Big Data Analytics

Détection d'Anomalies

Détection d'Anomalies

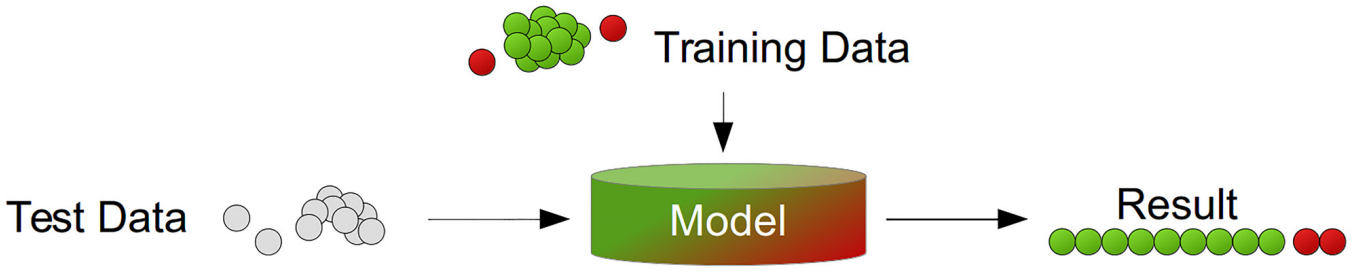
Détection :

- de Fraude
- d'Intrusion/Fuite (physique ou électronique)
- Santé (biologique, geologique, machine, ...)

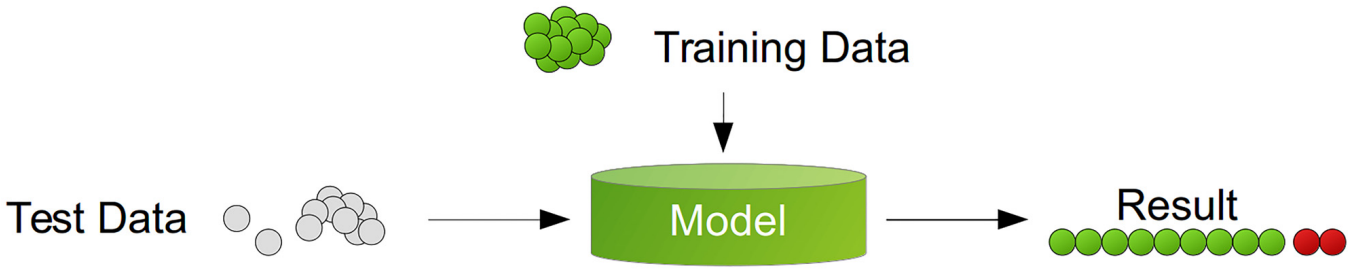
Définition

- une anomalie diffère de la norme par ses features
- les anomalies sont rares comparées aux instances normales

Modes de détection d'anomalie



(a) Supervised anomaly detection



(b) Semi-supervised anomaly detection

(c) Unsupervised anomaly detection

Détection d'Anomalies : Supervisé

Problème de classification normal.

Réseaux de neurones et SVM très performants.

Détection d'Anomalies : Semi-Supervisé

Détection de nouveauté.

Pas traité ici.

