Classe Master 1 SIAD

Cours en fouille de données

D. Boukraâ , MCA– 2018/2019. Université de Jijel

Chapitre 5

Arbres de décision

1. Introduction

Arbres de décision : moyen d'apprentissage supervisé qui permet de séparer des individus dans des groupes selon des règles ou de prévoir la valeur d'une variable continue (cible) à partir de variables en entrée.

Apprentissage supervisé: les classes ou variables à prévoir sont prédéfinis.

Objectif : induire un ensemble de règles à appliquer à une nouvelle instance afin de déterminer sa classe d'appartenance ou à connaître la valeur de la variable.

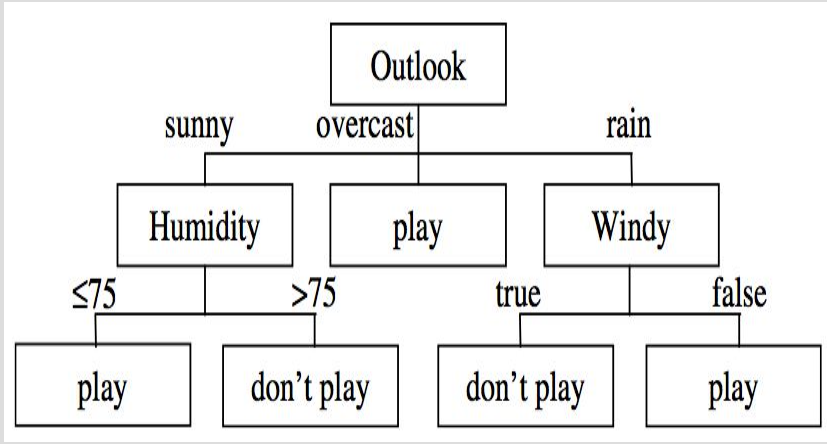
Les règles s'enchaînent de sorte que chaque règle correspond à un test donnant lieu à un nœud et les alternatives de réponse au test donnent lieu aux branches.

2. Exemples d'utilisation

- Connaître le degré risque d'un demandeur de prêt ?
- Connaître le risque qu'une population attrape une maladie
- Reconnaître n objet détecté par un radar (véhicule, individu, immeuble, arbre).
- Etablir le degré de ressemblance de personnes à un criminel recherché
- Savoir si un message est un spam
- …

3. Exemple illustratif

Outlook	Temp (°F)	Humidity (%)	Windy	Class
sunny	75	70	true	play
sunny	80	90	true	don't play
sunny	85	85	false	don't play
sunny	72	95	false	don't play
sunny	69	70	false	play
overcast	72	90	true	play
overcast	83	78	false	play
overcast	64	65	true	play
overcast	81	75	false	play
rain	71	80	true	don't play
rain	65	70	true	don't play
rain	75	80	false	play
rain	68	80	false	play
rain	70	96	false	play



4. Formulation du problème

- **Données nécessaires à l'induction de règles de décision**
    - Ensemble d'apprentissage (training set).
    - Se présentent sous une forme tabulaire *individu*/*attribut*
    - peuvent être qualitatives (e.g. Outlook, Windy, Class) ou numériques (Temp et Humidity).
- **Une variable à prédire ou expliquer, endogène** (ex: Class avec deux modalités : *play* et *don't play*.)
- **Plusieurs variables d'explication, exogènes** (ex: temperature)
- **Problème:** élaborer un arbre où chaque nœud (appelé nœud de décision) consiste en un test sur la valeur d'un attribut.

5. Algorithme de base

Principe

- A partir d'un nœud initial, choisir à chaque étape un attribut pour la division (split) de l'ensemble d'apprentissage en sous-ensembles
- Chaque sous-ensemble correspond à l'attribut sélectionné et donne lieu à un nœud fils.
- Chaque branche du père vers le fils correspond à une valeur de l'attribut sélectionné.
- L'algorithme s'arrête lorsqu'il n'y a plus d'attributs de sélection.

