Reduction de dimension en apprentissage non supervisé

Mourer Alex[1], Baron Alexis[1]

mourer.alex@gmail.com,alexis.b.baron@gmail.com

[1] M2 Techniques d'Information et de Décision dans l'Entreprise
SAMM
Université Paris 1 Panthéon-Sorbonne

Anglet 2017 Journées R

Outline

Outline

L'algorithme DIVCLUS-T

En entrée :

- une matrice de données $\mathcal{X}$ de dimension $n * p$ quantitative, qualitative ou mixte.
- un nombre K indiquant le nombre de classes de la dernière partition (par défaut la partition des singletons).

Répétition des deux étapes suivantes :

1. division d'une classe en deux : la bipartition doit optimiser un critère W. L'énumération complète est évitée par la contrainte monothétique.
2. choix de la classe à diviser : la nouvelle partition doit optimiser le critère W.

En sortie :
une hiérarchie indicée (un dendrogramme) qui se lit comme un arbre de décision.

Division d'une classe

Le principe

Sélectionner parmi toutes les bipartitions induites par toutes les questions binaires (définies sur une unique variable) possibles, celles de plus petite inertie intra-classe W (définie sur toutes les variables).

Pour une variable quantitative X_j, une question binaire est notée $[X_j \leqslant c]$?

- nombre infini de questions binaires possibles sur X_j mais au plus $n - 1$ bipartitions d'une classe C à n observations.
- tri des valeurs des observations de X_j et les valeurs de coupures sont les milieux de deux observations consécutives.

Choix de la classe à diviser

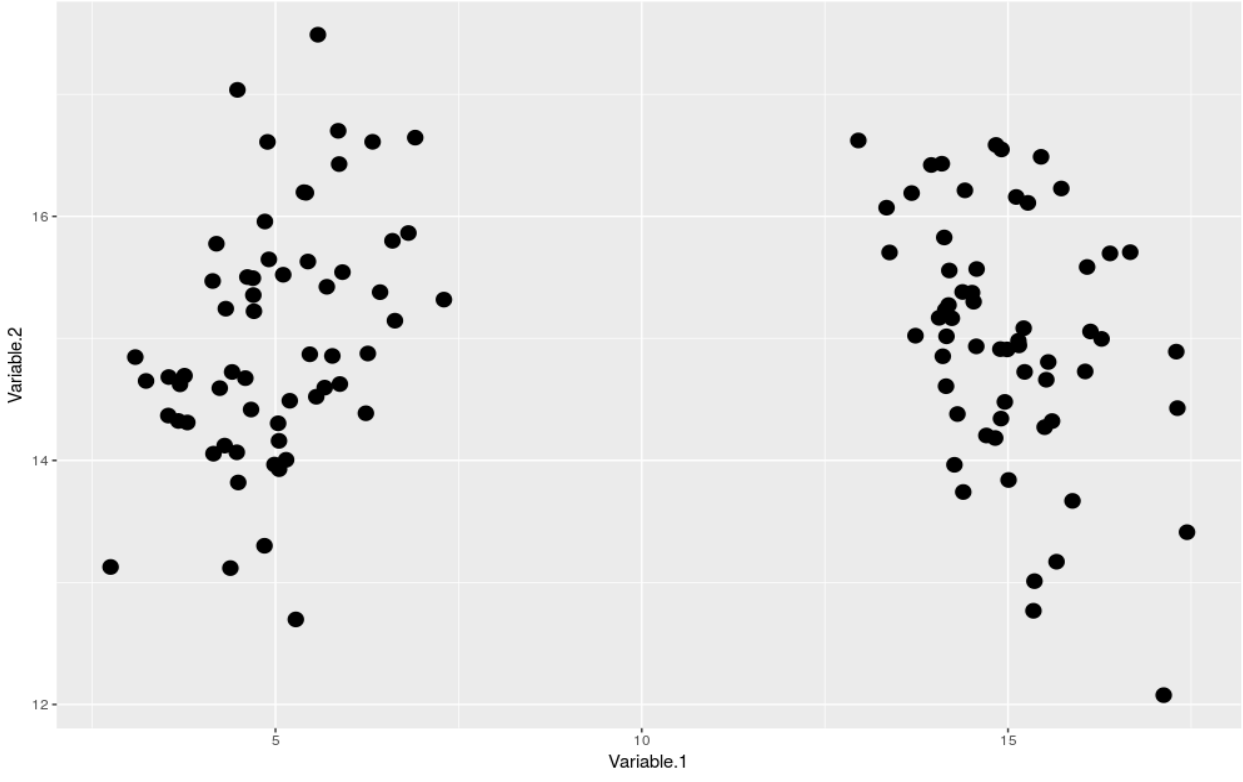
Le principe

Choisir la classe $C_l = A_l \cup \bar{A}_l$ de la partition :
$P_k = \{C_1, \ldots, C_{l-1}, A_l, \bar{A}_l, C_{l+1}, \ldots, C_k\}$ qui permet d'obtenir la partition de plus petite inertie intra-classe $W(P_{k+1})$.

