Fouille de données

▷ **Arbres de décision**

Philippe Lenca et Romain Billot

prenom.nom@imt-atlantique.fr

Telecom Bretagne

2016-2017

Plan

Plan

ECD - Rappels

Fouille de données (Fayyad & al.)

Processus complexe permettant l'identification, au sein des données, de motifs valides, nouveaux, potentiellement intéressants et les plus compréhensibles possible.

Questions

- maximiser le nombre de données concernées ?
- minimiser le nombre de contre-exemples ?
- garantir une fiabilité du modèle ?
- lisibilité des résultats ?

ECD - Rappels

Diverses problématiques

- supervisé *vs.* non-supervisé
- variables quantitatives *vs.* qualitatives
- données volumineuses *vs.* restreintes
- données structurées *vs.* non structurées
- ...

↪ Nécessité de disposer de différentes stratégies

ECD - Rappels

Quelques méthodes d'analyse

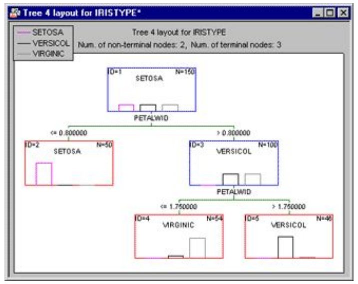
- analyse factorielle
- analyse discriminante à base de noyaux
- réseaux neuronaux
- arbres de décision
- cartes de Kohonen
- règles d'association
- ...

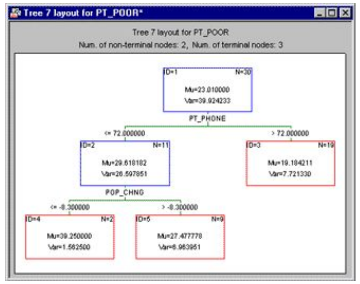
Arbres de décision

Quelques caractéristiques

- apprentissage supervisé
- tous types de variables explicatives
- variable cible
  - qualitative (arbres de segmentation)
  - quantitative (arbres de régression)
- modèles interprétables

Arbres de décision

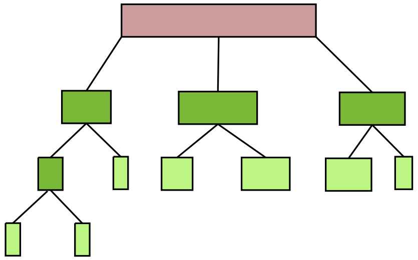




http://www.statsoft.com/

Qu'est-ce qu'un arbre ?

- une racine
- des branches
- des nœuds intermédiares
- des nœuds terminaux (des feuilles)



Arbres de classement

Prédire un attribut particulier `Y` (la classe, l'attribut cible)

- en fonction d'une liste $(\mathtt{X}_1, \ldots, \mathtt{X}_m)$ d'attributs prédictifs

↪ Présupposé : données semi-structurées (tableau croisé individus×caractéristiques).

	$\mathtt{X}_1$	...	$\mathtt{X}_m$	$\mathtt{Y}$
i_1	$\mathtt{X}_{11}$		$\mathtt{X}_{1m}$	$\mathtt{Y}_1$
i_2	$\mathtt{X}_{21}$		$\mathtt{X}_{2m}$	$\mathtt{Y}_2$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
i_n	$\mathtt{X}_{n1}$		$\mathtt{X}_{nm}$	$\mathtt{Y}_n$

Trouver `f` reliant `Y` aux $\mathtt{X}_i$

$$\mathtt{Y}_i = \mathtt{f}(\mathtt{X}_{1i}, \ldots, \mathtt{X}_{mi})$$

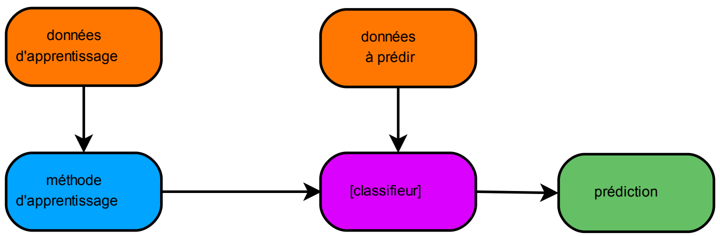
- minimiser le nombre de $\mathtt{X}_i$ utilisés ?
- qualité de `f` (couverture, erreurs) ?
- etc.

Apprentissage supervisé

Classifieur

- résume ce que l'on sait sur les données d'apprentissage

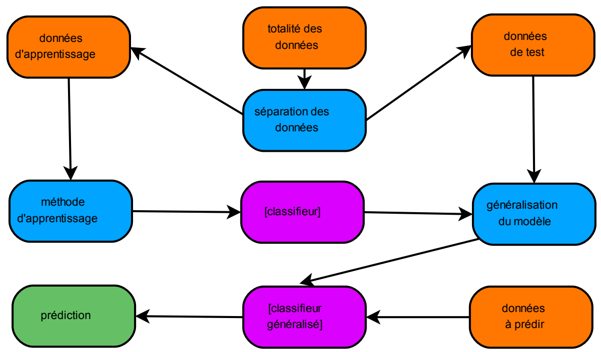
↪ Doit classer correctement les nouveaux individus.



