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

Exercice 14 février 2018

ETheme 2 Arbres de décision *Solution*

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 ?
   Pour la discrimination on considère des variables explicatives quantitatives ou qualitatives et une variable à expliquer qualitative (m modalités). Pour la régression seule la nature de la variable à expliquer est différente elle est quantitative réelle.

2. Par quoi est défini le noeud d'un arbre ?
   Par le choix conjoint d'une variable explicative et d'une division qui induit une partition en deux classes.

3. Qu'est-ce qu'une division ?
   C'est la répartition en deux classes selon une valeur seuil de la variable retenue si celle-ci est quantitative ou selon un partage en deux groupes de modalités si elle est qualitative.

4. Qu'est-ce qu'une division admissible ?
   C'est une division dans laquelle les deux noeuds descendants sont non vides.

5. Quelles sont les propriétés d'un critère d'homogénéité d'un noeud ?
   Doit être nul en cas d'homogénéité et maximal si les valeurs de la variable à expliquer sont équiprobables ou très dispersées.

6. Comment un noeud devient-il une feuille ?
   Lorsque le noeud est homogène (il n'existe plus de partition admissible) ou si le nombre d'observations qu'il contient et inférieur à une valeur seuil prédéfinie (entre 1 et 5).

7. Comment est affectée la valeur d'une feuille si la variable à prédire Y est qualitative ?
   Soit

- à la classe qui a le plus grand nombre de représentant dans la feuille
- à la classe à postériori la plus probable si l'on a des probabilités à priori on applique alors Bayes.
- à la classe la moins coûteuse si l'on a connaissance des coûts de mauvais classements.

8. Quels sont les critères d'homogénéités utilisables pour Y qualitative ?
   L'entropie, la concentration de Gini ou une statistique de test du χ^2.
   Le critère d'entropie est le plus souvent utilisé.

9. Quelles pénalisation est introduite afin de déterminer une séquence d'arbres emboîtés ?
   Une pénalité sur la complexité de l'arbre qui est mesurée par le nombre de noeuds terminaux de l'arbre.

10. Que faut-il minimiser pour rechercher l'arbre optimal de la séquence ?
    On minimise la qualité de discrimination de l'arbre $D(A_k)$.

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 $X=j$

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

4. Entropie conditionnelle de Y sachant X:

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

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 ?

$$\begin{aligned}
GI(X;Y) &= H(X) - H(X|Y) \\
&= -\sum_{x=1}^{k} p(X=x) \log p(X=x) + \sum_{y=1}^{k} p(Y=y) H(X|Y=y) \\
&= -\sum_{x=1}^{k} p(X=x) \log p(X=x) + \sum_{y=1}^{k} p(Y=y) \sum_{x=1}^{k} p(X=x|Y=y) \\
&\quad \log p(X=x|Y=y) \\
&= -\sum_{x=1}^{k} p(X=x) \log p(X=x) + \sum_{y=1}^{k} \sum_{x=1}^{k} p(X=x, Y=y) \log p(X=x|Y=y)
\end{aligned}$$

En sustituant $P(X=x) = \sum_{y=1}^{k} P(X=x, Y=y)$ (loi des probabilités totales) dans l'expression précédente, on obtient

$$\begin{aligned}
GI(X;Y) &= -\sum_{x=1}^{k} \sum_{y=1}^{k} p(X=x, Y=y) \log p(X=x) + \sum_{y=1}^{k} \sum_{x=1}^{k} p(X=x, Y=y) \\
&\quad \log p(X=x, Y=y) \\
&= -\sum_{x=1}^{k} \sum_{y=1}^{k} p(X=x, Y=y)(\log p(X=x|Y=y) - \log p(X=x)) \\
&= -\sum_{x=1}^{k} \sum_{y=1}^{k} p(X=x, Y=y)(\log p(X=x|Y=y) - \log p(Y=y) - \log p(X=x)) \\
&= -\sum_{x=1}^{k} \sum_{y=1}^{k} p(X=x, Y=y) \log \frac{p(X=x, Y=y)}{p(Y=y)p(X=x)}
\end{aligned}$$

On a donc

$$GI(X;Y) \quad = \quad -\sum_{x=1}^{k} \sum_{y=1}^{k} p(X=x, Y=y) \log \frac{p(X=x, Y=y)}{p(Y=y)p(X=x)} \quad (1)$$

Puisque cette dernière expression est symétrique relativement à X et Y, les interchanger ne change rien. Ainsi on a bien $GI(X;Y) = GI(Y;X)$ et donc le gain en information est bien symétrique.