- l'inertie intra-classe est un critère additif : $W(P_k) = \sum_{l=1}^{k} I(C_l)$.
- équivalent de choisir C_l qui maximise la variation de l'inertie : $h(C_l) = I(C_l) - I(A_l) - I(\bar{A}_l)$.

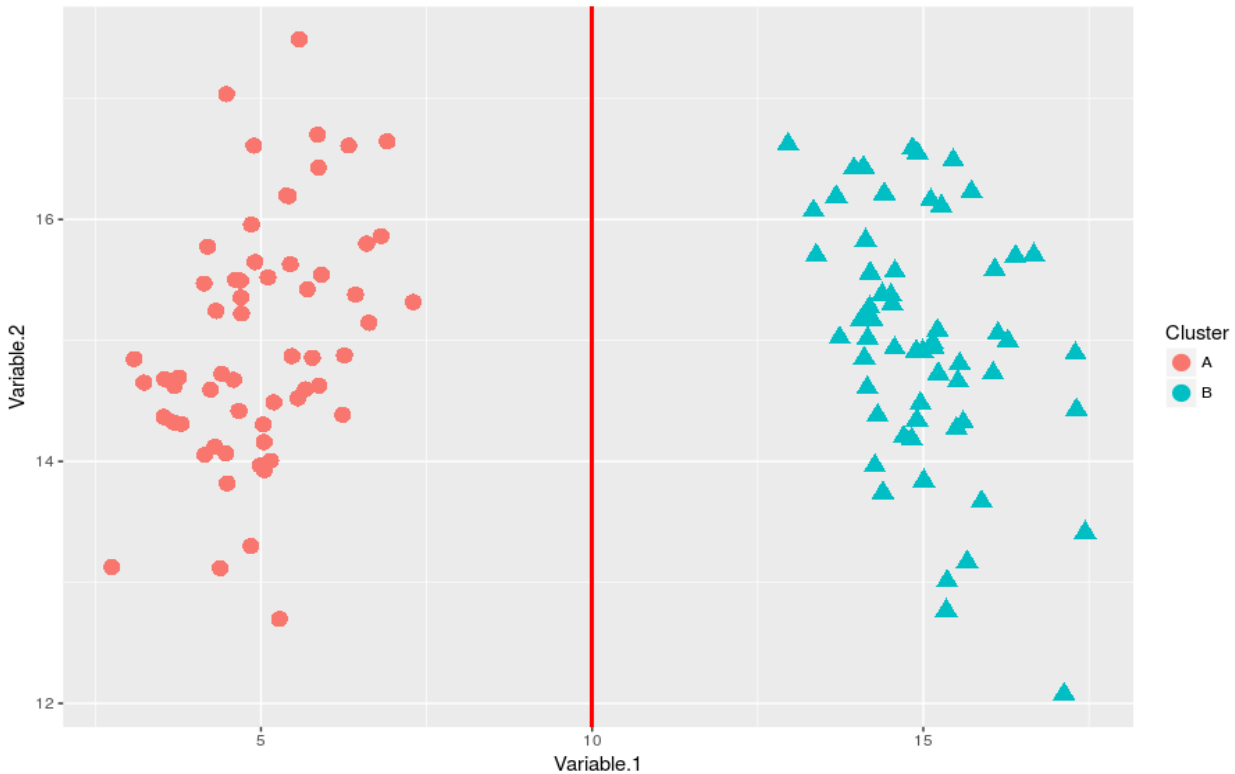
Exemples

- Avant clustering :



Exemples

- Après clustering :



La question binaire posée est : $[X_j \leqslant 10]$?

