Détection d'anomalies

Module 6

Objectifs

Objectifs

- comprendre les différents modes de détection d'anomalie
- détecter des anomalies locales et globales
- utiliser les méthodes applicables aux grands datasets

Introduction

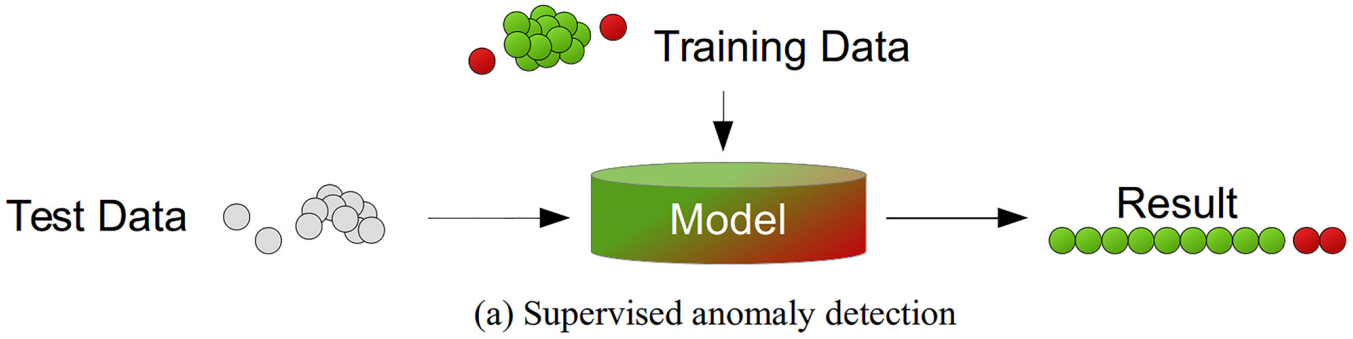
Introduction

- avant, beaucoup utilisée en préprocessing. Pourquoi ?
- récemment beaucoup moins le cas. Pourquoi ?
- très utilisée en :
  - détection d'intrusion
  - détection de fraude
  - prévention de fuite de données
  - monitoring de patients

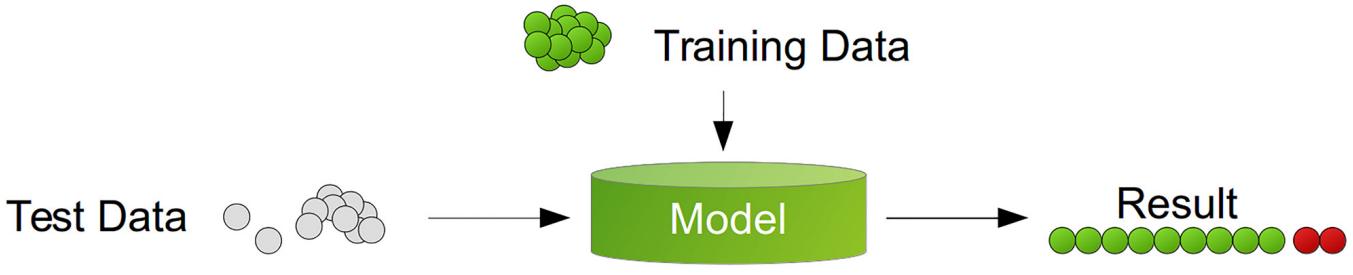
Définition

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

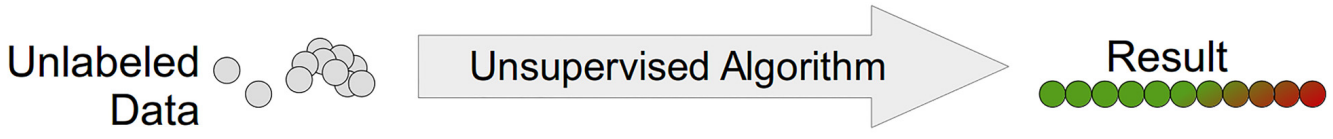
Modes de détection d'anomalie



(a) Supervised anomaly detection



(b) Semi-supervised anomaly detection



(c) Unsupervised anomaly detection

Mode supervisé

Problème de classification normal. Réseaux de neurones et SVM très performants.

Qu'en est-il des arbres ?

Modes semi-supervisé

Détection de nouveauté. Pas traité ici. One-class SVM très utilisé.

