Caractérisation, Classification
Arbre de Décision

Marie Beurton-Aimar, Nicolas Parisey

beurton@labri.fr

November 17, 2020

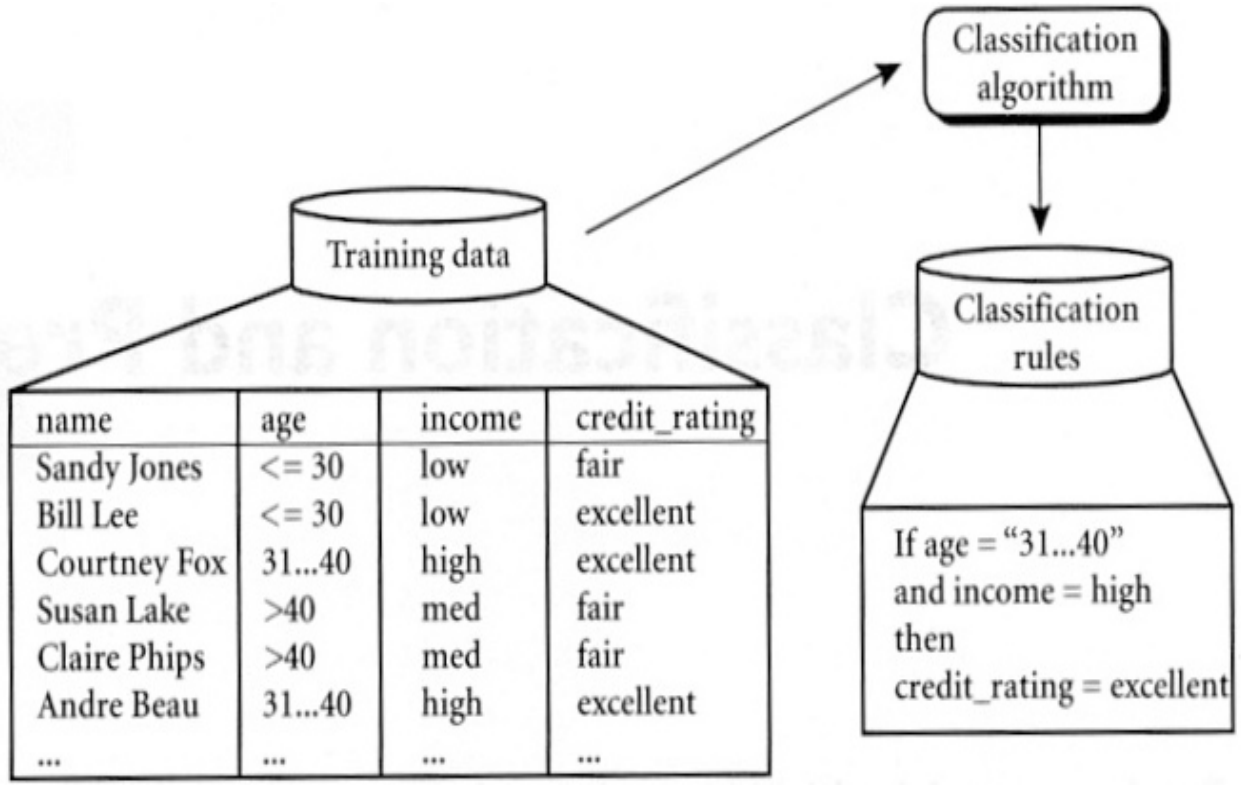
Classification vs. Prédiction

- Classification:
  - prédit la catégorie d'un objet
  - construit un modèle basé sur un jeu d'apprentissage et des valeurs (nom des catégories) et l'utilise pour classer des données nouvelles
- Prédiction:
  - modélise des données numériques pour prédire des valeurs inconnues ou manquantes
- Applications:
  - diagnostic médical
  - accord pour un crédit
  - marketing ciblé

