ENSAI École nationale de la statistique et de l'analyse de l'information	**2ème Année** **Méthodes de discrimination** **ELEMENTS DE CORRECTION TD/TP 2** DISCRIMINATION PAR ARBRES

PARTIE 1 : partie TD

1. **On désire appliquer la méthode CART (discrimination par arbre) pour détecter les champignons non comestibles. Quels sont les grands principes de cette méthode ?**

Principes généraux de la segmentation

Construire un *arbre* à l'aide de divisions successives d'un ensemble d'individus appartenant à un échantillon.

Chaque *division* (ou *scission*) conduit à deux (ou plus) *nœuds* (ou *segments*) :
- le nœud divisé est appelé *nœud-parent*,
- les nœuds générés par la division s'appellent *nœuds-enfants*.
- Les nœuds sont des groupes d'individus le plus homogènes possible par rapport à une *variable à expliquer* (ou *variable cible*) Y, qui peut être nominale, ordinale, quantitative.

Les divisions s'opèrent à partir de *variables explicatives* (ou *prédicteurs*) $X_1 \ldots X_j \ldots X_J$, qui peuvent être nominales, ordinales, quantitatives.

Résultats obtenus en général sous la forme d'un arbre inversé :
la racine (en haut de l'arbre) représente l'échantillon à segmenter, les autres nœuds sont soit des *nœuds intermédiaires* (encore divisibles), soit des *nœuds terminaux*.

L'ensemble des nœuds terminaux constitue une **partition** de l'échantillon en classes homogènes relativement à la variable Y.

Principes spécifiques de la méthode CART :

1. construire un arbre maximal
2. élaguer
3. définir l'arbre (optimal) fiable

Les divisions sont binaires uniquement. Les règles d'arrêt de la procédure de division sont telles qu'on obtient un arbre maximal de grande taille.

2. **Quelles sont les autres méthodes envisageables ?**

On peut aussi utiliser d'autres méthodes de segmentation par arbre comme la méthode CHAID….et KNN ou Classifieur Bayésien Naïf (ou régression logistique que vous verrez plus tard).

Principe de construction de l'arbre CHAID

1.L'échantillon complet (hors échantillon test) constitue la **racine de l'arbre**.

2.Pour chaque variable explicative X_j, on effectue un **regroupement "optimal" de modalités**
3.On cherche parmi les J variables explicatives X_j, leurs modalités ayant été regroupées, celle qui donne la meilleure partition du nœud au sens d'un certain critère :

Le nœud est alors scindé en un nombre de nœuds enfants égal au nombre de modalités de la variable sélectionnée.

On retourne à l'étape 2 pour chaque nœud ainsi constitué.

Deux algorithmes utilisés : un pour construire l'arbre et un pour regrouper les modalités des variables explicatives

D'autres méthodes sont envisageables comme l'analyse discriminante, le modèle logistique, voire les méthodes neuronales...et il existe d'autres méthodes d'apprentissage supervisé (SVM par ex).

3. **L'échantillon total constitué de 8124 observations pourrait être divisé en trois parties :**
   Echantillon d'apprentissage
   Echantillon de validation
   Echantillon test

   Quel serait le rôle de chacun de ces trois échantillons dans la mise en œuvre de CART?

La détermination de la bonne taille de l'arbre s'effectue par post-élagage : l'arbre est dans un premier temps complètement développé avec le critère de pureté sur un premier échantillon (échantillon apprentissage => arbre maximal + séquence d'élagage) , puis, dans un second temps, il est réduit de manière à optimiser le taux de mauvais classement calculé sur un second échantillon dit de validation (choix de l'arbre optimal dans la séquence d'élagage par validation simple). Lors de cette seconde phase, il est possible d'introduire une matrice de coût de mauvais classement. Enfin, l'échantillon-test donne une bonne estimation du coût de l'arbre pour le comparer à d'autres méthodes (autre arbre, régression logistique, analyse discriminante).

