Exercices-Ethème 3

Thomas Laurent

11/02/2018

Question 1

Question 1.1)

Pour les arbres de discrimination et arbres de régression, n considère respectivement des variables qualitatives et des variables quantatives.

Question 1.2)

Un noeud est défini par une variable choisie parmi les variables explicatives et une partition en deux classes.

Question 1.3)

Une division est définie par une valeur seuil pour une variable quantitative ou une partition en deux groupes des modalités pour une variable qualitative.

Question 1.4)

Une division est admissible si l'effectif dans le segment terminal est supérieur ou égal à une valeur seuil.

Question 1.5)

L'homogénéité d'un noeud est nulle si le noeud est homogène, et maximale lorsque les valeurs de la variable réponse sont équiprobables.

Question 1.6)

Un noeud devient une feuille lorsque celui-ci est homogène, c'est-à-dire que qu'il n'existe plus de partition admissible.

Question 1.7)

La valeur Y d'une feuille peut être affectée à une classe des manières suivantes:

- On choisit la classe la plus représentée dans la feuille
- Si on utilise des probabilités a priori, on choisit la classe la plus probable au sens bayésien
- On choisit la classe la moins couteuse (il faut définir des coûts de mauvais classement)

Question 1.8)

Pour une variable qualitative Y, le critère d'homogénéité peut être de la nature suivante: entropie, critère de concentration de Gini ou la statistique du χ^2

Question 1.9)

Le coefficient de pénalisation, γ est introduit pour prendre en compte la complexité de l'arbre et ainsi la complexité de l'arbre se formule de la manière suivante:

$C(A) = D(A) + \gamma K$

Plus le nombre de feuilles de l'arbre est grand plus la complexité de l'arbre est grande car $\gamma > 0$.

Question 1.10)

Pour trouver l'arbre optimal de séquence, on cherche la valeur de γ minimisant l'erreur de prévision. A cette valeur de γ correspondra l'arbre optimal dans la séquence estimée sur tout l'échantillon d'apprentissage.

Question 2

Question 2.1

$GI(X,Y) = -\sum_{x=1}^{k}\sum_{y=1}^{m} P(X=x,Y=y)logP(X=x) - \sum_{x=1}^{k} P(Y=y)H(X/Y=y))$

or $-\sum_{y=1}^{m} P(Y=y)H(X/Y=y) = -\sum_{y=1}^{m} P(Y=y)[-\sum_{x=1}^{k} P(X=x/Y=y)logP(X=x/Y=y)]$

$= \sum_{y=1}^{m}\sum_{x=1}^{k} P(X=x,Y=y)log\frac{P(X=x,Y=y)}{P(Y=y)}$

donc $GI(X,Y) = \sum_{y=1}^{m}\sum_{x=1}^{k} P(X=x,Y=y)log(\frac{P(X=x,Y=y)}{P(Y=y)P(X=x)})$

La formule est symmétrique car on peut démontrer que l'on obtient la même formule pour $GI(Y,X)$

Ainsi $GI(X,Y) = GI(Y,X)$

On remarque donc que la définition du gain d'information est symmétrique.

Question 2.2

$H(X/Y)$

$= \sum_{y=1}^{m} P(Y=y)H(X/Y=y)$

$= \sum_{y=1}^{m} P(Y=y)[-\sum_{x=1}^{k} P(X=x,Y=y)log(P(X=x/Y=y))]$

$= -\sum_{y=1}^{m}\sum_{x=1}^{k} P(X=x,Y=y)log(P(X=x,Y=y)) + \sum_{y=1}^{m}\sum_{x=1}^{k} P(X=x,Y=y)log(P(Y=y))$

$= -\sum_{y=1}^{m}\sum_{x=1}^{k} P(X=x,Y=y)log(P(X=x,Y=y)) + \sum_{y=1}^{m} P(Y=y)log(P(Y=y))$

$= -H(Y) + H(X,Y)$

Ainsi $GI(X,Y) = H(X) - H(X/Y)$

$= H(X) - H(X,Y) + H(Y)$

Question 2.3

$H(X,Y) - H(Y/X)$

$= -\sum_{x=1}^{k}\sum_{y=1}^{m} P(X=x,Y=y)log(P(X=x,Y=y)) - \sum_{x=1}^{k} P(X=x)H(Y/X=x)$

$= \sum_{x=1}^{k}\sum_{y=1}^{m} P(X=x,Y=y)log(P(X=x,Y=y)) - \sum_{x=1}^{k}\sum_{y=1}^{m} P(X=x)P(Y=y/X=x)log(P(Y=y/X=x))$

$= -\sum_{x=1}^{k}\sum_{y=1}^{m} P(X=x,Y=y)log(\frac{P(X=x,Y=y)}{P(Y=y/X=x)})$

$= -\sum_{x=1}^{k}\sum_{y=1}^{m} P(X=x,Y=y)logP(X=x)$

$= -\sum_{x=1}^{k} P(X=x)logP(X=x) = H(X)$

Ainsi, on trouve la formule suivante: $GI(X,Y) = H(X) - H(X/Y) = H(X,Y) - H(Y/X) - H(X/Y)$

