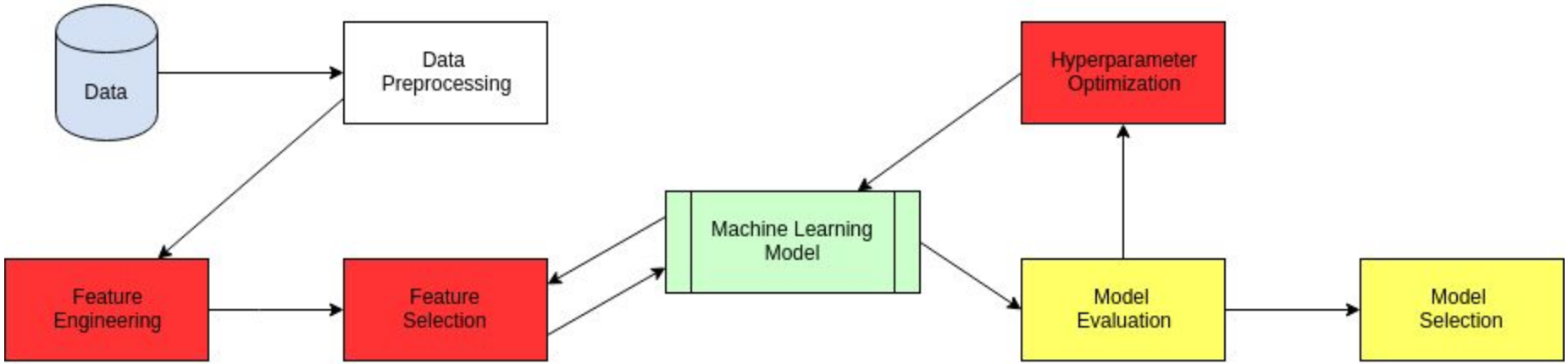
Machine Learning
& Intelligence Artificielle

Manuel Simoes
manuel.simoes@cpc-analytics.fr

- Machine Learning -
&
Démarche générale

Machine learning



Le processus d'apprentissage automatique (processus d'automatisation de base en rouge, processus auxiliaires en jaune)

Préparer et Analyser vos données

- Autant que possible afficher vos données
- Vérifier la compatibilité entre les valeurs des données et ce qu'elles représentent.
- Vérifier la présence d'artefact, d'extremum, NaN...
- Nettoyer les données
- Séparer les données en plusieurs ensembles
    - Un ensemble de données pour l'apprentissage [Training Set]
    - Un ensemble de données pour les tests de qualité de l'algo [Testing Set]
    - En fonction des données et de leur quantité, il est également possible de constituer un ensemble de données utilisées pour ajuster certains paramètres de l'algo [Calibration Set]

Chaque jeux de données est "unique" et des décisions sont à prendre au cas par cas.

Nettoyer les données

Ce qu'il faut (entre-autres) pouvoir identifier dans les données

- Les colonnes aux valeurs identiques / constante
- Vérifier les lignes identiques (uniquement si elle correspondent à un artefact)
- La proportion sur les lignes et les colonnes des valeurs manquantes NaN (**N**ot **a** **N**umber) par rapport aux valeurs connues.
- Les outlier (correspondent-ils à une certaine réalité ?)
- Corrélation forte entre 2 ou plusieurs colonnes

Gérer les valeurs manquantes NaN

- Calculer la proportion de NaN sur les colonnes ou/et sur les lignes.
- Évaluer s'il faut les supprimer ou les remplacer (Imputation).
- Différentes méthodes d'imputation existent
    - Remplacer par une valeur constante, la moyenne de la colonne, une moyenne glissante
    - Autres méthodes (voir les librairies...)

Analyse de l'erreur

- **Commencer avec un algo simple et facile** à mettre en place

- **Utiliser la courbe d'apprentissage** pour décider s'il faut plus de données, features…

- **Analyse de l'erreur :** Faites des calcul avec différentes hypothèses et identifiez les exemples pour lequel vous obtenez de mauvais résultats.

