Chapitre 5 - Pré-traitement des données

Neila Mezghani

(Hiver 2025)

COLLÈGE DE
BOIS-DE-BOULOGNE

Plan du cours

Table of Contents

Introduction au pré-patraitement des données

Pourquoi pré-traiter ?

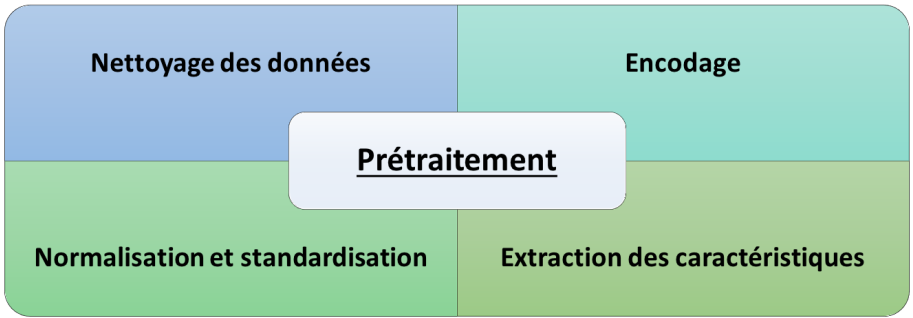
- Dans le monde réel, les données proviennent de plusieurs sources d'acquisition et de collecte de données et peuvent contenir des anomalies ou des valeurs incorrectes qui compromettent leurs qualités
- Les problèmes de qualité de données les plus fréquents :
  - Caractère incomplet : des valeurs ou des attributs sont manquants.
  - Bruit : les données contiennent des individus erronés ou des aberrations.
  - Redondance : les données sont de grande dimension et contiennent des informations inutiles.

Pourquoi pré-traiter ?

- La bonne qualité des données est essentielle pour obtenir des systèmes d'IA de bonne performance.
- Pour éviter de développer des modèles avec des données de mauvaise qualité, il faut impérativement les analyser, détecter les anomalies et déterminer les étapes de prétraitement et de nettoyage appropriées.

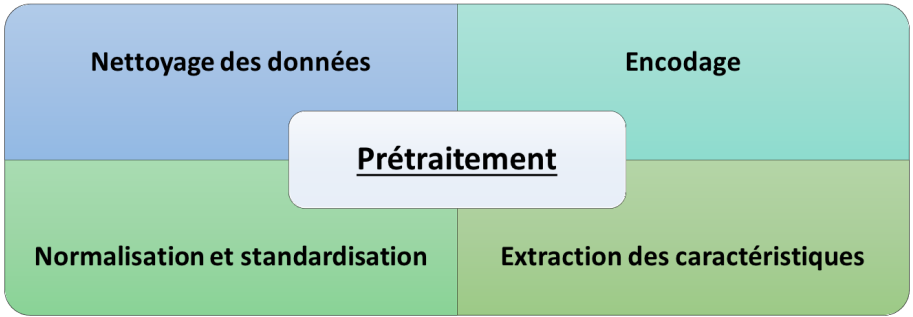
Quelles sont les principales opérations de pré-traitement des données ?

- Nettoyage des données (valeurs aberrantes & valeurs manquantes).
- Transformation des données (normalisation & standardisation).
- Encodage des données
- Réduction de la dimension (extraction des caractéristiques & sélection des caractéristiques)


Nettoyage des données
Encodage
Prétraitement
Normalisation et standardisation
Extraction des caractéristiques

Données manquantes
Données aberrantes

Encodage one-hot
Label encoding



Normalisation
Standardisation

Table of Contents

Données manquantes

Qu'est ce qu'une donnée manquante ?

- Les données manquantes (*missed data*) sont des données incomplètes, c'est-à-dire des données pour lesquelles certaines variables sont inconnues.
- Exemples :
  - Un patient peut être retirer d'une étude clinique si sa condition n'est pas bien contrôlée — on perd la suite.
  - Un problème d'acquistion de données survient lors de la collecte et donc une ou plusieurs variables ne peuvent pas être mesurées.
- `pandas.DataFrame.isnull` ou bien `pandas.DataFrame.isna`

