Apprentissage automatique supervisé

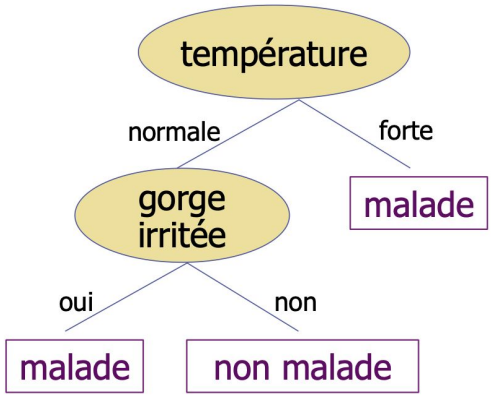
Amira Barhoumi

amira.barhoumi@univ-grenoble-alpes.fr

Année universitaire : 2025-2026

Arbre de décision

- Méthode d'arbre de décision
- Méthode d'apprentissage automatique supervisé
- Objectif : prédire la classe d'un nouvel exemple
- Principe : utiliser le corpus d'apprentissage (exemples déjà connus)
- Représentation graphique d'une procédure de décision sous forme d'arbre
- Représentation compréhensive sous forme de règles

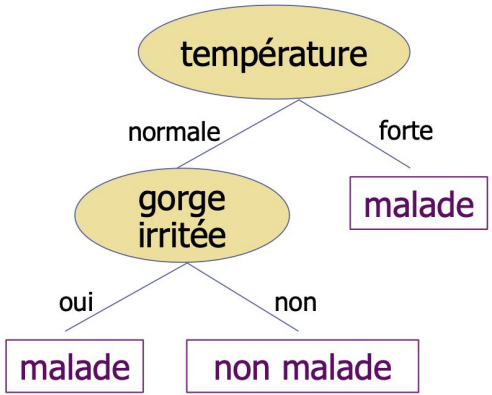


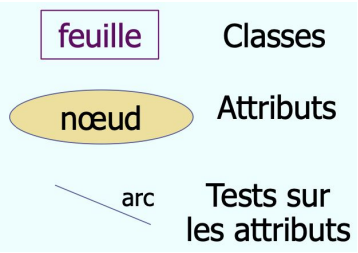
Arbre de décision

- Un arbre de décision est une représentation graphique d'une procédure de classification
  - Une feuille de l'arbre de décision est associée à toute description complète (branche)
  - Cette association est définie en commençant de la racine de l'arbre et en descendant dans l'arbre selon les réponses aux tests qui étiquettent les arcs des noeuds internes
  - La classe attribuée est alors la classe par défaut associée à la feuille qui correspond à la description
  - L'arbre de décision obtenu admet une traduction immédiate en terme de règles de décision

Arbre de décision

- Un arbre de décision est une représentation graphique d'une procédure de classification
  - Les noeuds internes sont appelés des noeuds de décision : appliquer pour tester chaque instance du corpus
  - Chaque test examine la valeur d'un unique attribut
  - Les réponses possibles au test correspondent aux labels des arcs issus du noeud

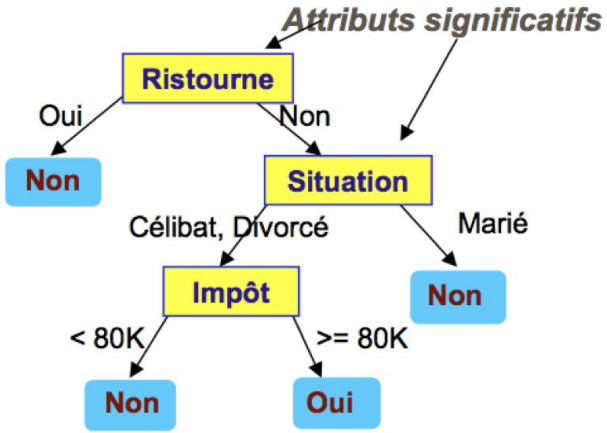




Arbre de décision

- Une règle est générée pour chaque branche de l'arbre (de la racine à une feuille)
- Les feuilles représentent la classe prédite
- Pour chaque noeud, on choisit l'attribut explicatif qui permet une meilleure séparation des individus

Id	Ristourne	Situation famille	Impôt revenu	Fraude
1	Oui	Célibat.	125K	Non
2	Non	Marié	100K	Non
3	Non	Célibat.	70K	Non
4	Oui	Marié	120K	Non
5	Non	Divorcé	95K	Oui
6	Non	Marié	60K	Non
7	Oui	Divorcé	220K	Non
8	Non	Célibat.	85K	Oui
9	Non	Marié	75K	Non
10	Non	Célibat.	90K	Oui

