

A review of outliers detection

Titouan Vayer

May 17, 2017

1 Introduction

- Qu'est-ce qu'un outlier ?
- Comment détecter un outlier ?
- Comment traiter un outlier

2 Les modèles statistiques

3 Les modèles basés sur des distances

4 Les modèles paramétriques

5 Les modèles semi paramétriques

Les travaux suivants sont basés essentiellement sur le papier de
[Victoria J.Hodge : A Survey of Outlier Detection Methodologies]

Qu'est-ce qu'un outlier ?

Deux définitions possibles :

- (Grubbs, 1969) Un outlier est une observation qui semble dévier fortement par rapport aux autres observations du sample dans lequel il se situe
- (Barnett & Lewis, 1994) Une observation (ou un ensemble d'observations) est un outlier s'il apparaît comme étant contradictoire avec le reste des données.

Comment mettre en place une stratégie de détection ?

- **Unsupervised clustering** : Déterminer l'outlier sans à priori sur les données. Suppose d'avoir un dataset suffisamment fourni
- **Supervised classification** : Modéliser la normalité et la normalité. Requier d'avoir des données labélisées.
- **Semi-supervised classification** : Modéliser seulement la normalité ou alors l'anormalité. Utilisable pour des données figées et non figées, apprend et s'améliore au fur et à mesure que les données arrivent.

Comment traiter un outlier ?

On a deux approches possibles pour traiter un outlier :

- **Le diagnostique** : on analyse les outliers et on les enlève ou pas
- **L'accomodation** : on garde les outliers quoiqu'il se passe



Victoria J.Hodge (2017)

A Survey of Outlier Detection Methodologies

http://book.itep.ru/depositary/security/anomaly/Hodge+Austin_OutlierDetection_AIRE381.pdf