Jeu de données ozone_complet

Source : Laboratoire de mathématiques appliquées de l'Agrocampus Ouest

Le jeu de données contient 1464 observations (journalières, du 01/04/1995 au 30/09/2002, à Rennes), des 24 variables suivantes :

- date : date (au format *aaaammjj*),
- maxO3 : teneur maximale en ozone observée sur la journée (en $\mu gr.m^{-3}$),
- T6, T9, T12, T15, T18 : température observée à 6h, 9h, 12h, 15h et 18h (en $^{\circ}C$),
- Ne6, Ne9, Ne12, Ne15, Ne18 : nébulosité observée à 6h, 9h, 12h, 15h et 18h (en *octas*),
- Vdir6, Vdir9, Vdir12, Vdir15, Vdir18 : direction du vent observée à 6h, 9h, 12h, 15h et 18h,
- Vvit6, Vvit9, Vvit12, Vvit15, Vvit18 : vitesse du vent observée à 6h, 9h, 12h, 15h et 18h (en $m.s^{-1}$),
- Vx : composante est-ouest du vent observée (en $m.s^{-1}$),
- maxO3v : teneur maximale en ozone observée la veille (en $\mu gr.m^{-3}$).

Jeu de données MNIST

MNIST : Modified National Institute of Standards and Technology
http://yann.lecun.com/exdb/mnist/

Le jeu de données est organisé en 4 fichiers :

- train-images-idx3-ubyte.gz : nuances de gris (sur $\{0, \ldots, 255\}$, 0 pour blanc, 255 pour noir) des 28×28 pixels des 60 000 images du jeu de données d'apprentissage,
- train-labels-idx1-ubyte.gz : étiquettes (sur $\{0, \ldots, 9\}$) des 60 000 images du jeu de données d'apprentissage,
- t10k-images-idx3-ubyte.gz : nuances de gris (sur $\{0, \ldots, 255\}$, 0 pour blanc, 255 pour noir) des 28×28 pixels des 10 000 images du jeu de données de test,
- t10k-labels-idx1-ubyte.gz : étiquettes (sur $\{0, \ldots, 9\}$) des 10 000 images du jeu de données de test.

Les fichiers sont au format IDX.

Jeu de données spam

Source : https://archive.ics.uci.edu/ml/datasets/spambase

Le jeu de données contient 4601 observations relatives à des e-mails, avec 57 variables :

- word_freq_make, word_freq_address, word_freq_all, word_freq_3d, word_freq_our, word_freq_over, word_freq_remove, word_freq_internet, word_freq_order, word_freq_mail, word_freq_receive, word_freq_will, word_freq_people, word_freq_report, word_freq_addresses, word_freq_free, word_freq_business, word_freq_email, word_freq_you, word_freq_credit, word_freq_your, word_freq_font, word_freq_000, word_freq_money, word_freq_hp, word_freq_hpl, word_freq_george, word_freq_650, word_freq_lab, word_freq_labs, word_freq_telnet, word_freq_857, word_freq_data, word_freq_415, word_freq_85, word_freq_technology, word_freq_1999, word_freq_parts, word_freq_pm, word_freq_direct, word_freq_cs, word_freq_meeting, word_freq_original, word_freq_project, word_freq_re, word_freq_edu, word_freq_table, word_freq_conference : « sacs de mots » (fréquences des mots dans les e-mails) de *make*, *address*, *all*, *3d*, *our*, *over*, *remove*, *internet*, *order*, *mail*, *receive*, *will*, *people*, *report*, *addresses*, *free*, *business*, *email*, *you*, *credit*, *your*, *font*, *000*, *money*, *hp*, *hpl*, *george*, *650*, *lab*, *labs*, *telnet*, *857*, *data*, *415*, *85*, *technology*, *1999*, *parts*, *pm*, *direct*, *cs*, *meeting*, *original*, *project*, *re*, *edu*, *table* et *conference* (en *%*),
- char_freq_semicolon, char_freq_leftbrac, char_freq_leftsquarebrac, char_freq_exclaim, char_freq_dollar, char_freq_pound : fréquence des caractères spéciaux *;*, *{*, *[*, *!*, *$* et *£* (en *%*),
- capital_run_length_average : longueur moyenne des séquences ininterrompues en majuscules,
- capital_run_length_longest : longueur de la plus grande séquence ininterrompue en majuscules,
- capital_run_length_total : nombre total de majuscules,
- spam : label de classe (1 si considéré comme spam, 0 sinon).