Data Mining 2: Modélisation statistique et apprentissage **G. Cohen**

Exercice 03 février 2018

ETheme 3 Arbres de décision

1 Questions

1. Quel type de données (c'est-à-dire quelle nature de données) considère-t-on lorsqu'on réalise un arbre de discrimination ? un arbre de régression ?
2. Par quoi est défini le noeud d'un arbre ?
3. Qu'est-ce qu'une division ?
4. Qu'est-ce qu'une division admissible ?
5. Quelles sont les propriétés d'un critère d'homogénéité d'un noeud ?
6. Comment un noeud devient-il une feuille ?
7. Comment est affectée la valeur d'une feuille si la variable à prédire Y est qualitative ?
8. Quels sont les critères d'homogénéités utilisables pour Y qualitative ?
9. Quelles pénalisation est introduite afin de déterminer une séquence d'arbres emboîtés ?
10. Que faut-il minimiser pour rechercher l'arbre optimal de la séquence ?

2 Gain d'information et entropie

1. Entropie de X

$$H(X) = -\sum_{x=1}^{k} p(X=x) \log p(X=x)$$

2. Entropie de X, Y

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

3. Entropie de Y conditionnée par $X = j$

$$H(Y|X = x) \quad = \quad -\sum_{y=1}^{k} p(Y = y|X = x) \log p(Y = y|X = x)$$

4. Entropie conditionnelle de Y sachant X:

$$H(Y|X) \quad = \quad \sum_{x=1}^{k} p(X = x) H(Y|X = x)$$

5. Gain d'information (ou information mutuelle) entre X et Y:

$$GI(X;Y) \quad = \quad H(X) - H(X|Y)$$

En utilisant ces définitions,

1. Montrer que $GI(X;Y) = GI(Y;X)$. Qu'est-ce que cela vous dit sur le gain d'information ?
2. Montrer que $GI(X;Y) = H(X) + H(Y) - H(X,Y)$
3. Montrer que $GI(X;Y) = H(X,Y) - H(X|Y) - H(Y|X)$

3 Index de Gini

Rappel:

$$GINI(X) \quad = \quad -\sum_{i \neq j}^{k} p(X = x_i) p(X = x_j) = \frac{1}{2}[1 - \sum_{j} p^2(x_j)]$$

Ex: lorsque $C = 2, GINI(X) = p(X = x_1) p(X = x_2)$

On considère l'ensemble d'apprentissage de la table 1 pour un problème de classification binaire

1. Calculez l'index de Gini pour l'ensemble d'apprentissage.
2. Calculez l'index de Gini pour l'attribut **Id Film**.
3. Calculez l'index de Gini pour l'attribut *Format*.

Id Film	Format	Catégorie	Classe
1	DVD	Loisirs	C_0
2	DVD	Comédie	C_0
3	DVD	Documentaire	C_0
4	DVD	Comédie	C_0
5	DVD	Comédie	C_0
6	DVD	Comédie	C_0
7	En ligne	Comédie	C_0
8	En ligne	Comédie	C_0
9	En ligne	Comédie	C_0
10	En ligne	Documentaire	C_0
11	DVD	Comédie	C_1
12	DVD	Loisirs	C_1
13	En ligne	Loisirs	C_1
14	En ligne	Documentaire	C_1
15	En ligne	Documentaire	C_1
16	En ligne	Documentaire	C_1
17	En ligne	Documentaire	C_1
18	En ligne	Loisirs	C_1
19	En ligne	Documentaire	C_1
20	En ligne	Documentaire	C_1

Figure 1: Jeu de données.

4. Calculez l'index de Gini pour l'attribut *Categorie*.

5. Lequel des trois attributs a l'index de Gini le plus bas?

6. Lequel des trois attributs allez-vous utiliser pour le découpage au noeud racine? Expliquer brièvement votre choix.

4 Construction d'arbres de décision avec le gain d'information

On considère le jeu de données de la table 2 comprenant 3 attributs binaires. En utilisant ce jeu de données, répondre aux questions suivantes:

1. Vous voulez construire un arbre de décision qui prédise le taux d'accidents sachant le temps et le traffic routier. Votre premier travail consiste à décider quel attribut mettre à la racine. Sans rien calculer expliquer

pourquoi vous devez utiliser le gain d'information pour décider entre le découpage selon le temps ou le traffic routier.

2. Déterminez l'attribut racine. Montrez vos calculs.

3. Supposez maintenant que le jeu de données comprenne un quatrième attribut, température, qui prend des valeurs continues. Lorsqu'un arbre de décision découpe selon un attribut continu X, il divise les données en exemples selon $X \leq a$ et $X > a$, pour un seuil a choisi. Si le jeu de données a K valeurs uniques pour la température, comment déterminez-vous le seuil optimum a?

4. Les jeux de données réels sont rarement parfait- certains contiennent des erreurs systématiques comme des attributs dupliqués ou des attributs avec une seule valeur. De plus, tous les algorithmes de datamining ne sont pas adaptés pour traiter de telles erreurs. Par exemple, le classifieur Naive Bayes à de mauvaises performances lorsque les attributs sont dupliqués. Lorsque l'on utilise des arbres de décision, que se passe t-il lorsque les attributs sont dupliqués? Que se passe t-il avec des attributs à une valeur? Expliquer vos réponses en terme de gain d'information.

Temps	Traffic routier	Taux d'accidents	Nombre
Ensoleillé	Chargé	Elevé	17
Ensoleillé	Chargé	Bas	22
Ensoleillé	Léger	Elevé	13
Ensoleillé	Léger	Bas	31
Pluvieux	Chargé	Elevé	20
Pluvieux	Chargé	Bas	5
Pluvieux	Léger	Elevé	12
Pluvieux	Léger	Bas	11

Figure 2: Temps journalier, traffic routier et taux d'accident.