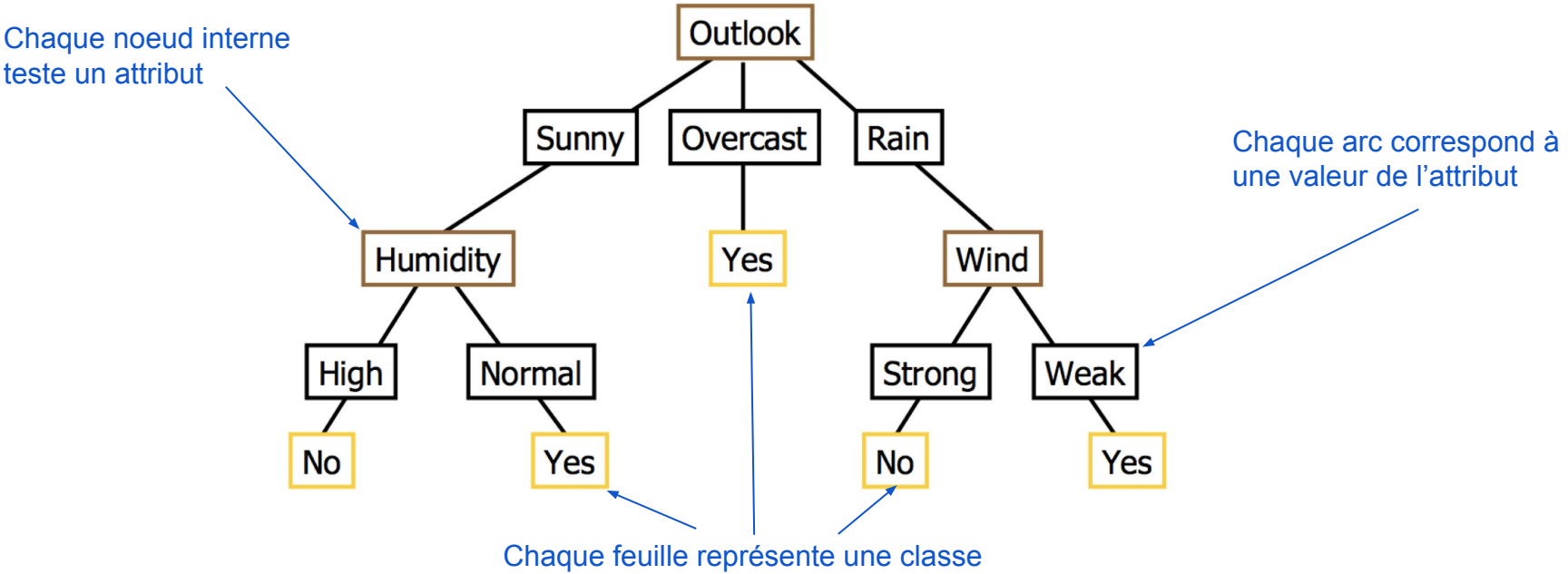


L'attribut significatif à un noeud est déterminé en fonction de l'indice de Gini

(Ristourne = Non et Situation = Divorcé et impôt = 100k) → Oui

Arbre de décision

- Représentation graphique



- Représentation sous forme de règles ?

Arbre de décision

- Division successive de l'ensemble de données en sous-groupes
- Il s'agit de sélectionner parmi les variables explicatives celle qui est la plus liée à la variable à expliquer. Cette variable fournit une première division de l'échantillon en plusieurs sous-ensembles appelés segments. Puis, on réitère cette procédure à l'intérieur de chaque segment en cherchant la deuxième meilleure variable explicative, et ainsi de suite…

  ⇒ C'est une classification descendante basée de sur la sélection de variables: chaque sous-groupe doit être le plus homogène possible vis à vis de la classe

- Génération de l'arbre de décision :
  - Construire de l'arbre (taille élevée)
  - Elaguer l'arbre (Pruning) : identifier et enlever les branches qui représentent du "bruit" afin d'améliorer le taux d'erreur

Arbre de décision

- Construction de l'arbre de décision :
  - 1- Au départ, toutes les instances d'apprentissage sont à la racine de l'arbre
  - 2- Sélectionner un attribut et choisir un test de séparation (*split test*) qui sépare le mieux les instances

    La sélection de l'attribut significatif est basée sur une heuristique
  - 3- Partitionner les instances entre les noeuds fils suivant la satisfaction des tests
  - 4- Traiter récursivement chaque noeud fils (itérer les étapes 1-, 2- et 3-)
  - 5- Répéter jusqu'à ce que tous les noeuds soient des terminaux.