Réparation des données manquantes

Réparation des données manquantes

- Les données manquantes ne peuvent pas être automatiquement traitées selon l'une ou l'autre des deux formes $\Longrightarrow$ Leur traitement dépend de leur proportion par rapport à l'ensemble des données.
- La réparation d'un ensemble de données contenant des données manquantes peut prendre principalement deux formes :
  - Retirer les données manquantes = Suppression
  - Remplacer les valeurs manquantes par des valeurs artificielles = Imputation

Suppression des données manquantes

- Les données manquantes sont retirées ou supprimées si leur proportion dans l'ensemble des données est faible.
- La supression permet de ne prendre en compte que les individus qui ne possèdent aucune valeur manquante dans les variables sélectionnées pour l'apprentissage machine.
- Inconvénient : perte d'échantillons de données surtout lorsque la base de données est de taille réduite.
- `DataFrame.dropna`

Imputation des données manquantes

- L'imputation de données manquantes réfère au fait de remplacer les valeurs manquantes dans l'ensemble de données par des valeurs artificielles.
- Idéalement, ces remplacements ne doivent pas conduire à une altération sensible de la distribution et la composition de l'ensemble des données.
- Approches d'imputation :
  - Imputation par une valeur statistique
  - Imputation par les k plus proche voisin
  - Imputation itérative

Imputation par une valeur statistique

- Dans le cas de données quantitatives, l'imputation de données manquantes peut se faire par :
  - Une constante = imputation par règle
  - La moyenne de la variable
  - La médiane de la variable
- Dans le cas de données qualitatives, l'imputation de données manquantes peut se faire par le mode.
- `sklearn.impute.SimpleImputer`

Exemple : Traitement des données manquantes (1/6)

Soit un ensemble de données `carsPreprocessing.xlsx` décrivant 21 types de voitures à travers 7 variables :

- Prix de la voiture (`prix`)
- Taille du moteur (`cylindree`)
- Puissance (`puissance`)
- poids (`poids`).
- Consommation (`conso`, liter per 100 km).
- Pays d'origine `Origine`
- Abbréviation du pays `Abb`

Exemple : Traitement des données manquantes (2/6)

type	prix	cylindree	puissance	poids	conso	origine	abb
Daihatsu Cuore	11600	846	32	650	5,7	Japon	JP
Suzuki Swift 1.0 GLS	12490	993	39	790	5,8	Japon	JP
Fiat Panda Mambo L	10450	899	29	730	6,1	Italie	IT
VW Polo 1.4 60	17140	1390	44	955	6,5	Allemagne	DE
Opel Corsa 1.2i Eco	14825	1195	33	895	6,8	Allemagne	
Toyota Corolla	19490	1331	55	1010	7,1	Japon	JP
Mercedes S 600	183900	5987	300	2250	18,7	Allemagne	DE
Maserati Ghibli GT	92500	2789	209	1485	14,5	Italie	IT
Opel Astra 1.6i 16V	25000	1597	74	1080	7,4	Allemagne	DE
Peugeot 306 XS 108	22350	1761	74	1100	9	France	
Renault Safrane 2.2. V	36600	2165	101	1500	11,7	France	FR
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9,5	Espagne	ES
VW Golt 2.0 GTI	31580	1984	85	1155	9,5	Allemagne	DE
Citroen ZX Volcane	28750	1998	89	1140	8,8	France	FR
Fort Escort 1.4i PT	20300	1390	54	1110	8,6	États-Unis	USA
Honda Civic Joker 1.4	19900	1396	66	1140	7,7	Japon	JP

Exemple : Traitement des données manquantes (3/6)

L'analyse de l'ensemble des données indique la présence de 3 données manquantes

```
display(df1[df1.isna().any(axis=1)].style.highlight_null('red'))
NaN_Rows_Ids=df1[df1.isna().any(axis=1)].index
```

	Id	type	prix	cylindree	puissance	poids	conso	origine	abb
4	5	Opel Corsa 1.2i Eco	14825.000000	1195	33	895	6.800000	Allemagne	nan
9	10	Peugeot 306 XS 108	22350.000000	1761	74	1100	9.000000	France	nan
26	27	Subaru Vivio 4WD	nan	658	32	740	6.800000	Japon	JP

Exemple : Traitement des données manquantes (4/6)

Imputation des données manquantes de la variable `prix` par la moyenne (moyenne = 30165)

