EISTI : Cycle Ingénieurs : Première Année - GM

Examen de Datamining 1

4 mai 2020

Durée 2H

Modalités : Vous devez rendre un document électronique contenant les réponses aux questions.

NOM Prénom :

Exercice 1.1	Exercice 1.2	Exercice 1.3	Exercice 1.4	Exercice 1.5	
Exercice 2.1	Exercice 2.2	Exercice 3.3	Exercice 2.4	Exercice 2.5	
Exercice 3.1	Exercice 3.2	Exercice 3.3	Exercice 3.4	Exercice 3.5	Exercice 3.6

Notations

Notes globales :

1 Clustering

Dans cet exercice, nous nous intéressons au clustering ascendant hiérarchique (CAH). Nous noterons d la distance entre les individus, C_i les clusters, G_i (respectivement n_i) leurs centres de gravité (resp. les nombres d'individus qu'ils contiennent).

Exercice 1.1 Rappelez pourquoi le clustering hiérarchique a besoin, en plus de la distance d, d'une "distance" entre clusters.

Exercice 1.2 Rappelez la définition de l'inertie **intra**-clusters et celle de l'inertie **inter**-clusters et donnez leur valeur à l'initialisation et à la fin de l'algorithme.

Exercice 1.3 En remarquant que durant chaque itération, l'augmentation de chacune de ces interties est égale à la baisse de l'autre, donnez la formule de la hausse de l'inertie intra-clusters.

Exercice 1.4 Déduisez-en que lorsque la distance d est la distance euclidienne, cette hausse prend la forme suivante :

$$\mathrm{h} = \frac{n_i * n_j}{n_i + n_j} d^2(G_i, G_j)$$

Exercice 1.5 En déduire qu'en utilisant la distance de Ward dans le CAH, nous assurons à chaque itération la fusion entre clusters la plus "**optimale**".

2 Motifs fréquents et motifs rares

Soit l'ensemble d'items suivant $\{A, B, C, D, E\}$. Soit la base de transactions suivante [1] :

	A	B	C	D	E
1	X		X	X	
2		X	X		X
3	X	X	X		X
4	X				X
5	X	X	X		X

Nous supposons que le support minimal est fixé à $minSupp = \frac{3}{5}$ pour les exercices suivants :

Exercice 2.1 Appliquer l'algorithme Apriori pour trouver les sous-ensembles d'items fréquents (ou les motifs fréquents). Indiquer pendant les étapes de l'application de l'algorithme comment vous utilisez la proprieté de la fermeture des sous-ensemble fréquents (Downward closure property) pour optimiser la recherche.

1. La transaction numéro 1 se lit : ACD, etc.

Exercice 2.2 Extraire les règles d'association à partir des sous-ensembles d'items fréquents de longueurs égales à 2. Associer à chaque règle sa mesure de confiance.

Exercice 2.3 Nous nous intéressons maintenant aux motifs (sous-ensembles d'items) **rare**. Un sous ensemble d'items est rare ssi son support est strictement inférieur à $minSupp$. En vous inspirant de l'algorithme Apriori, proposer un algorithme efficace qui permet de trouver les sous-ensemble d'items (motifs) rares. Essayer d'utiliser la même propriété utilisée dans Apriori pour l'optimisation de la recherche.

Exercice 2.4 Appliquer l'algorithme que vous venez de proposer sur la base de transactions en indiquant comment la recherche est optimisée.

Exercice 2.5 Extraire les règles d'association à partir des sous-ensembles d'items rares de longueurs égales à 4. Associer à chaque règle sa mesure de confiance.

3 Arbre de décision

Le jeu de données Cars (en annexe) représente 60 voitures caractérisées par 5 variables :

Price : a numeric vector giving the list price in US dollars of a standard model.

Mileage : fuel consumption miles per US gallon, as tested.

Type : a factor with levels : *Compact,Large,Medium,Small,Sporty,Van*

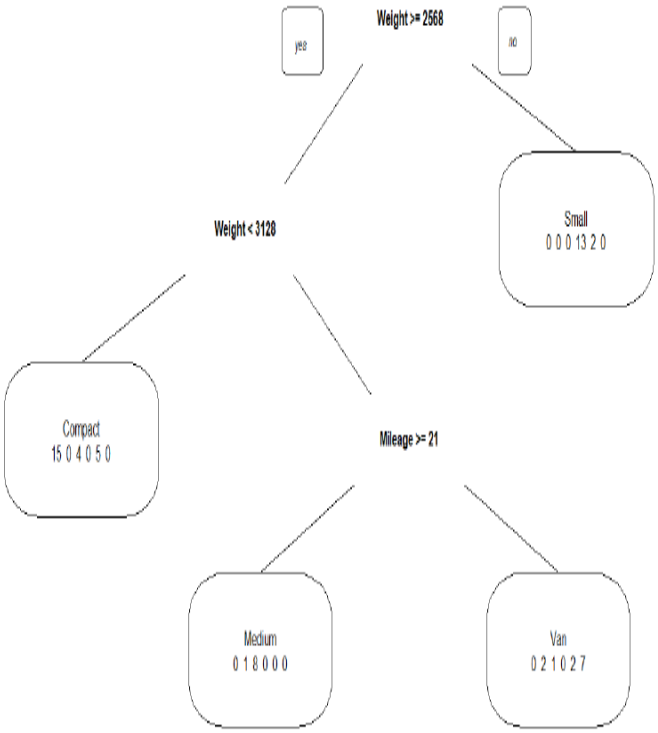
Weight : kerb weight in pounds.