    Un noeud est un terminal si :
    - il n'y a plus d'attributs disponibles
    - le noeud est pur, i.e. toutes les instances appartiennent à une seule classe
    - le noeud est presque pur, i.e. la majorité des instances appartiennent à une seule classe
    - nombre minimum d'instances par branche (Ex: C5 évite la croissance de l'arbre, k=2 par défaut)
  - 6- Étiqueter le noeud terminal par la classe majoritaire

Algorithme 3 CONSTRUIRE-ARBRE(D : ensemble de données)

1: Créer nœud N
2: **Si** tous les exemples de D sont de la même classe C alors
3: Retourner N comme une feuille étiquetée par C ;
4: **Si** la liste des attributs est vide alors
5: Retourner N Comme une feuille étiquetée de la classe de la majorité dans D ;
6: Sélectionner l'attribut A du meilleur Gain dans D ;
7: Etiqueter N par l'attribut sélectionné ;
8: Liste d'attributs ← Liste d'attributs - A ;
9: **Pour** chaque valeur V_i de A
10: Soit D_i l'ensemble d'exemples de D ayant la valeur de $A = V_i$;
11: Attacher à N le sous arbre généré par l'ensemble Di et la liste d'attributs
12: **FinPour ;**
13: **Fin ;**

Arbre de décision

- Comment choisir l'attribut qui sépare le mieux l'ensemble de données? Quelle variable de segmentation utilisée?
- Comment choisir les critères de séparation d'un ensemble selon l'attribut choisi? et comment ces critères varient en fonction du type de l'attribut (catégorique ou numérique)?
- Quel est le nombre optimal du nombre de critère qui minimise la taille de l'arbre et maximise la performance?
- Quels sont les critères d'arrêts de segmentation (partitionnement)?

Arbre de décision

- Construction de l'arbre de décision :
  - Construction récursive de manière "diviser pour régner"
  - Construction descendante
- Plusieurs variantes d'arbre : ID3, C4.5, CARD, CHAID
- Différence principale entre les variantes : la mesure de sélection d'un attribut (critère de branchement) à chaque noeud
  - choix de l'attribut (noeud)
  - choix des critères pour l'attribut (arcs)

- Toute nouvelle donnée peut être classée en testant ses valeurs d'attributs l'un après l'autre en commençant de la racine jusqu'à atteindre une feuille ⇒ une décision

Arbre de décision

- Mesure de sélection d'attributs :
  - Gain d'information (ID3, C4.5)
  - Indice Gini (CARD)
  - Table de contingence statistique (CHAID)

Arbre de décision ID3

- Choix de la variable de segmentation
  - Parmi les attributs, choisir l'attribut qui sépare le mieux les données de point de vue "classe"
  - Calculer pour chaque attribut une valeur appelée Gain qui dépend des différentes valeurs prises par cet attribut
  - Sélectionner l'attribut avec le plus grand gain (attribut significatif)

- Choix de la variable de segmentation

  Soit X une partition d'objets étiquetés de classe c $\in$ C

  - L'entropie H(X) mesure l'impureté de la partition X :

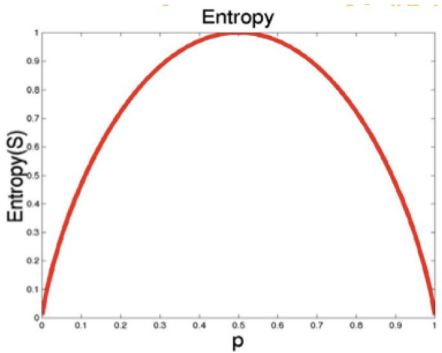
$$H(X) = -\sum_{c \in C} p_c \log_2(p_c) \quad ; \quad avec \quad 0 \leq H(X) \leq 1$$

Par exemple, soit X un ensemble d'objets classés en positif (+) et négatif (-)

$$H(X) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

Si p_+ =0 ou p_- = 0, alors H(X) = 0

Si p_+ = p_- = 0.5, alors H(X) = 1



Arbre de décision ID3

- Choix de la variable de segmentation

  Soit X une partition d'objets étiquetés de classe c $\in$ C

  - L'entropie H(X) mesure l'impureté de la partition X :

$$H(X) = -\sum_{c \in C} p_c \log_2(p_c) \quad ; \quad avec \quad 0 \leq H(X) \leq 1$$

  - Le Gain obtenu avec l'attribut a_j sur la population X :

$$Gain(X, a_j) = H(X) - \sum_{v \in valeurs(a_j)} \frac{|X_{a_j=v}|}{|X|} H(X_{a_j=v})$$

Où $X_{a_j=v}$, est l'ensemble des exemples dont l'attribut considéré a_j prend la valeur v, et la notation $|X|$ indique le cardinal de l'ensemble X.