5. Algorithme de base

Algorithme

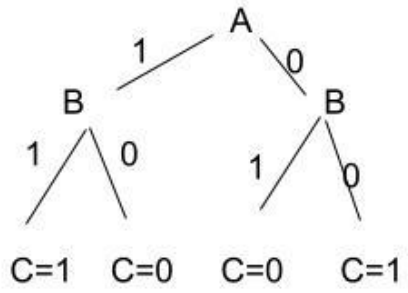
1. A □ l'ensemble d'apprentissage E. Créer un nœud de A.
2. Si tous les exemples de E ont la valeur positive pour la classe de prédiction, alors créer un nœud P sous la racine et s'arrêter.
3. Si tous les exemples de E ont la valeur négative, alors créer un nœud N sous la racine et s'arrêter.
4. Sélectionner un attribut X ayant les modalités v_1, v_2, …v_N et partitionner E en N sous-ensembles E_i, chacun correspondant à une modalité. Pour chaque sous-ensemble, créer un nœud ayant comme libellé la modalité v_i.
5. Pour chaque sous-ensemble E_i, E □ E_i, aller à ii.

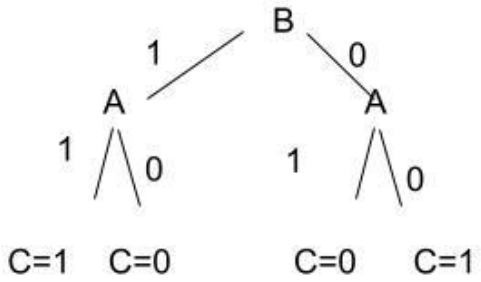
6. Problème de sélection des attributs

- Le choix peut être aléatoire, ou basé sur un critère particulier.

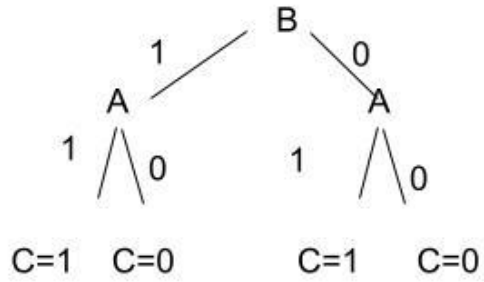
- Les arbres obtenus peuvent être différents en profondeur. Le choix du classifieur détermine l'efficacité de l'induction à partir de l'arbre de décision.

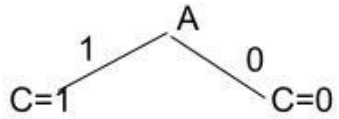
A	B	**C**
1	1	1
1	0	0
0	1	0
0	0	1





A	B	**C**
1	1	1
1	0	1
0	1	0
0	0	0





7. Sélection à base d'entropie / Algorithme ID3 (iterative dichotomiser)

- Entropie (Shannon): elle mesure la quantité d'information attendue lors de l'envoi d'un message. Elle est liée à l'incertitude par rapport à un phénomène donné.
- L'entropie est maximale lorsque les modalités ont la même probabilité.
- L'entropie est nulle lorsqu'on possède une certitude sur une modalité.
- L'entropie est utilisée dans l'estimation du gain information (ou réduction d'incertitude) par rapport à une valeur de la classe à prédire.
- L'attribut choisi à une étape est celui qui minimise l'incertitude par rapport aux valeurs de la classe à prédire.
- La réduction d'incertitude (d'entropie) est également appelée le gain informationnel.

7. Sélection à base d'entropie / Algorithme ID3 (iterative dichotomiser)

- Utilisation de l'entropie de Shannon pour le choix de la variable de découpage
- **Entropie**: quantité d'information attendue lors de l'envoi d'un message, liée à l'incertitude par rapport à un phénomène donné.
    - maximale lorsque les modalités ont la même probabilité.
    - nulle lorsqu'on possède une certitude sur une modalité.
- **Utilisation** : estimation du gain information (ou réduction d'incertitude) par rapport à une valeur de la classe à prédire.
- L'attribut choisi à une étape est celui qui minimise l'incertitude par rapport aux valeurs de la classe à prédire.
- La réduction d'incertitude (d'entropie) est également appelée le gain informationnel.