	Id	type	prix	cylindree	puissance	poids	conso	origine	abb
4	5	Opel Corsa 1.2i Eco	14825.0	1195	33	895	6.8	Allemagne	NaN
9	10	Peugeot 306 XS 108	22350.0	1761	74	1100	9.0	France	NaN
26	27	Subaru Vivio 4WD	30165.0	658	32	740	6.8	Japon	JP

Exemple : Traitement des données manquantes (5/6)

Imputation des données manquantes de la variable `prix` par la médiane (médiane = 44756.5)

	Id	type	prix	cylindree	puissance	poids	conso	origine	abb
4	5	Opel Corsa 1.2i Eco	14825.0	1195	33	895	6.8	Allemagne	NaN
9	10	Peugeot 306 XS 108	22350.0	1761	74	1100	9.0	France	NaN
26	27	Subaru Vivio 4WD	44756.5	658	32	740	6.8	Japon	JP

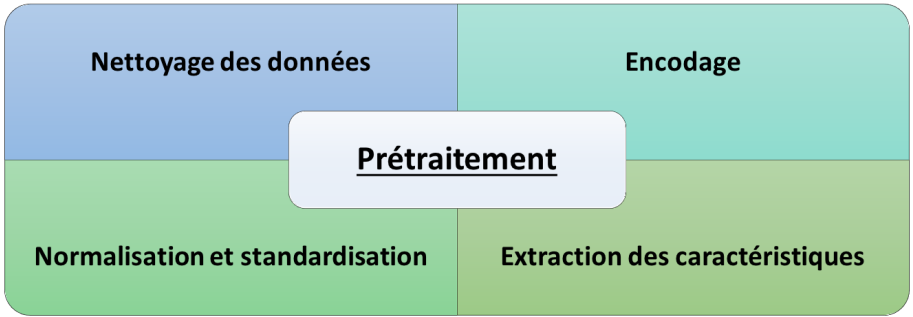
Exemple : Traitement des données manquantes (6/6)

Imputation des données manquantes de la variable `abb` par le mode (mode = JP)

	Id	type	prix	cylindree	puissance	poids	conso	origine	abb
4	5	Opel Corsa 1.2i Eco	14825.0	1195	33	895	6.8	Allemagne	JP
9	10	Peugeot 306 XS 108	22350.0	1761	74	1100	9.0	France	JP
26	27	Subaru Vivio 4WD	NaN	658	32	740	6.8	Japon	JP

Données manquantes
Données aberrantes

Encodage one-hot
Label encoding



Normalisation
Standardisation

Données aberrantes (1/2)

- Les données aberrantes (*outliers*) sont des valeurs extrêmes par rapport à l'ensemble des données à analyser.
- Elles sont souvent causées par une erreur commise lors de leur acquisition ou de leur transcription.
- Cependant, dans certains cas, elles peuvent correspondre à des observations réelles mais particulières.

Détection des valeurs aberrantes (2/2)

Plusieurs méthodes de détection des valeurs aberrantes

- Analyse graphique
- Détection selon la distance interquartile
- Détection basée sur l'écart type

Analyse graphique

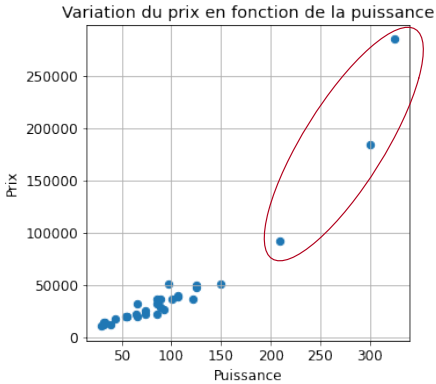
- En analyse univariée (x est un scalaire), les méthodes graphiques telle que le diagramme de dispersion des individus permettent de détecter les valeurs aberrantes
- Le diagramme de dispersion (*Scatterplot*) est un nuage de point qui permet de représenter une variable numérique en fonction d'une autre variable numérique :

$$Y \sim X$$

- Chaque point représente un individu
- Les positions sur l'axe X (horizontal) et Y (vertical) représentent les valeurs des 2 variables.
- `pyplot.scatter`