Arbre de décision ID3

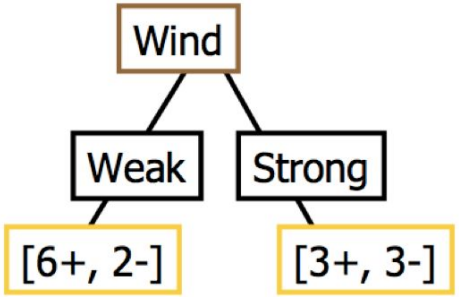
- Choix de la variable de segmentation (Exemple) :

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rain	mild	high	weak	yes
D5	rain	cool	normal	weak	yes
D6	rain	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rain	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rain	mild	high	strong	no

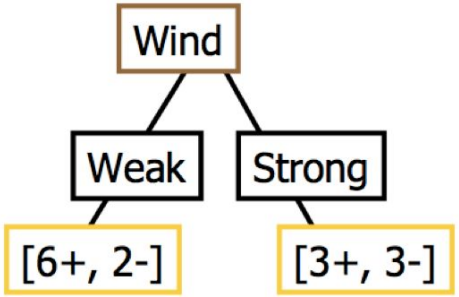
Construire l'arbre de décision ID3

Arbre de décision ID3

- Choix de la variable de segmentation (Exemple) :

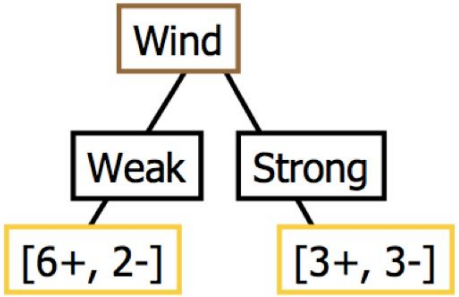


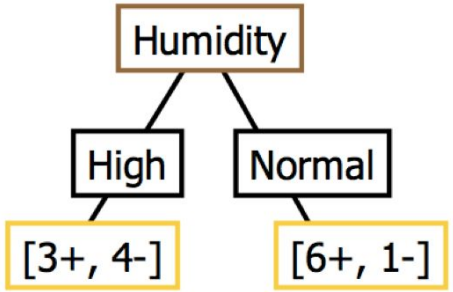
- Choix de la variable de segmentation (Exemple) :



Gain(S,Wind)
= 0.940-(8/14)*0.811
– (6/14)*1.0
= 0.048

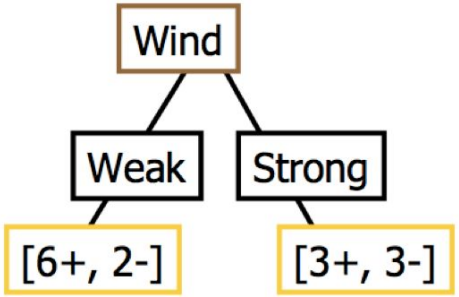
- Choix de la variable de segmentation (Exemple) :

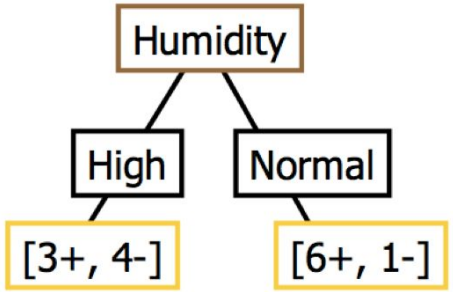




Gain(S,Wind)
= 0.940-(8/14)*0.811
– (6/14)*1.0
= 0.048

- Choix de la variable de segmentation (Exemple) :



