Exploration statistique multidimensionnelle

Résumé

Statistique, fouille ou Science des Données, les appellations changent le volume et la diversité des données explosent, les technologies se succèdent, les modèles et algorithmes se complexifient. L'estimation devient un apprentissage, la prévision remplace l'explication.

Le parcours pour devenir data scientist *est structuré en quatre parties :*

Retour *à l'introduction générale*

Saison 1 *(L3) Statistique élémentaire, descriptive* vs. *inférentielle.*

Saison 2 *(M1) Statistique Exploratoire multidimensionnelle et apprentissage non supervisé.*

Saison 3 *Apprentissage Statistique / Machine supervisé.*

Saison 4 *(M2) Technologies pour la Science des (grosses) Données.*

plus des réflexions sur : Statistique et Déontologie scientifique.

1 Objectifs

Avec la taille des données, le nombre des variables observées augmentent et des outils adaptés sont nécessaires pour en analyser et mieux comprendre les structures d'un point de vue global ou multidimensionnel. Les objectifs sont de résumer, représenter graphiquement, réduire la dimension, regrouper. Les outils présentés, multidimensionnels, sont à utiliser à la suite de ceux uni et bidimensionnels de la saison 1, sans chercher à bruler les étapes.

Les méthodes de *Statistique exploratoire multidimensionnelle* se décomposent en deux grands groupes selon l'objectif fixé.

1.1 Méthodes factorielles

Le premier groupe concerne les méthodes dites *factorielles* de décomposition sur une base adaptée : les facteurs sur lesquels projeter projeter les données pour des *représentations graphiques* en dimension réduite. Les principales méthodes se différencient selon le type (quantitatif, qualitatif) des variables considérées.

1.2 Classification non supervisée ou *clustering*

Le deuxième groupe concerne les méthodes ou algorithmes visent la recherche de classes, ou regroupements des individus, se ressemblant au mieux ou les plus proches au sens d'une mesure de distance. Ce groupe de méthodes est référencée sous l'appellation de *classification non supervisée* dans la communauté de l'*apprentissage machine*.

Attention : Ne pas confondre la *classification non supervisée*, en anglais *clustering*, avec la *classification supervisée*, en anglais *classification*, qu'il est préférable de traduire en français par le terme : *discrimination*, moins ambigu.

En classification non supervisée les classes ne sont pas connues *a priori* mais déterminées à partir des données. En classification supervisée ou discrimination, objet de la saison 3, les classes sont connues, observées, apprises, pour être prévues sur de nouvelles observations.

2 Méthodes factorielles

2.1 Historique

Les bases théoriques de ces méthodes sont anciennes et sont principalement issues de "psychomètres" américains : Spearman (1904) et Thurstone (1931, 1947) pour l'Analyse en Facteurs, Hotteling (1935) pour l'Analyse en Composantes Principales et l'Analyse Canonique, Hirschfeld (1935) et Guttman (1941, 1959) pour l'Analyse des Correspondances. Pratiquement, leur emploi ne s'est généralisé qu'avec la diffusion des moyens de calcul dans le courant des années 60. Sous l'appellation *"Multivariate Analysis"* elles poursuivent des objectifs sensiblement différents à ceux qui apparaîtront en France. Un individu ou unité statistique n'y est souvent considéré que pour l'information qu'il apporte sur la connaissance des liaisons entre variables au sein d'un échantillon statistique

dont la distribution est le plus souvent soumise à des hypothèses de normalité.

En France, l'expression *"Analyse des Données"* recouvre les techniques ayant pour objectif la *description statistique des grands tableaux* (n lignes, où n varie de quelques dizaines à quelques milliers, p colonnes, où p varie de quelques unités à quelques dizaines). Ces méthodes se caractérisent par une utilisation *intensive* de l'ordinateur, leur objectif *exploratoire* et une absence quasi systématique d'hypothèses de nature *probabiliste* au profit des propriétés et résultats de géométrie euclidienne. Elles insistent sur les représentations graphiques en particulier de celles des individus qui sont considérés au même titre que les variables.