4. **Pourrait-on se passer de créer ces trois sous-échantillons ? Si oui, quelle modification de la méthode en découlerait ?**

Oui, on pourrait avoir envie de sélectionner le sous-arbre optimal dans la séquence d'élagage grâce à la validation croisée au lieu d'un échantillon de validation (validation simple). En général, on réserve la validation croisée aux petits échantillons.

Méthode de la validation croisée :

On détermine la séquence de sous-arbres emboîtés de coût-complexité minimum de T_{max} à partir de l'échantillon **total** (identifié par la modalité « base » de la variable ***echantillon*** dans la table *muschroom.txt* pour la partie TP)

$$S_T = \{T_1, T_2 .... T_l\}$$

A cette séquence correspondent les 2 autres séquences :

$$S_C = \left\{ \hat{C}(T_1), \hat{C}(T_2), ..., \hat{C}(T_h), ..., \hat{C}\ (\{r\}) \right\}$$

$$S_\alpha = \left\{ \alpha_1 = 0 < \alpha_2 < ... < \alpha_h < ... < \alpha_{\{r\}} \right\}$$

I
On va estimer la séquence de coûts par validation croisée
L'échantillon total (hors test) d est divisé en V sous-ensembles (souvent V=10) ce qui permet de former V couples d'échantillons (d_V , d^V) avec v= 1 à V.

d_1 …d_V sous-échantillons de taille = card (d)/V (exemple = 1/10 de d)
Et V sous- échantillons complémentaires : d^1 … d^V / d^v=d-d_V (de taille = 9/10 de d)

Pour chaque échantillon d^V (9/10 de d), v= 1 à V
On construit un arbre maximal. On élague selon la méthode du coût-complexité.
On obtient la séquence

$$S_T^v = \left\{ T_1^v, T_2^v ... T_{lv}^v \right\}$$

On estime le coût de chaque sous-arbre de cette séquence sur chaque échantillon de validation d_V (1/10 de d)

$$S_C^v = \left\{ \hat{C}_1^v, \hat{C}_2^v .... \hat{C}_{lv}^v \right\}$$

Objectif de la validation croisée :
estimer les coûts associés aux sous-arbres de S_T à partir des coûts estimés des sous-arbres des V séquences S_{TV}, v=1 à V

5. **Quel critère de division d'un nœud utilise-t-on pour construire l'arbre maximal ?**
   Quelles sont les fonctions d'impureté les plus souvent utilisées ?

Le critère utilisé est celui de l'impureté. Les deux fonctions d'impureté les plus utilisées sont d'une part l'indice d'entropie de SHANNON, d'autre part celui de GINI :

SHANNON : $-\sum_r P\left(G_r / t\right) \ln\left(P\left(G_r / t\right)\right)$

GINI : $\sum_{r \neq s} P\left(G_r / t\right) P\left(G_s / t\right) = 1 - \sum_{r=1}^{K} \left[P\left(G_r / t\right) \right]^2$

Rappel : ces fonctions sont concaves, ce qui entraîne que toute division d'un noeud conduit à une réduction (au sens large, i.e. avec possible égalité) de l'impureté.

6. **La variable à expliquer Y étant binaire, elle définit une partition de la population en deux groupes r avec r=1 ou r=2. Rappeler l'expression de la probabilité a posteriori d'appartenance au groupe Gr pour les éléments d'un nœud t. Comment l'estime-t-on ?**

D'après le théorème de BAYES : $P(G_r / t) = \frac{\pi_r P(t / G_r)}{P(t)}$ avec $P(t) = \sum_s \pi_s P(t / G_s)$.

$P(t / G_s)$ est estimé par $\frac{n_s(t)}{n_s}$ proportion des observations du nœud t dans le groupe s.

Si $\pi_s = \frac{n_s}{n}$ alors $P(t) = \sum_s \frac{n_s}{n} \frac{n_s(t)}{n_s} = \sum_s \frac{n_s(t)}{n} = \frac{n(t)}{n}$ et $P(G_r / t) = \frac{n_r(t)}{n(t)}$.