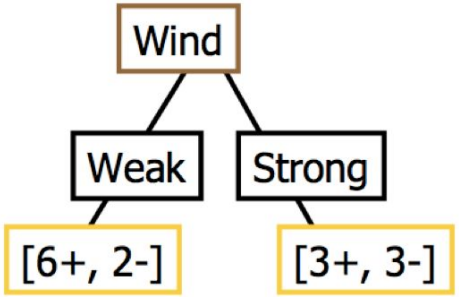


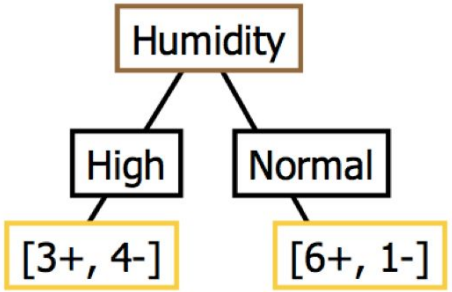
Gain(S,Wind)
= 0.940-(8/14)*0.811
– (6/14)*1.0
= 0.048

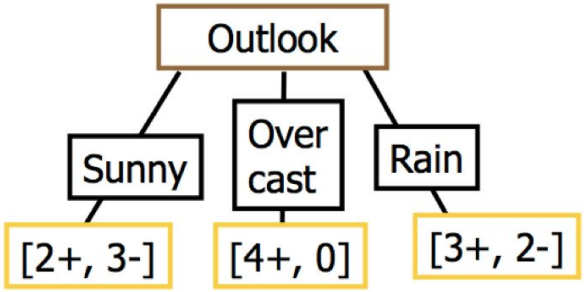
Gain(S,Humidity)
= 0.940-(7/14)*0.985
– (7/14)*0.592
= 0.151

Arbre de décision ID3

- Choix de la variable de segmentation (Exemple) :



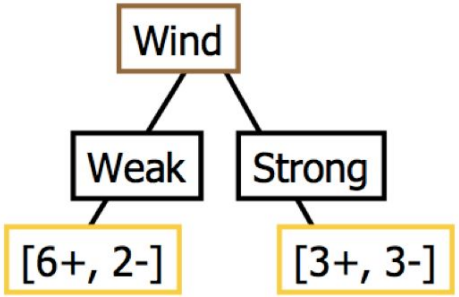


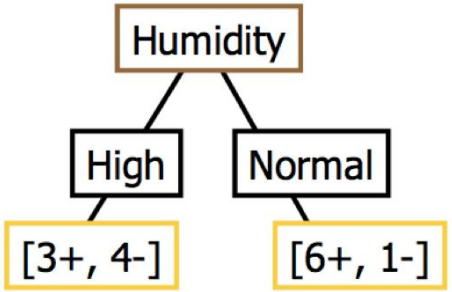


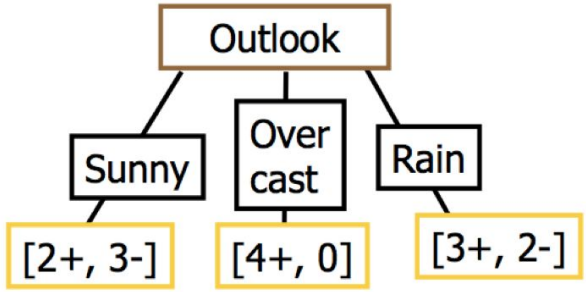
Gain(S,Wind)
= 0.940-(8/14)*0.811
– (6/14)*1.0
= 0.048

Gain(S,Humidity)
= 0.940-(7/14)*0.985
– (7/14)*0.592
= 0.151

- Choix de la variable de segmentation (Exemple) :







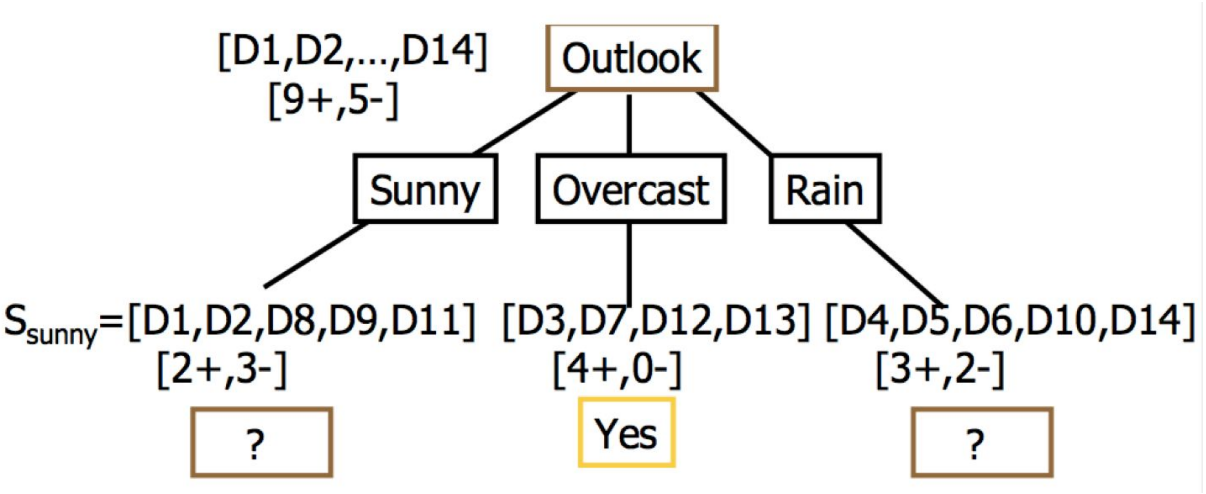
Gain(S,Wind)
= 0.940-(8/14)*0.811
– (6/14)*1.0
= 0.048

