# Paramétrage de l'Apprentissage

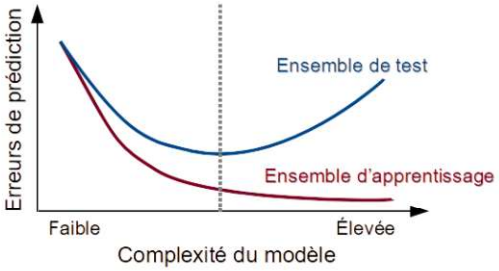
Nicolas PASQUIER
Université Côte d'Azur
Département Informatique
Laboratoire I3S (UMR-7271 UCA/CNRS)
http://www.i3s.unice.fr/~pasquier

## Évaluation des Classifieurs

- Pertinence d'un classifieur
  i. Qualité des prédictions : capacité à prédire correctement la classe
  ii. Faculté de généralisation (extrapoler les relations pertinentes) : capacité à s'appliquer avec la même efficacité à d'autres données
- Qualités des prédictions
  - Probabilités associées aux prédictions faites
  - Résultats sur l'ensemble de test
  - Outils de calcul automatique (courbe ROC, indice AUC, etc.)
- Faculté de généralisation
  - Déterminée par l'adéquation de la complexité du classifieur relativement à la complexité du problème
  - Exemple : profondeur de l'arbre de décision (nombre maximal de tests), nombres de nœuds internes (tests de variables), effectifs des nœuds feuilles (nombre d'exemples support de la prédiction)

## Problème du Sur-Apprentissage

- Problème du sur-ajustement ou *over-fitting*
- Classifieur trop spécifique aux exemples de l'ensemble d'apprentissage
  - Ex : seuls les exemples de l'ensemble d'apprentissage (combinaison précises de valeurs) sont représentées dans l'arbre de décision
  - Perte de sa capacité de généralisation, i.e. extrapoler uniquement les relations réellement utiles
    - Un exemple différent de ceux appris n'aura pas de correspondance




- Principe de *Parcimonie* (statistiques)
  - Simplicité du modèle de prédiction implique stabilité
  - Ex : arbre de décision de taille minimale parmi ceux dont la performance est maximale (ou ayant la performance requise)



## Paramétrage de l'Apprentissage

- Objectif : définir la configuration algorithmique utilisée pour l'apprentissage afin d'adapter au mieux le modèle de prédiction au problème traité
- Un classifieur est défini par la configuration algorithmique utilisée
  - Choix de l'algorithme implémentant une approche de classification (arbres de décision, random forests, etc.)
  - Choix des valeurs des paramètres de l'algorithme utilisé pour l'apprentissage
- Exemple : apprentissage d'arbres de décision
  - Choix de la mesure de sélection d'attribut (*feature selection*)
  - Objectif : sélectionner l'attribut qui partitionne le mieux les exemples vis à vis des classes

i. Minimiser le nombre de tests qui resteront nécessaires pour classer les exemples dans les partitions résultantes

ii. Obtenir des partitions dont le désordre est minimal (distribution des classes) afin de maximiser la simplicité de l'arbre résultant

## Mesures de Sélection d'Attribut

- Mesures qui quantifient l'importance de chaque attribut pour la distinction des classes
- Deux mesures centrales sont communes aux différentes implémentations
- Gain d'information (algorithme ID3) ou Gain Ratio (algorithmes C4.5 et C5.0)
  - Critère heuristique
  - Mesure la réduction de l'Entropie (quantité d'information à fournir pour classer un exemple) apportée par le choix de l'attribut
- Coefficient de Gini (algorithmes CART, SLIT, SPRINT)
  - Critère statistique
  - Mesure l'impureté des nœuds résultant du choix de l'attribut
- D'autres mesures reposant sur des principes similaires sont proposées par certaines implémentations (ReliefF, Mutual Information, Deviance, etc.)