Exemple

Formule de l'entropie: Entropie (C) = $\sum_{i=1}^{c} - p_i \log_2 p_i$

- Pour chaque attribut candidat aux découpage (fils du nœud courant), on calcule l'entropie de chaque modalité, les probabilités étant les fréquences de chaque modalité de la variable à expliquer
- On somme les entropies des modalités de chaque attribut fils et on les soustrait de l'entropie globale du nœud père, on obtient le gain informationnel

$$\text{Gain}\ (S,\ A)\ =\ \text{Entropy}\ (S)\ -\ \sum_{v\ \in\ \text{Values}\ (A)} \frac{|S_v|}{|S|}\ \text{Entropy}\ (S_v)$$

- L'attribut choisi pour le découpage est celui qui maximise le gain informationnel

Exemple

- L’ensemble S initial contient 14 jours.
- Les attributs *Temp* et *Humidity* sont continus
    - Classer le premier dans 3 classes : hot (>= 80), mild (<=70) et cool (70 < temp < 80)
    - Classer le second dans deux classes : normal (<85) et high (>=85).
- Coder les valeurs de l’attribut Class en Yes ou Play et No autrement

Exemple

Outlook	Temp	Humidity	Windy	Class
Sunny	Cool	Normal	True	Y
Sunny	Hot	High	True	N
Sunny	Hot	High	False	N
Sunny	Cool	High	False	N
Sunny	Mild	Normal	False	Y
Overcast	Cool	High	True	Y
Overcast	Hot	Normal	False	Y
Overcast	Mild	Normal	True	Y
Overcast	Hot	Normal	False	Y
Rain	Cool	Normal	True	N
Rain	Mild	Normal	True	N
Rain	Cool	Normal	False	Y
Rain	Mild	Normal	False	Y
Rain	Mild	High	False	Y

Exemple

- Etape 1: choix du premier attribut : on calcule le gain informationnel des choix Outlook, Temp, Humidity, Windy.
- Soit S l'ensemble initial. Valeurs (Outlook) : *Sunny*, *Overcast*, *Rain*.
- Gain (S,Outlook) = Entropie (S) -SOMME $_{v \text{ dans } \{sunny,\ overcast,\ rain\}}$($|S_v|/|S|$ * Entropie(S_v)) = Entropie (S) - (5/14)*entropie (S_{sunny}) –(4/14)*entropie($S_{overcast}$) – (5/14)*entropie(S_{rain}) = 0.94 - (5/14)*0,97-0-(5/14)*0,97= **0.25.**
- On trouve aussi Gain (S, Temp) = **0.04** , Gain(S, Humidity)=**0.005** et Gain(S, Windy) = **0.05**
- On choisit OUTLOOK comme attribut de découpage

Exemple

- Nouveaux sous-tableaux obtenu après le découpage par Outlook

Temp	Humidity	Windy	Class
Cool	Normal	True	Y
Hot	High	True	N
Hot	High	False	N
Cool	High	False	N
Mild	Normal	False	Y

Outlook *Sunny*

emp	Humidity	Windy	Class
Cool	High	True	Y
Hot	Normal	False	Y
Mild	Normal	True	Y
Hot	Normal	False	Y

Outlook *Overcast*

Temp	Humidity	Windy	Class
Cool	Normal	True	N
Mild	Normal	True	N
Cool	Normal	False	Y
Mild	Normal	False	Y
Mild	High	False	Y

Outlook *Rain*

Refaire le travail

8. Démarche de construction d'un AD

1. Définir le critère de choix de la variable de séparation (Chi2, Gini, Entropie…)
2. Définir un critère d'arrêt dans la construction de l'arbre
3. Construire l'arbre complet ou en respectant le critère d'arrêt
4. Elaguer l'arbre (construit complètement ou ayant une grande profondeur)
5. Tester différents arbres sur un échantillon d'instances classées et choisir le meilleur modèle selon le critère d'évaluation (voir plus loin)
6. Appliquer le modèle sur des instances non classées