Apprentissage supervisé

Améliorer la robustesse du classifieur

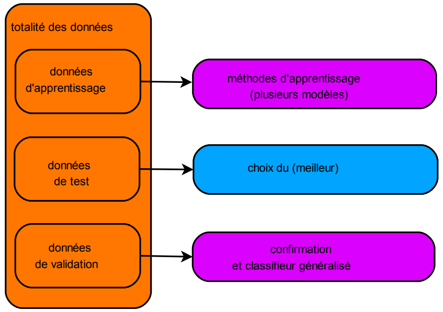
- ensemble d'apprentissage
- ensemble de test



Apprentissage supervisé

Améliorer la robustesse du classifieur

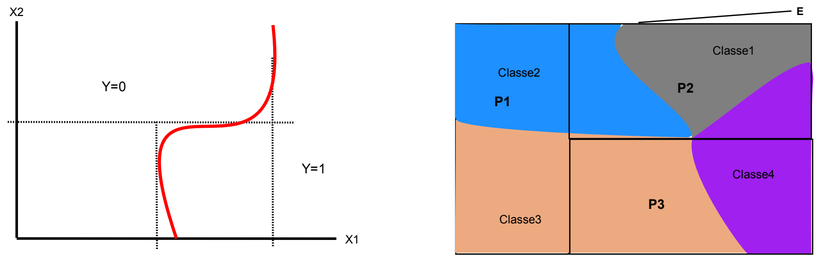
- ensemble d'apprentissage
- ensemble de validation
- ensemble de test



Apprentissage supervisé

Classification supervisée ou discriminante (variable cible qualitative)

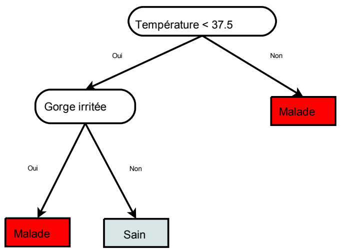
- règles de production
- segmentation de l'espace selon des coupes orthogonales



$\hookrightarrow$ **Si** `condition` **Alors** `conclusion`.

Exemple

Arbre :



Règles de production :

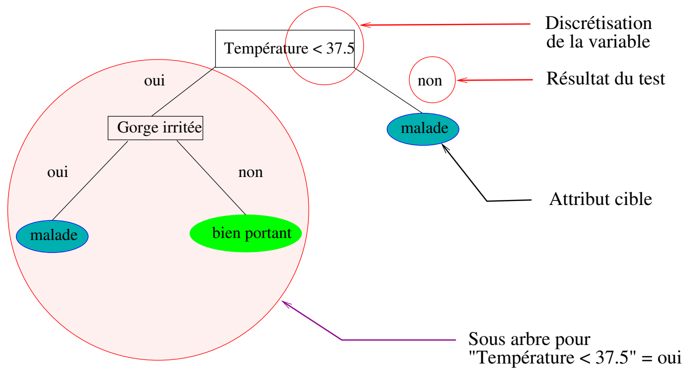
Si `température < 37.5 et Gorge irritée = oui` **alors** `malade`

Si `température < 37.5 et Gorge irritée = non alors sain`

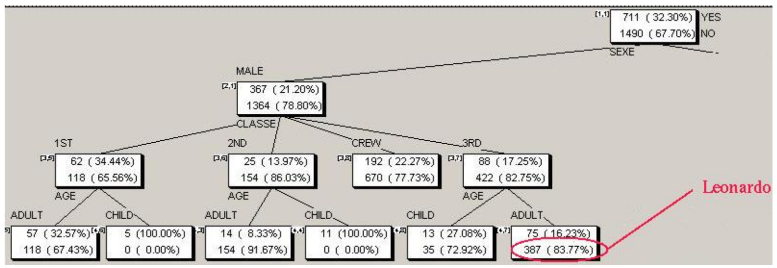
Si `température` $\geq$ `37.5 alors malade`

Liste de décision : on ordonne les règles.

Exemple



Exemple - Titanic



Arbre de décision

Avantages

- bon pouvoir prédictif
- intelligible (si l'arbre n'est pas trop complexe)

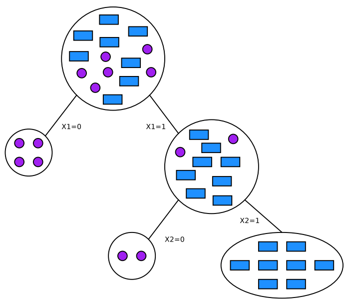
Inconvénients

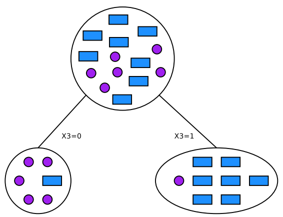
- sélection d'un seul attribut à chaque nœud
- éventuelle explosion combinatoire

Arbre de décision

Qualités requises (intuitivement)

- minimiser le nombre d'attributs considérés
- minimiser la complexité
- maximiser la taille des nœuds terminaux





Arbre de décision

Choisir le meilleur arbre

En pratique, impossibilité de générer tous les arbres possibles

- nombre important d'attributs prédictifs (m)
- grand nombre de valeurs possibles (nombre moyen v)

m	v	nb arbres
4	2	30
6	2	72385
8	2	$18\ 10^{18}$

$\hookrightarrow$ Nécessité d'heuristiques.

Arbre de décision

Objectifs

- s'approcher au mieux de la *meilleure* partition
- mais également chercher à produire un classifieur simple, qui prédit avec un minimum d'erreur l'attribut cible

L'heuristique doit donc permettre :

- de choisir rapidement l'attribut le plus intéressant
- les valeurs séparatrices de cet attribut
- sous diverses contraintes

Algorithme

Diviser pour régner

Initialement :

- arbre vide
- $\mathtt{E}$ ensemble d'apprentissage (individus $\times$ attributs)

Itération : nœud courant = terminal ?

- si oui, alors lui **affecter** une classe
- sinon, **sélectionner** un attribut $\mathtt{X_i}$,
  et partitionner $\mathtt{E}$ en $\mathtt{E_1}, \ldots, \mathtt{E_n}$ en fonction d'un **test** sur $\mathtt{X_i}$
- construire les sous-arbres $\mathtt{E_1}, \ldots, \mathtt{E_n}$

Algorithme

Préciser

- comment choisir l'attribut de partitionnement
  $\Rightarrow$ mesure de qualité locale de la subdivision
- le critère d'arrêt (attention à l'apprentissage par cœur !)
- comment choisir la classe à affecter à chaque feuille

Remarques

- algorithme de type glouton, sans retour-arrière
- la subdivision est effectivement réalisée sur la base du test retenu, d'où la nécessité de choisir le *meilleur*

$\hookrightarrow$ Le résultat va dépendre de l'heuristique de partitionnement, du critère d'arrêt et de la règle d'affectation de classe.