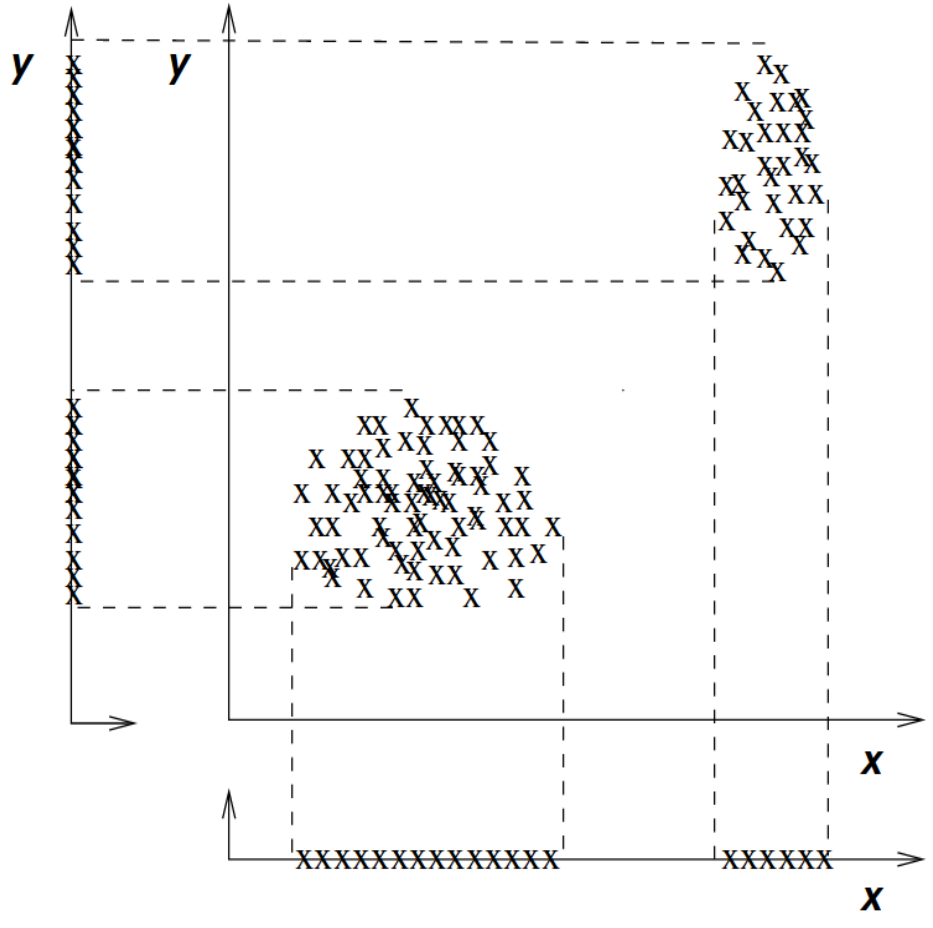
Outline

Motivation

Dans beaucoup de domaines d'apprentissage, les variables potentiellement utiles ont besoin d'être identifiées. Dans le cas où certaines variables ne sont pas pertinentes, choisir un sous-groupe de variables conduira souvent à de meilleurs résultats.

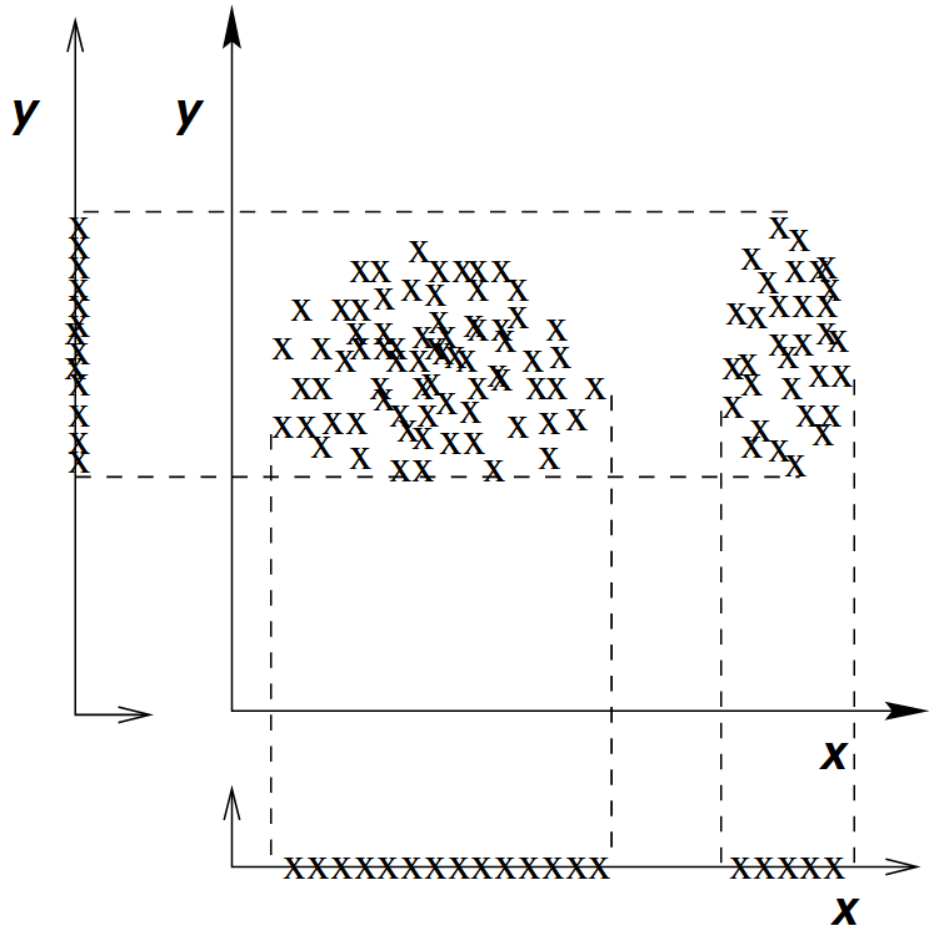
- certaines variables ne sont pas pertinentes, sont redondantes ou peuvent amener à un mauvais clustering.
- réduire le nombre de variables améliore la lisibilité et la compréhension des résultats.
- certains algorithmes peuvent s'avérer inconsistants en grande dimension.