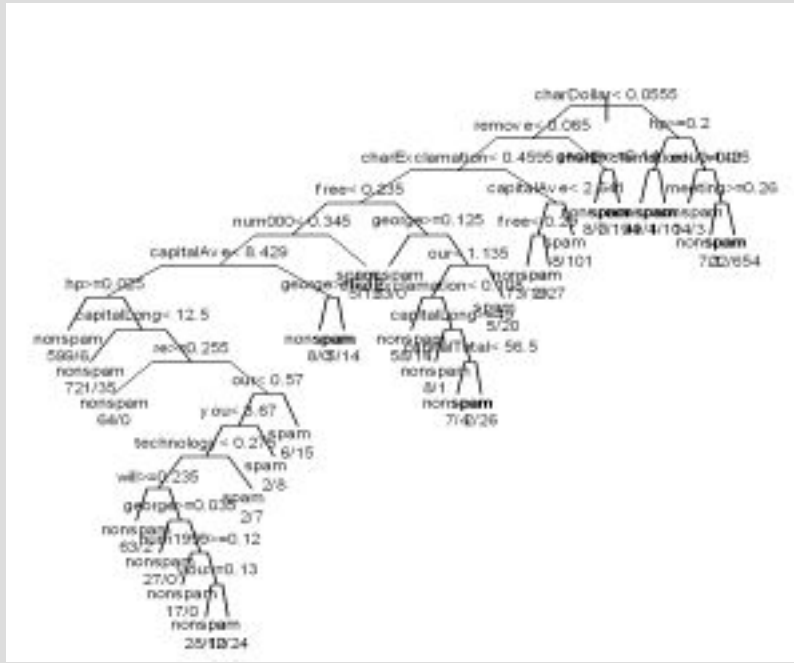
8. Démarche de construction d'un AD

Critères d'arrêt dans la construction d'un arbre

- Un arbre peut être construit complètement puis élagué (post-pruning) mais la construction peut être stoppée (pré-élagage)

- Objectif: éviter le phénomène de sur-apprentissage:

  Aboutir à des noeuds spécifiques à très peu d'individus, avec peu de signification statistique



8. Démarche de construction d'un AD

Critères d'arrêt dans la construction d'un arbre (pré-élagage)

- Atteinte d'un limite maximale de profondeur de l'arbre fixée
- Atteindre un nombre de feuilles maximal fixé
- Atteindre un effectif d'un noeud inférieur à une nombre d'instance fixé
- Atteindre une qualité suffisante de l'arbre (ex: taux d'erreur acceptable)
- Stabilisation de la qualité de l'arbre
- ...

9. Evaluation d'un modèle d'arbre de décision

Mesures de la qualité d'un arbre de décision

- Utiliser un échantillon de test comportant des instances déjà classées
- Pour une seule classe
    - **Précision** : Nombre d'instances bien classées / nombre d'instances attribuées à la classe
    - **Rappel** : nombre d'instances bien classées / nombre d'instances de la classe.
- Pour *n* classes (tout l'arbre)
    - **Précision**: moyenne de précisions
    - **Rappel**: moyenne des rappels

9. Evaluation d'un modèle d'arbre de décision

Mesures de la qualité d'un arbre de décision

- Exemple
  - Arbre

1	C1
2	C2
3	C1
4	C1
5	C1
6	C2
7	C2

Test

1	C1
2	C1
3	C1
4	C2
5	C2
6	C2
7	C2

Précision C1: 2/4
Rappel C1: 2/3

Précision C2: 2/3
Rappel C2: 2/4

Précision globale: 0,58
Rappel global: 0,58

9. Evaluation d'un modèle d'arbre de décision

Mesures de la qualité d'un arbre de décision

- Remarque: pour l'évaluation d'un ou de plusieurs arbres de décision, les données de test peuvent être
    - Séparées des données d'apprentissage si les données sont suffisantes
    - Font partie des données d'apprentissage; on parle alors de validation croisée
        - Découper l'échantillon d'apprentissage en dix avec neuf pour l'apprentissage et un seul pour le test (dix possibilités)
        - Calculer le taux d'erreur (proportion des instances mal classées) de chaque possibilité
        - En déduire le taux d'erreur moyen

9. Evaluation d'un modèle d'arbre de décision

Post-élagage

- S'effectue sur un arbre construit complètement et ayant une profondeur élevée
- On déduit l'ensemble des sous-arbres de l'arbre complet
- Selon la taille des données
    - Taille suffisante: on utilise des données de test séparées pour tester chaque sous-arbre et choisir le meilleur
    - Taille insuffisante: utiliser la validation croisée et choisir le meilleur sous-arbre

10. Avantages / inconvénients des AD

- **Avantages**
    - Règles explicites sous forme de comparaisons de valeurs
    - Facilité de programmation des arbres
    - Prise en compte des valeurs manquantes (ex: comme une modalité à part)
- **Inconvénients**
    - Nécessité de grand nombre d’instances pour éviter le surapprentissage
    - Dépendance entre les nœuds parents / fils