Apprentissage Supervisé : Évaluation et Sélection d'Attributs

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

Objectifs de l'Évaluation et la Sélection d'Attributs

1. Estimer l'importance de chaque variable pour la reconnaissance des classes
2. Identifier les variables primordiales pour la reconnaissance des classes
3. Résoudre des problèmes d'applicabilité des algorithmes :
   - Réduction de la dimensionalité du problème par sélection de variables
   - Optimisation de la représentation des informations (e.g. si discrétisation nécessaire des variables numériques continues)

- Techniques d'évaluation des attributs
  - Variables discrètes
    - Tests de dépendance entre variables
    - Graphique en mosaïque des dépendances entre valeurs
  - Variables continues : mesures de corrélation et covariance
  - Variables hétérogènes : mesure d'utilité prédictive

Variables Discrètes : Tests de Dépendance

- Tests du χ^2 et de Fisher : les valeurs des deux variables sont-elles significativement éloignées de l'indépendance?
- Mesure de p-value
  - *Probabilité que les données observées soient aussi ou plus éloignées que celles observées si l'hypothèse nulle (indépendance) est satisfaite*
  - Interprétation : dépendance entre variables si p-value < 0.05 (5 %)

Variables Habitats et Produit

```
Pearson's Chi-squared test

data:  Habitat and Produit
X-squared = 3.7906, df = 3, p-value = 0.285
```

```
Fisher's Exact Test for Count Data

data:  Habitat and Produit
p-value = 0.2851
alternative hypothesis: two.sided
```

Variables Enfants et Produit

```
Pearson's Chi-squared test

data:  Enfants and Produit
X-squared = 99.164, df = 3, p-value < 2.2e-16
```

```
Fisher's Exact Test for Count Data

data:  Enfants and Produit
p-value < 2.2e-16
alternative hypothesis: two.sided
```


Valeurs de Variables Discrètes : Mesure de Dépendance

- Calcul à partir des répartitions des classes pour chaque valeur

Habitat	Produit: Non	Produit: Oui	Produit: Total
Banlieue	45.2 %	54.8 %	100 %
Centre_ville	54.3 %	45.7 %	100 %
Petite_ville	59.0 %	41.0 %	100 %
Rural	52.1 %	47.9 %	100 %
Total	54.3 %	45.7 %	100 %

Enfants	Produit: Non	Produit: Oui	Produit: Total
0	63.5 %	36.5 %	100 %
1	18.5 %	81.5 %	100 %
2	59.0 %	41.0 %	100 %
3	80.9 %	19.1 %	100 %
Total	54.3 %	45.7 %	100 %

- *χ^2-residuals* : résidus du test du χ^2

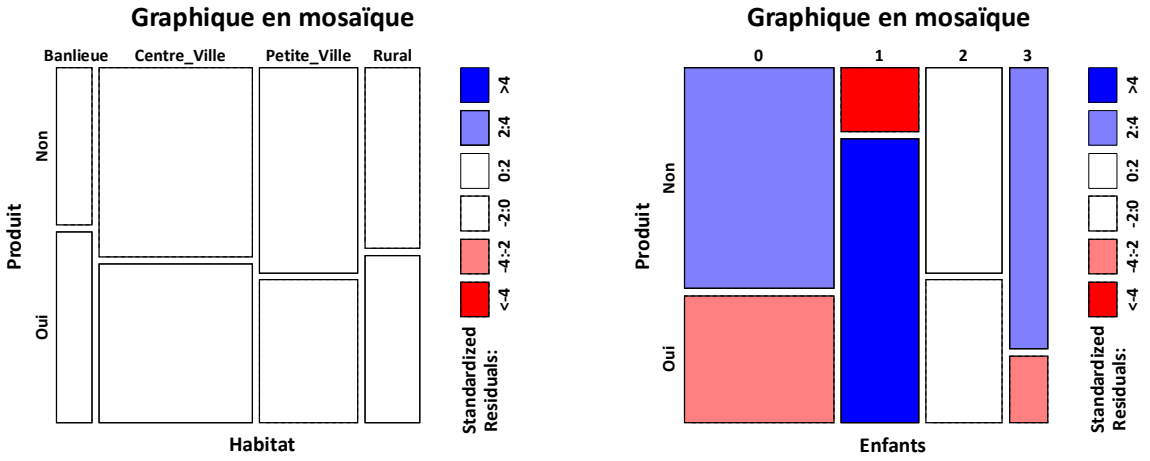
Habitat	Produit: Non	Produit: Oui	Produit: Total
Banlieue	-1.1	1.1	0.0
Centre_ville	0.2	-0.2	0.0
Petite_ville	0.8	-0.9	0.0
Rural	-0.1	0.1	0.0
Total	0.2	-0.2	0.0

Enfants	Produit: Non	Produit: Oui	Produit: Total
0	1.1	-1.2	0.0
1	-4.9	5.5	0.0
2	0.5	-0.6	0.0
3	3.5	-3.8	0.0
Total	-0.1	0.1	0.0