Plan

Heuristiques de partitionnement/mesures de qualités d'une partition

Idée

- comparer les différents choix possibles
- mesurer par une fonction h le degré de mélange des exemples dans les différentes partitions possibles
- choisir le meilleur éclatement

Intuitivement, h doit prendre

- minimum lorsque tous les exemples sont dans une même classe
- maximum lorsque les exemples sont équirépartis

Propriétés souhaitables de h

Soit Y une variable catégorielle a q modalités de fréquences $p = (p_1, \ldots, p_q)$

h est une fonction réelle positive de p dans $\mathbb{R}$ telle que :

- $h(p)$ est invariante par permutation des modalités de Y
- $h(p)$ atteint son maximum quand la distribution de Y est uniforme (chaque modalité de Y a une fréquence de $1/q$)
- $h(p)$ atteint son minimum quand la distribution de Y est certaine (centrée sur une modalité de Y, les autres étant de fréquence nulle)
- $h(p)$ est une fonction strictement concave

Mesures de qualité de partitionnement

Les classiques

- indice de d'impureté (ou de diversité) – coefficient de Gini
  choix : variable qui maximise la réduction d'impureté (CART [BFOS84])
- indice de pureté – coefficient de corrélation
  choix : variable qui maximise son coefficient de corrélation avec `Y`
- écart à l'indépendance – le lien du χ^2
  choix : variable qui maximise son lien avec `Y` (ChAID, 1980)
- gain informationnel – entropie de Shannon
  choix : variable qui maximise la gain d'information (ID3 [Qui86], C4.5 [Qui93])

Entropie de Shannon (Shannon, 1948)

Idée

Information : un ou plusieurs événements, parmi un ensemble fini d'événements possibles.
Cherchant un document dans une pile de dossiers on nous informe qu'il est dans un dossier rouge :

- information d'autant plus intéressante que le nombre de dossiers rouges est restreint, i.e. qu'elle réduira de beaucoup l'espace de recherche
- si l'on ajoute qu'il est dans un petit dossier, on peut réduire encore l'espace ...

$\hookrightarrow$ L'intérêt d'une information augmente avec la diminution du nombre de possibilités ultérieures.

Entropie – idée

- codage de l'information
  - soit n messages équiprobables
  - probabilité p_i de chaque message : $1/n$
  - information transportée par chaque message : $-\log(p_i) = \log(n)$
- soit la distribution $p = (p_1, \ldots, p_n)$
  - information transportée par p (entropie de p) est : $I(p) = -(p_1 \log(p_1) + \ldots + p_n \log(p_n))$

Application aux partitions

$\mathtt{E}$ ensemble
$\mathtt{E_1}, \ldots, \mathtt{E_n}$ sous-ensembles de $\mathtt{E}$ formant une partition de $\mathtt{E}$.
La quantité d'information liée à $\mathtt{E_i}$ est :

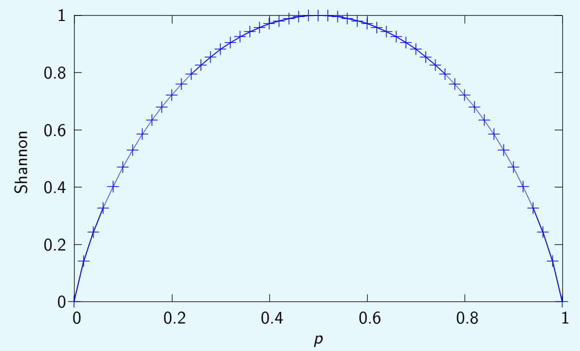
$$\mathtt{I(E_i)} = -\sum_{\mathtt{c \in classe(E_i)}} \mathtt{p_i(c)} \log_2(\mathtt{p_i(c)})$$

L'entropie moyenne de la partition est alors :

$$\mathtt{I(E_1, \ldots, E_n)} = -\sum_{\mathtt{i}} (\mathtt{p_i} \sum_{\mathtt{c}} \mathtt{p_i(c)} \log_2(\mathtt{p_i(c)}))$$

$h(p) = -\sum_{j=1}^{q} p_j \log_2 p_j$ est une fonction réelle positive de p dans [0..1]

- $h(p)$ est invariante par permutation des modalités de Y
- $h(p)$ atteint son maximum $\log_2(q)$ quand la distribution de Y est uniforme (chaque modalité de Y a une fréquence de $1/q$)
- $h(p)$ atteint son minimum 0 quand la distribution de Y est certaine (centrée sur une modalité de Y, les autres étant de fréquence nulle
- $h(p)$ est une fonction strictement concave



↪ Critère d'éclatement des nœuds en fils homogènes : minimiser l'entropie.

Entropie de Shannon

Mesures fondées sur l'entropy de Shannon

- le gain d'entropie gain (Quinlan, 1986) : $h(Y) - h(Y/X)$;
- le coefficient u (Theil, 1970), gain relatif de l'entropie de Shannon's entropy i.e. le gain d'entropie normalisé sur l'entropie a priori de Y : $\frac{h(Y)-h(Y/X)}{h(Y)}$;
- le gain-ratio (Quinlan, 1993) qui normalise le gain d'entropie de X par l'entropie de X de façon à pénaliser les attributs ayant de nombreuses modalités : $\frac{h(Y)-h(Y/X)}{h(X)}$;
- le coefficient de Kvalseth (Kvalseth, 1987), qui normalise le gain d'entropie par la moyenne des entropies de X et Y : $\frac{2(h(Y)-h(Y/X))}{h(X)+h(Y)}$;
- ...

