Rapport de projet

Le modèle ARMA

Par Sara POIREL, Kplola Ryan AGBOZOH et Morgane LE BOEDEC

HETIC

Sommaire

Présentation du projet

Selon la définition des Nations Unies, le changement climatique est défini comme « les variations des températures et des conditions météorologiques sur le long terme ». Avec des températures record attendues ce week-end en Espagne et au Portugal, dépassant les 30°c, le problème du réchauffement climatique semble de plus en plus palpable.

Ces changements causent d'importants problèmes pour les écosystèmes, les cultures mais aussi pour la santé des Hommes. Les canicules, de plus en plus récurrentes, sont une menace notamment pour les personnes âgées.

A travers ce projet, on souhaite prédire la température globale en 2023 afin de déterminer l'avancée du réchauffement climatique.

Traitement et nettoyage du dataset

Présentation du dataset

Le dataset utilisé dans le cadre de ce projet est tiré de « Observation météorologique historiques France (SYNOP) ». J'ai décidé de me concentrer sur la commune Athis-Mons, située dans l'Essonne en région Ile-de-France, sonde ORLY. Le dataset est composé de plus de 52 colonnes et de 38 843 lignes. Les observations sont réalisées entre 2010 et début 2023.

Le code SYNOP, ou synoptique, est un codage de données, adopté par l'Organisation météorologique mondiale, et employé pour diffuser par radiotélétype, ou autre moyen, les observations d'une station météorologique terrestre, à intervalles réguliers de 3 heures, dites synoptiques.

Traitement du dataset

Dans un premier temps, je m'attache à éliminer les colonnes dont je n'ai pas besoin. Pour cela, j'utilise Excel car ayant beaucoup de colonnes à supprimer, cela me semble plus optimal. Au final, après ce premier traitement, j'ai identifié 6 colonnes pertinentes soient :

- **Date** (objet) : 1e janvier 2010 au 25 avril 2023
- **Type de tendance barométrique** (float) : 1 (basse pression) à 8 (haute pression)
- **Température en degré celsius** (float)
- **Humidité** (float)
- **Mois** (float)

Ensuite, je suis directement passée sur Python pour traiter les différentes colonnes et procéder à une analyse exploratoire du dataset.

Le modèle ARMA

Le modèle ARMA (AutoRegressive Moving Average) est un modèle statistique utilisé pour modéliser et prédire des séries temporelles. Il combine deux modèles :

- **Le modèle autorégressif (AR) :** représente la dépendance temporelle des observations c'est-à-dire que les valeurs passées permettent de prédire les valeurs actuelles

- **La moyenne mobile (MA) :** représente les erreurs de prédiction précédentes c'est-à-dire qu'elle utilise les résidus des prédictions précédentes pour ajuster la prédiction actuelle

Grâce à cette combinaison, le modèle ARMA capture à la fois la dépendance temporelle à court terme et les fluctuations aléatoires.

Le modèle AR

Définition

Le modèle auto-régressif (AR) est un modèle statistique faisant partie des séries temporelles. Il utilise les données passées pour prédire les valeurs futures.

Pour le comprendre, il faut s'intéresser aux séries temporelles. Elles permettent de modéliser une tendance linéaire ou saisonnière et de prédire des données futures. Elles sont utilisées dans plusieurs domaines tels que la finance pour prédire le taux de change d'une monnaie ou la météorologie pour prédire les températures.

Le modèle AR consiste à combiner linéairement des observations passés avec un certain bruit blanc soit des données aléatoires. Il s'agit donc de prédire X_t à partir de X_{t-1}.

Description mathématique du modèle

Un processus auto régressif d'ordre p tel que p>0 peut s'écrit :

$$AR(p) : X_t = \sum_{k=1}^{p} \sigma_k X_{t-k} + \varepsilon_t$$

Avec :

$\varepsilon_t \sim \mathrm{WN}(0, \sigma_\varepsilon^2)$: erreur aléatoire, bruit blanc

$\sigma_k (k = 1, ..., p)$: paramètres du modèle soit le coefficient de régression

X_t : valeur que l'on veut prédire grâce à X_{t-k} valeur précédente

p représente le nombre de valeurs que le modèle prend en compte pour prédire la valeur actuelle.