2. Montrer que $GI(X;Y) = H(X) + H(Y) - H(X,Y)$
On sait par définition que $GI(X;Y) = H(X) - H(X|Y)$. Il est donc suffisant de montrer que $H(X|Y) = H(X,Y) - H(Y)$ ou que $H(X|Y) + H(Y) = H(X,Y)$. On a

$$\begin{aligned}
H(X|Y) + H(Y) &= -\sum_{y=1}^{k} p(Y=y) \sum_{x=1}^{k} p(X=x|Y=y) \log p(X=x|Y=y) + H(Y) \\
&= -\sum_{y=1}^{k}\sum_{x=1}^{k} p(X=x,Y=y) \log p(X=x|Y=y) - \sum_{y=1}^{k} p(Y=y) \\
&= \log p(Y=y) \\
&= -\sum_{y=1}^{k}\sum_{x=1}^{k} p(X=x,Y=y) \log p(X=x|Y=y) \\
&\quad -\sum_{y=1}^{k}\sum_{x=1}^{k} p(X=x,Y=y) \log p(Y=y) \\
&= -\sum_{y=1}^{k}\sum_{x=1}^{k} p(X=x,Y=y)[\log p(X=x|Y=y) + \log p(Y=y)] \\
&= -\sum_{y=1}^{k}\sum_{x=1}^{k} p(X=x,Y=y)[\log p(X=x|Y=y) + p(Y=y)] \\
&= -\sum_{y=1}^{k}\sum_{x=1}^{k} p(X=x,Y=y) \log p(X=x|Y=y) \\
&= H(X,Y)
\end{aligned}$$

Ainsi $GI(X;Y) = H(X) + H(Y) - H(X,Y)$

3. Montrer que $GI(X;Y) = H(X,Y) - H(X|Y) - H(Y|X)$
On utilise ici la définition de l'entropie jointe qui dit que $H(Y,X) = H(X,Y)$ par symétrie de la définition. On a déjà démontré précédemment que $H(X|Y) + H(Y) = H(X,Y)$ ou de façon équivalente que

$$H(Y) = H(X,Y) - H(X|Y) \qquad (2)$$

En remplaçant Y par X et X par Y dans l'équation (2) ci-dessus et en utilisant le résultat suivant $H(Y,X) = H(X,Y)$, on obtient que

$$H(X) = H(X,Y) - H(Y|X) \quad (3)$$

En substituant les résultats de l'équation (2) et (3) dans le résultat de la question précédente, on obtient

$$\begin{aligned} GI(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= [H(X,Y) - H(Y|X)] + [H(X,Y) - H(X|Y)] - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

Ainsi $GI(X;Y) = H(X,Y) - H(X|Y) - H(Y|X)$.

3 Index de Gini

Rappel:

$$\begin{aligned} GINI(X) &= \sum_j (p_j)(1-p_j) \\ &= 1 - \sum_j p_j^2 \\ &= \left\{ \sum_j p_j \right\}^2 - \sum_j p_j^2 \\ &= \sum_{i \neq j} (p_i)(p_j) \end{aligned}$$

où p_i est la probabilité d'avoir un exemple de la classe i, soit la fréquence relative de la classe i dans l'ensemble de données X. On utilise le fait que $\sum_j p_j = 1$.

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.

$$\begin{aligned} \text{Gini} &= 1 - (10/20)^2 + (10/20)^2 \\ &= 1 - 2 \times (1/2)^2 \\ &= 0.5 \end{aligned}$$

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.