Exemples



x et *y* sont des variables redondantes.

Exemples



Ici, *y* n'est pas pertinente.

Forêt aléatoire non supervisée

Le principe

Une forêt aléatoire non supervisée est un ensemble d'arbres non supervisé : chaque arbre étant construit sur un échantillon bootstrap (tirage aléatoire des individus avec remise) ainsi qu'un tirage sans remise de variables à chaque noeud de l'arbre.

L'importance d'une variable est ensuite calculée par rapport à son impact sur le clustering. Pour ce faire, on utilise le principe de permutation.

Forêt aléatoire non supervisée pour l'importance de variables

Pour un échantillon Bootstrap $b = 1, \ldots, B$:

1. identifier l'échantillon $\mathcal{L}_{OOB} = \mathcal{L} \backslash \mathcal{L}_b$; ce sont les observations qui n'ont pas été utilisées pour la construction de l'arbre.
2. récupérer les clusters associés aux observations $\mathcal{L}_{OOB}$ en suivant les différents chemins de l'arbre.
3. 1: **for all** $j = 1 \rightarrow p$ **do**
   2: permuter les valeurs de la variable j de l'échantillon $\mathcal{L}_{OOB}$.
   3: récupérer les nouveaux clusters associés à ces observations en suivant les différents chemins de l'arbre.
   4: compter le nombre d'observations $\mathcal{L}_{OOB}$ ayant changé de clusters.
   5: **end for**

Forêt aléatoire non supervisée pour l'importance de variables

Importance de variables par permutation

$$VI^{(t)}(x_j) = \frac{\sum_{i \in \mathcal{L}^t_{OOB}} \mathbb{1}(C_i \neq C_i\pi_j)}{|\mathcal{L}^t_{OOB}|}$$

C_i : cluster de l'observation i avant permutation.

$C_i\pi_j$: cluster de l'observation i après permutation.

$$\mathbb{1}(C_i \neq C_i\pi_j) = \begin{cases} 1 \text{ si l'individu i change de cluster après permutation} \\ 0 \text{ sinon} \end{cases}$$

$|\mathcal{L}^t_{OOB}|$: cardinal de l'ensemble $\mathcal{L}^t_{OOB}$.

Forêt aléatoire non supervisée pour l'importance de variables

Importance de variables par permutation

$$VI(x_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(x_j)}{ntree}$$

ntree : nombre d'arbres dans la forêt.

Note :

1. $VI^{(t)}(x_j) = 0$ par définition, si X_j n'est pas dans l'arbre t.
2. pour les données numériques, la complexité algorithmique de la forêt aléatoire non supervisée est $O(t(Kpn(log(n) + p)))$ avec K le nombre de clusters de la meilleure partition, p le nombre de variables, n le nombre d'objets et t le nombre d'arbres.

