Active Learning

October 15, 2019

L'active learning est un outil très puissant utilisé lorsque l'on manque de données étiquetées.

1 A quoi sert l'active learning ?

Petit exemple d'utilisation de l'active learning : mettre en place des limites de décisions, permettant de séparer deux catégories.

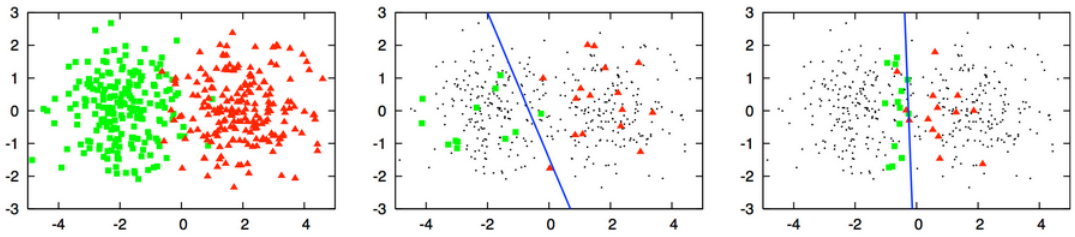

Figure 1: Exemple de datas où l'on veut appliquer des limites de décisions afin de mettre en place deux catégories.

Cependant on suppose qu'on ne connait pas les étiquettes des échantillons/données. Essayer de trouver l'étiquette de chaque point serait très couteux en temps. Pour cela on échantillonne un petit sous-ensemble de points et on trouve l'étiquette de ces données. Ces points, étiquetés, vont servir de classificateurs (données de formation).

⇒ Cette technique permet de récupérer un meilleur sous-ensemble d'échantillons afin de les rendre meilleur classificateur à partir de méthodes précises.

2 Définition et concepts

Si un algorithme d'apprentissage peut choisir les données dont il veut apprendre, il peut être plus performant que les méthodes traditionnelles avec beaucoup moins de données pour la formation. Ces méthodes sont sous la forme de tâches qui impliquent la collecte d'une grande quantité de données échantillonnées au hasard à partir de la distribution sous-jacente. De plus, l'utilisation de vaste ensemble de données pour former un modèle est mise en place afin d'effectuer une sorte de prédiction = apprentissage passif.

L'une des tâches les plus longues de l'apprentissage passif est la collecte de données étiquetées. Dans de nombreux contextes, il peut y avoir des facteurs limitatifs qui entravent la collecte de grandes quantités de données étiquetées.

Pour prédire par exemple si un patient aura un cancer du pancréas. On n'a pas forcément l'occasion de faire subir des examens à un petit nombre de patients aléatoires afin de recueillir des informations. Plutôt que de sélectionner des patients au hasard, on peut sélectionner les patients en fonction de certains critères. Le processus de sélection de ces patients (instances) sur la base des données que nous avons recueillies jusqu'à présent est appelé apprentissage actif.

Dans l'apprentissage actif, il y a 3 scénarios ou situations dans lesquels l'apprenant va interroger les étiquettes des instances.

3 Les différentes méthodes existentes

Dans l'apprentissage actif, on peut retrouver 3 situations dans lesquels l'apprenant va interroger les étiquettes des instances :

- Membership Query Synthesis : L'apprenant génère une instance (à partir d'une distribution naturelle sous-jacente). Par exemple, si les données sont des images de chiffres, l'apprenant crée une image similaire et cette image créée est ensuite est envoyée à l'oracle afin qu'il l'étiquette.

Figure 2: Membership Query Synthesis

- Stream-Based Selective Sampling : Ici, on fait l'hypothèse que l'obtention d'une instance non étiquetée est gratuit. A partir de cette hyptohèse, chaque instance non étiquetée sont sélectionnées une à la fois. L'apprenant décide s'il souhaite interroger l'étiquette de l'instance ou la rejeter en fonction de son caractère informatif. Pour déterminer ce caractère, il faut utiliser une stratégie de requête.

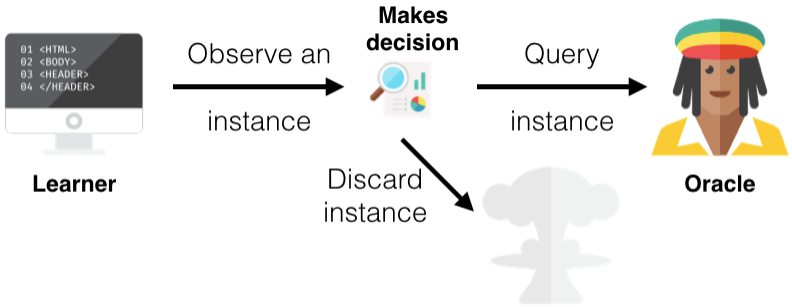

Figure 3: Stream-Based Selective Sampling

- Pool-Based sampling : On suppose ici qu'il y a un grand nombre de données non étiquetées comme dans le cas du "Stream-Based Selective Sampling". Selon l'information que contiennent ces données, des exemples sont tirés de ces données (les plus informatives sont sélectionnées). Ici, en reprenant l'exemple précédemment énoncé, toutes les images de chiffres non étiquetées sont classées puis la ou les plus informatives isntances sont sélectionnées et leurs étiquettes sont demandées à l'oracle.

  $\Rightarrow$ C'est la méthode la plus utilisée dans le cadre de l'apprentissage actif.

Figure 4: Pool-Based sampling