Mode non-supervisé

Méthodes très nombreuses :

- statistiques
- par voisinnage
- par réseaux de neurones
- par clustering
- par arbres

Souvent lourdes à calculer

Mode non-supervisé

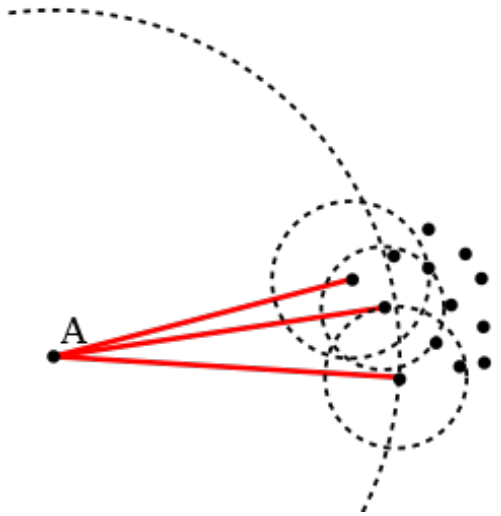
Programme :

- méthode classique (lente)
- méthode par clustering
- méthode statistique
- méthode par arbres
- méthode par réseaux de neurones

Local Outlier Factor

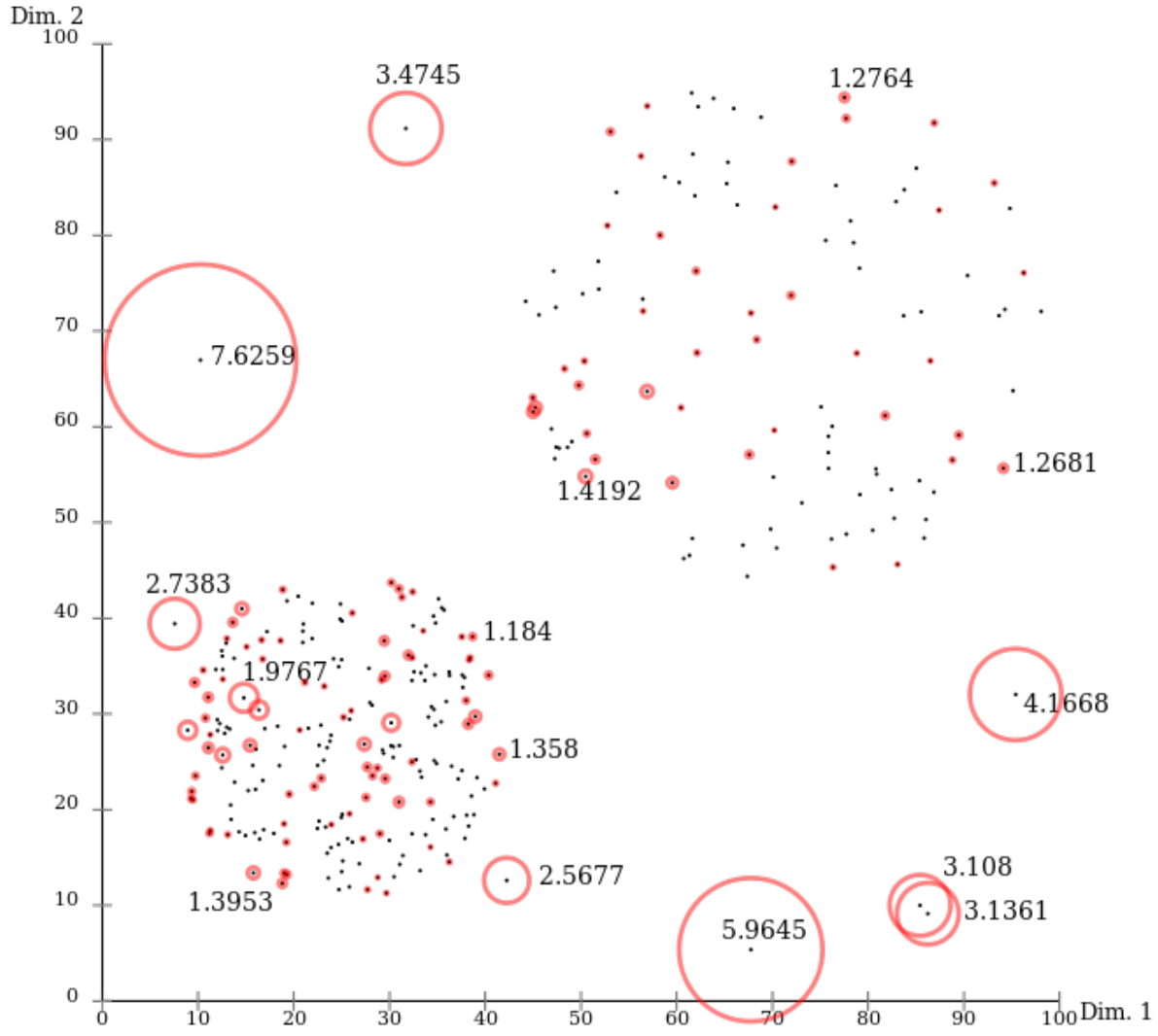
Local Outlier Factor

- 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



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

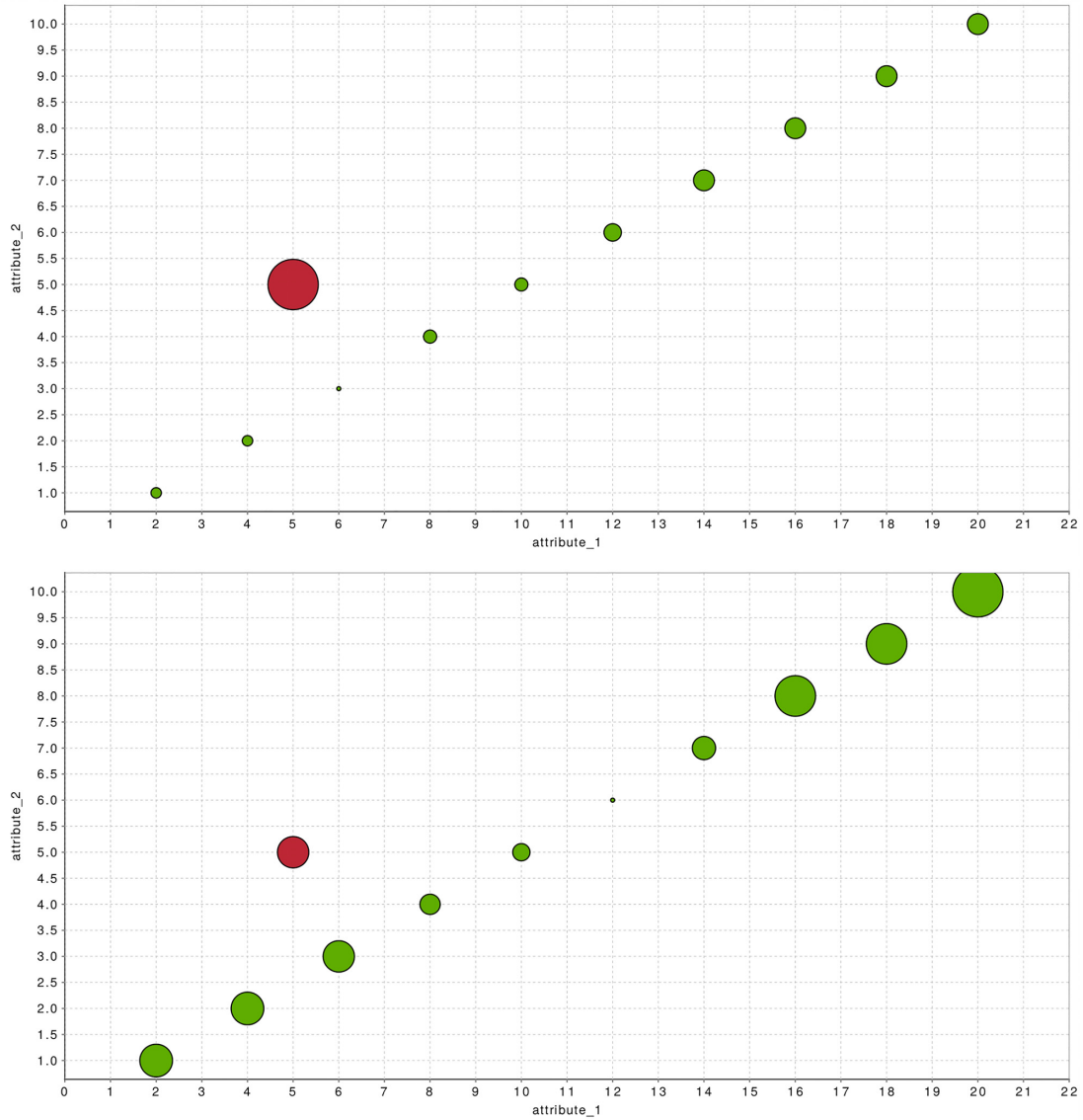
Local Outlier Factor