La probabilité a posteriori d'appartenir au groupe r pour une observation du nœud t est estimée par la proportion d'observations du nœud t appartenant au groupe r.

7. **Les probabilités a priori sont supposées proportionnelles aux effectifs dans l'échantillon. L'indice de diversité de GINI a été retenu comme fonction d'impureté. Quelle serait l'impureté initiale (segment racine t0), si par exemple parmi les 4882 champignons de l'échantillon d'apprentissage, 2531 étaient comestibles et 2351 étaient poisons?**

L'impureté initiale, en utilisant l'indice de GINI, est :

$i(t) = 2\,p(classe = "e" / t)\,p(classe = "p" / t)$,

soit ici 2 x (2531 x 2351)/ (4882^2) = 0,499

ou $i(t) = 1 - [\ p(classe = "e" / t)\ ]^2 - [\ p(classe = "p" / t)\ ]^2$

8. **Combien y a-t-il de divisions possibles pour le nœud racine ?**

On dispose de 6 variables nominales (**Odor, stalk-shape, stalk-root, stalk-color-above-ring, stalk-color-below-ring, spore-print-color**).
Une variable nominale à k modalités génère 2^{k-1} -1 divisions possibles.

odor = odeur : 9 modalités => 255 divisions possibles

stalk-shape : forme du pied 2 modalités => 1 division possible

stalk-root : racine 6 modalités=> 31 divisions possibles

stalk-color-above-ring : couleur de tige au-dessus de l'anneau
9 modalités=> 255 divisions possibles

stalk-color-below-ring : couleur de tige au-dessous de l'anneau
9 modalités=> 255 divisions possibles

spore-print-color : couleur des spores
9 modalités=> 255 divisions possibles

Au total : 255x4 + 1 + 31 = 1052 divisions possibles.

PARTIE 2 : partie TP avec mise en œuvre sous R

La table contenant les données s'intitule ***muschroom.csv***.

9. **Mettre en œuvre une première analyse sous R en supposant les probabilités a priori proportionnelles aux effectifs et les coûts de mauvais classement égaux.**

Voir document programme et listings R

10. **Par quelle variable et quelles modalités la racine est-elle divisée ? Comment sont définis les segments t_1 et t_2 ? Calculer la variation d'impureté due à cette division binaire (indicateur de Gini) dans l'échantillon de base**

La racine est divisée en 2 nœuds **t_1 et t_2** avec la variable odor et la coupure construite sur la division binaire des modalités :

- pour **t_1** les modalités odor = {a, l, n}
- pour **t_2** les modalités odor = {c,f,m,p,s,y}

Les segments t_1 et t_2 sont définis par les règles de décisions qui découlent de cette partition des modalités. Si odor dans {a, l, n} alors segment **t_1**, sinon segment **t_2**.

Calcul de la variation d'impureté due à cette première division binaire (indicateur de Gini) dans l'échantillon de base (voir page 2 des listings) :

p_1 = proportion d'observations du nœud t dans le nœud t_1 = $n(t_1) / n(t)$

p_2 = proportion d'observations du nœud t dans le nœud t_2 = $n(t_2) / n(t)$

$i(t0)$ = $i[p(Gr/t0)]$ avec r =1 à 2

$i(t_1)$ = $i[p(Gr/t_1)]$ avec r =1 à 2

$i(t_2)$ = $i[p(Gr/t_2)]$ avec r =1 à 2

$\Delta i(d,t) = i(t_0) - p_1\, i(t_1) - p_2\, i(t_2)$

$i(t0) = 2 \times (3153 \times 2940)/ (6093^2) = 0.499389$

$i(t_1) = 2 \times (3153 \times 84)/ (3237^2) = 0.05055311$

$i(t_2) = 0$

$\Delta i(d,t) = 0.499389 - (3237/6093) \times 0.05055311 = 0.4725319$

A noter, pour chaque nœud t non terminal, le logiciel R donne un indicateur « improve » qui est égal à $n(t) \times \Delta i(d,t)$