Ainsi, un processus AR(1) prend la forme suivante :

$$AR(1) : X_t = \sigma X_{t-1} + \varepsilon_t$$

Le paramètre $\sigma_k(k = 1, ..., p)$ permet de déterminer si la série est stationnaire c'est-à-dire si elle évolue avec le temps ou non. Ainsi, on a :

$$|\sigma| = \begin{cases} < 1 & : \text{Le processus est stationnaire} \\ = 1 & : \text{Le processus n'est pas stationnaire} \\ > 1 & : \text{Le processus est explosif} \end{cases}$$

Le modèle MA

Définition

Le modèle MA (Moving Average) est un modèle statistique basé sur la composante de la moyenne mobile qui cherche à capturer les variations aléatoires à court terme d'une série temporelle.
La prédiction d'une observation est basée sur une moyenne pondérée des erreurs de prédiction précédentes. L'ordre du modèle MA, noté q, indique le nombre de résidus passés pris en compte.

Description mathématique du modèle

Un modèle MA $(X_t)_{t\epsilon\mathbb{Z}}$ d'ordre q noté MA(q) se note :

$$X_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Avec :

ε_t : résidus
θ_i : réel

Pour un processus MA(q), les autocorrélations s'annulent à partir du rang q+1 ce qui signifie que les valeurs passées n'ont pas d'effet sur les valeurs futures une fois que le décalage dépasse q :

$$\begin{cases} p(q) \neq 0 \\ \forall h \epsilon \mathbb{N}, h \geq q+1 : p(h) = 0 \end{cases}$$

La fonction de coût

La fonction de coût est essentielle pour mesurer les performances du modèle. Il en existe plusieurs telles que la MAE (Mean Average Error), la MSE (Mean Squared Error)... Nous allons nous intéresser ici à la RMSE ou Root Mean Square Error.

La RMSE (Root Mean Square Error)

La RMSE, ou Root Mean Square Error, permet de mesurer l'erreur quadratique moyenne entre la réalité et la prédiction. Elle représente la racine carrée de l'erreur moyenne au carré. Ces erreurs entre les valeurs réelles et les valeurs prédites sont appelées les résidus.

La RMSE idéale serait égale à 0. Cependant, en pratique, une telle RMSE n'est jamais atteinte. Il n'y a pas réellement de meilleure RMSE qu'une autre dans la mesure où elle dépend des valeurs du dataset. Par exemple, si l'on prédit le prix de voitures de luxe (50 000 € à 100 000 €), une RMSE de 400 indiquerait une bonne performance du modèle. En revanche, si l'on prédit le prix d'enceintes bluetooth (80 € à 500 €), une même RMSE de 400 indiquerait que le modèle est très mauvais.

Description mathématique

LA formule de la RMSE est :

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - x_i)^2}{n}}$$

Avec :

n : ensemble de données
y_i : valeurs prédites
x_i : valeurs réelles

Sources

Opendatasoft, *Observation météorologiques historiques de France (SYNOP)*, https://public.opendatasoft.com/explore/dataset/donnees-synop-essentielles-omm/export/?sort=date&refine.nom_reg=%C3%8Ele-de-France

Wikipédia, *Processus autorégressif*, URL : Processus autorégressif — Wikipédia

Steven Fortier, *Les modèles MA, AR et ARMA multidimensionnels : estimation et causalité*, URL : Les modèles MA, AR et ARMA multidimensionnels : estimation et causalité

Code Cogs, *Online LaTeX Equation Editor*, URL : Online LaTeX Equation Editor - create, integrate and download

Medium, Jinit Shah, *ARIMA Model from Scratch in Python*, URL : ARIMA Model from Scratch in Python | by Jinit Shah | Analytics Vidhya | Medium

Open AI, *Chat GPT*, URL : https://chat.openai.com/

OpenClassrooms, Vincent Lefieux, *Analyser et modéliser les séries temporelles*, URL : Les processus AR, MA et ARMA - Analysez et modélisez des séries temporelles - OpenClassrooms