Désavantages

- lent (quadratique)
- a des priors sur la distribution des données

Prior



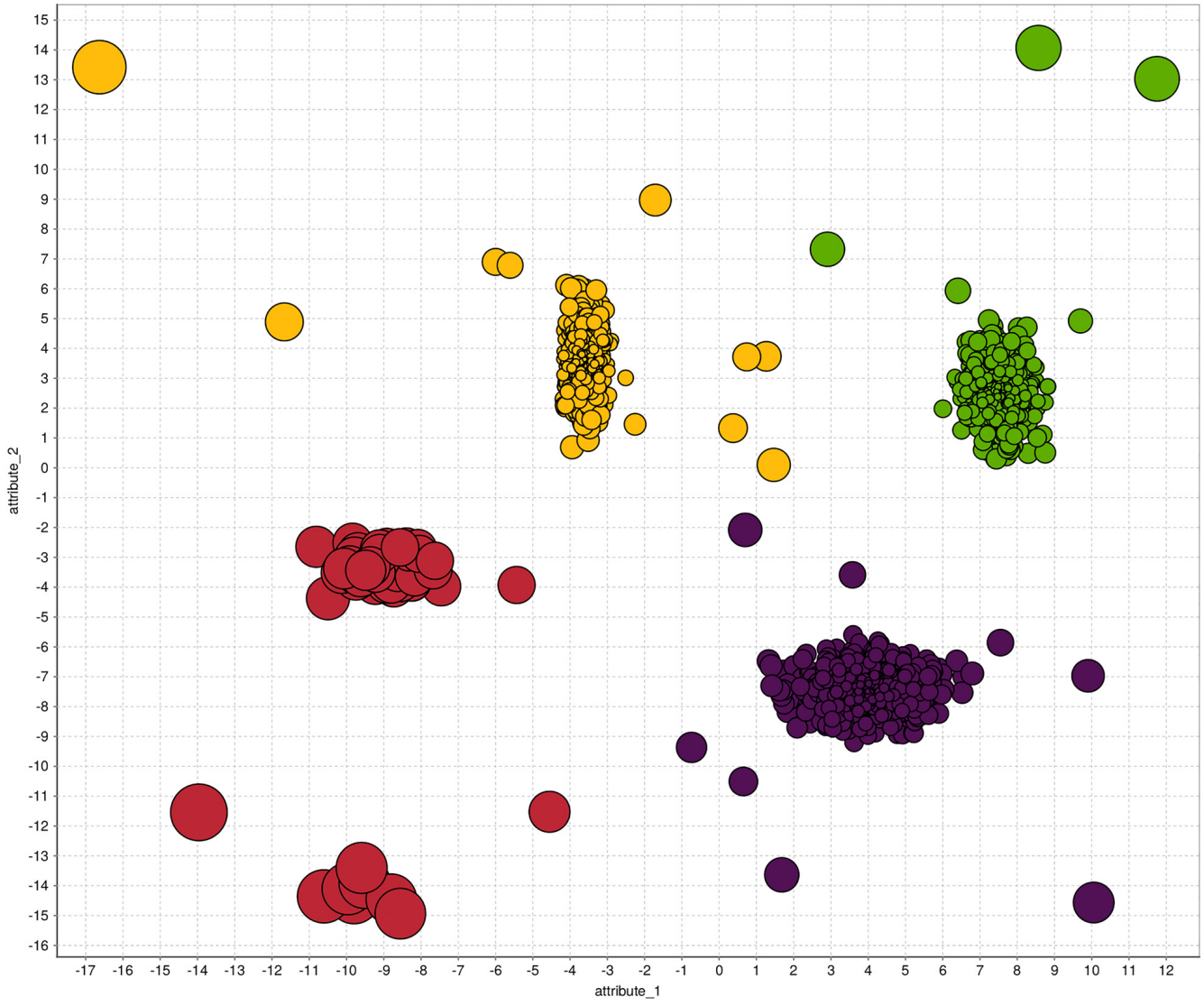
Unweighted Cluster-Based Outlier Factor

Unweighted Cluster-Based Outlier Factor

- anomalies globales
- fonctionne comme LOF mais voisinnage = cluster
- clusters calculés avec k-means