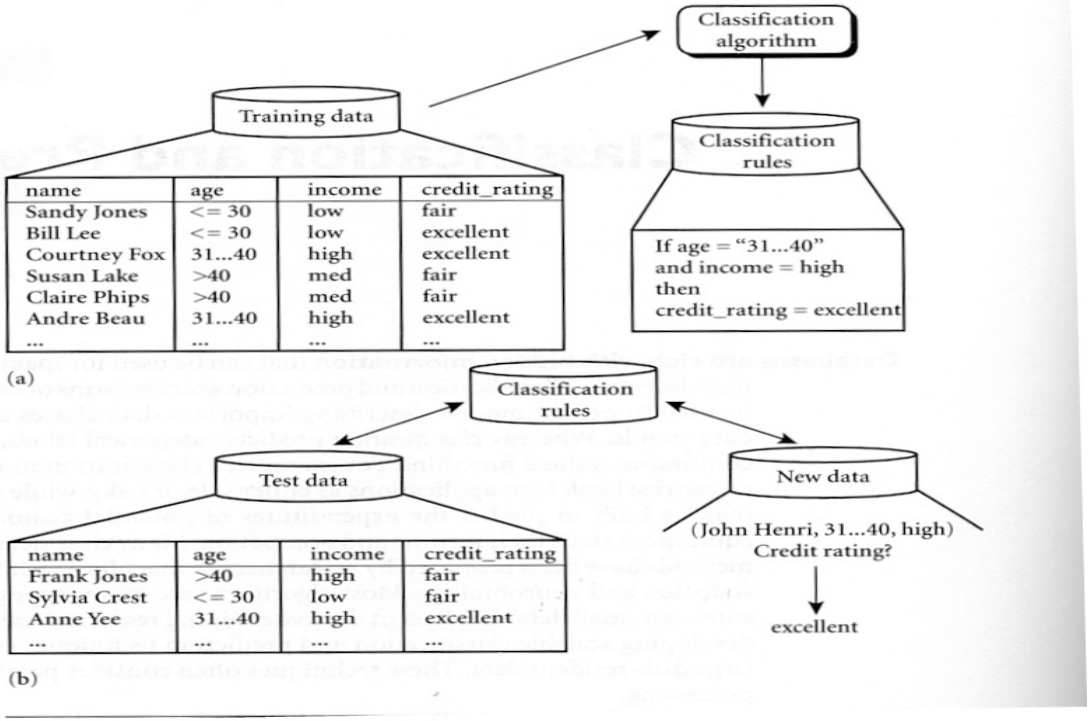
Classificateur: 2 étapes

- Construction du modèle
  - chaque objet appartient à une classe connue
  - jeu de données d'apprentissage: ensemble des objets utilisés pour la construction du modèle
- Utilisation du modèle pour classer des objets nouveaux ou inconnus
  - estimation de la précision du modèle
    - les classes connues du jeu d'apprentissage sont comparées à celles prédites
    - précision: pourcentage d'objets de jeu de test correctement classés
    - le jeu de test est indépendant du jeu d'apprentissage sinon risque de biais

Classification: construction du modèle



Classification: utilisation du modèle



name	age	income	credit_rating
Sandy Jones	<= 30	low	fair
Bill Lee	<= 30	low	excellent
Courtney Fox	31...40	high	excellent
Susan Lake	>40	med	fair
Claire Phips	>40	med	fair
Andre Beau	31...40	high	excellent
...	...	...	...

name	age	income	credit_rating
Frank Jones	>40	high	fair
Sylvia Crest	<= 30	low	fair
Anne Yee	31...40	high	excellent
...	...	...	...