Depuis la fin des années 1970, de nombreux travaux ont permis de rapprocher ou concilier les deux points de vue en introduisant, dans des espaces multidimensionnels appropriés, les outils probabilistes et la notion de *modèle*, usuelle en statistique *inférentielle*. Les techniques se sont ainsi enrichies de notions telles que l'estimation, la convergence, la stabilité des résultats, le choix de critères...

L'objectif essentiel de ces méthodes est l'aide à la compréhension de volumes de données souvent considérables. Réduction de dimension, représentation graphique optimale, recherche de facteurs ou variables latentes... sont des formulations équivalentes.

2.2 Méthodes

Les méthodes factorielles se classifient selon le type des variables à analyser (quantitatives et/ou qualitatives) :

- Analyse en Composantes Principales (p variables quantitatives),
- Analyse Factorielle Discriminante (p variables quantitatives, 1 variable qualitative),
- Analyse Factorielle des Correspondances simple (2 variables qualitatives) et Multiple (p variables qualitatives),
- Analyse Canonique (p et q variables quantitatives),
- *Multidimensional Scaling* (M.D.S.) ou positionnement multidimensionnel ou analyse factorielle d'un tableau de distances.
  Toutes les précédentes méthodes sont basées sur des outils classiques de géométrie euclidienne qui sont développés dans les rappels et compléments d'algèbre linéaire. Ils font appel à un algorithme de décomposition en valeurs singulières (SVD) d'une matrice rectangulaire.
- *Non negative Matrix Factorisation* ou NMF. Cette dernière approche de décomposition en facteurs sous des contraintes de non-négativité, contrairement à la SVD, peut être obtenue par différents algorithmes plus ou moins complexes, efficaces, selon les données à étudier.

3 Classification non supervisée

L'objectif d'une méthode de classification déborde le cadre strictement exploratoire. C'est la recherche d'une *typologie*, ou *segmentation*, c'est-à-dire d'une partition, ou répartition des individus en *classes*, ou catégories. Ceci est fait en optimisant un *critère* visant à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible.

3.1 Contraintes

Un calcul élémentaire de combinatoire montre que le nombre de partitions possibles d'un ensemble de n éléments croît plus qu'exponentiellement avec n ; le nombre de partitions de n éléments en k classes est le nombre de Stirling, le nombre total de partition est celui de Bell : $P_n = \frac{1}{e} \sum k = 1^{\infty} \frac{k^n}{k!}$.

Pour $n = 20$, il est de l'ordre de 10^{13}. Il n'est donc pas question de chercher à optimiser le critère sur toutes les partitions possibles. Les méthodes se limitent à l'exécution d'un algorithme itératif convergeant vers une "bonne" partition qui correspond en général à un optimum local. Même si le besoin de classer des objets est très ancien, seule la généralisation des outils informatiques en a permis l'automatisation dans les années 1970. Celeux et col. (1989)[1] décrivent en détail ces algorithmes.

3.2 Méthodes

Il n'existe donc pas de solution analytique du problème de classification et un très grand nombre d'algorithmiques ont été proposés pour atteindre cet objectif. En voici quelques uns parmi les plus utilisés, tous ne sont pas ou pas encore décrits.

- Classification ascendante hiérarchique,
- Algorithmes de réallocation dynamique,
- Cartes de Kohonen (réseaux de neurones),

- DBSCAN,
- Mélanges gaussiens.
- ...

4 Déroulement de la saison 3

Les apprentissages de cette saison nécessitent l'acquisition, en parallèle, de compétences plus approfondies en R, Python, éventuellement SAS. Approfondir successivement les différents tutoriels découpés en épisodes qui alternent la pratique des environnements logiciels et celles de l'exploration statistique. multidimensionnelle.

Références

[1] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier et H. Ralambondrainy, *Classification automatique des données*, Dunod, 1989.