$\rightarrow$ Anomalie si le ratio moyen d'atteignabilité est beaucoup plus faible que celui dans son cluster

Unweighted Cluster-Based Outlier Factor



Histogram-based Outlier Score

Histogram-Based Outlier Score

- calculer un histogramme pour chaque feature
- $HBOS(x) = \sum_{f \in F} \log(\frac{1}{hist_f(x)})$
- linéaire sur les données au train, instantané au test
- suppose l'indépendance des features (!!!)
- efficace sur les anomalies globales

Isolation forest

Isolation tree

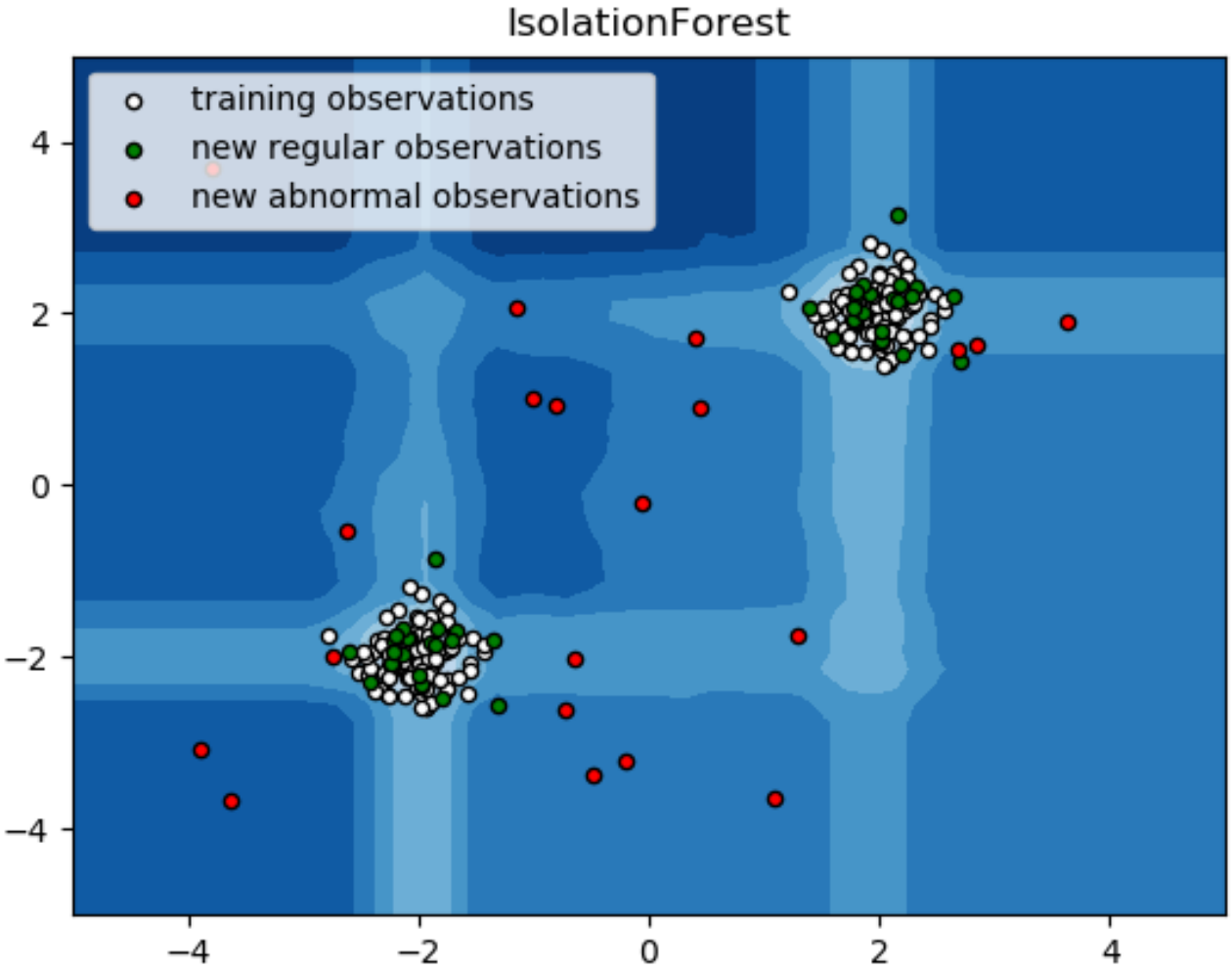
- arbre aléatoire (comme random forest mais le split est aléatoire, ExtraTree)
- but : isoler une anomalie plus vite qu'un exemple normal
- petit chemin pour arriver à une feuille : anomalie

