Datamining 2: Classification supervisée Partie 3 : Les arbres de décision

M2 STD

September 21, 2016

Les limites des méthodes précédentes

a) Les X_i sont forcément des variables quantitatives

b) On peut classer mais difficilement expliquer les classes, le probleme est résolu numériquement mais la solution n'apporte pas de compréhension générale

L'exemple historique

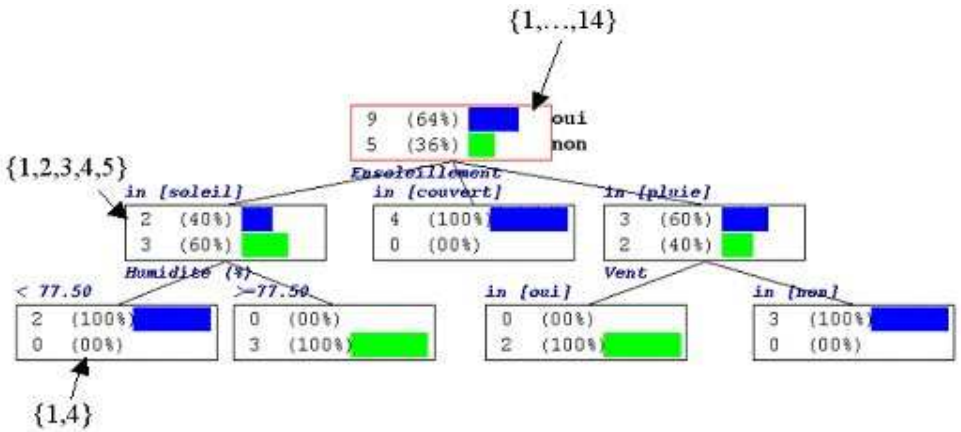
L'exemple initial était le suivant:

Numéro	Ensoleillement	Température (°F)	Humidité (%)	Vent	Jouer
1	soleil	75	70	oui	oui
2	soleil	80	90	oui	non
3	soleil	85	85	non	non
4	soleil	72	95	non	non
5	soleil	69	70	non	oui
6	couvert	72	90	oui	oui
7	couvert	83	78	non	oui
8	couvert	64	65	oui	oui
9	couvert	81	75	non	oui
10	pluie	71	80	oui	non
11	pluie	65	70	oui	non
12	pluie	75	80	non	oui
13	pluie	68	80	non	oui
14	pluie	70	96	non	oui

L'exemple historique

L'exemple initial était le suivant:

Numéro	Ensoleillement	Température (°F)	Humidité (%)	Vent	Jouer
1	soleil	75	70	oui	oui
2	soleil	80	90	oui	non
3	soleil	85	85	non	non
4	soleil	72	95	non	non
5	soleil	69	70	non	oui
6	couvert	72	90	oui	oui
7	couvert	83	78	non	oui
8	couvert	64	65	oui	oui
9	couvert	81	75	non	oui
10	pluie	71	80	oui	non
11	pluie	65	70	oui	non
12	pluie	75	80	non	oui
13	pluie	68	80	non	oui
14	pluie	70	96	non	oui



Algorithme

tant qu'un critère d'arret n'est pas satisfait:

1. Les anciennes "feuilles" actives sont les "racines"
2. Pour chaque racine on cherche la variable qui discrimine le mieux la discrimination qui en ressort nous donne de nouvelles feuilles
3. Si dans une feuille on n'observe qu'une seule classe ou bien un effectif de 1 la feuille deviens passive

On affecte une classe a chaque feuille de l'arbre.
On éllague l'arbre

Les critères d'arrets

Combinaison de:

a) la profondeur de l'arbre a atteint un certain seuil

b) le nombre de feuilles a atteint un certain seuil

c) l'effectif de chaque feuille est suffisament petit (variante: si on continuait d'une étape on aurait des feuilles d'effectif trop faible)

d) la qualité de l'arbre est suffisante (variante: si on continuait d'une étape on ne gagnerait pas beaucoup en qualité)

l'affectation d'une classe a une feuille

Notons $C_{i,j}$ le cout du mauvais classement suivant d'un individu de la (vrai) classe i dans la classe (estimée) j

1. bien sur $C_{i,i} = 0$
2. par défaut $C_{i,j} = 1$ des que $i \neq j$

Si dans une feuille donnée on observe $p_1, \ldots p_j, \ldots p_K$ ou p_j est la proportion de l'effectif de la feuille dans la classe j. Affecter toute la feuille a la classe j a un cout $\sum p_i C_{i,j}$. On affectera la feuille a la classe qui minimise son cout.