Algorithme ID3 (Quinlan, 1986)

$$\texttt{Gain(X,E)} = \texttt{Info(E)} - \texttt{Info(X)}$$

On recherche de la meilleure partition :

- maximiser le gain informationnel
- *ie.* minimiser l'entropie de la partition retenue

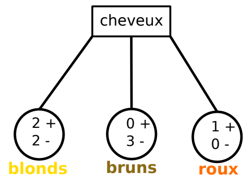
Algorithme ID3, exemple

nom	cheveux	taille	poids	écran solaire	coup de soleil
Sarah	blonde	moyenne	léger	non	oui (+)
Dana	blonde	grande	moyen	oui	non (-)
Alex	brune	petite	moyen	oui	non (-)
Annie	blonde	petite	moyen	non	oui (+)
Emily	rousse	moyenne	lourd	non	oui (+)
Pete	brun	grande	lourd	non	non (-)
John	brun	moyenne	lourd	non	non (-)
Katie	blonde	petite	léger	oui	non (-)

Critère d'arrêt : classes homogènes

Algorithme ID3, exemple

nom	cheveux	taille	poids	écran solaire	coup de soleil
Sarah	blonde	moyenne	léger	non	oui (+)
Dana	blonde	grande	moyen	oui	non (-)
Alex	brune	petite	moyen	oui	non (-)
Annie	blonde	petite	moyen	non	oui (+)
Emily	rousse	moyenne	lourd	non	oui (+)
Pete	brun	grande	lourd	non	non (-)
John	brun	moyenne	lourd	non	non (-)
Katie	blonde	petite	léger	oui	non (-)



Entropie moyenne : $\sum_b (\frac{n_b}{n_t} \times (\sum_c -\frac{n_{bc}}{n_b} \log_2 \frac{n_{bc}}{n_b}))$

$\frac{4}{8} \times (-\frac{2}{4}\log_2(\frac{2}{4}) - \frac{2}{4}\log_2(\frac{2}{4}))$
$+\frac{3}{8} \times (-\log_2(1))$
$+\frac{1}{8} \times (-\log_2(1))$
$= 0.50$

Algorithme ID3, exemple

nom	cheveux	taille	poids	écran solaire	coup de soleil
Sarah	blonde	moyenne	léger	non	oui (+)
Dana	blonde	grande	moyen	oui	non (-)
Alex	brune	petite	moyen	oui	non (-)
Annie	blonde	petite	moyen	non	oui (+)
Emily	rousse	moyenne	lourd	non	oui (+)
Pete	brun	grande	lourd	non	non (-)
John	brun	moyenne	lourd	non	non (-)
Katie	blonde	petite	léger	oui	non (-)

- cheveux : 0.5
- taille : 0.69 = $\frac{3}{8}(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})+\frac{3}{8}(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3})+\frac{2}{8}(-\log_2 1)$
- poids : $0.94=\frac{2}{8}(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})+\frac{3}{8}(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})+\frac{3}{8}(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})$
- écran solaire : $0.61=\frac{5}{8}(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5})+\frac{3}{8}(-\log_2 1)$

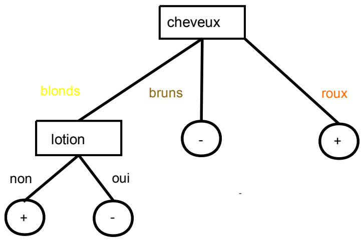
Choix de l'attribut : cheveux

Algorithme ID3, exemple

nom	cheveux	taille	poids	écran solaire	coup de soleil
Sarah	blonde	moyenne	léger	non	oui (+)
Dana	blonde	grande	moyen	oui	non (-)
Alex	brune	petite	moyen	oui	non (-)
Annie	blonde	petite	moyen	non	oui (+)
Emily	rousse	moyenne	lourd	non	oui (+)
Pete	brun	grande	lourd	non	non (-)
John	brun	moyenne	lourd	non	non (-)
Katie	blonde	petite	léger	oui	non (-)

- taille :
  $0.50=\frac{1}{4}(-\frac{1}{1}\log_2\frac{1}{1})+\frac{1}{4}(-\frac{1}{1}\log_2\frac{1}{1})+\frac{2}{4}(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})$
- poids :
  $1.0=\frac{2}{4}(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})+\frac{2}{4}(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})$
- écran solaire : 0.0

Algorithme ID3, exemple



Algorithme C4.5 (Quinlan, 1993)

Critères par défaut de C4.5

sélection du test basé sur un ratio de gains

$$GainRatio(\mathcal{E}|P)=\frac{GainInformationnel(\mathcal{E}|P)}{SplitInfo(\mathcal{E}|P)}$$

$$\text{avec } SplitInfo(\mathcal{E}|P)=-\sum_i p_i\log_2(p_i)$$

critère d'arrêt un nœud est terminal lorsque tous les individus appartiennent à la même classe, ou lorsqu'aucun test n'a pu être retenu

critère d'affectation de classe on affecte au nœud terminal la classe majoritaire des individus classés dans ce nœud

Entropie de Shannon

Particularité des coefficients fondés sur l'entropie de Shannon

La distribution prend sa valeur maximale quand elle est uniforme :

- la valeur de référence correspond à la distribution uniforme (situation d'indétermination)

Problèmes :

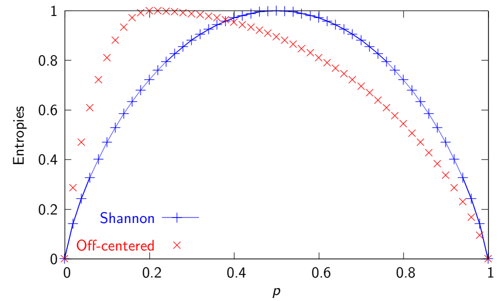
- classes déséquilibrées
- coûts des erreurs très différents

↪ Remarque : par construction avec les arbres de décision les classes sont toujours déséquilibrées.

Entropies non symétriques

Vers des entropies non symétriques

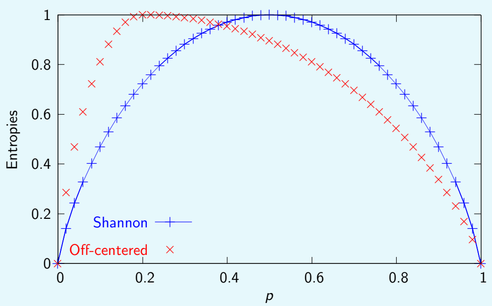
Evaluer $h(Y)$ et $h(Y/X = x_i)$ sur une échelle où la valeur de référence est centrée sur la situation d'indépendance i.e. sur la distribution à priori des classes.