$\rightarrow$ 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

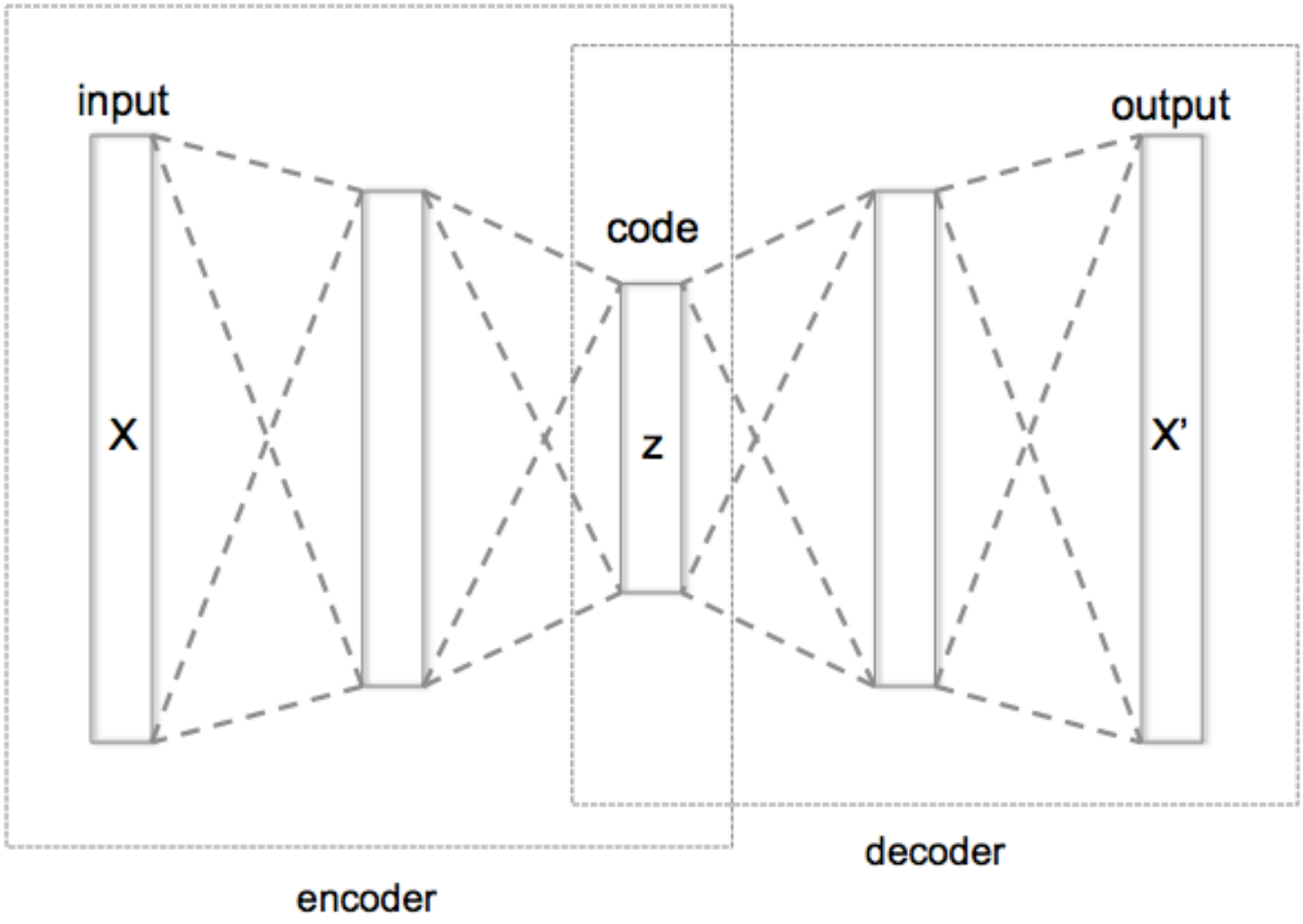


Auto-encodeurs

Introduction

- auto-encodeur = réseau de neurone
- input = output : le réseau apprend à reproduire
- pénalisé quelque part pour éviter la copie
- anomalie si le réseau reproduit mal