L'ellagage

L'ellagage repose sur la capacité de généralisation de l'arbre. On apprend un arbre sur une base d'apprentissage. On mesure son "cout" total sur une base test en fonction de sa profondeur d'éllagage. On coupe les branches pour la profondeur minimisant le cout sur la base test. Cette procédure permet de rendre moins important le critère d'arret choisit (au pire on va jusqu'a des feuilles pures ou ne contenant qu'un individu et on ellague)

Rappel

tant qu'un critère d'arret n'est pas satisfait:

1. Les anciennes "feuilles" actives sont les "racines"
2. **Pour chaque racine on cherche la variable qui discrimine le mieux la discrimination qui en ressort nous donne de nouvelles feuilles**
3. Si dans une feuille on n'observe qu'une seule classe ou bien un effectif de 1 la feuille deviens passive

On affecte une classe a chaque feuille de l'arbre.
On éllague l'arbre

Selection de la variable discrimante et discrimination

Cela dépend de la méthode choisie (principalement CART ou CHAID)

CHAID (historiquement le plus ancien)

On considère que les variables explicatives sont QUALITATIVES (si ce n'est pas le cas on regroupe les variable quantitatives en plusieurs classes "a la main" au préalable).
Le critère de base est le crière du χ^2.

CHAID : organisation

A chaque noeud on répete:

1. Pour chaque variable discriminante
   - 1-a Fusion des modalités
   - 1-b Calcul du χ^2
2. On utilise la variable associée au meilleur χ^2 (i.e. celui associé a la plus petite $p-$value, éventuellement corrigée par Bonferroni) et on la scinde en les modalités obtenues a l'étape $1-a$

CHAID : étape fusion des modalité

Pour une variable discrinante donnée a k' modalités Pour un seuil fixé α (usuellement 5.10^{-2}) on itère. Tant qu'il existe des modalités non significativement discriminantes au seuil α.

1. On cherche le meilleur regroupement de modalités i et j i.e. celui dont le tableau de contingence

	$Y=1$	$\ldots$	$Y=k$
i	$n_{i,1}$	$\ldots$	$n_{i,k}$
j	$n_{j,1}$	$\ldots$	$n_{j,k}$

   a le plus petit χ^2
2. On regroupe les modalités i et j qui ont le plus petit χ^2 si celui ci est associé a une $p-$value $> \alpha$

CHAID : étape fusion des modalité

Deux détails

1. si les modalités sont “ordonées” on ne regroupe que des modalités “adjacentes” (attention a bien spécifier la nature des variables)
2. Si on obtient a la fin de la fusion une modalité d'effectif “trop” petit on la regroupe a sa modalité la plus proche (toujours en terme de χ^2) même si elle est “éloignée”.

CHAID : correction de Bonferroni

Pour chaque variable explicative X_i, on obtient un regroupement en k_i modalités, et un χ^2_i, associé a une $p-$value p_i. Le critère de correction de Bonferroni reviens a regarder plutot la variable associée à la plus petite valeur de $p_i.k_i$ (on pénalise par le nombre de modalités) sinon on recherche al variable associée a la plus petite valuer de p_i. Fondamentalement ca ne change pas grand choses

CART : organisation

A chaque noeud on répete:

1. Pour chaque variable discriminante on
   - 1-a recherche de sa meilleur sission en deux classes (selon le crière de GINI)
2. On discrimine suivant la variable donnant le “meilleur” indice de Gini

CART: recherche de la meilleur sission en 2

1. Variables qualitatives : toutes les possibilités (i.e $2^{k-1} - 1$ si la variable explicatives a k modalités).
2. Variables ordonnées : toutes les possibilités "raisonnables" (i.e. $k - 1$ si k modalités).
3. Variables quantitatives : toutes les possibilités du type $X \leq s$ (i.e. $n - 1$ si on a n individus).

CART : l'indice de GINI

GROS AVANTAGE : si on a les couts de mauvaises affectation on peut les insérer pour avoir quelque chose d'optimal. L'indice de Gini d'un noeud est $G(\text{noeud}) = \sum_{i,j} C_{i,j} f_i f_j$ ou f_i est le pourcentage d'individu dans la classe i. Plus le noeud est pur plus l'indice de Gini est petit. Dans le cas idéal on a 100pourcent des individus dans une classe et personne ailleurs donc $G = 0$. La valeur maximal (equirépartition) dépend du nombre de classes (exercice facile)

CART : meilleur séparation suivant l'indice de Gini

A un noeud fixé, on teste toutes les séparations en 2 suivant toutes les variables. A chaque séparation on a n_1 individus qui partent dans un nouveau noeud 1 et n_2 dans un nouveau noeud 2. On garde la séparation qui minimise

$$\frac{n_1}{n_1+n_2} G(\text{noeud}_1) + \frac{n_1}{n_1+n_2} G(\text{noeud}_2)$$

CART ou CHAID ?

CART produit des arbres moins larges et plus profond (car il produit des arbres binaires) en revanche il y a bien moins de parametres a regler (au pire on va jusqu'a des noeuds purs et on ellague a posteriori) alors que dans CHAID il faut regler le parametre α

Les forets aléatoires

On réalise beaucoup d'arbres de décisions complets (on s'arrette lorsque chaque feuille comprend une seule observation) L'aléatoire provient de deux points :

1. des bases d'apprentissages différents tirées aléatoirement
2. A chaque étape de l'arbre on tire aléatoirement quelques variables "a considérer"

Au final pour une nouvelle observation on affecte la classe obtenue le plus souvent