Gain(S,Humidity)
= 0.940-(7/14)*0.985
– (7/14)*0.592
= 0.151

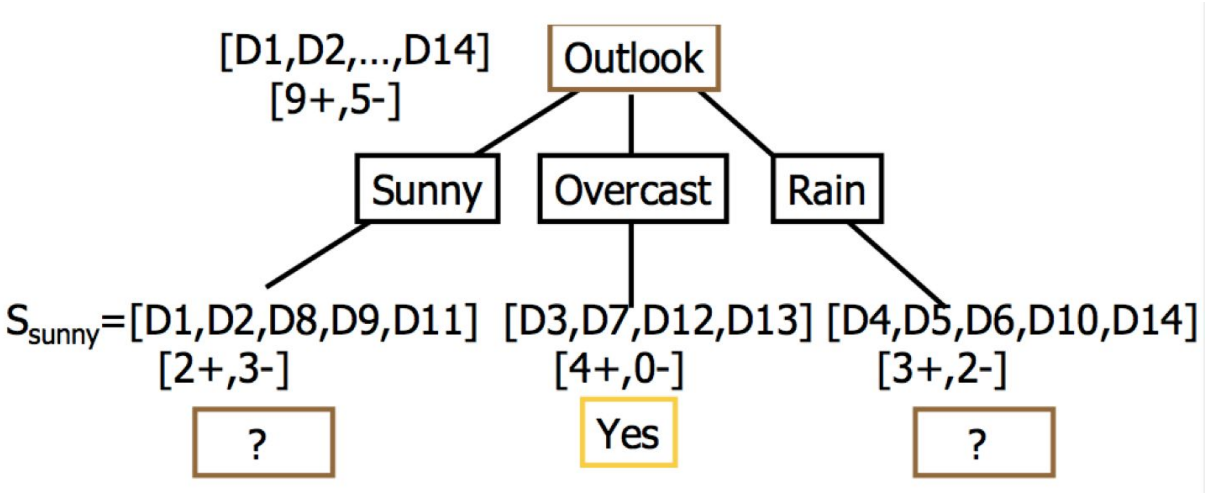
Gain(S,Outlook)
=0.940-(5/14)*0.971
-(4/14)*0.0 – (5/14)*0.0971
=0.247

Arbre de décision ID3

- Choix de la variable de segmentation (Exemple) :



- Choix de la variable de segmentation (Exemple) :



