Détection d'anomalies

Qu'est-ce que la détection d'anomalie ?

La détection d'anomalie est une composante de la **fouille de données** qui cherche à déterminer si un « point extrême » est une anomalie ou du bruit, c'est à dire un point inhabituel.
La détection d'anomalies se fait à l'aide de **modèles**.
Chaque modèle résout un type de problème différents. Nous allons ici présenter les différent modèles existants.

Comment créer une IA pour détecter les anomalies ?

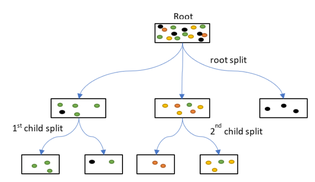
Pour écrire une IA qui va travailler sur des données **annotées**, on va utiliser les méthodes de **recherche supervisée**. Notre IA va « apprendre » à distinguer les anomalies sur un **sous-ensemble** des données et restituer ses connaissances quand nous l'utiliserons. Un **paramétrage** différent donnera un **score** différent représenté par une **matrice de confusion**.

Le modèle bayes naïf

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

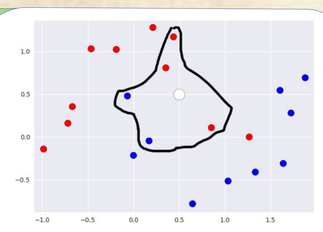
Le modèle est basé sur le **Théorème de Bayes**. La stratégie est de considérer des « classes de valeurs » pour lesquelles le classificateur saura prédire si oui ou non la valeur appartient à la classe de données.

Le modèle de l' arbre de décision



L' arbre de décision est une représentation des **choix** en arborescence qui conduit à une des classes de données.

Le modèle des K plus proches voisins



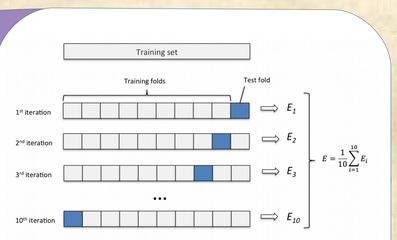
Ce modèle consiste en un regroupement des k points les plus proches du point à tester. Notre point prend alors la classe la plus présente parmi ses voisins.

La technique du « Voting »

$$\sum_{i=1}^{N}$$

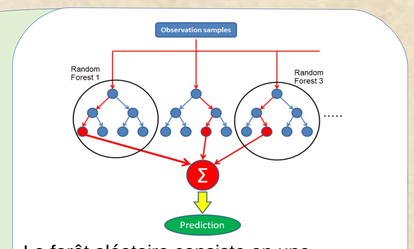
Le voting consiste à comparer différents modèles entre eux (comme le «Bagging») et à donner un poids à chacun d'entre eux selon leurs performances. Ensuite on choisit au hasard une des méthodes pour donner le résultat.

La technique de la Validation croisée



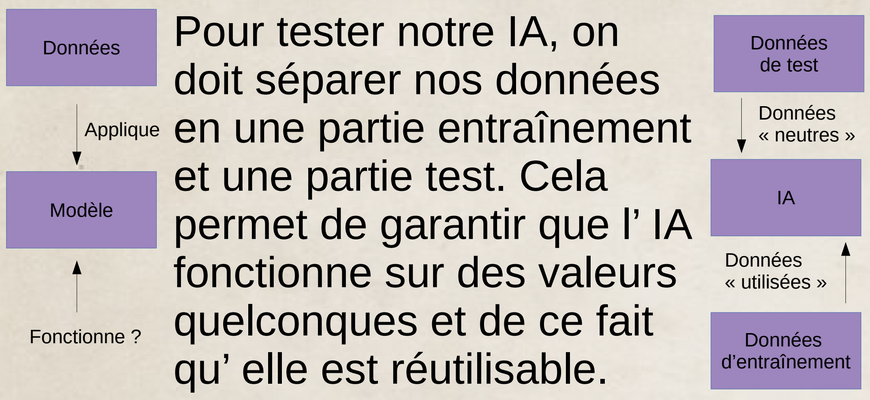
La validation croisée consiste à créer des sous-ensembles et lancer des tests en parallèle pour obtenir un résultat final de meilleure **précision**

Le modèle de la forêt aléatoire



La forêt aléatoire consiste en une simulation de plusieurs arbres de décision dont on prend la moyenne des résultats comme valeur finale

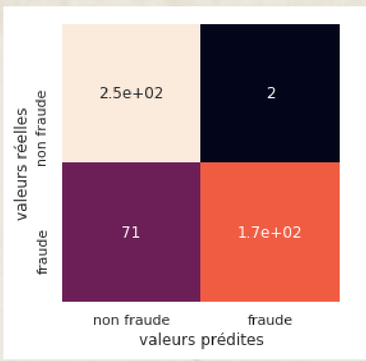
Comment tester son IA ?



Pour tester notre IA, on doit séparer nos données en une partie entraînement et une partie test. Cela permet de garantir que l' IA fonctionne sur des valeurs quelconques et de ce fait qu' elle est réutilisable.

Comment donner un score ?

Le résultat de notre exécution donne une **matrice de confusion**. On la transforme alors en score qui sera renvoyé comme résultat de notre modèle.



Comment paramétrer correctement ?

L' utilisation d'un modèle demande un paramétrage qui modifie grandement la puissance du modèle.

Pour aller plus loin

Les méthodes décrites ci-dessus ne correspondent qu'à un apprentissage **supervisé.** Il existe d'autres méthodes pour le cas non-supervisé mais aussi des cas plus complexes où les données ne sont pas indépendantes et où il faut traiter ces données différemment.

RENAULT Alexis - PITROU-FAIZEAU Adrien
BREGEON Dany - MONET Loïs

Master 1 Informatique IMIS Université Orléans
Année 2019