Veille Machine Learning

Aurélien GUILLEMOT

26/05/2021

Problématique : Peut on prédire qu'une personne est dépressive en fonction de son mode de vie ?

Afin d'estimer le risque de dépression chez une personne, je vais utiliser des algorithmes de **classification**. Je peux, grâce à mon jeu de données, observer les différents facteurs de dépression afin de savoir si le contexte social, familial, culturel, financier … influent sur la santé mentale.

Pour classifier ces données, je peux utiliser les éléments suivants :

1. L'arbre de décision
2. La forêt d'arbre aléatoire
3. Les K plus proches voisins
4. La classification Naive Bayésienne
5. La matrice de confusion

Pour le choix de l'algorithme, je suis allé étudier lequel de « la forêt d'arbres aléatoire » et du « K plus proches voisins » (voir les sources).

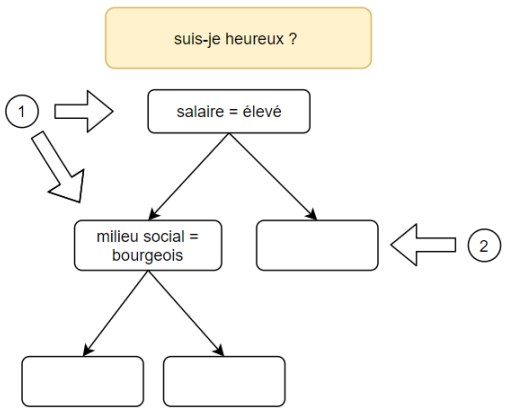
Choix de l'algorithme

J'ai décidé d'utiliser la forêt d'arbre aléatoire.

Explication de l'algorithme :

La forêt d'arbre aléatoire est un ensemble d'arbres de décisions. L'arbre de décision fonctionne de la manière suivante :

1. « Sur l'arbre, chaque question correspond à un **nœud** »
2. « Les **feuilles** sont les Nœuds terminaux de l'arbre »



L'objectif d'un arbre de décision est d'observer des données en fonction des règles défini. Autrement dit, on parcourt un arbre pour mettre une observation dans une boite.

Pour en revenir sur la forêt ; c'est un ensemble d'observations faite sur les arbres de décision. Au lieu de n'avoir qu'un seul parcours au travers duquel on essaye de répondre à notre problématique, on multiplie les arbres pour être plus précis. Attention en revanche à ne pas faire une forêt trop grande dans l'objectif de correspondre à 100% au jeu de données. L'intérêt est de définir un modèle viable, même sur un nouveau jeu de données. Si la forêt est trop précise, on fait de « l'overfitting », c'est-à-dire qu'on risque de rendre le modèle difficile à appliquer à un autre jeu.

Application de l'algorithme sur mon projet :

C'est une classification, car je ne peux répondre que par « oui » ou « non ». Les *feuilles* de mon arbre auront donc soit la valeur « oui » ou « non ».

Dans le cadre de mon étude, la forêt d'arbre aléatoire me permettra de mettre en perspective les différents facteurs pouvant définir si une personne est dépressive ou non. Il est impossible de définir un simple schéma qui permettra de mettre les gens dans une case. Cela est dû au trop grand nombre de paramètres nécessaires pour le définir : il y a l'âge, le niveau d'éducation, le salaire, le sexe, le mariage …

On entrainera donc notre jeu de test pour travailler avec N arbre afin de tester le jeu de données sur l'ensemble. L'objectif est ensuite de savoir quel arbre est le plus représenté.

Dans le cadre de mon projet, je possède une vingtaine de données explicatives. Je pense donc utiliser une forêt de 100 arbres. Ce choix est motivé par la valeur choisi lors du « projet fraude ». Nous avions entre 10 et 15 données explicatives et nous avons choisi de faire 100 arbres, nous sommes sur le même intervalle.

Source :

Global :

- https://datakeen.co/8-machine-learning-algorithms-explained-in-human-language/
- Intro_ML_H3.pdf

Forêt d'arbre aléatoire :

- https://fr.wikipedia.org/wiki/For%C3%AAt_d%27arbres_d%C3%A9cisionnels
- http://cedric.cnam.fr/vertigo/cours/ml2/coursForetsAleatoires.html
- https://datascientest.com/random-forest-definition
-

K plus proches voisins :

- https://fr.wikipedia.org/wiki/M%C3%A9thode_des_k_plus_proches_voisins
- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html