## Définition de l'Entropie

- (C. Shannon, 1948)
- Objectif : quantifier l'information requise (e.g. nombre de tests) pour prédire un événement
- *Entropie* d'un ensemble d'exemples de données *E*

$$Entropie\ (E) = -\sum_{i=1}^{i=M} P(C_i) \times log_2 P(C_i)$$

- Valeur de la variable de classe $C_i \in \{C_1, \ldots, C_M\}$
- $M$ : nombre de classes
- $P(C_i)$ : proportion d'exemples de $E$ appartenant à la classe $C_i$

# Calcul de l'Entropie : Exemple

- Prédiction des classes Buyer=True et Buyer=False

**Ensemble d'Apprentissage *EA*** → Dénombrement des classes

| ID | Age | Income | Student | Credit | Buyer |
|---|---|---|---|---|---|
| 1 | ≤ 30 | High | No | Intermediate | False |
| 2 | ≤ 30 | High | No | Excellent | False |
| 3 | 31..40 | High | No | Intermediate | True |
| 4 | ≥ 41 | Medium | No | Intermediate | True |
| 5 | ≥ 41 | Low | Yes | Intermediate | True |
| 6 | ≥ 41 | Low | Yes | Excellent | False |
| 7 | 31..40 | Low | Yes | Excellent | True |
| 8 | ≤ 30 | Medium | No | Intermediate | False |
| 9 | ≤ 30 | Low | Yes | Intermediate | True |
| 10 | ≥ 41 | Medium | Yes | Intermediate | True |
| 11 | ≤ 30 | Medium | Yes | Excellent | True |
| 12 | 31..40 | Medium | No | Excellent | True |
| 13 | 31..40 | High | Yes | Intermediate | True |
| 14 | ≥ 41 | High | No | Intermediate | True |

Dénombrement des classes

**Nombre d'Exemples**

| Buyer=True | Buyer=False | Total |
|---|---|---|
| 9 | 5 | 14 |

Calcul de l'Entropie

$$Entropie\ (EA) = -\frac{9}{14} \times log_2\left(\frac{9}{14}\right) - \frac{5}{14} \times log_2\left(\frac{5}{14}\right) = 0.940$$

- Quantifie l'information requise pour classer un exemple de *EA*



# Mesure du Gain d'Information

- *Information Gain* (S. Kullback & R.A. Leibler, 1951)
- Gain d'Information : quantifie l'information acquise (vs. entropie) en choisissant la variable *V* dans l'objectif de classer un exemple de *EA*

$$Gain_V(EA) = Entropie(EA) - Info_V(EA)$$

- Variable *V* : variable à *N* valeurs $V = \{V_1, \ldots, V_N\}$
- $Info_V(EA)$ : information encore requise pour classer un exemple après avoir utilisé *V* pour diviser l'ensemble *EA* en *N* partitions

$$Info_V(EA) = \sum_{j=1}^{j=N} \frac{|EA_{V_j}|}{|EA|} \times Entropie(EA_{V_j})$$

- $EA_{V_j}$ : exemples de *EA* pour lesquels la variable *V* prend la valeur $V_j$

# Calcul du Gain d'Information : Exemple de l'Attribut Age

- $Info_{Age(EA)}$ : mesure la quantité d'information (vs. entropie) qui sera encore requise pour classer une exemple de *EA* si on choisi *Age*

**Ensemble d'Apprentissage *EA***

| ID | Age | Income | Student | Credit | Buyer |
|---|---|---|---|---|---|
| 1 | ≤ 30 | High | No | Intermediate | **False** |
| 2 | ≤ 30 | High | No | Excellent | **False** |
| 3 | 31..40 | High | No | Intermediate | **True** |
| 4 | ≥ 41 | Medium | No | Intermediate | **True** |
| 5 | ≥ 41 | Low | Yes | Intermediate | **True** |
| 6 | ≥ 41 | Low | Yes | Excellent | **False** |
| 7 | 31..40 | Low | Yes | Excellent | **True** |
| 8 | ≤ 30 | Medium | No | Intermediate | **False** |
| 9 | ≤ 30 | Low | Yes | Intermediate | **True** |
| 10 | ≥ 41 | Medium | Yes | Intermediate | **True** |
| 11 | ≤ 30 | Medium | Yes | Excellent | **True** |
| 12 | 31..40 | Medium | No | Excellent | **True** |
| 13 | 31..40 | High | Yes | Intermediate | **True** |
| 14 | ≥ 41 | High | No | Intermediate | **True** |