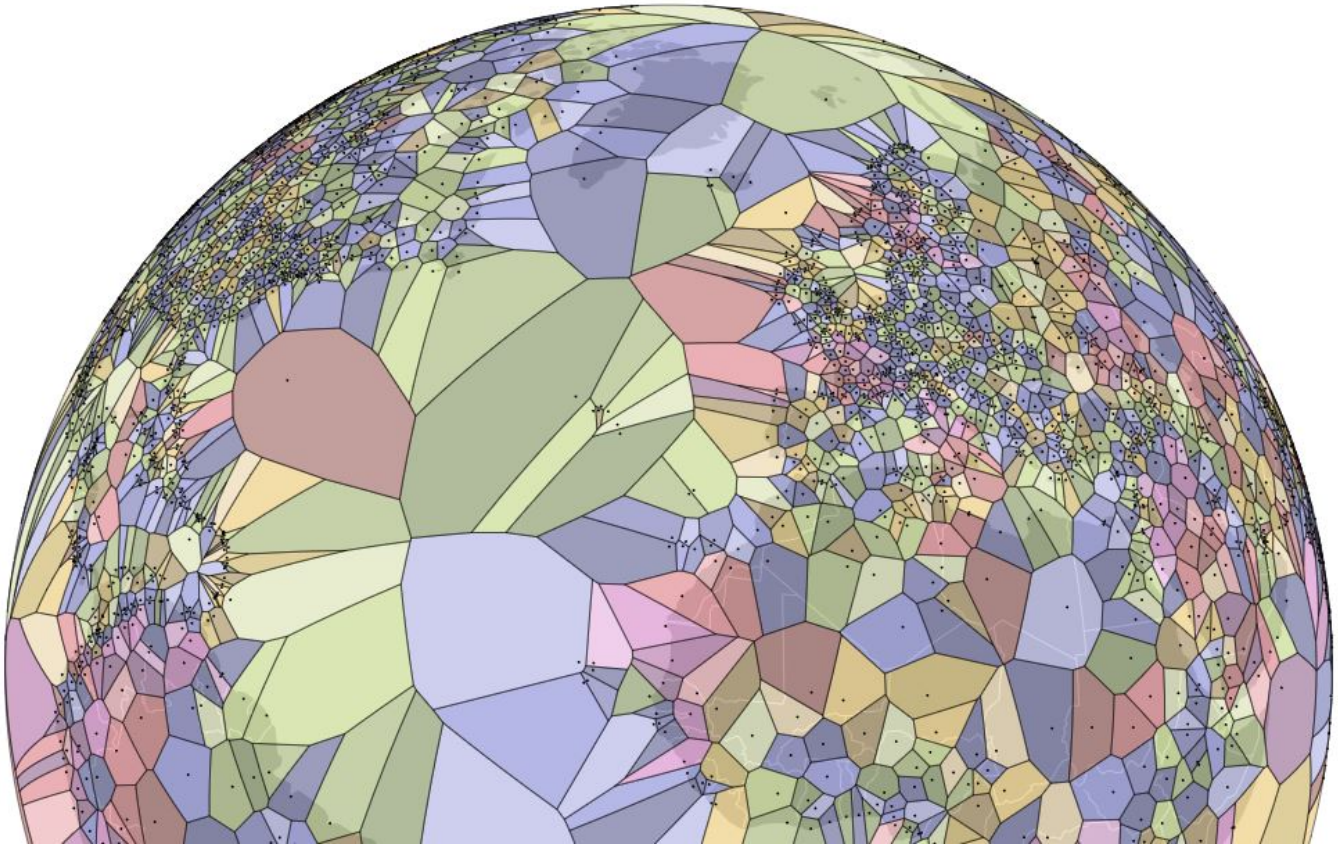
One-class SVM très utilisé.

Détection d'Anomalies : Non-Supervisé

De nombreuses méthodes :

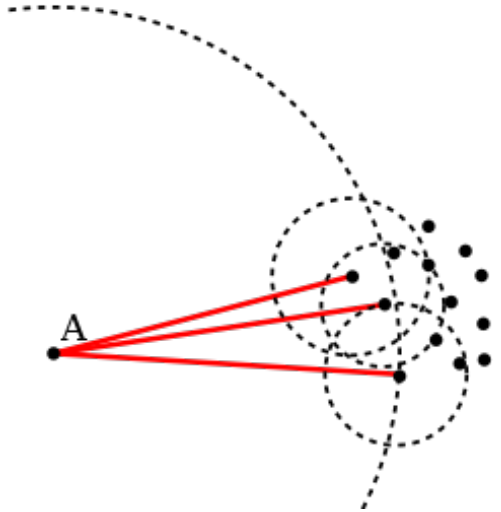
- Local Outlier Factor (LOF)
- Unweighted Cluster-Based Outlier Factor
- Isolation Forest
- Autoencoder
- ...