- Écart à l'indépendance significatif : résidu > 2 ou < -2

Graphique en Mosaïque des Dépendances entre Valeurs

- *Mosaic Plot* de représentation des χ^2-residuals
  - Coloration des cooccurrences dont l'écart à l'indépendance est significatif (résidu > 2 ou < -2)
  - Bleu : dépendance positive / Rouge : dépendance négative
  - Largeur et hauteur des boîtes : proportionnelles au nombre d'exemples



Variables Numériques Continues : Mesures de Corrélation

- Corrélation entre deux variables numériques : Coefficients de Pearson, Spearman (rho) et Kendall (tau)

```
    Pearson's product-moment correlation

data:  Age and Revenus
t = 28.18, df = 598, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.7186707 0.7876923
sample estimates:  cor  0.7552679
```

```
    Spearman's rank correlation rho

data:  Age and Revenus
S = 8586631, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:  rho  0.7614818
```

```
    Kendall's rank correlation tau

data:  Age and Revenus
z = 20.607, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:  tau  0.5681503
```

Interprétation des Valeurs

Corrélation	Négative	Positive
Faible	[−0.5 : 0.0]	[0.0 : 0.5]
Forte	[−1.0 : −0.5]	[0.5 : 1.0]

p-value < 0.05 : rejet de l'hypothèse d'indépendance

Évaluations Basées sur l'Entropie

- « *Information Entropy* » (C. Shannon, 1948)
  - Valeur numérique indiquant la quantité d'information requise pour prédire la valeur d'une variable $V \in \{V_1, \ldots, V_m\}$ dans un ensemble de données

$$Entropie(V_1, \ldots, V_m) = -\sum_{i=1}^{i=m} \left(P(V_i) \times log_2\left(P(V_i)\right)\right) \quad \in [0.0, +\infty]$$

    ou $P(V_i)$ est la proportion d'exemples possédant la valeur V_i

- Exemples : ensemble de données *Data Produit*
  - Entropie de la variable *Produit* dans la sous-matrice *Produit* = "Oui"

```
> entropy(produit[produit$Produit=="Oui", "Produit"])
[1] 0.0
```

  - Entropie de la variable *Produit* dans la matrice complète

```
> entropy(produit$Produit)
[1] 0.9945751
```


Évaluations Basées sur l'Entropie : Information Gain

- Entropie de la variable de classe $C \in \{C_1, \ldots, C_n\}$: quantifie l'information requise (valeurs des variables prédictives) pour reconnaître la classe
  - L'apprentissage d'un modèle de prédiction sera d'autant plus difficile que la valeur obtenue est élevée
- Information Gain : quantifie la valeur fournie par une variable $V \in \{V_1, \ldots, V_m\}$ à l'Entropie
  - Calcul : *Entropie de la variable de classe C – Somme des entropies de chaque valeur $V_{j=1..m}$ pondérée par la proportion d'exemples qu'elle représente pour chaque classe $C_{i=1..n}$*

$$IG(V) = \sum_{i=1}^{i=n} \left(P(C_i) \times log_2\left(P(C_i)\right)\right) - \sum_{j=1}^{j=m} \left(\sum_{i=1}^{i=n} \left(P(V_j) \times log_2\left(P(V_j|C_i)\right)\right)\right)$$

  - Plus la valeur est élevée, plus la variable *V* contribue à reconnaître la classe

Évaluations Basées sur l'Entropie : Gain Ratio

- Information Gain
  - Incapacité à différencier des variables de cardinalités différentes qui ont la même Entropie vs. les classes
  - Un algorithme d'apprentissage d'arbre de décision sélectionnera au hasard
- Gain Ratio
  - Objectif : favoriser les variables de plus faibles cardinalités
  - Meilleures propriétés de généralisation pour l'apprentissage d'un arbre de décision, i.e. moins de risque de surajustement
  - Calcul : Information Gain de la variable *V* / Entropie de la variable *V* seule

$$Gain(V) = \frac{\sum_{i=1}^{i=n}\left(P(C_i) \times log_2\left(P(C_i)\right)\right) - \sum_{j=1}^{j=m}\left(\sum_{i=1}^{i=n} P(V_j) \times log_2\left(P(V_j|C_i)\right)\right)}{\sum_{j=1}^{j=m}\left(P(V_j) \times log_2\left(P(V_j)\right)\right)}$$

Mesure de l'Index Gini

- Gini Index : quantifie l'utilité d'une variable $V \in \{V_1, \ldots, V_m\}$ pour prédire la valeur d'une variable de classe $C \in \{C_1, \ldots, C_n\}$ selon la « pureté » du résultat
  - Calcul : *Somme des poids de chaque valeur $V_{j=1..m}$ pondérée par les probabilités combinées pour chaque classe $C_{i=1..n}$ de choisir un exemple avec la valeur $V_{j=1..m}$ qui soit de la mauvaise classe (non majoritaire)*

$$Gini(V) = \sum_{j=1}^{j=m}\left(P(V_j) \times \left(1 - \sum_{i=1}^{i=n} P(V_j|C_i)^2\right)\right) \qquad \in [0.0, 1.0[$$