11. **Une autre division de la racine aurait peut-être pu donner une réduction d'impureté presque aussi bonne. Comment qualifie-t-on cette autre division ? Que peut nous apporter le fait de s'intéresser à cette autre division ? Donner cette autre division pour le nœud t1.**

Division équi-réductrice (concurrente) : « primary splits »

Lors de la construction de l'arbre, division qui correspond à la valeur du critère de sélection des divisions, la plus proche de celle de la meilleure division d* (selon X_j) , avec la variable $X_{j'}$ (j'≠j).

Intérêt : choisir lors de la construction une variable alternative qui soit plus pertinente (d'un point de vue médical par ex), ou moins difficile à collecter

Pour le nœud racine (nœud n°1 pour R) on a page 4 du listing : d* grâce à odor puis la 2ème division équi-réductrice est donnée par spore_print_color.

12. Quel est le nombre de segments terminaux de l'arbre optimal ?

Trois (page 3 des listings) : on définit ainsi une nouvelle partition des observations.

13. Quel est le principe d'affectation d'un noeud terminal ?

Règle pour une matrice de coûts unitaires: Affectation du nœud t à G_r si **$p(G_r/t) > p(G_S/t)$**

$\forall r,s = 1$ à K avec $s \neq r$

Donc si on choisit les probabilités a priori égales aux fréquences empiriques $\pi_r = n_r /n$, la règle s'écrit : **$p(G_r/t) = n_r(t)/n(t) > p(G_S/t) = n_S(t)/n(t)$**

$\Rightarrow$affectation du segment t au groupe le plus représenté dans t.

14. Construire la matrice de confusion et calculer le taux d'erreur sur l'échantillon test.

```
predtest
     e    p
e 1055    0
p   12  964
```

On voit ici 12 faux négatifs si on considère que l'événement c'est « poison ».

Taux d'erreur = 0.0059 soit 0.59% « seulement » …mais si ça représente des décès par empoisonnement c'est trop !

15. Si l'on vous apporte un nouveau champignon qui présente les caractéristiques suivantes : odeur = amande, forme du pied s'élargissant, racine en forme de massue, couleur de tige au-dessus de l'anneau = cannelle, couleur de tige au-dessous de l'anneau = chamois, couleur des spores = chamois. Le classerez-vous en catégorie poison ou comestible ?

On applique les règles de décisions = une succession de tests = on fait « parcourir » l'arbre à cette nouvelle observation. On arrive alors au nœud terminal où on lui attribue sa valeur : e comestible.

16. Mettre en œuvre une seconde analyse CART. On fait désormais varier le coût de mauvais classement selon la matrice de coût indiquée : C(comestible/poison)=1000 et C(poison/comestible)=1. En quoi cette modification est-elle pertinente ?

Cette modification est pertinente dans la mesure où on préfère largement prendre le risque de jeter un champignon comestible plutôt que de conserver un champignon qui risque de tuer quelqu'un.

17. Quel est le principe d'affectation d'un noeud terminal avec cette nouvelle matrice de coût°?

Coût d'erreur d'affectation d'un segment

On définit les coûts d'erreur de classement a priori, pour un individu :

c(s/r) = coût de l'affectation d'un individu au groupe s alors qu'il appartient au groupe r

c(s/s)=0

On définit le **coût d'affectation du segment t** au groupe G_S

$$CM_s(t) = \sum_{r=1}^{K} c(s/r)\, p(G_r/t)$$

Nouveau critère d'affectation = minimisation du coût d'affectation et non plus maximisation de $p(G_r/t)$.

18. Tester une autre méthode de segmentation par arbre vue en cours.

On va essayer la méthode CHAID. Remarque : impossible avec cette méthode d'avoir des coûts d'erreurs différents contrairement à la méthode CART = pas les mêmes critères de construction de l'arbre.
On peut comparer la matrice de confusion à celle obtenue avec CART sur échantillon test.