2. Calculez l'index de Gini pour l'attribut **Id Film**.
   L'index de Gini pour chaque valeur **Id Film** est 0:

$$\begin{aligned} \text{Gini}_{Id=i} &= 1-(1/1)^2-(0/1)^2=0 \text{ pour } i=1,\ldots,10 \text{ et} \\ \text{Gini}_{Id=i} &= 1-(0/1)^2-(1/1)^2=0 \text{ pour } i=11,\ldots,20. \end{aligned}$$

   Donc, Le Gini global **Id Film** est 0:

$$\sum_{i=1}^{2} 01/20 \times 0 = 0$$

3. Calculez l'index de Gini pour l'attribut *Format*.
   L'index de Gini pour DVD est:

$$\begin{aligned} \text{Gini}_{\text{DVD}} &= 1-(2/8)^2-(6/8)^2 \\ &= 1-0.25^2-0.75^2=0.375 \end{aligned}$$

L'index de Gini pour **En ligne** est:

$$\begin{aligned}\text{Gini}_{\text{En ligne}} &= 1-(4/12)^2-(8/12)^2 \\ &= 1-(1/3)^2-(2/3)^2 = 0.4444\end{aligned}$$

Donc, le Gini global pour *Format* est:

$$\begin{aligned}\text{Gini}_{Global} &= n_{\text{DVD}}/n \times \text{Gini}_{\text{DVD}} + n_{\text{en ligne}}/n \times \text{Gini}_{\text{En ligne}} \\ &= 8/20 \times 0.375 + 12/20 \times 0.4444 \\ &= 0.4 \times 0.375 + 0.6 \times 0.4444 = 0.4166\end{aligned}$$

4. Calculez l'index de Gini pour l'attribut *Categorie*.
   L'index de Gini pour *Loisirs* vaut:

$$\begin{aligned}\text{Gini}_{\text{Loisirs}} &= 1-(2/8)^2-(6/8)^2 \\ &= 1-(1/4)^2-(3/4)^2 = 0.375\end{aligned}$$

$$\begin{aligned}\text{Gini}_{\text{Comédie}} &= 1-(2/8)^2-(6/8)^2 \\ &= 1-(7/8)^2-(1/8)^2 = 0.2188\end{aligned}$$

   et Documentaire vaut:

$$\begin{aligned}\text{Gini}_{\text{Documentaire}} &= 1-(2/8)^2-(6/8)^2 \\ &= 1-(2/8)^2-(6/8)^2 = 0.375\end{aligned}$$

   Le gini global vaut alors:

$$\begin{aligned}\text{Gini}_{Global} &= n_{\text{Loisirs}}/n \times \text{Gini}_{\text{Loisirs}} + n_{\text{Comédie}}/n \times \\ &\quad \text{Gini}_{\text{Comédie}} + n_{\text{Documentaire}}/n \times \text{Gini}_{\text{Documentaire}} \\ &= 4/20 \times 0.375 + 8/20 \times 0.2188 + 8/20 \times 0.375 \\ &= 0.2 \times 0.375 + 0.4 \times 0.2188 + 0.4 \times 0.375 = 0.31252\end{aligned}$$

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

6. Lequel des trois attributs allez-vous utiliser pour le découpage au noeud racine? Expliquer brièvement votre choix.
   Catégorie. Bien que **Id Film** ai la plus petite valeur d'index de gini, c'est clairement juste un attribut d'identification et un arbre de décision utilisant **Id Film** ne généralisera pas bien.

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

On considère le jeu de données de la table 3 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 l'attribut **temps** ou l'attribut **traffic routier**.
   La procédure pour l'utilisation du gain d'infomation GI consiste à décider si il faut diviser sur le temps ou sur le traffic routier:
   - Calcul de l'entropie de la distribution du taux d'accidents au noeud racine.
   - Pour diviser selon l'attribut temps, partitionner les données basées sur la valeur de l'attribut temps.
   - Calcul de l'entropie pondérée de la distribution des taux d'accidents aux noeuds fils, avec le poids donné par le nombre d'exemples dans le noeud.
   - Calcul du gain d'information du au temps comme la différence entre l'entropie à la racine et l'entropie pondérée des fils.
   - Répéter le processus pour la division selon le traffic routier afin de déterminer le gain d'information selon le traffic routier.
   - Choisir l'attribut qui donne le plus grand gain d'information.
2. Déterminez l'attribut racine. Montrez vos calculs.