Exemple : Analyse graphique

L'analyse graphique démontre la présence de 3 données aberrantes



Détection selon la distance interquartile

- Selon la définition classique d'un boxplot, un point est affiché comme aberrant si :

$$x < Q_{25\%}(x) - 1.5 \times IQR_{25\%75\%}(x)$$

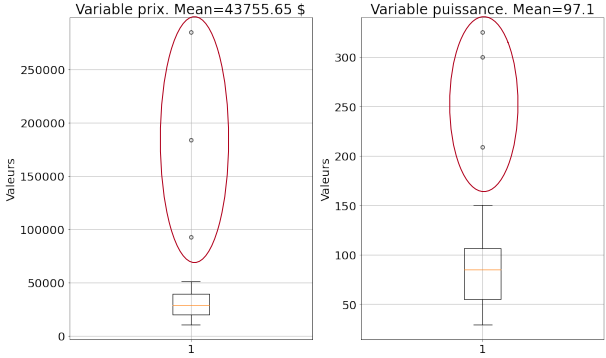
ou

$$x > Q_{75\%}(x) + 1.5 \times IQR_{25\%75\%}(x)$$

où Q_a est le quartile pour la probabilité a
IQR_{ab} est la distance entre les quartiles a et b $(b > a)$

Exemple : Détection selon la distance interquartile

L'analyse du diagramme en boite démontre la présence de 3 données aberrantes pour chacune des variables `prix` et `puissance`



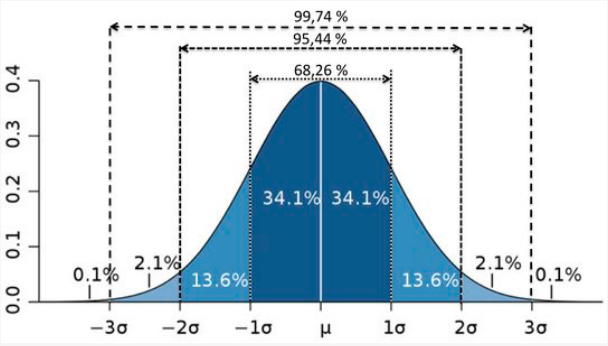
Détection basée sur l'écart type (1/2)

- Consiste à utiliser la variation de la variable autour de la moyenne et à exclure les valeurs exceptionnellement éloignées de cette moyenne, selon un certain intervalle compris entre deux seuils centrés autour de la moyenne
- Les seuils sont mesurée en terme de nombre η d'écart-type :

$$[\mu - \eta\sigma, \mu + \eta\sigma]$$

  μ est la moyenne, σ est l'écart type et η le seuil
- Inconvénient : seuil de η est arbitraire

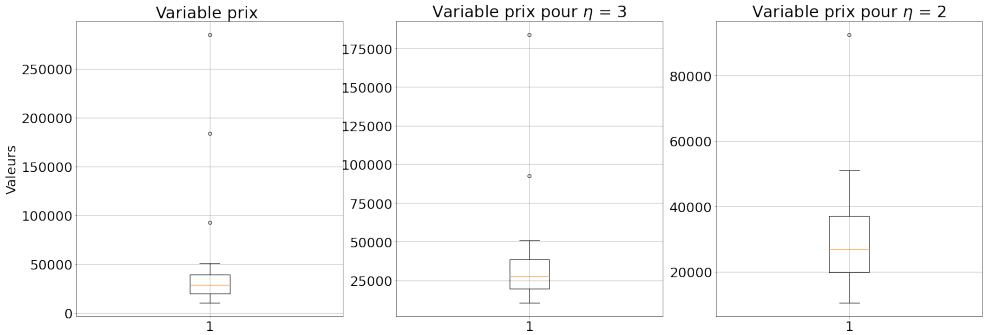
Détection basée sur l'écart type (2/2)



- Si une valeur est située à 3 écarts-type de la moyenne $\Longrightarrow$ On pourra détecter les valeurs aberrantes à partir de 3σ

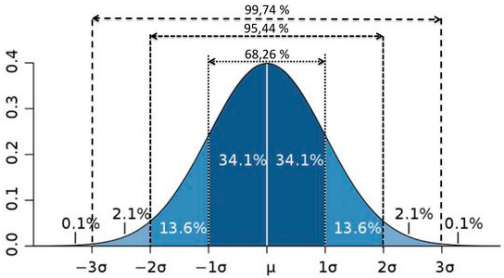
Exemple : Détection basée sur l'écart type (1/2)

L'analyse graphique démontre la suppression d'une valeur aberrante pour $\eta = 3$ et la suppression de deux valeurs aberrantes pour $\eta = 2$



Exemple : Détection basée sur l'écart type (2/2)