Dénombrement des classes pour chaque valeur de *Age*

**Matrice de contingence**

| | Buyer=True | Buyer=False | Total |
|---|---|---|---|
| **Age ≤ 30** | 2 | 3 | 5 |
| **Age = 31..40** | 4 | 0 | 4 |
| **Age ≥ 41** | 3 | 2 | 5 |
| **Total** | 9 | 5 | 14 |

Calcul de l'information restant à calculer si *Age* est choisi

$$Info_{Age}(EA) = \frac{5}{14} \times Entropie\left(EA_{Age \leq 30}\right) + \frac{4}{14} \times Entropie\left(EA_{Age=31..40}\right) + \frac{5}{14} \times Entropie\left(EA_{Age \geq 41}\right) +$$



# Calcul du Gain d'Information : Exemple de l'Attribut Age

**Matrice de contingence**

| | Buyer=True | Buyer=False | Total |
|---|---|---|---|
| **Age ≤ 30** | 2 | 3 | 5 |
| **Age = 31..40** | 4 | 0 | 4 |
| **Age ≥ 41** | 3 | 2 | 5 |
| **Total** | 9 | 5 | 14 |

Calcul de l'information restant à fournir pour classer un exemple si *Age* est choisi

$$Info_{Age}(EA) = \frac{5}{14} \times \left(-\frac{2}{5} \times log_2\left(\frac{2}{5}\right) - \frac{3}{5} \times log_2\left(\frac{3}{5}\right)\right) + \frac{4}{14} \times \left(-\frac{4}{4} \times log_2\left(\frac{4}{4}\right) - \frac{0}{4} \times log_2\left(\frac{0}{4}\right)\right) + \frac{5}{14} \times \left(-\frac{3}{5} \times log_2\left(\frac{3}{5}\right) - \frac{2}{5} \times log_2\left(\frac{2}{5}\right)\right)$$

$$= 0.694$$

## Calcul du Gain d'Information : Exemple

- Le Gain d'Information quantifie l'information « gagnée » en choisissant la variable
- Choix de la variable de gain maximal
  - $Gain_{Age}(EA)$ = 0.940 – 0.694 = **0.246**
  - $Gain_{Income}(EA)$ = 0.940 – 0.911 = 0.029
  - $Gain_{Student}(EA)$ = 0.940 – 0.789 = 0.151
  - $Gain_{Credit}(EA)$ = 0.940 – 0.892 = 0.048
- Le Gain d'Information peut dans certains cas favoriser l'éclatement en un grand nombre de petites partitions très pures
- La mesure du Gain Ratio pénalise l'éclatement : $GainRatio_V(EA) = \frac{Gain_V(EA)}{SplitInfo_V(EA)}$ avec $SplitInfo_V(EA) = \sum_{j=1}^{j=N} \frac{|EA_{V_j}|}{|EA|} \times log_2\left(\frac{|EA_{V_j}|}{|EA|}\right)$ mais peut parfois entrainer des partitions très déséquilibrées (effectifs)



## Mesure du Coefficient de Gini

- *Gini coefficient* (C. Gini, 1912)
- Objectif : quantifier numériquement l'impureté des partitions résultantes vis-à-vis des classes

$$Gini_V(EA) = \sum_{j=1}^{j=N}\left(\frac{|EA_{V_j}|}{|EA|} \times \left(1 - \sum_{i=1}^{i=M} P\left(C_i|EA_{V_j}\right)^2\right)\right)$$