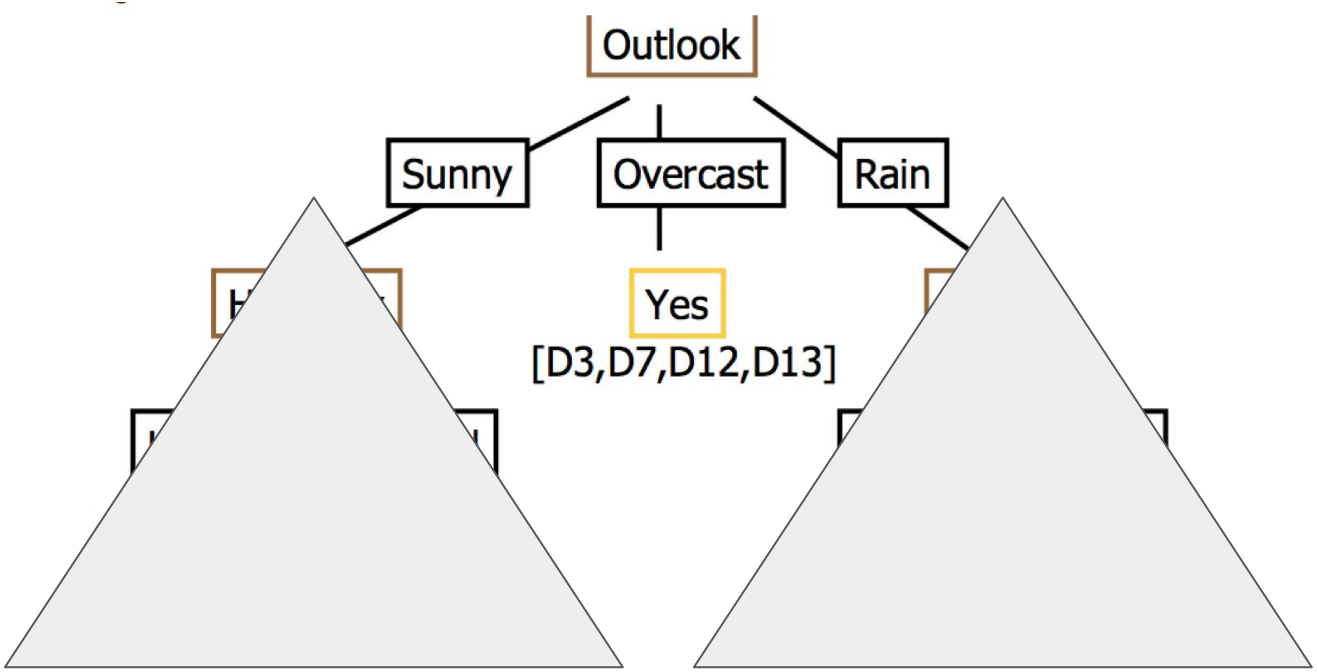
$Gain(S_{sunny}, Humidity) = 0.970 - (3/5)0.0 - 2/5(0.0) = 0.970$