Apprentissage supervisé ou non

- Apprentissage supervisé (classification)
  - supervision: le jeu de données d'apprentissage fournit les classes des objets
  - les nouveaux objets sont classés en fonction du jeu d'apprentissage
- Apprentissage non supervisé (clustering)
  - pas de classes définies
  - étant donné un ensemble de mesures, d'observations, etc, essayer d'établir l'existence de classes ou de clusters dans les données

Considérations pour la classification

Préparation des données pour la classification et la prédiction:

- Nettoyage des données
  - pré-traiter les données pour réduire le bruit et gérer les valeurs manquantes
- Analyse de pertinence
  - supprimer les attributs non pertinents ou redondants
- Transformation des données
  - généraliser ou normaliser les données

Comparaison des méthodes de classification:

- Précision de la prédiction
- Efficacité et mise à l'échelle
  - pour construire le modèle et pour l'utiliser
- Robustesse
  - tolérance au bruit et aux données manquantes
- Interprétabilité
  - compréhension des données via le modèle

Entropie et gain d'information

- S contient S_i tuples de classe C_i pour i$= \{1..m\}$
- L'information mesure la quantité d'information nécessaire pour classer un objet

$$I(S_1, S_2, \ldots, S_m) = -\sum_{i=1}^{m} \frac{S_i}{S} \log_2 \frac{S_i}{S}$$

- Entropie d'un attribut A ayant pour valeurs $\{a_1, \ldots, a_v\}$

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + \ldots + S_{mj}}{S_j} I(S_{1j}, S_{2j}, \ldots, S_{mj})$$

- Gain d'information en utilisant l'attribut A

$$Gain(A) = I(S_1, S_2, \ldots, S_m) - E(A)$$

Analyse de pertinence pour l'attribut pour la description des concepts

1. Collecte des données (classe cible et classe contrastante)
2. Analyse de la pertinence préliminaire en utilisant l'induction orientée attribut (obtention de la relation candidate)
3. Suppression des attributs peu ou pas pertinents
4. Génération de la description du concept en utilisant l'induction orientée attribut

Exemple: caratérisation analytique

- Tâche:
  - fouiller les caractéristiques générales qui décrivent les étudiants de 3ème cycle en faisant une caractérisation analytique
- Etant donné:
  - attributs: *name*, *gender*, *major*, *birth_place*, *birth_date*, *phone#* et *gpa*

Exemple: caractérisation analytique

- Données:
  - classe cible: étudiants 3ème cycle
  - classe contrastante: étudiants 1er et 2ème cycle
- Généralisation analytique
  - suppression d'attributs
    - supprime *name et phone#*
  - généralisation d'attributs
    - généralise *major, birth_place, birth_date et gpa*
    - *accumule les effectifs*
  - relation candidate: *gender, major, birth_country, age_range et gpa*

Exemple: caractérisation analytique

Relation initiale:

name	*gender*	*major*	*birth_place*	*birth_date*	*residence*	*phone#*	*gpa*
Jim Woodman	M	CS	Vancouver, BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...

Exemple: caractérisation analytique

Relation candidate pour la classe cible: 3ème cycle ($\sum = 120$):

gender	major	birth_country	age_range	gpa	count
M	Science	Canada	21-25	Very_good	16
F	Science	Foreign	25-30	Excellent	22
M	Engineering	Foreign	25-30	Excellent	18
F	Science	Foreign	25-30	Excellent	25
M	Science	Canada	21-25	Excellent	21
F	Engineering	Canada	21-25	Excellent	18

Relation candidate pour la classe contrastante: 1er et 2ème cycles ($\sum = 130$):

gender	major	birth_country	age_range	gpa	count
M	Science	Foreigna	<20	Very_good	18
F	Business	Canada	<20	Fair	20
M	Business	Canada	<20	Fair	22
F	Science	Canada	20-25	Fair	24
M	Engineering	Foreign	20-25	Very_good	22
F	Engineering	Canada	<20	Excellent	24

Exemple: caractérisation analytique

Analyse de pertinence:

- Calcul de l'information nécessaire pour classer un objet

$$I(s_1, s_2) = I(120, 130) = -\frac{120}{250}\log_2\frac{120}{250} - \frac{130}{250}\log_2\frac{130}{250} = 0.9988$$