Entropies non symétriques

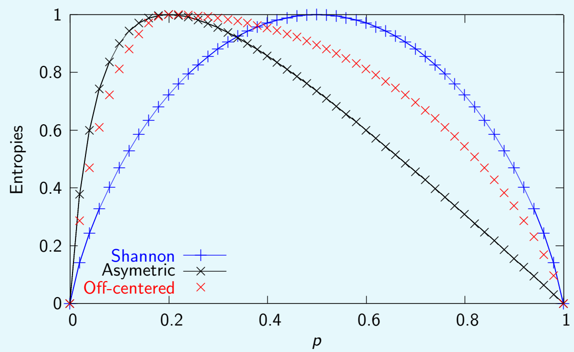
Illustration

Avec une distribution initiale $p = (0.2, 0.8)$ où la classe minoritaire est la classe positive, il semble intéressant d'atteindre une répartition $(0.5, 0.5)$ car il y a 2.5 fois plus de cas positifs.



Entropies non symétriques

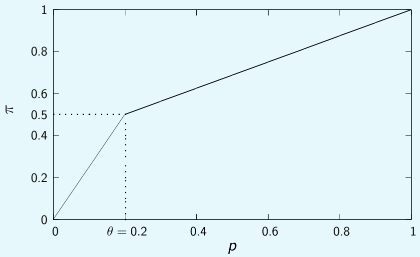
Différentes façon d'atteindre la situation d'indépendance



Off-centered entropies

Construction de l'Off-Centered Entropy [LLDP08]

- $(1-p, p)$ est plongé dans $(1-\pi, \pi)$ de telle façon que :
  - $\pi = \frac{p}{2\theta}$ si $0 \leq p \leq \theta$ (π croît de 0 à 0.5)
  - $\pi = \frac{p+1-2\theta}{2(1-\theta)}$ si $\theta \leq p \leq 1$ (π croît de 0.5 à 1)

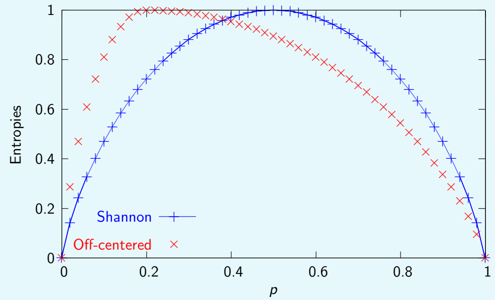


Off-centered entropies

Définition de OCE

OCE $\eta_\theta(p)$ est alors définie comme $h((1-\pi, \pi))$:

$$\eta_\theta(p) = -\pi \log_2 \pi - (1-\pi)\log_2(1-\pi)$$



Off-centered entropies

Propriétés

- maximale quand $p = \theta$, θ étant fixé par l'utilisateur (et non nécessairement égale à 0.5)
- $\eta_\theta(p) = h(\pi)$ est l'entropie de la distribution transformée $(1-\pi, \pi)$ et en possède les *mêmes* caractéristiques
- évidemment l'invariance par permutation des modalités de Y est abandonnée et $\eta_\theta(p)$ est maximale pour $p = \theta$ i.e. pour $\pi = 0.5$

$\hookrightarrow$ Extensions à plus de deux classes et pour un cadre général de décentrage des entropies (off-centered generalized entropies)

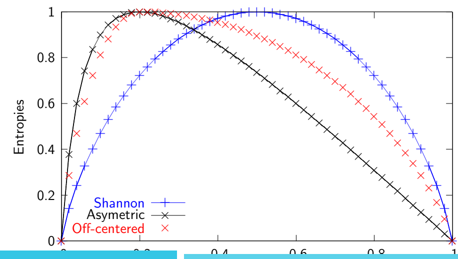
Asymmetric entropy

Asymmetric entropy

- asymmetric entropy pour le cas booléen [MZR06]

$$h_\theta(p) = \frac{p(1-p)}{(1-2\theta)p + \theta^2}$$

- extensions à plus de deux classes [ZMR07]



Stratégie adaptative [LLDP08]

Replacer les entropies usuelles par des entropies décentrées

- même si les données ne sont pas déséquilibrées, un arbre de décision peut être amené à traiter des classes déséquilibrées dans chaque nœud
- intérêt des entropies décentrées repose justement sur leur capacité à prendre leur valeur maximale sur la distribution à priori des des classes dans n'importe quel nœud
- utiliser cette potentialité pour une stratégie adaptative de construction des arbres

Plan

Expérimentations [LLDP08]

Comparaison sur 20 bases déséquilibrées

n°	Base	Nb. case	Nb. dim	Class min	Class max	Validation
1	Opticdigits	5620	64	10%(0)	90%(rest)	trn-tst
2	Tictactoe	958	9	35%(1)	65%(2)	10-fold
3	Wine	178	13	27%(3)	73%(rest)	loo
4	Adult	48842	14	24%(1)	76%(2)	trn-tst
5	20-newsgrp	20000	500	5%(1)	95%(rest)	3-fold
6	Breast	569	30	35%(M)	65%(B)	10-fold
7	Letters	20000	16	4%(A)	96%(rest)	3-fold
...	...	...	...	...	...	...

Expérimentations

Comparaison sur 25 bases déséquilibrées

n°	Base	Nb. case	Nb. dim	Class min	Class max	Validation
1	Opticdigits	5620	64	10%(0)	90%(rest)	trn-tst
2	Tictactoe	958	9	35%(1)	65%(2)	10-fold
3	Wine	178	13	27%(3)	73%(rest)	loo
4	Adult	48842	14	24%(1)	76%(2)	trn-tst
5	20-newsgrp	20000	500	5%(1)	95%(rest)	3-fold
6	Breast	569	30	35%(M)	65%(B)	10-fold
7	Letters	20000	16	4%(A)	96%(rest)	3-fold
...	...	...	...	...	...	...

Voir les sites de l'UCI, Statlog, ..., situations très diverses.
Protocole de transformation des cas multi-classes, de validation, ...