L'analyse graphique démontre la suppression d'une valeur aberrante pour $\eta = 3$ et la suppression de deux valeurs aberrantes pour $\eta = 2$



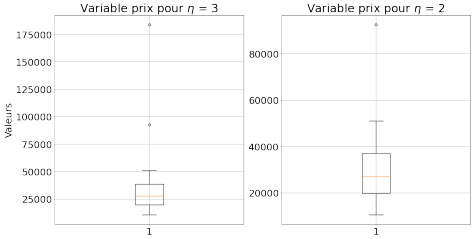
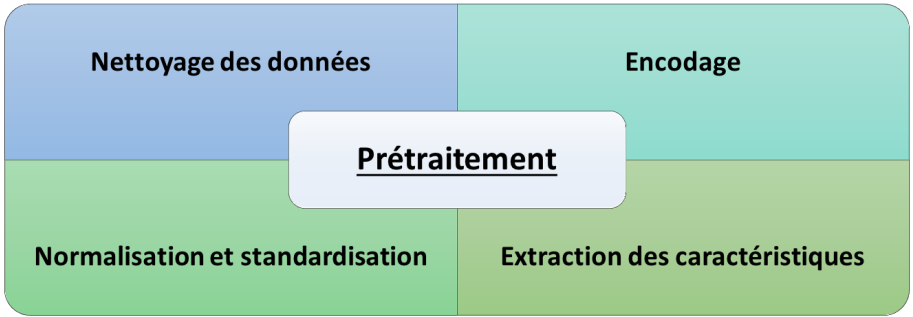


Table of Contents

Données manquantes
Données aberrantes

Encodage one-hot
Label encoding



Normalisation
Standardisation

Normalisation & Standardisation

Pourquoi normaliser & standardiser ?

- La plupart du temps, les données proviennent de source de données différentes $\Longrightarrow$ des ordres de grandeurs différents.
- Cette différence d'échelle peut conduire à des performances médiocres en apprentissage machine.
- On applique deux traitements préparatoires pour rendre les données "homogènes" qui comprennent, entre autres, la **Normalisation** et la **Standardisation** .

La normalisation

- **La mise en échelle min-max** transforme chaque valeur numérique x en une autre valeur $x_e \in [0, 1]$ en utilisant la valeur minimale et la valeur maximale dans les données.
- Cette normalisation conserve la distance proportionnelle entre les valeurs d'une caractéristique.

$$x_e = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

avec x_{min} la valeur minimale de la variable x et x_{max} la valeur maximale.

- `sklearn.preprocessing.MinMaxScaler`

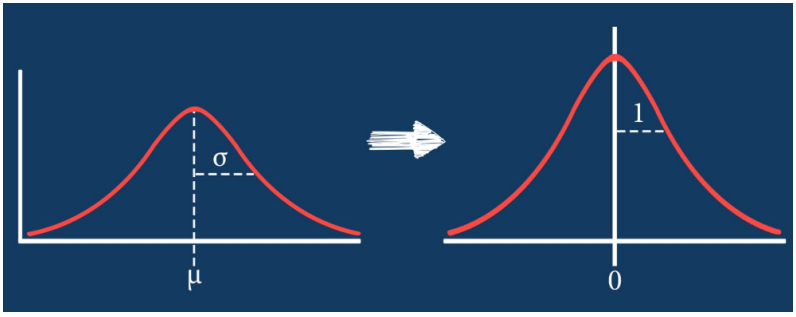
La standardisation

- **La standardisation** peut-être appliquée quand la variable répond aux critères d'une distribution normale (Distributions Gaussiennes).
- La standardisation est le processus de transformer une variable en une autre qui répondra à la loi normale (Gaussian Distribution) $X \sim \mathcal{N}(\mu, \sigma)$ avec : $\mu = 0$ La moyenne de la loi de distribution
  $\sigma = 1$ est l'écart-type (Standard Deviation)
- La standardisation consiste alors à transformer x en :

$$x_s = \frac{x - \mu}{\sigma} \tag{2}$$

- `sklearn.preprocessing.StandardScaler`

Exemple : La standardisation



https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

Table of Contents

Pourquoi l'encodage des données ?

Pourquoi l'encodage des données ?

- La présence de variable qualitative (qui prend des modalités) complique souvent les algorithmes d'apprentissage machine
- La plupart des algorithmes d'apprentissage prennent des valeurs numériques en entrée $\Longrightarrow$ il faut trouver une façon de transformer ces modalités en données numériques.