Remarque:

On utilise parfois la déviance qui est égale à l'entropie à un facteur multiplicatif près.
Déviance : Soit un modèle probabiliste dans lequel au noeud i d'un arbre, la distribution de probabilité des classes est p_{ik}. Chaque cas est éventuellement affecté à une feuille, et donc à chaque feuille, on a un échantillon aléatoire n_{ik} provenant de la distribution multinomiale p_{ik}. Avec les variables observées x_i dans l'ensemble d'apprentissage, on connait les nombres n_i affectés à chaque noeud de l'arbre, en particulier les feuilles. La vraisemblance conditionnelle est alors proportionnelle à

$$\prod_{feuille\ i} \prod_{classes\ k} p_{ik}^{n_{ik}}$$

La déviance(-2 log-vraisemblance décalée en zéro pour le modèle parfait) est

$$D_i = -2 \sum_k n_{ik} log p_{ik}$$

pour chaque feuille, et on en fait la somme sur toutes les feuilles pour obtenir la déviance total de l'arbre:

$$D = \sum_i D_i$$

Pour une variable à prédire binaire k=2 on a la déviance pour le noeud i (-2 log vraisemblance):

$$\begin{aligned}
D_i &= -2log\left[p_{i1}^{n_{i1}}(1-p_{i1})^{n_{i0}}\right] \\
&= -2\left[n_{i1}log(p_{i1}) + n_{i0}log(1-p_{i1})\right] \\
&\quad \text{on utilise l'estimation de } p_{ik} : \hat{p}_{ik} = n_{ik}/n_i \\
&= -2n_i\left[\frac{n_{i1}}{n_i}log(\hat{p}_{i1}) + \frac{n_{i0}}{n_i}log(1-\hat{p}_{i1})\right] \\
&= -2n_i\left[\hat{p}_{i1}log(\hat{p}_{i1}) + \hat{p}_{i0}log(\hat{p}_{i0})\right] \\
&= -2n_i\sum_{k=0}^{1}\hat{p}_{ik}log(\hat{p}_{ik})
\end{aligned}$$

L'entropie elle vaut :

$$E_i = -\sum_k \hat{p}_{ik}log(\hat{p}_{ik})$$

Il y a donc entre les deux un facteur $2/n_i$

A la racine, la distribution du taux d'accidents est $[62_{\text{élevé}}, 69_{\text{bas}}]$

$$\begin{aligned} H(\text{racine}) &= H[62_{\text{élevé}}, 69_{\text{bas}}] \\ &= -\frac{62}{131}\log_2\frac{62}{131} - \frac{69}{131}\log_2\frac{69}{131} \\ &= 0.998 \end{aligned}$$

Si on divise sur le temps avec ensoleillé ou pluvieux, les deux noeuds fils que l'on obtient on une distribution du taux d'acccidents de $[30_{\text{élevé}}, 53_{\text{bas}}]$ pour ensoleillé et $[32_{\text{élevé}}, 16_{\text{bas}}]$ pour pluvieux. On a alors

$$\begin{aligned} H(\text{ensoleillé}) &= -\frac{17+13}{17+22+13+31}\log_2\left(\frac{17+13}{17+22+13+31}\right) \\ &\quad -\frac{22+31}{17+22+13+31}\log_2\left(\frac{22+31}{17+22+13+31}\right) \\ &= 0.5306 + 0.4132 = 0.9438 \\ &= 0.944 \end{aligned}$$

et

$$\begin{aligned} H(\text{pluvieux}) &= -\frac{20+12}{20+12+5+11}\log_2\left(\frac{20+12}{20+12+5+11}\right) \\ &\quad -\frac{5+11}{20+12+5+11}\log_2\left(\frac{5+11}{20+12+5+11}\right) \\ &= 0.39 + 0.528 = 0.918 \end{aligned}$$

et on peut calculer GI(temps):

$$\begin{aligned} GI(\text{temps}) &= H(\text{racine}) - \left(\frac{83}{131}H(\text{ensoleillé}) + \frac{48}{131}H(\text{pluvieux})\right) \\ &= 0.998 - 0.934 \\ &= 0.064 \end{aligned}$$