- Interprétation : mesure avec quelle fréquence un exemple serait mal classé si sa classe était sélectionnée aléatoirement depuis la distribution des classes dans les partitions résultantes
  - Gini = 0.0 si dans chaque partition résultante on a une classe unique
- Exemple : apprentissage de l'arbre de décision de prédiction de Buyer
  - Pour la sous-matrice *Age* ≤ 30 nous avons $Gini_{Student}$(EA)= 0.0 car dans les partitions *Student* = *Yes* et *Student* = *No* nous avons une unique classe

# Calcul du Coefficient de Gini : Exemple de l'Attribut Age

**Ensemble d'Apprentissage *EA***

| ID | Age | Income | Student | Credit | Buyer |
|---|---|---|---|---|---|
| 1 | ≤ 30 | High | No | Intermediate | **False** |
| 2 | ≤ 30 | High | No | Excellent | **False** |
| 3 | 31..40 | High | No | Intermediate | **True** |
| 4 | ≥ 41 | Medium | No | Intermediate | **True** |
| 5 | ≥ 41 | Low | Yes | Intermediate | **True** |
| 6 | ≥ 41 | Low | Yes | Excellent | **False** |
| 7 | 31..40 | Low | Yes | Excellent | **True** |
| 8 | ≤ 30 | Medium | No | Intermediate | **False** |
| 9 | ≤ 30 | Low | Yes | Intermediate | **True** |
| 10 | ≥ 41 | Medium | Yes | Intermediate | **True** |
| 11 | ≤ 30 | Medium | Yes | Excellent | **True** |
| 12 | 31..40 | Medium | No | Excellent | **True** |
| 13 | 31..40 | High | Yes | Intermediate | **True** |
| 14 | ≥ 41 | High | No | Intermediate | **True** |

Dénombrement des classes pour chaque valeur de *Age*

**Matrice de contingence**

| | Buyer=True | Buyer=False | Total |
|---|---|---|---|
| **Age ≤ 30** | 2 | 3 | 5 |
| **Age = 31..40** | 4 | 0 | 4 |
| **Age ≥ 41** | 3 | 2 | 5 |
| **Total** | 9 | 5 | 14 |

Calcul du Coefficient de Gini pour *Age*

$$Gini_{Age}(EA) = \frac{5}{14} \times \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{4}{14} \times \left(1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right) + \frac{5}{14} \times \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) = 0.343$$



# Calcul du Coefficient de Gini : Exemple

- Le Coefficient de Gini évalue la « difficulté » de classement d'un exemple dans les sous-branches (nœuds descendants) si on choisi la variable
- Choix de la variable de Coefficient de Gini minimal
  - $Gini_{Age}(EA)$ = **0.343**
  - $Gini_{Income}(EA)$ = 0.441
  - $Gini_{Student}(EA)$ = 0.367
  - $Gini_{Credit}(EA)$ = 0.428
- Le Coefficient de Gini peut parfois favoriser l'éclatement en petites partitions très pures

# Références et Bibliographie

- Principales Librairies R
  - rpart : arbres de décision CART
  - tree : variante des arbres de décision CART
  - party : sélection de variables et critère d'arrêt statistiques
  - C50 : arbres de décision C5.0
  - RWeka : variantes J4.8 de l'algorithme C5.0 et algorithme M5
- Bibliographie
  - R and Data Mining - Examples and Case Studies. Chapter 4 (Decision Trees and Random Forest) and Chapter 13 (Case Study II: Customer Response Prediction and Profit Optimization). Yanchang Zhao. Academic Press, Elsevier, 2012. ISBN 978-0-123-96963-7
  - Data Classification: Algorithms and Applications. Chapter 4 (Decision Trees: Theory and Algorithms). Charu C. Aggarwal. Chapman and Hall/CRC, 2014. ISBN 978-1-466-58674-1