

Projet Machine Learning pour la Prévision: introduction

Yannig Goude, yannig.goude@edf.fr

Objectifs du cours:

- ▶ manipuler des données réelles
- ▶ appréhender les méthodes récentes d'apprentissage statistique et leur implémentation en R
- ▶ maîtriser R et son environnement (Rstudio, Rmarkdown)
- ▶ construire un modèle prédictif et le valider
- ▶ restituer les résultats: “rapport” en Rmarkdown (html ou pdf), package R associé correctement documenté, présentation orale de votre travail

Choix des données

- ▶ choix parmi différents jeux de données “open”
- ▶ vous pouvez proposer un jeu de données, à valider ensemble
- ▶ type de données: série temporelle “numériques” et /ou régression (dépendance temporelle ou non)

$$(X_i, Y_i)_{i=1,\dots,n}, X_i \in \mathbb{R}^k, Y_i \in \mathbb{R}$$

et l'objectif est de prévoir $Y = f(X)$

Exemple de jeux de données

► Smarter Mobility Challenge

<https://codalab.lisn.upsaclay.fr/competitions/7192>

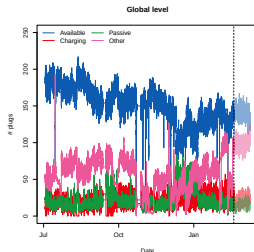


Figure 1: Total number of plugs for each state of Paris charging station.

voir <https://data.mlr.press/assets/pdf/v01-16.pdf> et
<https://gitlab.com/smarter-mobility-data-challenge/tutorials>

Exemple de jeux de données

Challenge data ens:

- ▶ prix de l'électricité :
<https://challengedata.ens.fr/challenges/140>
- ▶ affluence transilien :
<https://challengedata.ens.fr/challenges/149>
- ▶ précipitations à court terme:
<https://challengedata.ens.fr/challenges/88>

Exemple de jeux de données

Traffic forecast:

- ▶ <https://github.com/iarai/NeurIPS2022-traffic4cast>
- ▶ https://proceedings.neurips.cc/paper_files/paper/2023/hash/ee57cd73a76bd927ffca3dda1dc3b9d4-Abstract-Datasets_and_Benchmarks.html

Financial Data:

- ▶ <https://numerai.fund/>
- ▶ <https://www.kaggle.com/competitions/optiver-trading-at-the-close>

Meteo Data:

- ▶ <https://donneespubliques.meteofrance.fr/>
- ▶ <https://cran.r-project.org/web/packages/riem/index.html>
- ▶ Climate: <https://openreview.net/forum?id=3z9YV29Ogn>

Exemple de jeux de données

Catastrophes naturelles:

- ▶ Wildfire: <https://openreview.net/forum?id=RgdGkPRQ03>;
<https://openreview.net/forum?id=VH1vxapUTs>
- ▶ Typhoon: <https://openreview.net/forum?id=9gLnjw8DfA>

Revues/Repos de data:

- ▶ DMLR: Data Centric Machine Learning Research
- ▶ Data in Brief:
<https://www.sciencedirect.com/journal/data-in-brief>
- ▶ Monash Time Series : <https://forecastingdata.org/>
- ▶ NeurIPS Datasets and Benchmarks Track

Exemple de jeux de données

- compétitions GEFCOM 2014

{<https://www.crowdanalytix.com/contests/global-energy-forecasting-competition-2014-probabilistic-electric-load-forecasting>}

- prévoir les quantiles de la consommation électrique à 24h d'horizon
- variables explicatives: température (plusieurs localisations), températures lissées, variables calendaires, lag...