Expérimentations

Comparaison sur 20 bases déséquilibrées

OCE vs. SE	Tree size	Acc.	MinClass acc.	MajClass acc.
Mean (OCE-SE)	-9.900	0.76%	1.94%	0.44%
Mean Std. dev. (OCE-SE)	6.318	0.47%	0.53%	0.53%
Student ratio	-1.567	1.621	3.673	0.830
p-value (Student)	Non sign.	Non sign.	0.0016	Non sign.
OCE wins	12	16	18	7
Exaequo	3	1	1	5
SE wins	5	3	1	8
p-value (sign test)	Non sign.	0.0044	0.0000	Non sign.

- OCE améliore la précision de MinClass 18 fois sur 20 ($\sim +2\%$)
- la reconnaissance de MajClass n'est pas significativement améliorée, mais la précision globale est améliorée 16 sur 20
- les arbres produits par OCE sont souvent plus petits

Expérimentations

Comparaison sur 20 bases déséquilibrées

AE vs. SE	Tree size	Acc.	MinClass acc.	MajClass acc.
Mean (AE-SE)	-1.750	0.25%	1.04%	-0.01%
Mean Std. dev. (AE-SE)	7.500	0.14%	0.37%	0.14%
Student ratio	-0.233	1.746	2.808	-0.048
p-value (Student)	Non sign.	0.0970	0.0112	Non sign.
AE wins	8	14	15	8
Exaequo	2	1	1	4
SE wins	10	5	4	8
p-value (sign test)	Non sign.	Non sign.	0.0192	Non sign.

- AE a des résultats légèrement moins significatifs
- AE améliore 15 fois sur 20 la précision de MinClass ($\sim +1\%$)
- l'amélioration de la précision globale n'est pas significative, performances comparables pour MajClass et de la taille des arbres

Expérimentations

Comparaison sur 20 bases déséquilibrées

OCE vs. AE	Tree size	Acc.	MinClass acc.	MajClass acc.
Mean (OCE- AE)	-8.150	0.51%	0.90%	0.45%
Mean Std. dev. (OCE- AE)	4.563	0.38%	0.49%	0.44%
Student ratio	-1.786	1.330	1.846	1.014
p-value (Student)	0.0901	0.1991	0.0805	0.3234
OCE wins	8	11	11	8
Exaequo	6	5	3	4
AE wins	6	4	6	8
p-value (sign test)	Non sign.	Non sign.	Non sign.	Non sign.

- une légère mais non significative supériorité de OCE pour chaque critère
- notons un gain de 1 point sur le taux d'erreur de MinClass et de 0.5 point sur le taux global d'erreur

Expérimentations

Comparaison sur 20 bases déséquilibrées

- les entropies décentrées surclassent, particulièrement OCE, l'entropie de Shannon
- les entropies décentrées améliorent significativement la précision de la classe minoritaire, sans pénaliser la classe majoritaire et la taille des arbres
- une supériorité non significative d'OCE sur AE pour chaque critère

Plan

Retour sur le partitionnement

Création d'une partition

La partition est réalisée sur la base des valeurs prises par l'attribut sélectionné. On considère trois types de tests :

- test sur attribut nominal
- test sur attribut ordinal
- test sur attribut continu

Possibilités de tests suivant les attributs :

- tests d'égalité : création d'autant de sous-arbres que de valeurs possibles (variantes : regroupement de valeurs, notamment pour les arbres binaires)
- tests d'infériorité : nécessitent une opération de discrétisation (*attention, fortes conséquences sur la qualité du modèle induit*)

Discrétisation

Discrétiser X de domaine de définition $V(X)$.

Découper $V(X)$ en k intervalles I_j :

$I_1 = [d_0,\ d_1[$

$\vdots$

$I_j = [d_{j-1},\ d_j[$

$\vdots$

$I_k = [d_{k-1},\ d_k[$

- quelle est la valeur de k (nombre de d_j) ?
- où se trouvent les d_j ?

Une fois les d_j déterminés, la variable quantitative est remplacée par une variable qualitative prenant ses valeurs dans $(1, \ldots, k)$.

$\hookrightarrow$ La discrétisation d'une variable peut également être réalisée pendant l'étape de préparation des données.

Plan

Critères d'arrêt

Critères d'arrêt initiaux

- absence d'apport informationnel des attributs prédictifs
- homogénéité totale de la partition construite

↪ Toutes les variables peuvent être introduites unes à unes (obtention de l'arbre maximum) : sur-apprentissage, l'un des écueils majeurs en induction. Très problématique en présence de données bruitées

Critères d'arrêt

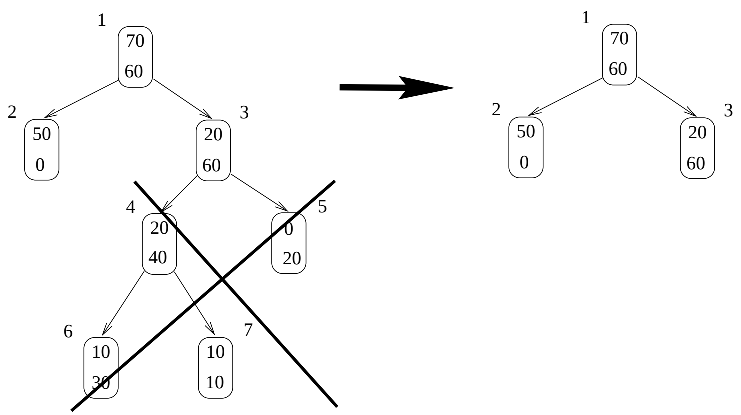
Idéalement

- arrêter la croissance de l'arbre au bon moment
- *mais...* critère inconnu
- le risque d'arrêter trop tôt est plus grand que d'arrêter trop tard

↪ Deux grandes stratégies :

- pré élagage
- post élagage

Critères d'arrêt



Critères d'arrêt

Exemple de critères

- homogénéité des sous-arbres (critère de confiance)
- effectifs des sous-arbres (critère de support)
- tests statistiques d'indépendance

Critère de support

Imposer une taille minimale à un sommet :

- les règles sont extraites des sommets terminaux
- les groupes correspondants doivent avoir un cardinal suffisamment important.