Si maintenant on divise sur l'attribut traffic routier Chargé ou Léger, les deux noeuds fils que l'on obtient on une distribution du taux d'accidents qui est $[37_{\text{élevé}}, 27_{\text{bas}}]$ pour la modalité Chargé et $[25_{\text{élevé}}, 42_{\text{bas}}]$ pour Léger. On a donc $H(\text{chargé}) = 0.982$ et $H(\text{léger}) = 0.953$ et l'on peut alors calculer $GI(\text{traffic})$:

$$\begin{aligned} GI(\text{traffic}) &= H(\text{racine}) - \left(\frac{64}{131}H(\text{chargé}) + \frac{67}{131}H(\text{léger})\right) \\ &= 0.998 - 0.967 \\ &= 0.031 \end{aligned}$$

Les calculs montrent que $GI(\text{temps}) > GI(\text{traffic})$. On divisera donc sur l'attribut temps en premier.

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?

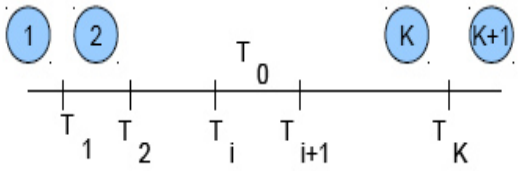

Figure 2: Echelle des valeurs de températures triées. Les cercles bleus indiquent les $K + 1$ régions dans lesquelle les valeurs de températures partitionnent les données.

La Figure (2) montre comment les K valeurs de températures triées divisent la ligne réelle et les données correspondantes. Cela divise la ligne en $K + 1$ régions, indiquées par les cercles pleins bleus. En mettant un seuil dans la région la plus à gauche ou la plus à droite on induit une partition triviale des données. On a donc seulement besoin de considérer les seuils appartenant aux $K-1$ régions étiquetées $2, 3, \ldots, K$. Considérons une région particulière entre 2 et K, disons la région $[T_i, T_{i+1})$. Si T_0 indique un seuil de cette région, il est alors clair que pour tout $T_0 \in [T_i, T_{i+1})$ la même partition est induite sur les données d'apprentissage. On va dire par souci de simplicité que le seuil d'intérêt de la région $[T_i, T_{i+1})$ est $\frac{T_i+T_{i+1}}{2}$. Si les valeurs triées de la variable température sont $T_1, T_2, \ldots, T_K$, alors on a seulement besoin de considérer les $K-1$ valeurs $\frac{T_1+T_2}{2}, \ldots, \frac{T_{K-1}+T_K}{2}$ comme seuils pour des divisions potentielles.

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.
Un arbre de décision n'est pas affecté par des variables dupliqués ou monovalués.

- attribute monovalué. On suppose un découpage avec un attribut monovalué à n'importe quelle étape de la construction de l'arbre. Dans ce cas, toutes les données seront associées à un seul noeud fils qui sera identique au noeud racine pour le découpage. Le gain d'information sera donc égal à 0 pour cette division, rendant l'attribut monovalué le candidat le moins probable pour un découpage.
- attribut dupliqué. On suppose que l'attribut A1 est dupliqué en l'attribut A2. Si aucun des deux ne sont dans l'arbre de décision et sont considérés pour une division, on peut voir que de par leur nature de duplicat il vont induire la même partition sur les données après le découpage. Et comme résultat le gain d'information sera le même. On peut de plus choisir l'un ou l'autre pour la division. Si A1 est déja dans l'arbre et A2 est considéré pour un découpage, le noeud racine du découpage contient déja une valeur (ou un rang de valeurs) de A1. Si A1 prend une seule valeur à la racine, alors A2 doit également prendre une seule valeur et du cas précédent, on sait que le gain d'information du à A2 vaudra 0. Si A1 prend une gamme de valeurs au noeud racine du découpage, alors A2 aura un gain d'information non-négatif et pourra permettre une amélioration de l'erreur d'apprentissage de l'arbre de décision.

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