2 Normalisation

2. Normalisation : pourquoi ?

- Mettre sur une même échelle toutes les données quantitatives
- Rendre les plages cohérentes entre les variables
- Avoir une influence similaire des variables sur les modèles
- Conserver les rapports de distance
- Améliorer les performances et stabiliser le modèle

2. Normalisation : comment ?

Scaler	Caractéristiques
Min Max Scaler sklearn.preprocessing.MinMaxScaler	Std faible, distrib. non normale, [0,1] Sensible oultiers
Standard Scaler sklearn.preprocessing.StandardScaler	Std 1, distrib. Normale centrée en 0 Sensible oultiers (Z-Score)
Robust Scaler sklearn.preprocessing.RobustScaler	Std faible, distrib. non gaussienne Insensible oultiers
Max Abs Scaler sklearn.preprocessing.MaxAbsScaler	Distrib. Éparse,centrée en 0, [-1,1] Insensible oultiers
Power Transformer Scaler sklearn.preprocessing.PowerTransformer	trouve le facteur d'échelle optimal pour stabiliser la variance et minimiser l'asymétrie grâce à l'estimation du maximum de vraisemblance. rendre les données plus gaussiennes.
Quantile Transformer Scaler sklearn.preprocessing.QuantileTransformer	1. Calcule la fonction de distribution cumulative de la variable 2. Il utilise ce cdf pour faire correspondre les valeurs à une distribution normale 3. Établit une correspondance entre les valeurs obtenues et la distribution de sortie souhaitée à l'aide de la fonction quantile associée Distrib. Normale, large dataset Insensible oultiers
Unit Vector Scaler sklearn.preprocessing.normalize	Mise à l'échelle se fait en considérant que le vecteur d'élément entier est de longueur unitaire. [0,1]
Normalizer Scaler sklearn.preprocessing.Normalizer	Normalise selon la norme L1 (manhatan) ou L2 (euclidienne)

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

$$x_{new} = \frac{x - \mu}{\sigma}$$

$$\frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$$

$$x' = \frac{x}{||x||}$$

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \sum_{i=1}^{n}|x_i - y_i|$$

2. Normalisation : algorithme ?

Calculer un vecteur de poids

- Scoring

Descente de gradient

- Logistic Regression
- Linear Regression
- Réseaux de neurones

Calculer des distances pour déduire le degrée de similarité de deux items

- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)
- K-Means (clustering…)
- Principal Component Analysis (PCA)

Tree-base algorithm :

- Gradient Boosted Decision Trees
- Regression Tree
- Classification Trees
- Random Forests

Linear Discriminant Analysis(LDA)

Naive Bayes