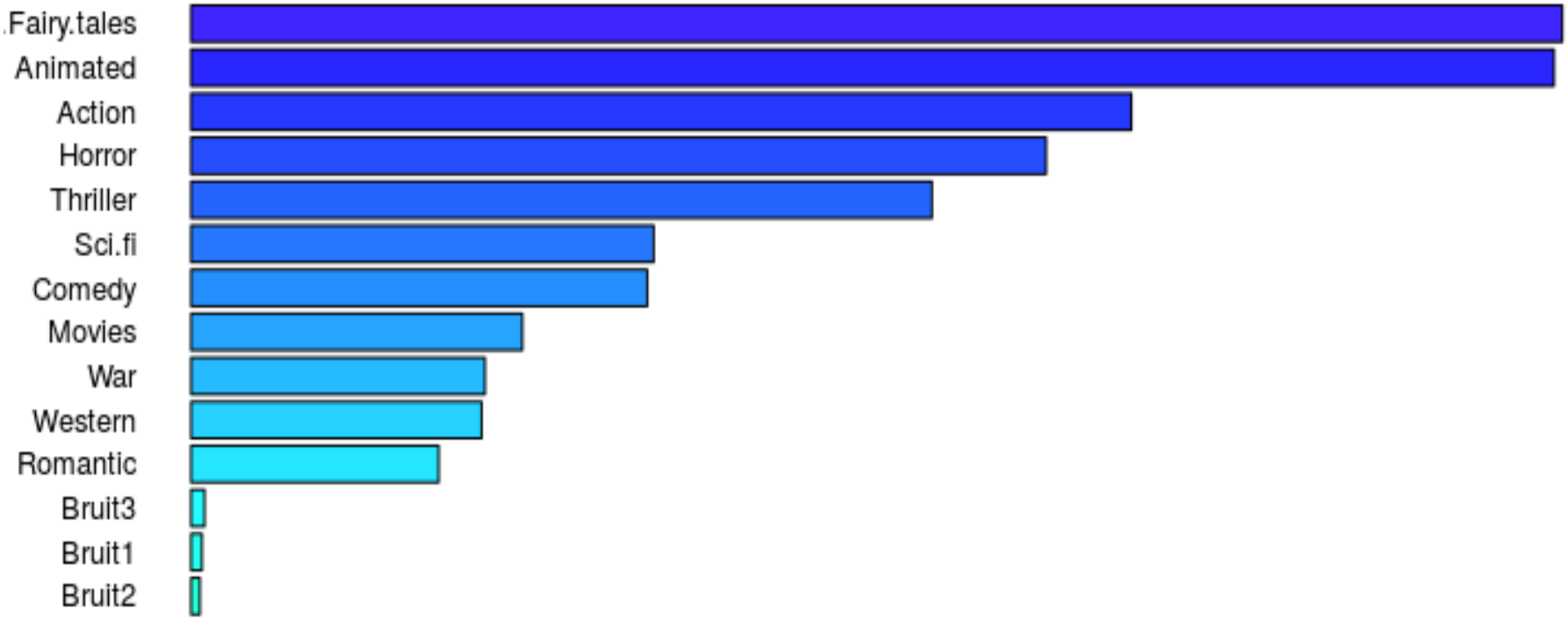
Exemples

- jeu de données tiré d'une enquête sociologique réalisée auprés de 1010 jeunes slovaques agés de 18 à 30 ans.
- 150 variables qui correspondent à 150 questions sur les peurs, les habitudes, ou encore les goûts de ces jeunes.
- les variables sont rangées par thématique.
- chaque observation correspond à une réponse, et chaque réponse est donnée par une note comprise entre 1 et 5.
- pour l'exemple qui suit, on s'intéresse uniquement à la thématique "MOVIES".
- on introduit 3 variables suivant une loi uniforme dans les données. Elles ne sont donc d'aucune utilité dans le clustering.

Exemples

Name	Importance
Fantasy.Fairy.tales	15.8804224
Animated	15.7841039
Action	10.8896975
Horror	9.9022546
Thriller	8.5801941
Sci.fi	5.3581621
Comedy	5.2825342
Movies	3.8363299
War	3.4032534
Western	3.3647260
Romantic	2.8674372
Bruit3	0.1562500
Bruit1	0.1262842
Bruit2	0.1020263

Exemples


.Fairy.tales
Animated
Action
Horror
Thriller
Sci.fi
Comedy
Movies
War
Western
Romantic
Bruit3
Bruit1
Bruit2

Outline

Conclusion

- l'importance des variables peut être calculée à l'aide de la forêt aléatoire en apprentissage non supervisée.
- l'algorithme peut traiter des données de type mixte : quantitatives et qualitatives (cette dernière n'a pas été introduite dans cette présentation).
- cela permet, par exemple, d'identifier les questions les plus pertinentes dans des données issues d'un questionnaire.

References

Marie Chavent, Yves Lechevallier, Olivier Briant. DIVCLUS-T: a monothetic divisive hierarchical clustering method. Computational Statistics and Data Analysis, Elsevier, 2007, 52 (2)

https://www.kaggle.com/miroslavsabo/young-people-survey

Breiman, L. (2001). Random forests. Machine Learning 45(1),5–32.