Introduction au data mining
L3 MIS - STA 1616 - 2010
V. Monbet
Arbres de décision

L'objectif de ce TD est de comprendre comment on construit un arbre de décision.

1 A la main...

Considérons le tableau de données suivant portant sur l'idée suivante : *Imagine you only ever do four things at the weekend : go shopping, watch a movie, play tennis or just stay in. What you do depends on three things : the weather (windy, rainy or sunny) ; how much money you have (rich or poor) and whether your parents are visiting. You say to your yourself : if my parents are visiting, we'll go to the cinema. If they're not visiting and it's sunny, then I'll play tennis, but if it's windy, and I'm rich, then I'll go shopping. If they're not visiting, it's windy and I'm poor, then I will go to the cinema. If they're not visiting and it's rainy, then I'll stay in.*

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

1. Construire et tracer un arbre de décision permettant de prédire la variable décision. On se basera sur le critère d'entropie pour déterminer les noeuds de l'arbre.

2. Estimer le support des règles correspondant aux feuilles dont la modalité prédite est `Cinema`.

2 Avec le logiciel R

Dans cette section nous considérons un jeux de données (`cmc1.data`) qui est un sous ensemble de l'enquête menée en 1987 en Indonésie et concernant les moyens de contraception utilisés par les femmes. Les individus interrogés sont des femmes mariées qui ne sont pas enceintes. Le problème consiste à proposer un modèle afin de prédire la méthode de contraception choisie par une femme (no use, long-term methods, or short-term methods) à partir de certaines de ses caractéristiques démographiques et socio-économiques.

Les variables renseignées sont les suivantes :

1. Age de la femme (numérique)
2. Education de la femme (catégorielle) - 1=faible, 2, 3, 4=élevé
3. Education du mari (catégorielle) - 1=faible, 2, 3, 4=élevé
4. Nombre d'enfants déjà nés (numérique)
5. Religion de la femme (binaire) - 0=Non-Islam, 1=Islam
6. La femme travaille-t'elle ? (binaire) 0=Oui, 1=Non
7. Occupation du mari (catégorielle) - 1, 2, 3, 4
8. Indice de niveau de vie (catégorielle) - 1=faible, 2, 3, 4=élevé
9. Exposition aux médias (binaire) - 0=Bon, 1=Mauvais
10. Méthode contraceptive utilisée (attribut de classe) - 1=Pas de contraception, 2=Contraception à long terme, 3=Contraception à court terme

Nous proposons dans un premier temps de construire un arbre de décision selon l'algorithme CART. Et nous utiliserons le package R intitulé `rpart`.

1. Lire les données sous R.
2. Faire une analyse descriptive rapide du jeux de données (histogrammes, ACM).
3. La fonction `rpart` du package permet de construire et d'élaguer un arbre de décision.

```
A <- rpart(Method ~ age + education + husb_educ + nb_childs + religion
+ working + husn_occup + liv_index + media_exp , data = cmc, method="class")
printcp(A)
plotcp(A) # Gives a visual representation of the cross-validation
```

```
          # results in an rpart object.
summary(A)
```

4. La fonction `predict` permet de prédire en utilisant l'arbre obtenu par la fonction `rpart`. Construire un échantillon d'apprentissage et un échantillon de test (utiliser par exemple la fonction `sample`). Estimer le risque de Bayes correspondant à l'arbre obtenu. Répéter plusieurs fois la génération des échantillons de test et d'apprentissage. Quel est le risque moyen obtenu ?
5. Regrouper les modalités de l'attribut de classe de façon à ne conserver que 2 classes (Pas de contraception/Contraception). Et reprendre les questions précédentes.
6. Utiliser maintenant l'algorithme apriori pour extraire les règles contenant l'item `Contraception=Oui`. Puis interpréter les règles ayant une confiance et un support important.

3 Avec le logiciel SAS

Dans SAS, les algorithmes de construction d'arbres de décision ne sont implémentés que dans SAS/Enterprise Miner. Construire un diagramme pour répondre aux mêmes questions qu'avec le logiciel R. Obtient-on les mêmes résultats ?