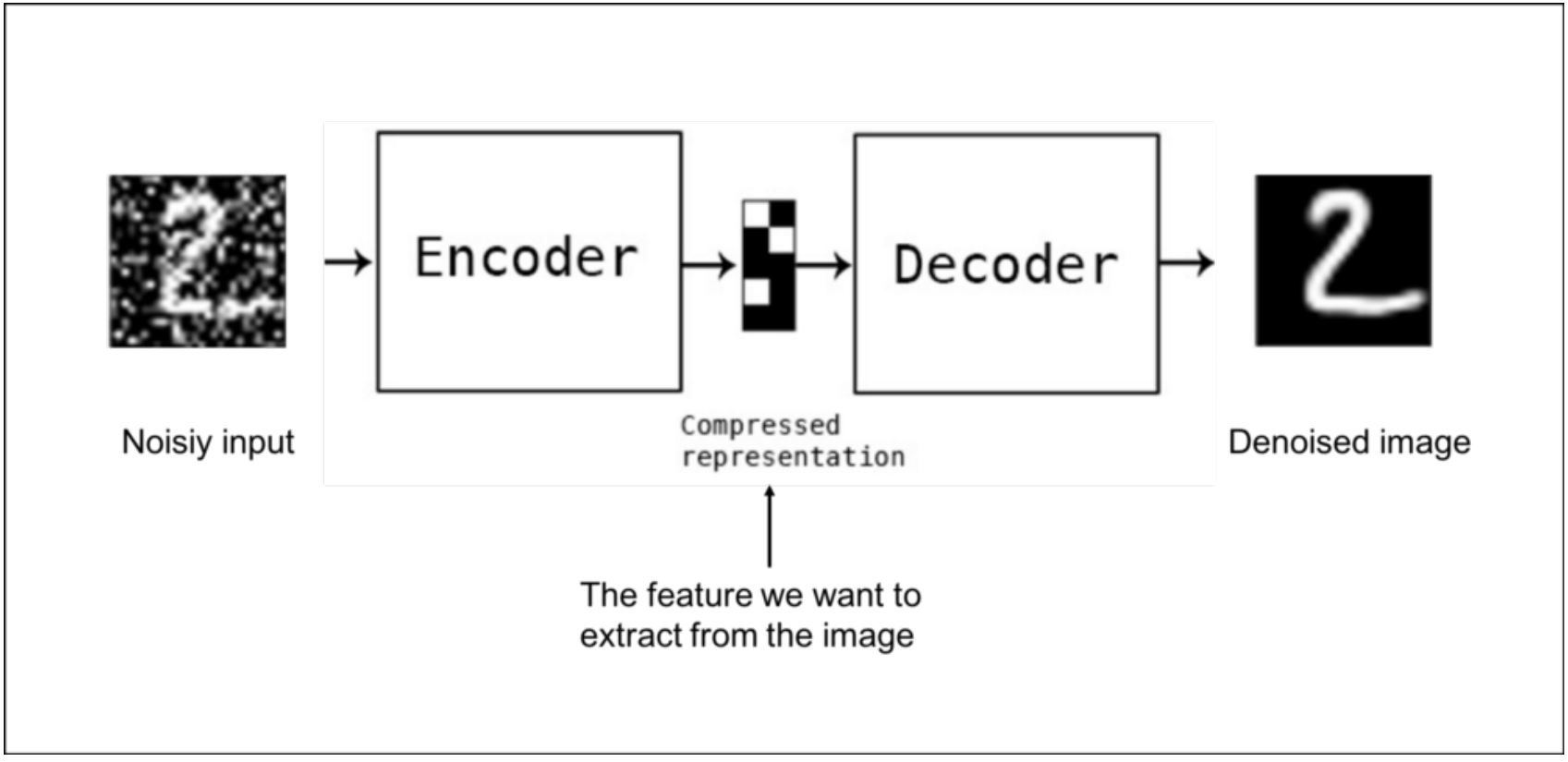
Introduction



Auto-encodeur « standard »

