



Module 1 - Introduction à l'apprentissage machine

Exercices - Consignes

Cette première série d'exercices à deux objectifs. Le premier est de se familiariser avec l'utilisation des données. Le deuxième est une prise en main du logiciel R et certaines notions de base en apprentissage machine.

Exercice 1

1. Définissez de manière concise l'apprentissage machine.
2. Donnez un exemple concret d'apprentissage machine supervisé pour un domaine de votre choix.
3. Donnez un exemple concret d'apprentissage machine non supervisé pour un domaine de votre choix.

NB : Les exemples doivent être autre que ceux fournis dans le texte de référence du Module 1.

Exercice 2*

Déterminez le type de chacune des variables suivantes (variable quantitative, qualitative, nominale ordinale, continue, discrète). Expliquez votre réponse.

- Age.
- Poids.
- Sexe.
- Citoyenneté.
- Tour de taille.
- Couleur des yeux.
- Nombre de langues parlées.
- Nombre d'enfants dans la famille.

Exercice 3*

Nous disposons d'une base de données issue d'une étude sur la surveillance de la qualité de l'air et des conditions météorologiques en France durant l'été 2001. Cette base de données décrit des mesures journalières des 13 variables suivantes :

- Le maximum de concentration d'ozone (MaxO3) observé sur la journée en $\mu\text{gr}/\text{m}^3$.
- La température observée à 9h, 12h et 15h (respectivement T9, T12 et T15).
- Les valeurs de nébulosité Ne9, Ne12 et Ne15 qui correspondent à la fraction de ciel couvert par les nuages, respectivement, à 9h, 12h et 15h.
- La teneur maximum en ozone observée la veille (MaxO3v).
- L'orientation du vent à 12h (Vent).
- La possibilité ou non de précipitations (Pluie).

Ces données se trouvent dans le fichier `DonneesOzone.xls`. Elles constituent une adaptation d'un corpus de données disponible sur le site du laboratoire de mathématiques appliquées de l'Agrocampus Ouest (<http://math.agrocampus-ouest.fr/infoglueDeliverLive/enseignement/support2cours/donnees>).

Afin d'explorer ces données, nous vous proposons de réaliser les étapes suivantes :

1. Chargez dans R le contenu de la base de données en utilisant la commande `xlsread`. Consignez et commentez les lignes de code dans le fichier R.
2. Déterminez les différentes variables et leurs types. Justifiez votre réponse.
3. Déterminez le nombre d'individus. Consignez et commentez les lignes de code à utiliser dans le fichier R.
4. Pour chacune des variables quantitatives, calculez les caractéristiques de tendances centrales et les caractéristiques de dispersion. Présentez les résultats sous la forme d'un tableau et commentez les valeurs.
5. Nous désirons faire une analyse graphique de la température et de la nébulosité observée à 9h, 12h et 15h.

- (a) Représentez graphiquement le diagramme en boîtes des variables **T9**, **T12**, **T15**, **Ne9**, **Ne12** et **Ne15**. Analysez ces graphiques.
 - (b) Consignez et commentez les lignes de code dans le fichier R.
 - (c) Déterminer, s'il y a lieu les valeurs aberrantes pour les variables **Ne9**, **Ne12** et **Ne15**. Expliquez votre réponse.
6. Nous désirons faire une analyse graphique des variables **Vent** et **Pluie**.
- (a) Représentez graphiquement le diagramme circulaire des variables **Vent** et **Pluie**.
 - (b) Consignez et commentez les lignes de code dans le fichier R.
 - (c) Analysez ces graphiques.

Exercice 4*

Nous disposons d'un jeu de données qui décrit un ensemble de 30 individus dans le fichier **JeuDonnees.xls**. Pour chaque individu nous disposons de la classe réelle et de la classe prédite.

1. Chargez dans R le contenu de la base de données en utilisant la commande **xlsread**.
2. Représentez la matrice de confusion en utilisant R.
3. Calculez à l'aide de R le taux de bonne classification par classe et le taux de classification global.