Détection d'Anomalies



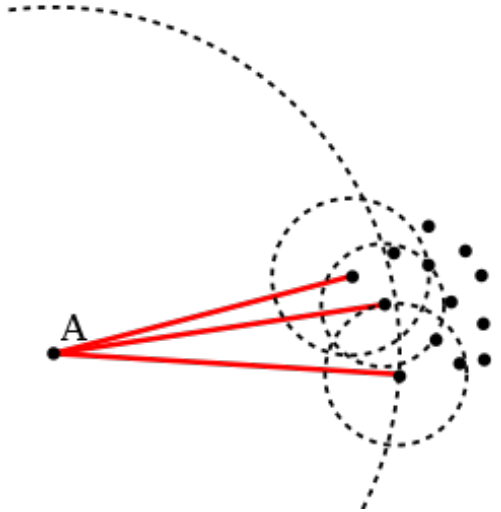
Local Outlier Factor

- anomalies locales



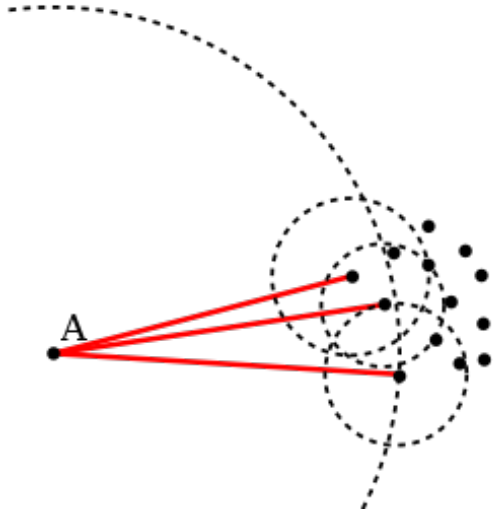
Local Outlier Factor

- anomalies locales
- basé sur les k voisins du point



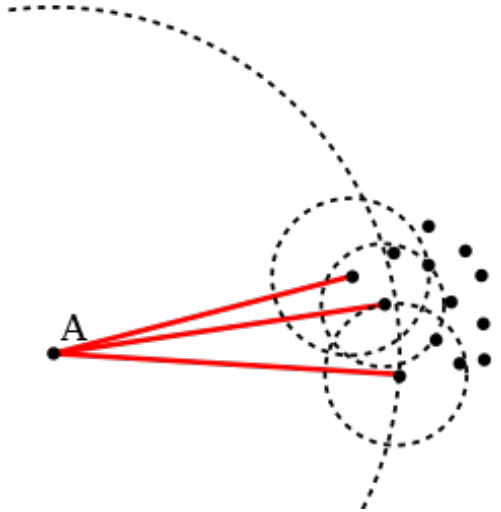
Local Outlier Factor

- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins



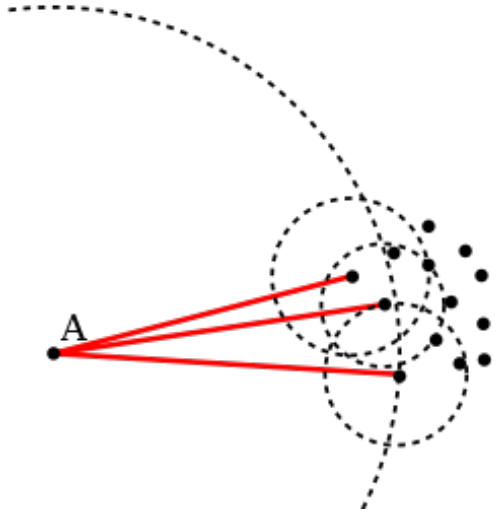
Local Outlier Factor

- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins
- calcule un ratio moyen d'atteignabilité du point et de ses voisins