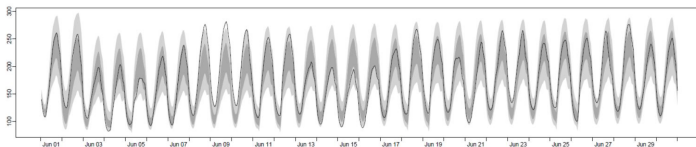


Figure 2: Electricity consumption and probabilistic forecasts

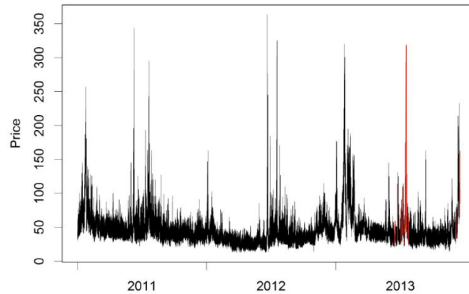


Figure 3: Electricity Price

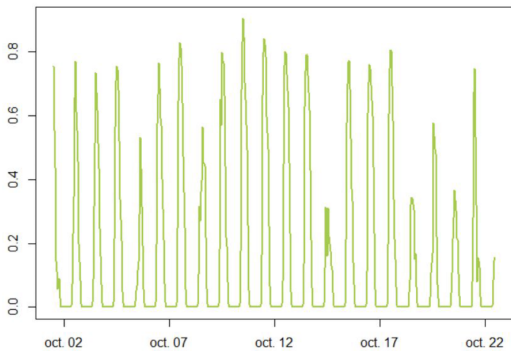


Figure 4: PV production

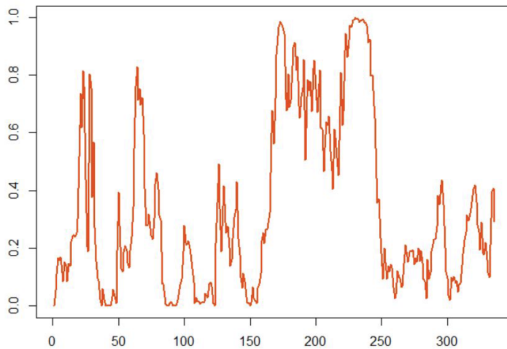


Figure 5: Wind farm production

compétitions GEFCOM 2012

<https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting>

- ▶ prévoir la consommation électrique à 24h d'horizon

Autres exemples de sources de données:

- ▶ production solaire: <https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest>
- ▶ consommation/production électrique en europe:
<https://www.entsoe.eu>,
<https://rte-opendata.opendatasoft.com/>
- ▶ macro-économiques/sociales:
<https://data.oecd.org/fr/energie.htm>, insee
- ▶ fréquentation de station vélib. à Paris & Lyon:
<https://maxhalford.github.io/blog/openbikes-challenge/>
- ▶ météo: <https://www.ncdc.noaa.gov/cdo-web/>
- ▶ pollution atmosphérique:
<https://www.epa.gov/outdoor-air-quality-data>,
<http://www.openair-project.org>
- ▶ open data gouvernemental: <http://www.data.gouv.fr/fr>
- ▶ transport: <http://www.data.gouv.fr/fr/datasets/trafic-annuel-entrant-par-station-2013/>
- ▶ financières: <https://cran.r-project.org/web/packages/tidyquant/index.html>

Méthodes d'apprentissage statistique

- ▶ régression linéaire & sélection de variable (ridge, lasso, elastic net, slope. . .)
- ▶ modèles additifs (GAM), projection poursuit
- ▶ arbres de régression (CART)
- ▶ random forest
- ▶ gradient boosting
- ▶ time series (exponential smoothing, SARIMA)
- ▶ online expert aggregation

Références utiles

Manipulation de données avec R

- ▶ R for data science, Hadley Wickham & Garrett Golemund, O' Reilly. <http://r4ds.had.co.nz>

R

- ▶ R packages, Hadley Wickham, O' Reilly
<http://r-pkgs.had.co.nz>
- ▶ Rmarkdown gallery <http://rmarkdown.rstudio.com/gallery.html>
- ▶ Introduction to Rmarkdown
http://rmarkdown.rstudio.com/articles_intro.html

Machine learning

- ▶ Breiman, L. Random Forests, 2001. <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>
- ▶ Classification and Regression Trees, Breiman, L, 1984, The Wadsworth Statistics/Probability Series.
- ▶ Boosting, Foundations and Algorithms, Robert E. Schapire and Yoav Freund, The MIT press.
- ▶ Generalized Additive Models: An Introduction with R, Simon Wood, Chapman and Hall/CRC, 2006
- ▶ Generalized additive models, Trevor Hastie and Robert Tibshirani, Chapman & Hall, 1990.
- ▶ Statistical Learning Tools for Electricity Load Forecasting, Antoniadis, Cugliari ,Fasiolo , Goude, Poggi, Springer 2024.

Machine learning

- ▶ The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer Series in Statistics, 2009. https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf
- ▶ Time Series Analysis, 2008, George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, John Wiley & Sons.
- ▶ Prediction, Learning, and Games, Nicolo Cesa-Bianchi, Gabor Lugosi, 2006. http://www.ii.uni.wroc.pl/~lukstafi/pmwiki/uploads/AGT/Prediction_Learning_and_Games.pdf

Packages

packages génériques

```
library(magrittr)
library(XML)
library(RCurl)
library(tidyverse)
library(lubridate)
library(xts)
```

Packages

pour la modélisation

```
library(gbm)
library(xgboost)
library(mgcv)
library(mboost)
library(randomForest)
library(ranger)
library(forecast)
library(rpart)
library(party)
library(tree)
library(opera)
```

Packages

pour la visualisation/description de données

```
library(RColorBrewer)
library(yarr)
library(mgcViz)
library(dygraphs)
library(ggplot2)
library(ggthemes)
library(gridExtra)
library(rpart.plot)
library(ALEPlot)
```

Packages

pour les données météo

```
library(riem)  
library(weathermetrics)
```

Organisation du cours

LMO, salle 1A14

Le lien vers le fichier pour proposer vos groupes de travail et sujets:

<https://drive.google.com/open?id=1zFOpdbYRQmdujaZIKkeN2nztifLEzschqbOZmEMXtCo>