



Module 1 - Introduction à la science des données

Exercices - Solutions

Exercice 1 :

L'apprentissage machine est une branche de la science des données.

Exercice 2 :

Par définition, les données massives se caractérisent par trois composantes : volume, variété et vélocité. Ceci dit que chaque composante représente un défi en soi. Bien qu'il existe des outils avancés à savoir les frameworks Spark et Hadoop pour faire face aux composantes volume et vélocité, la composante variété reste encore un défi important auquel il faudra faire face.

Exercice 3 :

La communication des résultats est l'une des tâches les plus importantes d'un scientifique de données. Cette tâche devient beaucoup plus difficile pour une audience non technique. Pour ce faire, le scientifique de données doit :

- Bien comprendre le problème. La compréhension du problème permet de simplifier les choses pour une audience non technique.
- Essayer de formuler le problème d'apprentissage automatique dans le même langage couramment utilisé par l'audience. Par exemple, si l'audience est composée des hommes d'affaires, le scientifique de données peut parler des coûts et bénéfices au lieu des classes.
- Faire une présentation pour chaque type d'audience et éviter d'utiliser la même présentation pour tout le monde.
- Éviter les termes trop techniques, par exemple optimisation des hyperparamètres d'un modèle d'apprentissage profond. Phrase très difficile à comprendre par une audience non technique.
- Décrire clairement les données, pourquoi et comment elles ont été collectées. Décrire aussi, si c'est possible, quelles données il serait utile de collecter et d'utiliser.
- À l'affichage des graphiques et tableaux, il faut toujours réitérer sur l'idée principale pour que l'audience soit capable de faire le lien entre le problème et le résultat.
- Utiliser toujours des exemples pratiques pour expliquer le modèle.

- Présenter clairement l'impact des résultats sur le domaine d'affaire en question.
- Ne pas quitter directement après la présentation et donner toujours du temps à l'audience pour poser des questions et formuler une rétroaction.