- Calcul de l'entropie pour chaque attribut: ex: *major*

  major = Science : $S_{11} = 84$ $S_{21} = 42$ $I(S_{11}, S_{21}) = 0.9183$
  major = Engineering : $S_{12} = 36$ $S_{22} = 46$ $I(S_{12}, S_{22}) = 0.9892$
  major = Business : $S_{13} = 0$ $S_{23} = 42$ $I(S_{13}, S_{23}) = 0$

  (nb: S_{1i}: nbr de 3ème cycle, S_{2i}: nbr de 1er et 2ème cycle)

Exemple: caractérisation analytique

- Calcul de l'information nécessaire pour classer un objet si S est partionné selon l'attribut

$$E(major) = \frac{126}{250} I(s_{11}, s_{21}) + \frac{82}{250} I(s_{12}, s_{22}) + \frac{42}{250} I(s_{13}, s_{23}) = 0.7873$$

- Calcul du gain d'information pour chaque attribut

$$Gain(major) = I(s_1, s_2) - E(major) = 0.2115$$

  - gain pour chaque attribut

Gain(gender)	=0.0003
Gain(birth_country)	=0.0407
Gain(major)	=0.2115
Gain(gpa)	=0.4490
Gain(age_range)	=0.5971

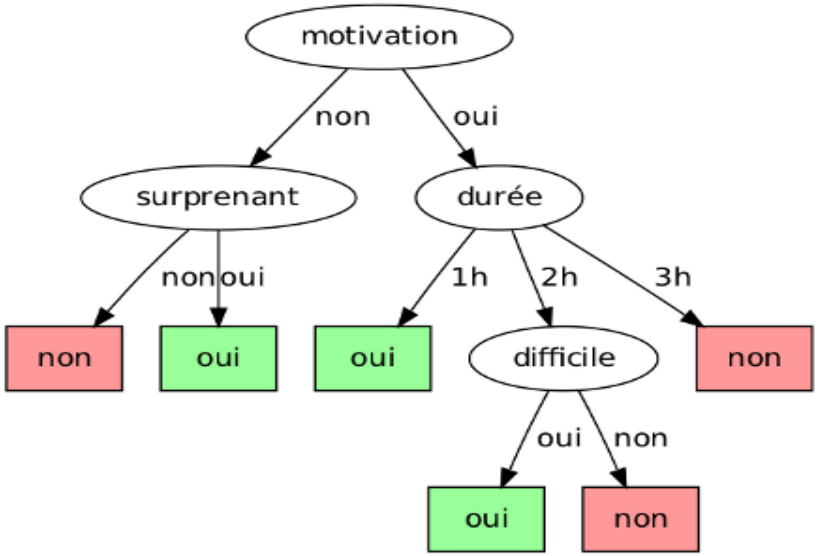
Exemple: caractérisation analytique

- Dérivation de RelationInitiale
  - seuil de pertinence R = 0.1
  - supprimer les attributs peu ou pas pertinents de la relation candidate: supprime *gender et birth_country*

major	*age_range*	*gpa*	*count*
Science	*20-25*	*Very_good*	*16*
Science	*25-30*	*Excellent*	*47*
Science	*20-25*	*Excellent*	*21*
Engineering	*20-25*	*Excellent*	*18*
Engineering	*25-30*	*Excellent*	*18*

Classification par arbre de décision

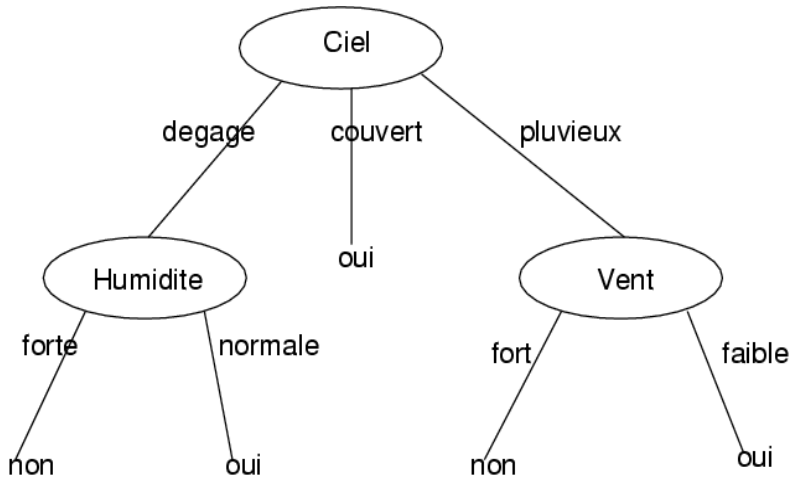
- Modélise une hiérarchie de tests sur les valeurs d'un ensemble de variables/attributs.
- Produit une valeur numérique - régression - ou choisit un élément dans un ensemble discret de conclusions - classification.