HP : the net horsepower of the vehicle.

Il s'agit d'un extrait d'un jeu de données plus vaste collecté en avril 1990 dans Consumer Reports.
On utilise un arbre de décision pour identifier le type de véhicule en fonction des autres variables.
Avec l'instruction R :

```
arbre=rpart(Type~.,Cars,parms=list(split="gini"))
```

on obtient l'arbre suivant :



Exercice 3.1 Expliquer sans calcul comment a été choisie la variable $Weight$ à la racine ?

Exercice 3.2 Calculer sur la variable $Weight$(uniquement) l'indicateur qui a permis de faire ce choix. Pour faire ce calcul, on a :

Compact	Large	Medium	Small	Sporty	Van	Tot
15	3	13	13	9	7	60

et le jeu de données Cars en annexe. Toutes les voitures au-dessus du trait rouge vérifient $Weigth < 2568$ et celles au-dessous vérifient $Weigth > 2568$.

Exercice 3.3 A quelle classe appartient l'observation définie par : $Price = 8400$, $Weight = 3200$, $Mileage = 19$ et $HP = 100$?

Exercice 3.4 Quelle est la probabilité qu'elle appartienne à la classe Van, à la classe Medium et à la classe Compact ? (les classes sont affichées par ordre alphabétique : Compact, Large, Medium, Small, Sporty, Van)

Exercice 3.5 Construire la diagonale de la matrice de confusion à partir du graphe.

Exercice 3.6 En déduire le taux d'erreur d'ajustement.

	Price	Mileage	Type	Weight	HP
Eagle Summit 4	6319	37	Small	1845	63
Ford Escort 4	5866	34	Small	1900	73
Ford Festiva 4	6488	35	Small	2075	78
Mazda Protege 4	6635	32	Small	2260	92
Mercury Tracer 4	7399	33	Small	2275	90
Nissan Sentra 4	8672	26	Small	2285	82
Pontiac LeMans 4	9599	25	Small	2295	90
Subaru Loyale 4	9995	26	Small	2330	100
Subaru Justy 3	7402	33	Small	2345	90
Toyota Corolla 4	7254	28	Small	2350	74
Toyota Tercel 4	8748	29	Small	2390	102
Volkswagen Jetta 4	6599	32	Small	2440	103
Dodge Daytona	8895	33	Small	2560	113
Honda Civic 4	9410	33	Sporty	2170	108
Chevrolet Camaro V8	13071	28	Sporty	2485	97
Ford Mustang V8	15930	24	Compact	2575	120
Ford Probe	10565	23	Compact	2640	110
Honda Civic CRX Si 4	9995	23	Compact	2645	110
Honda Prelude Si 4WS 4	10320	26	Compact	2655	95
Nissan 240SX 4	18900	27	Compact	2670	108
Audi 80 4	10989	25	Compact	2745	102
Buick Skylark 4	9483	24	Compact	2750	98
Ford Tempo 4	12459	24	Compact	2780	110
Nissan Stanza 4	12145	26	Compact	2920	125
Oldsmobile Calais 4	11650	21	Compact	2920	138
Peugeot 405 4	11588	27	Compact	2920	115
Subaru Legacy 4	11499	23	Compact	2935	130
Volvo 240 4	18450	23	Compact	2985	114
Acura Legend V6	10945	25	Compact	3065	141
Buick Century 4	17879	21	Compact	3110	142
Oldsmobile Cutlass Supreme	16145	23	Large	3325	165
Nissan Axxess 4	17257	20	Large	3850	150
Nissan Van 4	14525	18	Large	3855	170
Chevrolet Beretta 4	13150	21	Medium	2765	110
Mazda 626 4	13150	21	Medium	2880	110
Mitsubishi Sigma V6	9999	23	Medium	2885	110
Toyota Camry 4	12495	22	Medium	2975	150
Chrysler Le Baron Coupe	15350	22	Medium	3145	150
Eagle Premier V6	13195	22	Medium	3190	140
Ford Thunderbird V6	17899	22	Medium	3200	160
Hyundai Sonata 4	14495	21	Medium	3220	135
Mazda 929 V6	24760	20	Medium	3265	160
Buick Le Sabre V6	16342	22	Medium	3450	147
Chevrolet Caprice V8	23300	21	Medium	3480	158
Ford LTD Crown Victoria V8	21498	23	Medium	3480	190
Chevrolet Lumina APV V6	14980	23	Medium	3610	140
Plymouth Laser	11470	30	Sporty	2695	110
Subaru XT 4	13945	27	Sporty	2710	140
Chrysler Le Baron V6	13249	24	Sporty	2775	140
Honda Accord 4	10855	26	Sporty	2840	92
Mitsubishi Galant 4	9745	27	Sporty	2885	100
Nissan Maxima V6	12164	19	Sporty	3310	225
Oldsmobile Cutlass Ciera 4	11545	20	Sporty	3320	170
Chrysler New Yorker V6	13949	20	Van	3185	138
Ford Taurus V6	13995	18	Van	3195	110
Toyota Cressida 6	14929	20	Van	3415	107
Dodge Grand Caravan V6	12267	18	Van	3665	145
Ford Aerostar V6	14799	19	Van	3690	106
Mazda MPV V6	15395	18	Van	3735	150
Mitsubishi Wagon 4	14944	19	Van	3735	150