L'encodage one-hot

- **L'encodage one-hot** (ou encodage 1 parmi n) est courant en apprentissage automatique
- Consiste à encoder une variable à n étiquette sur n bits dont la modalité prise par la variable prend la valeur 1, les autres étant à 0.
- Avantage : facilité d'application
- Inconvénient : L'inconvénient est la taille de la variable en mémoire puisqu'il utilise autant de bits que de modalités
- `sklearn.preprocessing.OneHotEncoder`

Exemple : L'encodage one-hot

	abb_DE	abb_ES	abb_FR	abb_IT	abb_JP	abb_KR	abb_SE	abb_USA
0	0	0	0	0	1	0	0	0
1	0	0	0	1	0	0	0	0
2	1	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0
4	0	1	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0
6	0	0	0	0	0	0	0	1
7	0	0	0	0	0	1	0	0

L'encodage LabelEncoding

- **L'encodage LabelEncoding** (L'encodage des étiquettes) consiste à convertir les étiquettes (variables qualitatives) sous une forme quantitative afin de les rendre lisibles pa r la machine
- `sklearn.preprocessing.LabelEncoder`

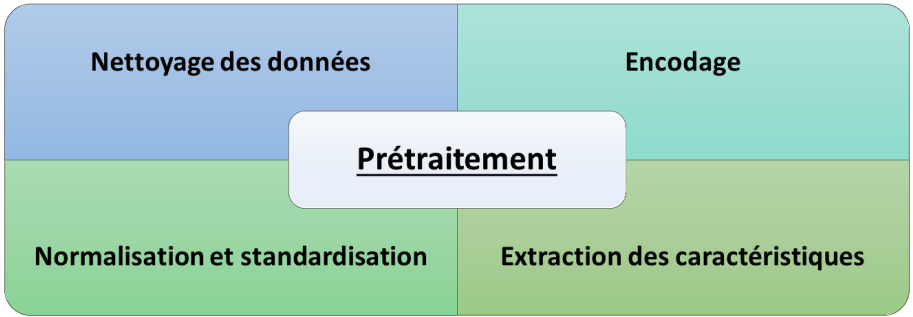
Exemple : L'encodage LabelEncoding

type	prix	cylindree	puissance	poids	conso	origine	abb
Daihatsu Cuore	11600	846	32	650	5,7	Japon	4
Suzuki Swift 1.0 GLS	12490	993	39	790	5,8	Japon	4
Fiat Panda Mambo L	10450	899	29	730	6,1	Italie	3
VW Polo 1.4 60	17140	1390	44	955	6,5	Allemagne	0
Opel Corsa 1.2i Eco	14825	1195	33	895	6,8	Allemagne	0
Toyota Corolla	19490	1331	55	1010	7,1	Japon	4
Mercedes S 600	183900	5987	300	2250	18,7	Allemagne	0
Maserati Ghibli GT	92500	2789	209	1485	14,5	Italie	3
Opel Astra 1.6i 16V	25000	1597	74	1080	7,4	Allemagne	0
Peugeot 306 XS 108	22350	1761	74	1100	9	France	2
Renault Safrane 2.2. V	36600	2165	101	1500	11,7	France	2
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9,5	Espagne	1

Table of Contents

Données manquantes
Données aberrantes

Encodage one-hot
Label encoding



Normalisation
Standardisation

Extraction et sélection des caractéristiques

Qu'est ce que l'extraction des caractéristiques ?

- Parmi les aspects importants de l'apprentissage automatique, on retrouve l'extraction et la sélection de caractéristiques.
- L'extraction des caractéristiques permet d'obtenir un ensemble de variables informatives.
- Par exemple, si nous voulons modéliser le temps (weather), la date, la température, l'humidité et la vitesse du vent sont informatives (elles sont liées au problème). En revanche, le résultat d'un match de football ne sera pas une caractéristique informative car il n'affecte pas le temps.

Qu'est ce que la sélection des caractéristiques ?

- Consiste à sélectionner un sous-ensemble de caractéristiques qui serviront à l'entrainement de l'algorithme d'apprentissage $\Longrightarrow$ les caractéristiques pertinentes
- Les caractéristiques sélectionnées seront utilisées et toutes les autres seront ignorées $\Longrightarrow$ réduction de la dimensionnalité.
- Dans l'exemple précédent, on peut se rendre compte qu'avec le changement climatique la date n'est plus une caractéristique pertinente et que seulement la température, l'humidité et la vitesse du vent sont des caractéristiques pertinentes.