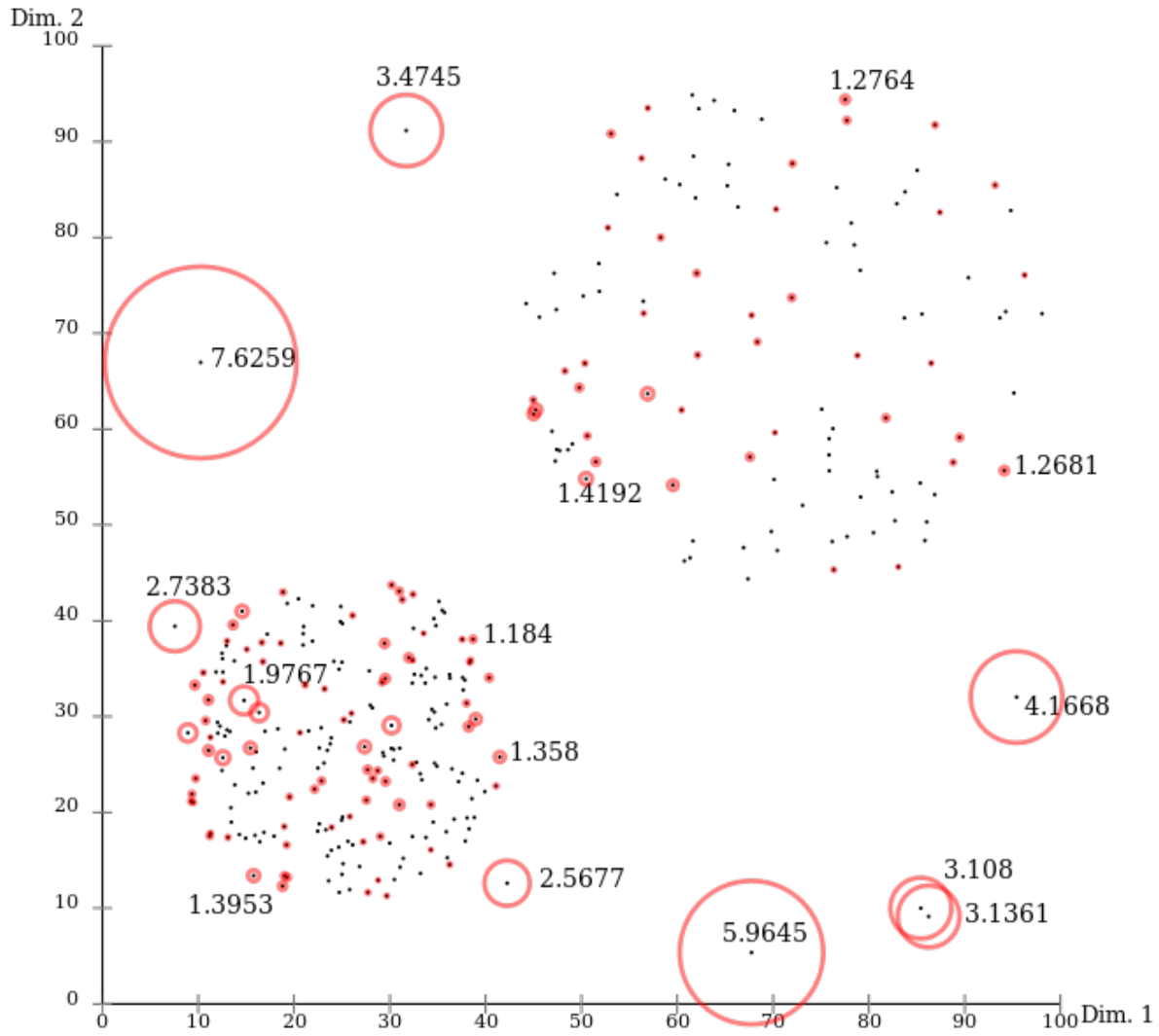
Local Outlier Factor

- anomalies locales
- basé sur les k voisins du point
- définit une « atteignabilité » par les distances de ces voisins
- calcule un ratio moyen d'atteignabilité du point et de ses voisins



$\rightarrow$ Anomalie si le ratio moyen d'atteignabilité est beaucoup plus faible que celui de ses plus proches voisins

Local Outlier Factor



Désavantages

- lent (quadratique)
- a des à priori sur la distribution des données

Isolation tree

- arbre aléatoire (comme random forest mais le split est aléatoire, ExtraTree)
- but : isoler une anomalie plus vite qu'un exemple normal
- petit chemin pour arriver à une feuille : anomalie

$\rightarrow$ Se sert du fait que les features des anomalies ne sont pas distribuées comme les autres.

Isolation forest

- forêt d'isolation trees
- construits sur des sous-échantillons sans replacement des données
- sous-échantillons plus petits que dans random forest typiquement, pour mieux isoler les anomalies
- converge souvent vite : 100 arbres souvent suffisants

Isolation forest