Classification par arbre de décision

Arbre de décision:

- nœuds internes: test sur un attribut
- branches: résultat d'un test / valeur de l'attribut
- feuilles: classe



Classification par arbre de décision

- Génération de l'arbre en 2 étapes:
  1. **Construction**
     - au départ, tous les exemples du jeu d'apprentissage sont à la racine
     - on partionne récursivement les exemples en sélectionnant des attributs
  2. **Elagage**
     - identification et suppression des branches correspondant à des exceptions ou du bruit
- Utilisation de l'arbre
  - tester les valeurs des attributs avec l'arbre de décision

Approche basée sur la théorie de l'information

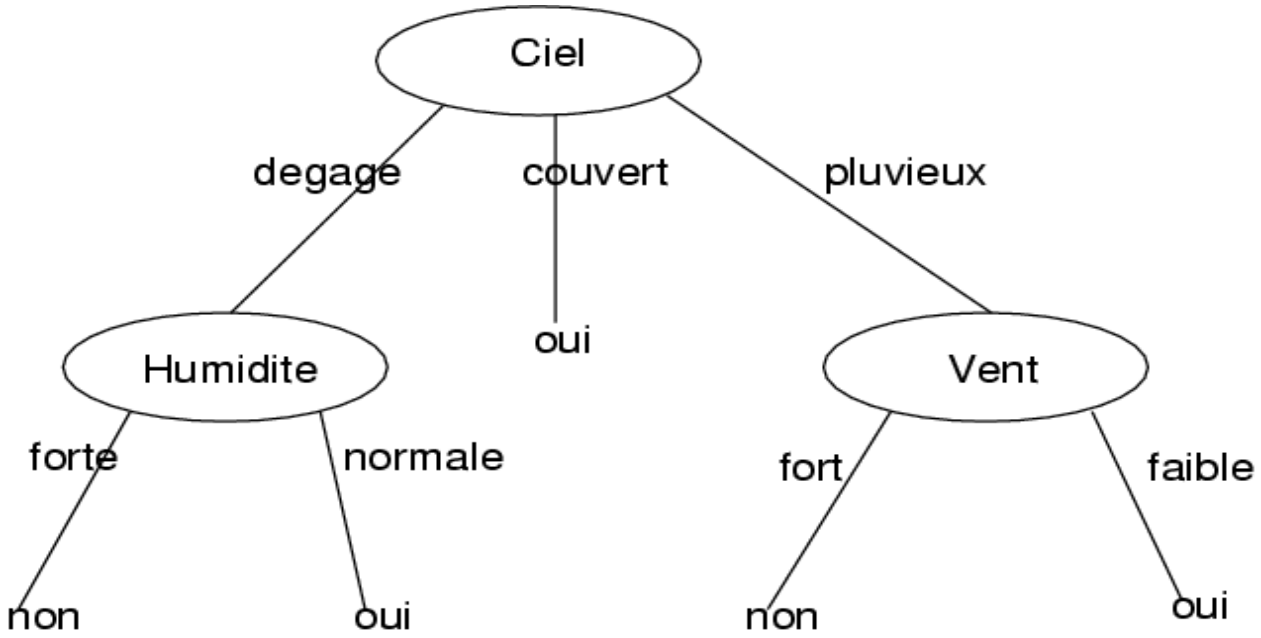
- Arbre de décision
  - les nœuds internes testent un attribut
  - les branches correspondent à une valeur d'attribut
  - les feuilles attribuent une classe
- Algorithme ID3
  - construit un arbre de décision basé sur un jeu de données d'apprentissage (les classes des objets sont connues)
  - ordonne les attributs selon la mesure de gain d'information
  - taille minimale
    - le moins de test possibles pour classer un objet

