ENSAI École nationale de la statistique et de l'analyse de l'information

TP 2 Apprentissage supervise 2A

Septembre 2019

On dispose d'un ensemble de données sur les champignons (source The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York : Alfred A. Knopf). Il est constitué de 8124 observations pour lesquelles diverses descriptions sont disponibles comme la surface, l'odeur, la couleur, etc, ainsi que l'information : comestible ou poison.

L'objectif de ce TD/TP est de construire un modèle prédictif capable de différencier les champignons comestibles des non-comestibles, grâce aux méthodes de segmentation par arbres.

Contexte

Variable cible :

Classe : comestible=e, poison=p

Variables explicatives :

odor = odeur : amande (almond) = a, anis (anise) = l, creosote (creosote) = c, poisson (fishy) = y, repugnant (foul) = f, moisi (musty) = m, aucune (none) = n, âcre (pungent) = p, épicé (spicy) = s

stalk-shape : forme du pied s'élargissant (enlarging) = e, se resserrant (tapering) = t

stalk-root : racine bulbeux (bulbous) = b, en forme de massue (club)=c, en forme de corolle (cup)=u, égales ou par paires (equal) = e, avec des rhizomes (rhizomorphs) =z, racines (rooted) = r

stalk-color-above-ring : couleur de tige au-dessus de l'anneau marron (brown)=n, chamois (buff)=b, cannelle (cinnamon) =c, gris (gray)=g, orange=o, rose (pink) = p, rouge (red) = e, blanc (white) = w, jaune (yellow) =y

stalk-color-below-ring : couleur de tige au-dessous de l'anneau marron (brown)=n, chamois (buff)=b, cannelle (cinnamon) =c, gris (gray)=g, orange=o, rose (pink) = p, rouge (red) = e, blanc (white) = w, jaune (yellow) =y

spore-print-color : couleur des spores noire (black) = k, marron (brown) = n, chamois (buff) = b, chocolat (chocolate) = h, verte (green) = r, orange=o, violette (purple) =u, blanche (white) = w, jaune (yellow) = y

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 ?

2. Quelles sont les autres méthodes envisageables ?

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 oeuvre de CART ?

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

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

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

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 (dans le segment racine), si par exemple, parmi les 4882 champignons de l'échantillon d'apprentissage, 2531 étaient comestibles et 2351 étaient poisons ?

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

PARTIE 2 : partie TP avec mise en oeuvre sous R

La table contenant les données s'intitule *mushroom.csv*. Elle se trouve dans le répertoire Apprentissage Supervisé dans moodle.

9. Mettre en oeuvre une première analyse sous R :
- en supposant les probabilités a priori proportionnelles aux effectifs et les coûts de mauvais classement égaux,
- en utilisant la validation croisée sur l'échantillon "base" (lignes identifiées par cette modalité avec la variable *echantillon* de la table *mushroom.csv*).

10. Par quelle variable et quelles modalités la racine t_0 est-elle divisée ? Comment sont définis les segments t_1 (noeud enfant gauche) et t_2 (noeud enfant droit) ? Calculer la variation d'impureté due à cette division binaire (indicateur de Gini) dans l'échantillon "base".

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 noeud t_1.

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

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

14. Calculer le taux d'erreur sur l'échantillon test.

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 ?

16. Mettre en oeuvre 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 ?

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

18. Tester une autre méthode de segmentation par arbre : CHAID. Permet-elle d'intégrer la notion de coûts différenciés des erreurs ?