  - Une valeur de 0.0 indique une répartition parfaite des classes si on partitionne les exemples selon les valeurs de la variable
    - Chaque valeur V_j est associée à une unique classe C_i
  - Plus la valeur est élevée, plus les classes seront mélangées dans les partitions résultantes $V_1, \ldots, V_m$
  - Plus la valeur est faible, plus la variable *V* contribue à reconnaître la classe

Algorithme de calcul du Relief

- Relief : génère un poids pour chaque variable prédictive selon une mesure de distance entre exemples dans l'espace des données

1. Sélection aléatoire d'un échantillon d'exemples
2. Pour chaque exemple de l'échantillon
    1. Identification des exemples dans le voisinage selon la mesure de distance
    2. Identification du voisin de même classe le plus proche (*Near Hit*) et du voisin de classe distincte le plus proche (*Near Miss*)
    3. Mise à jour des poids des variables prédictives selon ces distances

- Variantes (*ReliefF*) et paramétrages
    - Mesure de distance (Euclidienne, Manhattan) pour les variables numériques
    - Distance de voisinage
- Méthode étendue à la classification multi-classes par décomposition du problème en multiples classifications binaires

Minimum Description Length

- Minimum Description Length (MDL) (Kononenko, 1995)
- Principe : toute régularité dans un ensemble de données peut être utilisée pour compresser ces données
- Compresser des données : les décrire en utilisant moins de symboles qu'il n'en faut pour les décrire littéralement
- Davantage de régularités dans les données signifie davantage de compression de l'ensemble de données
- Selon ce principe, l'apprentissage peut être interprété comme la découverte de régularités dans les données
- Évaluation de la contribution des variables prédictives à l'apprentissage selon les régularités dans les distributions de leurs valeurs et des classes

Exemple : Variables de l'ensemble de données *Data Produit*

Variable	Information Gain
Revenus	0.033199
Enfants	0.028356
Marie	0.025961
Age	0.019366
ID	0.005616
Habitat	0.004556
Compte_Epargne	0.003813
Sexe	0.001583
Compte_Courant	0.000457
Emprunt	0.000422
Voiture	0.000258

Variable	Gain Ratio
Revenus	0.109689
Enfants	0.055612
Age	0.034135
Marie	0.028071
ID	0.025158
Compte_Epargne	0.004269
Habitat	0.002535
Sexe	0.001583
Compte_Courant	0.000573
Emprunt	0.000453
Voiture	0.000258

Variable	Relief
Enfants	0.26333
Revenus	0.10273
Age	0.06381
Marie	0.04833
Emprunt	0.04167
Compte_Epargne	0.01833
Voiture	0.00333
Compte_Courant	-0.00167
Sexe	-0.02667
ID	-0.04596
Habitat	-0.06833

Échelle de valeurs du Gini Index inversée (dépend de l'implémentation)

Variable	Gini Index (inv.)
Revenus	0.021670
Enfants	0.018019
Marie	0.017835
Age	0.013322
ID	0.003673
Habitat	0.003135
Compte_Epargne	0.002629
Sexe	0.001089
Compte_Courant	0.000314
Emprunt	0.000290
Voiture	0.000178

Variable	MDL
Revenus	0.02857
Enfants	0.02338
Marie	0.02051
Age	0.01427
ID	0.00188
Compte_Epargne	-0.00150
Sexe	-0.00392
Compte_Courant	-0.00468
Emprunt	-0.00497
Voiture	-0.00524
Habitat	-0.00961

Exemple : Variables de l'ensemble de données *Data Produit*

- Interprétation des résultats : similitudes et cas particuliers
- Variables Revenus et Enfants : évaluées comme les deux variables les plus utiles par toutes les mesures, Revenus majoritairement
- Variables Age et Marié : évaluées comme les 3ème et 4ème variables les plus utiles par toutes les mesures, Marié majoritairement
- Variable ID : dépendance fonctionnelle de conception (BD) ID –df–> Produit
  - À chaque valeur de ID est associée une valeur de Produit
  - Implique une dépendance de valeur
  - Mesure du Relief faible car la méthode est moins impactée (considère davantage la répartition aléatoire des classes par rapport à l'ID)
- Autres variables : valeurs d'évaluation significativement plus faibles
  - Utilité significativement moindre considérées individuellement

Références et Bibliographie

- Librairies R
  - stats (R Base) : fonctions de calculs de mesures de corrélation et covariance
  - regclass : graphiques en mosaïque des proportions de cooccurrences
  - questionr : tests de dépendance du χ^2 et de Fisher, calculs des résidus du χ^2
  - MASS : fonctions de calculs de statistiques descriptives et de visualisations graphiques de données
  - DescTools : calcul de l'entropie selon différentes bases (log, $\log_2$, etc.) et des coefficients de Pearson, Spearman et Kendall
  - CORElearn : calculs de mesures d'évaluation d'attributs (variantes des mesures Relief, Information Gain, Gain Ratio, Index Gini, MDL et DKM)
- Références
  - CRAN Task View : Multivariate Statistics. https://cran.r-project.org/web/views/Multivariate.html