Exemple

ciel	**température** (^oF)	**humidité**	**vent**	**joue?**
Dégagé	85	85	non	non
Dégagé	80	90	oui	non
Couvert	83	78	non	oui
Pluvieux	70	96	non	oui
Pluvieux	68	80	non	oui
Pluvieux	65	70	oui	non
Couvert	64	65	oui	oui
Dégagé	72	95	non	non
Dégagé	69	70	non	oui
Pluvieux	75	80	non	oui
Dégagé	75	70	oui	oui
Couvert	72	90	oui	oui
Couvert	81	75	non	oui
Pluvieux	71	90	oui	non

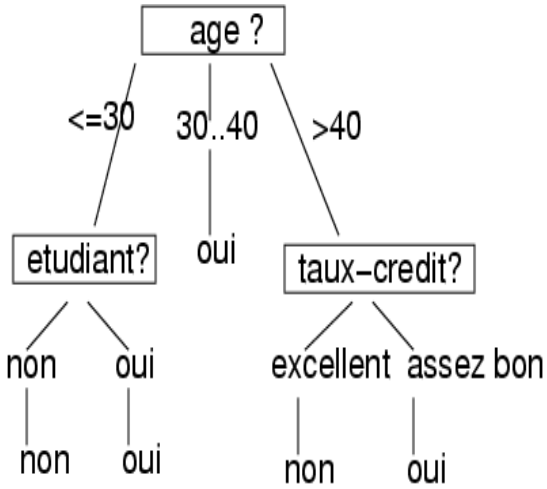
Induction descendante par arbre de décision

Attributs:{ Ciel, Température, Humidité, Vent }
Classes: joue au tennis { oui, non }



Exemple: achète un ordinateur

âge	revenus	étudiant	taux_crédit
≤ 30	élevé	non	assez bon
≤ 30	élevé	non	excellent
31..40	élevé	non	assez bon
> 40	moyen	non	assez bon
> 40	faible	oui	assez bon
> 40	faible	oui	excellent
31..40	faible	oui	excellent
≤ 30	moyen	non	assez bon
≤ 30	faible	oui	assez bon
> 40	moyen	oui	excellent
≤ 30	moyen	oui	excellent
31..40	moyen	non	excellent
31..40	élevé	oui	assez bon
> 40	moyen	non	excellent



Algorithme pour l'induction d'arbre de décision

- Algorithme glouton
  - approche descendante récursive *diviser pour régner*
  - au départ, tous les objets sont à la racine
  - attributs catégoriels (les valeurs continues sont discrétisées à l'avance)
  - les exemples sont partitionnés récursivement par la sélection d'attribut
  - les attributs sont sélectionnés sur la base d'une heuristique ou d'une mesure statistique
- Conditions d'arrêt
  - tous les exemples pour un nœud appartiennent à la même classe
  - plus d'attribut pour partionner; dans ce cas la classe attribuée correspond à celle la plus représentée
  - plus d'exemple à classer

Mesure pour la sélection d'attribut
Exemple: gain d'information (ID3 et c4.5)

- Sélectionne l'attribut qui a le gain le plus élevé
- Soient 2 classes `P` et `N`
  - soit un jeu d'apprentissage `S` qui contient `p` objets de classe `P` et `n` objets de classe `N`
  - La quantité d'information nécessaire pour décider si un objet de `S` appartient à `P` ou `N` se définie par

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

- Les valeurs de `A` partitionnent `S` en $\{S_1, \ldots, S_v\}$
  - Si S_i contient p_i exemples de `P` et n_i exemples de `N`, l'entropie (l'information attendue) nécessaire pour classer les objets dans les sous-arbres S_i est

$$E(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I(p_i, n_i)$$

- Le gain d'information de l'attribut `A` est

$$Gain(A) = I(p,n) - E(A)$$

Exemple: achète un ordinateur?