↪ En pratique : lors de la construction de l'arbre, toute décomposition engendrant au moins un groupe de cardinal inférieur à la taille limite est refusée.

Critère de support

Fixer la valeur limite :

- pas de règle véritable
- la plupart du temps, les différents auteurs préconisent une valeur de 5, un pourcentage de l'échantillon, etc.
- surtout, le choix dépend de l'effectif de l'échantillon initial et de la complexité du problème que l'on traite, notamment le nombre de modalités de la classe à étudier.

↪ En pratique : l'étape de compréhension des objectifs et des données doit aider à fixer la valeur limite.

Critères statistiques

Principe C4.5 [Qui93] :

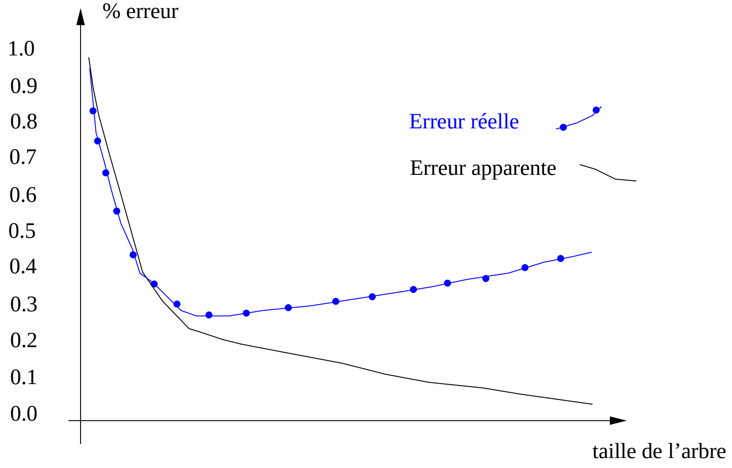
- observation selon laquelle toute partition engendre de l'information
- ce gain est-il statistiquement significatif, et non pas résultant tout simplement du hasard de l'échantillonnage ?
- solution simple : utiliser une valeur critique fondée sur la mesure d'information utilisée
- écart à l'indépendance du χ^2

Critères statistiques

Critiques :

- les feuilles étant étiquetées de telle façon qu'il y ait peu d'erreur, l'algorithme peut déterminer un arbre d'erreur apparente faible
- mais l'erreur réelle peut être importante
- arbre est bien adapté à l'échantillon d'apprentissage mais possède un pouvoir de prédiction faible

Critères d'arrêt



Généralités Heuristiques Exp. Dis. Arrêt Décision Evaluation Conclusion Bibliographie

Critères d'arrêt

Définition d'une double mesure de coût-complexité (CART [BFOS84])

- estimation du taux d'erreur par resubstitution est toujours optimiste
- construire un arbre aussi grand que l'on veut
- définir une séquence de sous-arbres imbriqués à l'aide d'une mesure de coût-complexité
- choisir celle qui minimise le taux d'erreur sur un échantillon à part, dit de test

Généralités Heuristiques Exp. Dis. Arrêt Décision Evaluation Conclusion Bibliographie

Critères d'arrêt

Taux d'erreur

Principe : scinder le jeu de données en deux parties (généralement 2/3 et 1/3)

- utiliser la première pour l'apprentissage
- utiliser la deuxième en test de généralisation
- supprimer les branches ayant trop d'erreurs

$\hookrightarrow$ Disposer de suffisamment de données.

Généralités Heuristiques Exp. Dis. Arrêt Décision Evaluation Conclusion Bibliographie

Critères d'arrêt

Validation croisée (CART)

Subdiviser l'échantillon de départ en v sous-échantillons

- répéter v fois l'analyse en prenant tour à tour chaque échantillon comme test
- déterminer α_v qui minimise le taux d'individus mal classés sur l'ensemble des v arbres
- utiliser α_v pour déterminer la *bonne* taille de l'arbre construit sur la totalité de l'échantillon

Critères d'arrêt

Taux d'erreur pessimiste (C4.5)

Basé sur l'hypothèse que la proportion d'individus mal classés sur un sommet suit une loi binomiale. Utilisation d'une base d'apprentissage et une base de test

Définition de TEP : borne haute de l'intervalle de confiance du taux d'erreur estimé

Mise en œuvre : vérifier pour chaque avant-dernier nœud si son TEP est inférieur à la moyenne pondérée des TEP de ses descendants directs

Plan

Classement d'un nouvel individu

Affecter une classe à un nouvel individu

Statut particulier du nœud terminal : correspond à un sous-groupe de l'ensemble de départ dans lequel il y a une présence *significativement élevée* d'une des modalités de la variable à expliquer.

Quelle est cette modalité ?

- la classe la plus fréquente
- sélection aléatoire
- ou tout simplement refuser de conclure
- etc.

Classement d'un nouvel individu

Trois grandes stratégies

- règles globales qui désignent la même classe à tous les individus de la feuilles
- règles individuelles qui utilisent une informations supplémentaires pour décider de la classe attribuée à chaque individu.
  Adaptée aux données déséquilibrées, mais au prix d'une perte de l'intelligibilité de la règle (même si l'arbre reste intelligible)
- règles agrégées fondées sur une combinaison de règles issues de classifieurs multiples.
  Intelligibilité de l'ensemble des classifieurs et des règles sont perdues.

Classement d'un nouvel individu

Exemples de stratégies

- règles globales : règle majoritaire, tirage aléatoire proportionnel aux classes [GK54], indice d'implication [RZM07]
- règles individuelles : la classe est désignée par un vote majoritaire sur les k-nearest neighbors de l'individu dans la feuille correspondante [PDLL08]
- règles agrégées : ensemble d'arbres & combiner les résultats (bagging : bootstrap aggregating) [Bre96]

Classement d'un nouvel individu

Prise en compte des coûts d'erreur

- cas d'un tirage aléatoire simple avec matrice de coûts symétrique : choisir la classe la plus fréquente (minimiser l'espérance de pertes), sélection aléatoire
- cas de coûts non symétriques : on peut être amené à sélectionner une conclusion ne se rapportant pas nécessairement à la classe majoritaire (assignation en fonction de coûts, de risques, ou refus de conclure. . .)

