Data Science / Analyse et Traitement de l'Information

Extra TP2: Time series prediction (50 pts).

You will find two csv data files in the data folder:

- `data_mintemp.csv` contains daily minimum temperature data with two columns: "Date" and "Temp".
- `data_cons.csv` contains daily electrical consumption, wind and solar electricity production with three columns: `"Date"`, `"Consumption"`, `"Wind+Solar"`.

Questions

1. Import the data contained in these files and create three datasets with the for- mat $[(t, x_t)]$ where t represents all possible times for the dataset ("Date"), and x_t could be either the corresponding `"Temp"`, `"Consumption"`, or `"Wind+Solar"`. Hint: you could use `pandas.read_csv` method with the right parameters for `path`, `header`, `index_col`, and `usecols`.

2. For each of these datasets, perform the following tasks (you could create a function for these tasks and apply it on each dataset):

   (a) For $k \in \{7, 30, 365\}$, create a new feature of the dataset named `"SMA_k"` which corresponds to the simple moving average with a window of size k defined as the simple mean of the k previous values:

   $$\text{SMA-k}_t = \frac{\sum_{i=0}^{k-1} x_{t-i}}{k}$$

   The dataset will now have 5 columns: `"Date"`, `"value"`, `"SMA_7"`, `"SMA_30"`, and `"SMA_365"`.

   (b) Plot the data along with their three simple moving averages and give an interpretation.

   (c) Create a new feature of the dataset named `"value-365"` (`"value"` being the correct considered data) which corresponds to the serie $[y_t] = [x_t - \text{SMA-365}_t]_t$

(d) Create a new feature of the dataset named `"SMA_30(value-365)"`, which corresponds to the simple moving average of the previous series `"value-365"`, with a window of size 30, this new serie is $[z_t]_t$.

(e) Create a new feature of the dataset named `"value-365-30"` (`"value"` being the correct considered data) which corresponds to the series $[y_t - z_t]_t$. The dataset will now have 8 columns.

(f) Plot in the same figure the series `"value"`, `"SMA_365"`, `"SMA_30(value-365)"`, and `"value-365-30"`. Explain the mathematical link between these series and give a clear interpretation to them (you should use the same terminology as in slides ATI.06).

3. For data, with only the original two columns, perform the following tasks: the first dataset only, which contains the daily minimum temperature

   (a) For k $\in \{365, 182, 91\}$, create a new feature column which is the serie of temperature lagged by k days. You could use the pandas function shift, and get the autocorrelation $\mathrm{Cor}(x_t, x_{t-k})$ between the original and lagged series. You can use the `pandas.plotting.autocorrelation_plot` function.

   (b) On the same figure, plot the three previous autocorrelation points and the global autocorrelation curve (for any possible k): the x-axis corresponds to k and the y-axis to the correlation.

   (c) Give an interpretation to this figure.

4. Define a function which takes as arguments (y_{t-1}, a, σ), and returns the predicted value y_t (scalar) defined in slide 17 of ATI.6 where:

   - y_{t-1} is a vector of length d: the d previous values of the serie,
   - a is a vector of length d: the Yule-Walker coefficients of deepness d,
   - σ is a scalar: the constant in the autoregressive formula.

5. Define a function which takes as arguments (`time_series`, `deep`, `n_pred`, `n_train`) and returns the predicted time series (of length `n_pred`), where:

   - `time_series` is a time series (i.e. a pandas dataframe or series) of length more than `n_train` + `n_pred`,

- `deep` is the deepness value for autoregression formula and for the Yule-Walker coefficients (it must be smaller than `n_train`),
- `n_pred` is the number of values of the series to predict: with indexes `n_train` $\rightarrow$ `n_train` + `n_pred - 1`,
- `n_train` is the number of the first values of the series to use for the training: with indexes 0 $\rightarrow$ `n_train - 1`.

You can use the previous function and the function

`statsmodels.api.regression.yule_walker`.

6. For the first dataset only, which contains the daily minimum temperature data, for each `deep` $\in \{7, 30, 365\}$ predict the last year of data

   (`length(time series) - 365` and `n_pred = 365`) and find the Mean Square Error (MSE) with the real values of the last year of the series. Comment on you results.

Submission

Please archive your report and codes in "Prénom Nom.zip" (replace "Prénom" and "Nom" with your real name), and upload to "`Upload Extra TP2: Time series prediction.`" on `https://moodle.unige.ch` before **Monday, December 11 2023, 21:59 PM**. Note, that the assessment is mainly based on your report, which should include your answers to all questions and the experimental results. *Importance is given on the mathematical explanations of your works and your codes should be commented*

Please use this overleaf template.