- Classe P: achète un ordinateur = oui
- Classe N: achète un ordinateur = non
- $I(p, n) = I(9, 5) = 0.940$

âge	p_i	n_i	$I(p_i, n_i)$
≤ 30	2	3	0.971
30..40	4	0	0
> 40	3	2	0.971

- Calcul de l'entropie:
  $E(age) = \frac{5}{14} I(2, 3) + \frac{4}{14} I(4, 0) + \frac{5}{14} I(3, 2) = 0.69$
- Gain d'information:
  $\text{Gain}(\text{âge}) = I(p, n) - E(age)$
- De même:
  $\text{Gain}(\text{revenus}) = 0.029$
  $\text{Gain}(\text{étudiant}) = 0.151$
  $\text{Gain}(\text{taux\_credit}) = 0.048$

Eviter de trop modéliser le jeu d'apprentissage (overfitting)

- L'arbre généré risque de trop refléter le jeu d'apprentissage
  - trop de branches, certaines peuvent représenter des anomalies
  - précision faible pour des données nouvelles
- 2 approches:
  - pré-élagage: arrêter la construction de l'arbre tôt = ne pas partitioner un nœud si la mesure de qualité dépasse un seuil
    - difficulté de fixer le seuil
  - post-élagage: supprimer des branches d'un arbre totalement construit = obtenir une séquence d'arbres prgressivement élagués
    - décider un jeu de données différentes pour décider du meilleur arbre élagué

Améliorations de l'algorithme

- Attributs définis sur des valeurs continues
  - définir dynamiquement les valeurs pour partitionner les données
- Tolérance aux données manquantes
  - attribuer la valeur la plus fréquente
  - attribuer une probabilité pour chaque valeur possible

Classification bayésienne

- Apprentissage probabiliste: calcule explicitement les probabilités des hypothèses, une des approches les plus pragmatiques pour certains types d'apprentissage
- Incrémental: chaque exemple met à jour la probabilité qu'une hypothèse est correcte. Des connaissances *a priori peuvent être combinées avec des données d'observation*
- *Prédiction probabiliste: prédit plusieurs hypothèses, pondérées par leur probabilité*

Classification bayésienne

- Le problème de classification peut être formulé en utilisant les probabilités *a posteriori*:
  - $P(H|X)$ = probabilité que l'objet X $= < x_1, \ldots, x_k >$ soit de la classe C
- Ex: $P(non|ciel)$ = (ciel=dégagé, vent=fort, ...)
- Idée: attribuer à l'objet X la classe qui maximise $P(H|X)$

Estimation des probabilités *a posteriori*

- Théorème de Bayes:

$$P(C|X) = P(X|C) \cdot \frac{P(C)}{P(X)}$$

- $P(X)$ est constant pour toutes les classes
- $P(C)$ = fréquence de la classe C

Classificateur bayésien naïf

- Hypothèse naïve: indépendance des attributs

$$P(x_1, \ldots, x_k|C) = P(x_1|C) \cdot \ldots \cdot P(x_k|C)$$

- si le i-ème attribut est catégoriel: $P(x_i|C)$ est estimée comme la fréquence des échantillons qui ont pour valeur x_i et qui sont de classe C
- si le i-ème attribut est continue: $P(x_i|C)$ est estimée avec une gaussienne
- calcul facile

Exemple

Ciel	Tempér.	Humidité	Vent	Classe
ensoleillé	chaud	élevé	faux	N
ensoleillé	chaud	élevé	vrai	N
couvert	chaud	élevé	faux	P
pluvieux	moyen	élevé	faux	P
pluvieux	frais	normal	faux	P
pluvieux	frais	normal	vrai	N
couvert	frais	normal	vrai	P
ensoleillé	moyen	élevé	faux	N
ensoleillé	frais	normal	faux	P
pluvieux	moyen	normal	faux	P
ensoleillé	moyen	normal	vrai	P
couvert	moyen	élevé	vrai	P
couvert	chaud	normal	faux	P
pluvieux	moyen	élevé	vrai	N

