Machine Learning
& Intelligence Artificielle

Manuel Simoes
manuel.simoes@cpc-analytics.fr

- Préparation des données -

Préparation des données : les types variables

Valeur Continue

Une donnée numérique, entière ou réelle qui peut prendre une infinité de valeurs possibles. Ces variables peuvent entrer dans les calculs numériques comme les opérations arithmétiques, la standardisation…

Variable catégorielle

C'est une données qui appartient à un ensemble fini de valeurs discrètes appelées catégories (ex. bleu/rouge/jaune). On peut encoder ces catégories par des valeurs numériques (ex. 0 pour bleu/ 1 pour rouge/ 2 pour jaune), mais pas leur appliquer de notions mathématiques, comme letri ou les opérations arithématiques.

Variable Ordinale

C'est une donnée à mi-chemin entre continue et catégorielle. Elle a des valeurs limites (des catégories), mais peut être triée (ex nombre d'étoile d'appréciation d'un film)

Préparation des données : Discrétisation des données

Discrétisation

La discrétisation des données fait référence au processus de partitionnement des données continue en des données discrètes ou nominales (données discrètes ordonnées).
Pour cela on découpe les données continues en N partition de même largeur appelées classes.

Exemple :

120, 113, 128 ,108 , 136, 138, 132, 120, 138, 139, 131

On crée 4 classes de largeur 10 :

La classe 0 [100, 109] contient 1 élément.
La classe 1 [110, 119] contient 1 élément.
La classe 2 [120, 129] contient 3 éléments.
La classe 3 [130, 139] contient 6 éléments.

Préparation des données : Binarisation des données

Binarisation

La binarisation des données consiste à ne donner à un attribut que 2 valeurs possibles. Les valeurs dépassant un seuil (qu'on fixe) ont une valeur 1 et le reste 0.

Utile (en *feature engineering*) entre autres pour créer de nouvelles variables prédictives.

Préparation des données Encodage one-hot

Encodage one-hot

En apprentissage automatique, les algorithmes ne savent généralement pas fonctionner avec des variables qualitatives non numériques. Pour cette raison, une étape d'encodage *one-hot* est nécessaire. Il s'agit de convertir une variable qualitative ou catégorie que ayant *N* classes (valeurs possibles) en *N* variable suivant une forme bien déterminée. Ces nouvelles variables prennent la valeur 1 pour la bonne classe et 0 sinon.

Exemple

Voiture	couleur
Peugeot	rouge
Renaud	bleu

Voiture	rouge	bleu	vert	jaune
Peugeot	1	0	0	0
Renaud	0	1	0	0