Plan

Matrice de confusion et mesures classiques

Matrice observation×prédiction de confusion (2 classes : N et P)

	N	P
N	TN	FP
P	FN	TP

- TN est le nombre/la proportion de prédictions correctes de N
- FP est le nombre/la proportion de prédictions incorrectes de P
- FP est le nombre/la proportion de prédictions incorrectes de N
- TP est le nombre/la proportion de prédictions correctes de P

Accuracy (précision)

Proportion de prédictions correctes.

	N	P
N	TN	FP
P	FN	TP

$$\frac{TN + TP}{TN + FP + FN + TP}$$

Recall ou taux de vrais positifs

	N	P
N	TN	FP
P	FN	TP

$$\frac{TP}{FN + TP}$$

Taux de faux positifs

	N	P
N	TN	FP
P	FN	TP

$$\frac{FP}{TN + FP}$$

Taux de vrais négatifs

	N	P
N	TN	FP
P	FN	TP

$$\frac{TN}{TN + FP}$$

Matrice de confusion et mesures classiques

Taux de faux négatifs

	N	P
N	TN	FP
P	FN	TP

$$\frac{FN}{FN + TP}$$

Matrice de confusion et mesures classiques

Precision/précision (taux de prédictions de positifs corrects)

	N	P
N	TN	FP
P	FN	TP

$$\frac{TP}{FP + TP}$$

Matrice de confusion et mesures classiques

F_β

$$F_\beta = (1+\beta^2)\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

i.e.

$$F_\beta = \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP}$$

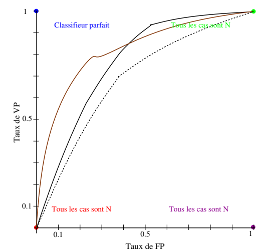
Moyenne harmonique, classique et équilibrée : $F_1 = 2\frac{recall \cdot precision}{recall + precision}$

F_2 privilégie recall, $F_{0.5}$ la precision, etc.

Evaluation graphique

Courbe ROC

- (essentiellement) pour les problèmes à deux classes
- indique la capacité du classifieur à placer les positifs devant les négatifs (graphique avec les taux de faux positifs en abscisse et les taux de vrais positifs en ordonnée)



Evaluation graphique

Area Under Curve ? Aire Sous la Courbe

Elle indique la probabilité d'un individu positif d'être classé devant un individu négatif.
Il existe une valeur seuil, si l'on classe les individus au hasard, l'AUC sera égale à 0.5.

Plan

Arbre et règles

Implantation d'arbres dans des SGB (lecture de la racine vers les feuilles)

Quelques propriétés :

- tous les éléments de l'ensemble d'apprentissage sont classifiés
- ils ne sont classifiés qu'une seule fois (*ie.* pour tout élément, il y a exactement une règle de prédiction de sa classe)
- tout nouvel élément sera également affecté à une feuille

Bilan sur les arbres de décision

Points faibles :

- Effectif important nécessaire
- Instabilité de l'arbre
- Pas de combinaison linéaire de variables explicatives
- Une variable explicative peut en cacher une autre
- Classification : difficile d'expliquer une variable à plus de 2 groupes

Points forts :

- Pas d'hypothèses sur les données
- Variable explicative de toutes natures
- Profusion des variables explicatives
- Hiérarchisation des variables explicatives
- Simplicité de lecture des règles
- Règles opérationnelles
- Robuste aux données aberrantes
- Interactions

Bilan sur les arbres de décision

Mais encore . . .

- nombreuses autres fonctions de séparation
- regroupement de variables
- critères d'arrêt
- arbres de régression
- méthodes ensemblistes (forêts d'arbres) & hybrides (arbres obliques)
- parallélisation
- gestion des valeurs manquantes
- etc.

Plan

Bibliographie I

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C.J. Stone.
Classification and Regression Trees.
Wadsworth International, 1984.

[Bre96] L. Breiman.
Bagging predictors.
Machine Learning, 24(2) :123–140, 1996.

[GK54] L. A. Goodman and W. H. Kruskal.
Measures of association for cross classifications, i.
JASA, I(49) :732–764, 1954.

[LLDP08] P. Lenca, S. Lallich, T.-N. Do, and N.-K Pham.
A comparison of different off-centered entropies to deal with class imbalance for decision trees.
In T. Washio, E. Suzuki, K. M. Ting, and A. Inokuchi, editors, *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'08)*, Lecture Notes in Computer Science, pages 634–643, Osaka, Japan, May 20-23 2008. Springer-Verlag.

[MZR06] S. Marcellin, D. A. Zighed, and G. Ritschard.
An asymmetric entropy measure for decision trees.
In *IPMU 2006*, pages 1292–1299, Paris, France, 2006.

[PDLL08] N.-K Pham, T.-N. Do, P. Lenca, and S. Lallich.
Using local node information in decision trees : Coupling a local decision rule with an off-centered entropy.
In R. Stahlbock, S. F. Crone, and S. Lessmann, editors, *The International Conference on Data Mining (DMIN'08)*, volume 1, pages 117–123, Las Vegas, Nevada, USA, July 14-17 2008. CSREA Press.

[Qui86] J. R. Quinlan.
Induction of decision trees.
Machine Learning, 1(1) :81–106, 1986.

[Qui93] J. Ross Quinlan.
C4.5 : Programs for Machine Learning.
Morgan Kaufmann, 1993.

Bibliographie II

[RZM07] G. Ritschard, D. A. Zighed, and S. Marcellin.
Données déséquilibrées, entropie décentrée et indice d'implication.
In *Rencontres Internationales Analyse Statistique Implicative*, pages 315–327, Castellón, Spain, 2007.

[ZMR07] D. A. Zighed, S. Marcellin, and G. Ritschard.
Mesure d'entropie asymétrique et consistante.
In *Extraction et Gestion des Connaissances*, pages 81–86, Namur, Belgium, 2007.