$Gain(S_{sunny}, Temp.) = 0.970 - (2/5)0.0 - 2/5(1.0) - (1/5)0.0 = 0.570$

$Gain(S_{sunny}, Wind) = 0.970 = -(2/5)1.0 - 3/5(0.918) = 0.019$

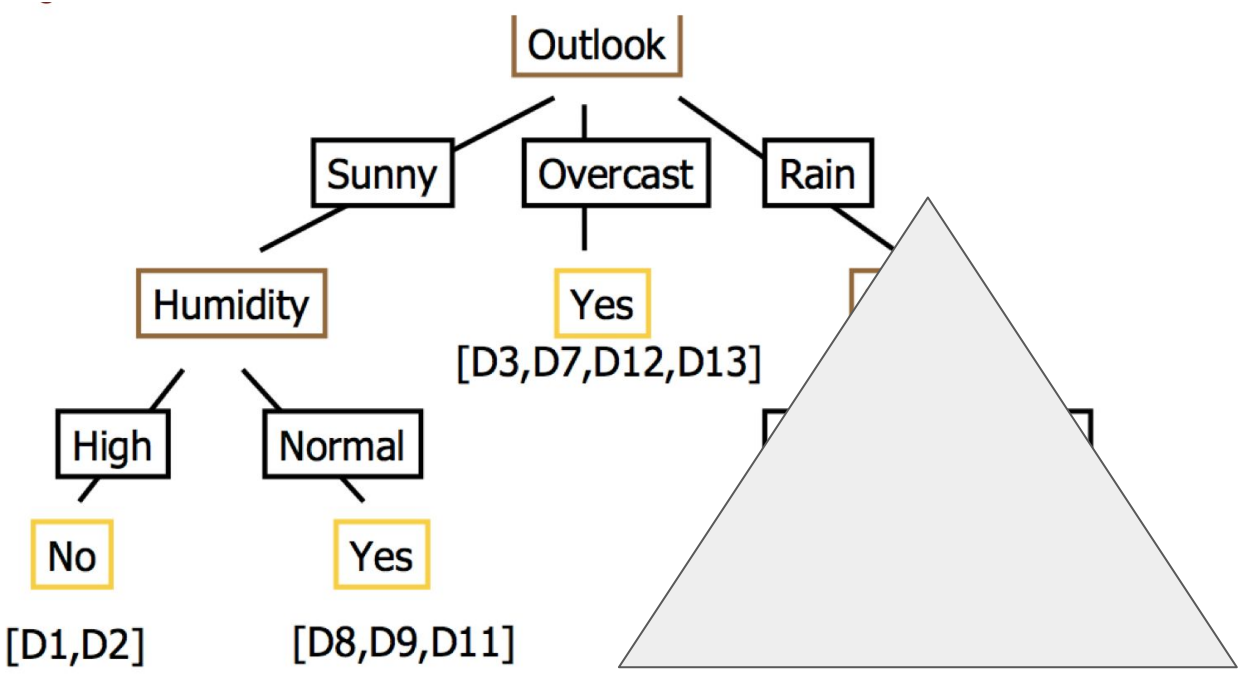
Arbre de décision ID3

- Choix de la variable de segmentation (Exemple) :



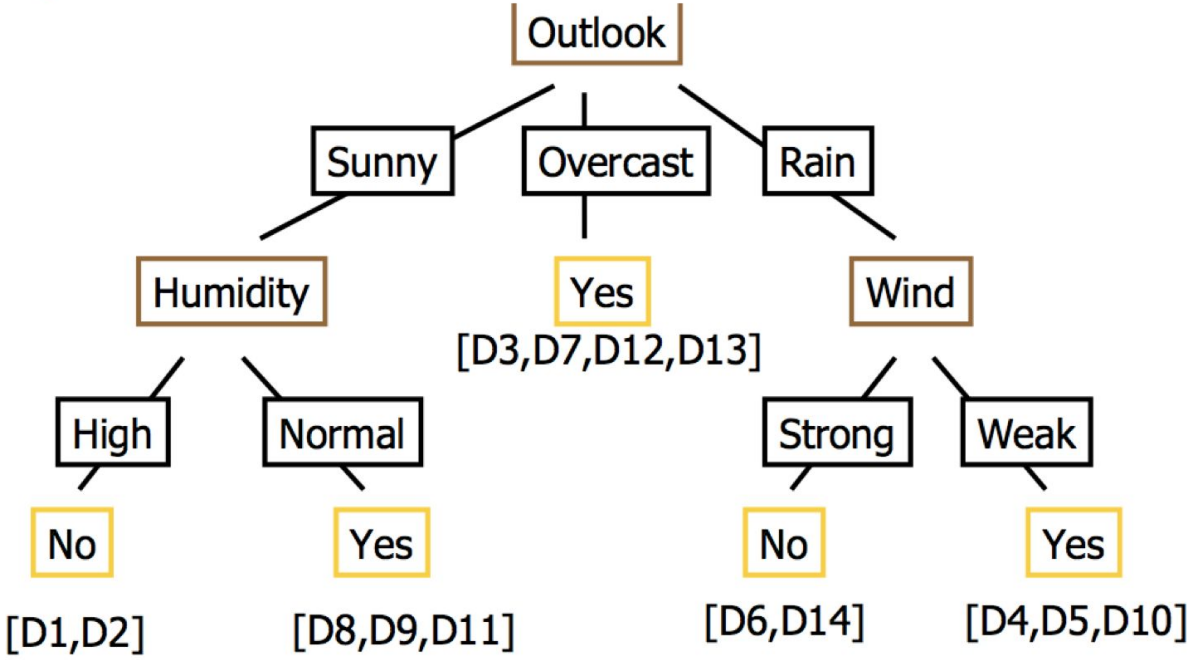
Arbre de décision ID3

- Choix de la variable de segmentation (Exemple) :



Arbre de décision ID3

- Choix de la variable de segmentation (Exemple) :



- Classer la nouvelle instance (sunny, cool, high, strong)

Arbre de décision ID3

- Exemple 2 : construire l'arbre de décision ID3

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Arbre de décision

- Elaguer l'arbre obtenu (pruning)
  - Supprimer les sous-arbres qui n'améliorent pas la classification => obtenir un arbre ayant un meilleur pouvoir de généralisation
  - Éviter le problème de sur-apprentissage (overfitting)
  - Raisons du sur-apprentissage :
    - peu de données d'apprentissage
    - bruits et exception dans les données

- Approches d'élagage :
  - Pre-élagage: arrêter de façon prématurée la construction d'arbre
  - Post-élagage:
    - supprimer des branches de l'arbre complet
    - convertir l'arbre en règles + manipuler les règles de façon indépendante (C4.5)

Arbre de décision

- Avantages :
  - Décision explicable
  - Classification rapide d'une nouvelle instance
  - Permet la sélection des attributs pertinents
- Inconvénients
  - Performances moins bonnes si beaucoup de classes