$P(p) = 9 / 14$
$P(n) = 5 / 14$

Ciel	
$P(\text{soleil} \mid p) = 2/9$	$P(\text{soleil} \mid n) = 3/5$
$P(\text{couvert} \mid p) = 4/9$	$P(\text{pluvieux} \mid n) = 0$
$P(\text{pluvieux} \mid p) = 3/9$	$P(\text{pluvieux} \mid n) = 2/5$
Température	
$P(\text{chaud} \mid p) = 2/9$	$P(\text{chaud} \mid n) = 2/5$
$P(\text{moyen} \mid p) = 4/9$	$P(\text{moyen} \mid n) = 2/5$
$P(\text{frais} \mid p) = 3/9$	$P(\text{frais} \mid n) = 1/5$
Humidité	
$P(\text{élevé} \mid p) = 3/9$	$P(\text{élevé} \mid n) = 4/5$
$P(\text{normal} \mid p) = 6/9$	$P(\text{normal} \mid n) = 2/5$
Vent	
$P(\text{vrai} \mid p) = 3/9$	$P(\text{vrai} \mid n) = 3/5$
$P(\text{faux} \mid p) = 6/9$	$P(\text{vrai} \mid n) = 2/5$

Exemple

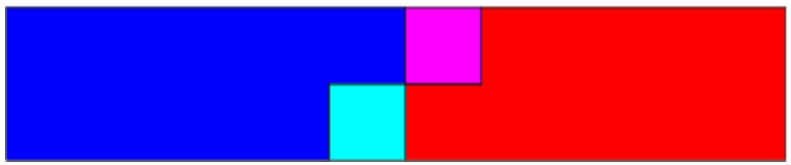
- Un objet X = < pluie, chaud, élevé, faux >
- $P(X|p) =$
  $P(pluie|p) \cdot P(chaud|p) \cdot P(eleve|p) \cdot P(faux|p)$
  $= \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} = 0.01646$
- $P(X|n) =$
  $P(pluie|n) \cdot P(chaud|n) \cdot P(eleve|n) \cdot P(faux|n) =$
  $\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{2}{5} = 0.0512$
- X est classé comme `n`

L'hypothèse d'indépendance ...

- rend le calcul possible
- mène à des classificateurs optimaux si elle est vérifiée
- mais c'est rarement le cas, car les attributs sont souvent corrélés
- tentative de pallier cette difficulté:
  - réseaux bayésiens: combinent le raisonnement bayésien avec la relation causale entre les attributs
  - arbres de décision: considère un seul attribut à la fois, en commençant par le plus important

Précision du classificateur

- Positif (P = **O**+**O**), Négatif (N= **O**+**O**),
  Prédit positif (PP= **O**+**O**), Prédit Négatif (PN=**O**+**O**),
  Vrai Positif (VP= **O**), Faux Positif (FP= **O**),
  Vrai Négatif (VN= **O**), Faux Négatif (FN= **O**)
- Sensibilité = VP/P
- Spécificité = VN/N
- Précision = VP/(VP+FP) = VP/PP
- Exactitude = sensibilité.P/(P+N) + spécificitée.N/(P+N)
  = (VP+VN)/(P+N)



Estimation du taux d'erreurs

- Partitionnement:
  - utilisation de jeux indépendants: apprentissage (2/3), test (1/3)
- Validation croisée:
  - diviser les données en k partitions
  - utilsier k-1 partitions pour l'apprentissage et la dernière pour le test
  - précision = nbr d'objets bien classés lors des k itérations / nbr d'objets
- Bootstrapping
  - tirage aléatoire avec remise des objets constituant le jeu d'apprentissage
  - validation croisée