Science des données

Thibaut Cantaluppi

November 7, 2024

Science des données

La **science des données** (*data science*) a pour objectif d'extraire de l'information à partir de données brutes.

Science des données

La **science des données** (*data science*) a pour objectif d'extraire de l'information à partir de données brutes.

Exemples :

- Données sur des fleurs : longueur et largeur des pétales et des sépales.

Science des données

La **science des données** (*data science*) a pour objectif d'extraire de l'information à partir de données brutes.

Exemples :

- Données sur des fleurs : longueur et largeur des pétales et des sépales.
- Données sur les clients d'une banque : âge, sexe, épargne, ...

Représentation des données

Pour pouvoir avoir une notion de distance entre deux données, **on représente chaque donnée comme un vecteur** de $\mathbb{R}^p$.

Représentation des données

Pour pouvoir avoir une notion de distance entre deux données, **on représente chaque donnée comme un vecteur** de $\mathbb{R}^p$.

Exemple : chaque donnée de fleur peut être représentée par un quadruplet de $\mathbb{R}^4$ correspondant à la longueur et largeur des pétales et des sépales.

Les composantes de ce vecteur sont appelées les **attributs**.

Représentation des données

Parfois il est moins évident de représenter une donnée par un vecteur :

Représentation des données

Parfois il est moins évident de représenter une donnée par un vecteur :

- Variable catégorielle (non numérique : genre, couleur, etc.) : on utilise souvent un vecteur avec un 1 et que des 0 (*one-hot vector*).

Fleur	c_Red	c_Purple	c_Blue	Couleur	Vector
Fleur 1	0	0	1	"Blue"	001
Fleur 2	0	1	0	"Purple"	010
Fleur 3	1	0	0	"Red"	100
Fleur 4	0	1	0	"Purple"	010

  En pratique, on utilise les fonctions `get_dummies()` de la bibliothèque `Pandas` ou `OneHotEncoder0` dans `Scikit-learn`.

- Image : On passe d'une matrice de pixels avec n lignes, p colonnes à un vecteur de taille np.

Représentation des données

Parfois il est moins évident de représenter une donnée par un vecteur :

- Variable catégorielle (non numérique : genre, couleur, etc.) : on utilise souvent un vecteur avec un 1 et que des 0 (*one-hot vector*).

Fleur	c_Red	c_Purple	c_Blue	Couleur	Vector
Fleur 1	0	0	1	"Blue"	001
Fleur 2	0	1	0	"Purple"	010
Fleur 3	1	0	0	"Red"	100
Fleur 4	0	1	0	"Purple"	010

  En pratique, on utilise les fonctions `get_dummies()` de la bibliothèque `Pandas` ou `OneHotEncoder0` dans `Scikit-learn`.

- Image : On passe d'une matrice de pixels avec n lignes, p colonnes à un vecteur de taille np.

- Son : Transformée de Fourier discrète.

Représentation des données

On représente classiquement l'ensemble des données (donc de vecteurs de $\mathbb{R}^p$) par une matrice $\mathtt{X}$ dont chaque ligne est une donnée et chaque colonne est un attribut.

Représentation des données

On représente classiquement l'ensemble des données (donc de vecteurs de $\mathbb{R}^p$) par une matrice `X` dont chaque ligne est une donnée et chaque colonne est un attribut.

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

Distance

On a besoin de savoir si deux données sont « proches » l'une de l'autre. Pour cela, on utilise une **distance sur les données**, c'est-à-dire sur $\mathbb{R}^p$.

Distance

On a besoin de savoir si deux données sont « proches » l'une de l'autre. Pour cela, on utilise une **distance sur les données**, c'est-à-dire sur $\mathbb{R}^p$.

On utilise souvent la **distance euclidienne** :

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Question

Écrire une fonction `d(x, y)` renvoyant la distance euclidienne entre deux vecteurs x et y.

Distance

On a besoin de savoir si deux données sont « proches » l'une de l'autre. Pour cela, on utilise une **distance sur les données**, c'est-à-dire sur $\mathbb{R}^p$.

On utilise souvent la **distance euclidienne** :

$$d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

Question

Écrire une fonction `d(x, y)` renvoyant la distance euclidienne entre deux vecteurs x et y.

On peut utiliser d'autres distances, par exemple la distance de Manhattan :

$$d(x, y) = \sum_{i=1}^{p} |x_i - y_i|$$

Nettoyage des données

En science des données, on travaille la plus part du temps à partir de données brutes (raw data). Ces données sont très rarement parfaites et contiennent des problèmes de validité ou d'uniformité et sont parfois incomplètes, incorrectes, inrrégulières ou inconsistantes.

Nettoyage des données

En science des données, on travaille la plus part du temps à partir de données brutes (raw data). Ces données sont très rarement parfaites et contiennent des problèmes de validité ou d'uniformité et sont parfois incomplètes, incorrectes, inrrégulières ou inconsistantes.

Il faut donc les nettoyer, c'est le rôle du data cleansing (ou data cleaning). On applique trois étapes successives aux données brutes :

- Analyse des données afin de détecter les potentiels problèmes
- Choix des transformations à effectuer
- Application ces transformations aux données

Nettoyage des données

Classe	Effectifs	Année	Etoilée
PC*	42	-26455632	`True`
PTSI 1	45	1	`True`
PCSI 3			`False`
BCPST2	38	2	`False`
PSI*	39	2	`True`
MP*	43	2	`False`

Question

Quelles transformations effectuer sur ces données brutes afin de les nettoyer ?

Standardisation

Quand les attributs n'ont pas la même échelle (par exemple, l'argent d'un client d'une banque peut être beaucoup plus élevé que son âge), un attribut peut avoir beaucoup plus d'importance qu'un autre dans les calculs de distance.

Standardisation

Quand les attributs n'ont pas la même échelle (par exemple, l'argent d'un client d'une banque peut être beaucoup plus élevé que son âge), un attribut peut avoir beaucoup plus d'importance qu'un autre dans les calculs de distance.

Pour que les attributs aient la même importance, on peut **standardiser** (ou : **normaliser**) les données, c'est-à-dire les modifier pour avoir une moyenne de 0 et un écart-type de 1.

Standardisation

Quand les attributs n'ont pas la même échelle (par exemple, l'argent d'un client d'une banque peut être beaucoup plus élevé que son âge), un attribut peut avoir beaucoup plus d'importance qu'un autre dans les calculs de distance.

Pour que les attributs aient la même importance, on peut **standardiser** (ou : **normaliser**) les données, c'est-à-dire les modifier pour avoir une moyenne de 0 et un écart-type de 1.

La plupart des algorithmes de science des données fonctionnent mieux avec des données standardisées.

Standardisation

Si les données sont des vecteurs $x_1, \ldots, x_n$ ayant chacuns p attributs, on calcule la moyenne μ_j et l'écart-type σ_j pour chacune des p features. Ensuite, pour tout les vecteurs x_i, on remplace chaque élément $x_{i,j}$ par $\dfrac{x_{i,j} - \mu_j}{\sigma_j}$.

Question

Écrire une fonction `standardiser(X)` qui renvoie la matrice obtenue en standardisant les données de la matrice `X`.