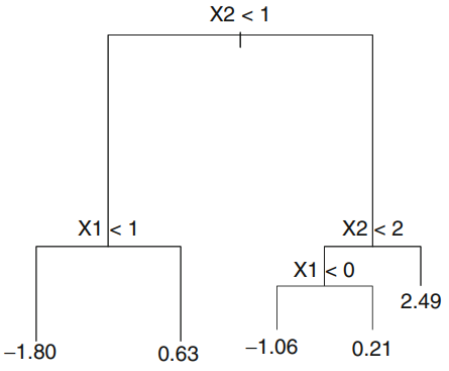

FIGURE 2 – Arbre à traduire en régions

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.

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.
2. Construire l'arbre de régression sous R