  Tentez de comprendre ces ensembles.

Attitude générale

La Data Science est une science empirique et vous pouvez faire vos propres expériences :

Prenez un jeu de donnée de référence et mesuré la qualité de vos résultats

Modifier votre jeu de données d'entrée et/ou modifier des paramètres de l'algo et/ou changer d'algo …
[Ne modifier qu'un paramètre à la fois et sur plusieurs valeurs]

Mesurez de nouveau la qualité de vos résultats

Faites votre conclusion

TD 1er contact avec Python / Panda

The Boston Housing Dataset

A Dataset derived from information collected by the U.S. Census Service concerning housing in the area of Boston Mass.

- Charger les données Boston
  - Charger les données sur les appartements dans le DataFrame df.
  - Charger les prix (boston.target) des appartements dans une nouvelle colonne de df (df["Prix"]).
- Visualiser les 15 premières lignes du tableau df.
- Tracer les features (l'objectif est de voir les dépendances des colonnes entre elles et surtout par rapport au prix).
  - Décrire les features/Colonne qui vous semble les plus importantes pour la détection du Prix dans jupyter (seaborn.pairplot)
- Séparer les données en 2 ensembles [Learning & Testing].
  Sklearn.model_selection.train_test_split
  - LinearRegression
- Estimer les prix des maisons à Boston [PRICE en fonction de ??] en utilisant une régression linéaire.

Dataset Naming

The name for this dataset is simply **boston**. It has two prototasks: **nox**, in which the nitrous oxide level is to be predicted; and **price**, in which the median value of a home is to be predicted

Miscellaneous Details

- **Origin**
  The origin of the boston housing data is **Natural**.
- **Usage**
  This dataset may be used for **Assessment**.
- **Number of Cases**
  The dataset contains a total of **506** cases.
- **Order**
  The order of the cases is **mysterious**.
- **Variables**
  There are **14** attributes in each case of the dataset. They are:
  1. CRIM - per capita crime rate by town
  2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
  3. INDUS - proportion of non-retail business acres per town.
  4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
  5. NOX - nitric oxides concentration (parts per 10 million)
  6. RM - average number of rooms per dwelling
  7. AGE - proportion of owner-occupied units built prior to 1940
  8. DIS - weighted distances to five Boston employment centres
  9. RAD - index of accessibility to radial highways
  10. TAX - full-value property-tax rate per $10,000
  11. PTRATIO - pupil-teacher ratio by town
  12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
  13. LSTAT - % lower status of the population
  14. MEDV - Median value of owner-occupied homes in $1000's
- **Note**
  Variable #14 seems to be censored at 50.00 (corresponding to a median price of $50,000); Censoring is suggested by the fact that the highest median price of exactly $50,000 is reported in 16 cases, while 15 cases have prices between $40,000 and $50,000, with prices rounded to the nearest hundred. Harrison and Rubinfeld do not mention any censoring.

Source : http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

Attitude générale

Pendant les TD

Comprendre les paramètres d'entrées des fonctions utilisées

- Faite rapidement une 1er boucle (des données à l'algorithme de ML)
- [cela n'empêche pas de faire **(au début)** des tests à l'aveugle ou à l'instinct pour appréhender les outils]
- Consulter les pages techniques et documentation [python, scikit-learn, pandas...]
- Ne pas hésiter à aller dans le code pour plus de documentation ou tout simplement comprendre le code utilisé.
- Google / StartPage / sayhello (https://www.startpage.com/, https://beta.sayhello.so/)
- StackOverflow (pour les problèmes technique: https://stackoverflow.com)

Au long court

- Continuer à se tenir au courant [blog, articles, congrès, rencontres...]
- Continuer à se former (publications, cours en ligne...)
- Quitter sa zone de confort (normalement il n'y en a pas beaucoup en Data Science...)