- apprend parce que $\mathbf{z}$ est plus petit que $\mathbf{X}$: compression
- dur à entrainer : éviter la mémorisation
- très faisable cependant avec un recherche d'hyperparamètres

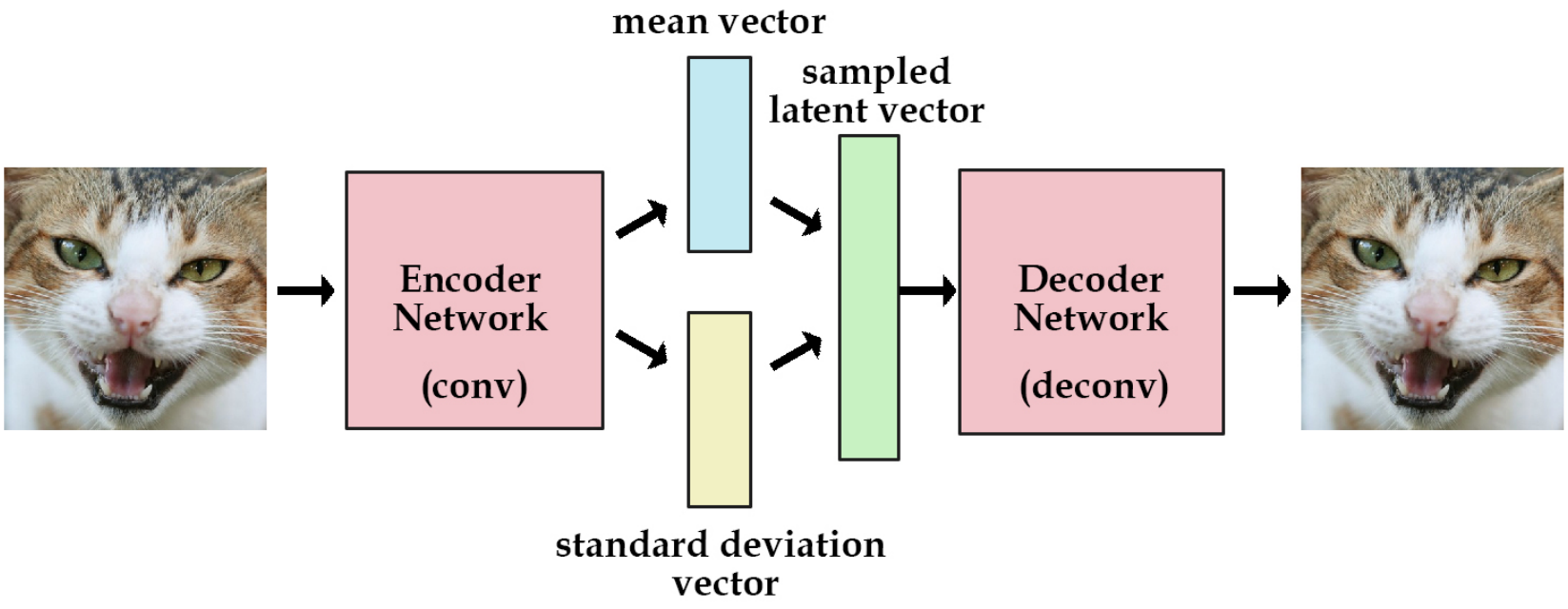
Auto-encodeur débruiteur



Auto-encodeur débruiteur

- apprend parce que $\mathbf{X}$ est bruité
- (apprend parce que $\mathbf{z}$ est plus petit que $\mathbf{X}$: compression)
- plus facile à entrainer : la mémorisation devient compliquée pour le réseau, en fonction du type de bruit

Auto-encodeur variationnel



Auto-encodeur variationnel

- apprend parce que $\mathbf{z}$ est une gaussienne : mémorisation dure
- (apprend parce que $\mathbf{z}$ est plus petit que $\mathbf{X}$: compression)
- intéressant : vecteurs des écart-types = très bonne information pour les anomalies

$\rightarrow$ Deux moyens de scorer une anomalie : erreur de reconstruction ou écarts-types élevés de $\mathbf{z}$

Conclusion

Conclusion

- plusieurs modes de détection d'anomalies
- méthodes globales ou locales
- état de l'art : isolation forest et auto-encodeurs variationnels