Sélection des caractéristiques : Formalisation

- Étant donné un ensemble de caractéristiques $F = \{f_1, f_2, ..., f_n\}$, la sélection des caractéristiques consiste à déterminer un sous-ensemble qui maximise la capacité du modèle à caractériser les formes $\Longrightarrow$ Déterminer le sous-ensemble F' qui maximise la fonction coût.

Pourquoi la sélection des caractéristiques ?

- Permet à l'algorithme d'apprentissage machine de s'entraîner plus rapidement.
- Réduit la complexité d'un modèle et le rend plus facile à interpréter.
- Améliore la précision d'un modèle si on choisi un sous-ensemble de caractéristiques pertinentes
- Réduit le sur-apprentissage

Exemple : Extraction des caractéristiques (1)

Extraction des caractéristiques

type	prix	cylindree	puissance	poids	conso	origine	abb
Daihatsu Cuore	11600	846	32	650	5,7	Japon	4
Suzuki Swift 1.0 GLS	12490	993	39	790	5,8	Japon	4
Fiat Panda Mambo L	10450	899	29	730	6,1	Italie	3
VW Polo 1.4 60	17140	1390	44	955	6,5	Allemagne	0
Opel Corsa 1.2i Eco	14825	1195	33	895	6,8	Allemagne	0
Toyota Corolla	19490	1331	55	1010	7,1	Japon	4
Mercedes S 600	183900	5987	300	2250	18,7	Allemagne	0
Maserati Ghibli GT	92500	2789	209	1485	14,5	Italie	3
Opel Astra 1.6i 16V	25000	1597	74	1080	7,4	Allemagne	0
Peugeot 306 XS 108	22350	1761	74	1100	9	France	2
Renault Safrane 2.2. V	36600	2165	101	1500	11,7	France	2
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9,5	Espagne	1

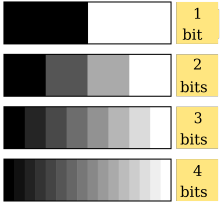
Exemple : Extraction des caractéristiques (2)

Extraction des caractéristiques

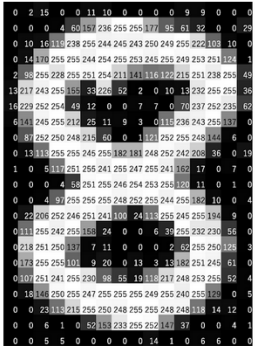
type	prix	cylindree	puissance	poids	conso	origine	abb
Daihatsu Cuore	11600	846	32	650	5,7	Japon	4
Suzuki Swift 1.0 GLS	12490	993	39	790	5,8	Japon	4
Fiat Panda Mambo L	10450	899	29	730	6,1	Italie	3
VW Polo 1.4 60	17140	1390	44	955	6,5	Allemagne	0
Opel Corsa 1.2i Eco	14825	1195	33	895	6,8	Allemagne	0
Toyota Corolla	19490	1331	55	1010	7,1	Japon	4
Mercedes S 600	183900	5987	300	2250	18,7	Allemagne	0
Maserati Ghibli GT	92500	2789	209	1485	14,5	Italie	3
Opel Astra 1.6i 16V	25000	1597	74	1080	7,4	Allemagne	0
Peugeot 306 XS 108	22350	1761	74	1100	9	France	2
Renault Safrane 2.2. V	36600	2165	101	1500	11,7	France	2
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9,5	Espagne	1

Exemple : Extraction des caractéristiques (3)

Image originale

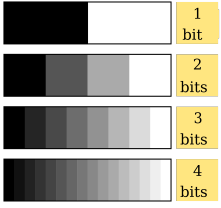


Matrice de nombre qui correspondent aux niveaux de gris



Exemple : Extraction des caractéristiques (4)

Image originale



Matrice de nombre qui correspondent aux niveaux de gris

0	2	15	0	0	11	10	0	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29	
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0	
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1	
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49	
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36	
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62	
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0	
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0	
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19	
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0	
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0	
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4	
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0	
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0	
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3	
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0	
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4	
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5	
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0	
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1	
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0	

Take-home message ....