Question 3

Question 3.1

Pour l'ensemble d'apprentissage, l'indice de Gini vaut:

$GINI = \frac{10}{20}\frac{10}{20} = \frac{1}{4}$

Question 3.2

Pour la partition en utilisant IdFilm, on obtient seulement une observation par feuille. Ainsi, l'indice GINI pour chaque feuilles est égal à 0 et l'indice GINI global pour la partition est :

$GINI = 0$

Question 3.3

Pour l'attribut format, il y a deux modalités:

$GINI_{DVD} = \frac{2}{8}\frac{6}{8} = \frac{3}{16}$ $GINI_{Enligne} = \frac{4}{12}\frac{8}{12} = \frac{2}{9}$

Finalement $GINI_{Format} = \frac{8}{20}GINI_{DVD} + \frac{12}{20}GINI_{Enligne}$

$= \frac{8}{20}\frac{3}{16} + \frac{12}{20}\frac{2}{9} = \frac{5}{24}$

Question 3.4

Pour l'attribut catégorie, il y a trois modalités:

$GINI_{Loisirs} = \frac{1}{4}\frac{3}{4} = \frac{3}{16}$ $GINI_{Comedie} = \frac{7}{8}\frac{1}{8} = \frac{7}{64}$ $GINI_{Documentaire} = \frac{2}{8}\frac{6}{8} = \frac{3}{16}$

Finalement $GINI_{Categorie} = \frac{4}{20}GINI_{Loisirs} + \frac{8}{20}GINI_{Comedie} + \frac{8}{20}GINI_{Documentaire}$

$= \frac{4}{20}\frac{3}{16} + \frac{8}{20}\frac{7}{64} + \frac{8}{20}\frac{3}{16} = \frac{5}{32}$

Question 3.5

On obtient un indice de GINI le plus bas pour l'attribut IdFilm ($GINI = 0$) ce qui parait normal puisque les feuilles sont homogènes mais ne contiennent qu'une seule observation chacune.

Question 3.6

L'attribut IdFilm présente l'indice GINI le plus faible. En revanche, le découpage est trop fin car on obtient seulement une seule observation dans chaque feuille. Il est donc nécessaire de trouver une autre variable explicative pour le découpage. Entre la variable Format et la variable Categorie, cette dernière présente l'indice GINI le plus faible (environ égal à 0.156). On retient donc cette dernière variable pour le découpage au noeud racine.

Question 4

Question 4.1

Les variables explicatives sont des variables binaires et présente par conséquent un nombre de modalités limitées. On ne rencontrera pas une situation d'overfitting liée à la sélection d'un attribut particulier. Ainsi, on utilise le gain d'information.

Question 4.2

On calcul l'entropie totale:

$H(Total) = -[\frac{62}{131}log(\frac{62}{131}) + \frac{69}{131}log(\frac{69}{131})]$

≈ 0.692

On s'intéresse tout d'abord à l'attribut traffic:

$Gain_{traffic} = H_{Total} - \sum_{j=1}^{2} P_j H(Traffic = j)$

$= 0.692 - [\frac{64}{131}H(Traffic = chargé) + \frac{67}{131}H(Traffic = léger)]$

or $H(Traffic = chargé) = -[\frac{37}{64}log(\frac{37}{64}) + \frac{27}{64}log(\frac{27}{64})] \approx 0.681$

$H(Traffic = léger) = -[\frac{25}{67}log(\frac{25}{67}) + \frac{42}{67}log(\frac{42}{67})] \approx 0.681$

Ainsi, $Gain_{traffic} = 0.692 - [\frac{64}{131}0.681 + \frac{67}{131}0.661]$

≈ 0.021

Ensuite, on calcule le gain pour l'attribut temps.

$Gain_{Temps} = 0.692 - [\frac{83}{131}H(Temps = ensoleillé) + \frac{48}{131}H(Temps = pluvieux)]$

$H(Temps = ensoleillé) = -[\frac{30}{83}log(\frac{30}{83}) + \frac{53}{83}log(\frac{53}{83})] \approx 0.654$

$H(Temps = pluvieux) = -[\frac{32}{48}log(\frac{32}{48}) + \frac{16}{48}log(\frac{16}{48})] \approx 0.637$

Ainsi, $Gain_{Temps} = 0.692 - [\frac{83}{131}0.654 + \frac{48}{131}0.637]$

≈ 0.044

Etant donné que le gain d'information est maximal pour l'attribut temps, on retient celui-ci pour la racine de l'arbre.

Question 4.3

Si l'on considère une variable supplémentaire quantitative X avec K valeurs uniques, pour évaluer le seuil optimale α de la variable, il faut procéder de la manière suivante:

- Réaliser les K-1 découpages possibles de la variable
- Calculer le gain d'information pour chaque découpage
- Sélectionner le découpage avec un gain maximum