Science des données

Quentin Fortier

April 5, 2023

Représentation des données

Pour pouvoir avoir une notion de distance entre deux données, on se ramène à $\mathbb{R}^p$:

- Variable catégorielle (non numériques : genre, couleur, etc.) : on utilise souvent un vecteur avec un 1 et que des 0 (*one-hot vector*).
- Image : On passe d'une matrice de pixels avec n lignes, p colonnes à un vecteur de taille np.
- Son : Transformée de Fourier discrète.

Dans la suite, on suppose que $X \subseteq \mathbb{R}^p$.

Représentation des données

On représente classiquement l'ensemble des données (donc de vecteurs de $\mathbb{R}^p$) par une matrice `X` dont les lignes sont les données et les colonnes sont les attributs (coordonnées des vecteurs).

Python Données	Matrice
`X[i]` $i$ème donnée	$i$ème ligne
`len(X)` nombre de données	nombre de lignes
`X[i][j]` $j$ème attribut de la $i$ème donnée	élément ligne i, colonne j
`len(X[0])` nombre d'attributs	nombre de colonnes

Représentation des données

Exemple :