GenHack3 - Simulation of the maize yield across the years in the context of the climate change

February 1, 2024

In Task 1, the goal was to build a generator on a specific subset of the data, corresponding roughly to a nationwide dry summer. In Task 2, we extend this problem to build a generator for multiple weather scenarios, mimicking a conditional generative problem.

1 Task 2: Yield generation for multiple meteorological scenarios

The objective in this second round is to build a generative model able to simulate yields distribution given 9 meteorological scenarios.

We define new random variables $\bar{T}$ and $\bar{R}$ being respectively the national average daily maximum temperature **from May to October** (i.e. period 1 to 9) and the national average of daily rainfall **in summer** (i.e. perdiod 4 to 6). Mathematically, for a given year $i \in \{1, 2, \ldots, 10000\}$:

$$\bar{T}_i = \frac{1}{4 \times 9} \sum_{j=1}^{4} \left(W_{j,i}^{(1)} + W_{j,i}^{(2)} + W_{j,i}^{(2)} + W_{j,i}^{(4)} + W_{j,i}^{(5)} + W_{j,i}^{(6)} + W_{j,i}^{(7)} + W_{j,i}^{(8)} + W_{j,i}^{(9)} \right),$$

$$\bar{R}_i = \frac{1}{4 \times 3} \sum_{j=1}^{4} (W_{j,i}^{(13)} + W_{j,i}^{(14)} + W_{j,i}^{(15)}),$$

where $W_{j,i}^{(k)}$ corresponds to the k-th component of the weather variable W_j at the j-th station during the i-th year.

Recall that for each station $j \in \{1, \ldots, 4\}$, the weather variable W_j is defined as

$$W_j := \begin{bmatrix} W_j^{(1)} \\ \vdots \\ W_j^{(9)} \\ W_j^{(10)} \\ \vdots \\ W_j^{(18)} \end{bmatrix} = \begin{bmatrix} \texttt{average}\text{ daily maximum temperature - period 1 - station j} \\ \vdots \\ \texttt{average}\text{ daily maximum temperature - period 9 - station j} \\ \texttt{average}\text{ daily rainfall - period 1 - station j} \\ \vdots \\ \texttt{average}\text{ daily rainfall - period 9 - station j} \end{bmatrix}.$$

The period definition can be found in the previous document.

Interpretation

The global variables we just defined are critical variables to determine crop growth:

- $\bar{T}$ is up to a linear transformation the approximated cumulative Growing degree-day (GDD). It is a heuristic to measure the heat accumulated by the crop during its growth period.
- $\bar{R}$ is up to a linear transformation the average accumulated rainfall during summer. The latter is the period when the hydric sensitivity of the maize is the most important.

Then, we average the variables over the four stations to study global scenarios and simplify the problem.

Then, we split the data according to 3 classes of national temperatures in °C

$$\mathcal{C}_{\mathrm{T}} := \{T_1 =]-\infty, 21.2], T_2 =]21.2, 22], T_3 =]22, +\infty[\} ,$$

and 3 classes of national rainfalls in $\mathrm{mm/m^2}$

$$\mathcal{C}_{\mathrm{R}} := \{R_1 =]-\infty, 1.8], R_2 =]1.8, 2.2], R_3 =]2.2, +\infty[\} .$$

Computing the Cartesian product of the two set:

$$\mathcal{C} = \mathcal{C}_{\mathrm{R}} \times \mathcal{C}_{\mathrm{T}} = \{(r, t) \mid r \in \mathcal{C}_{\mathrm{R}} \text{ and } t \in \mathcal{C}_{\mathrm{T}}\} ,$$

yield to 9 categories which we label by a vector $\mathbf{x}_k$ of standard base, where $\mathbf{x}_k$ denotes the vector with 1 in the k-th coordinate and 0's elsewhere. For example, in dimension 9,

$$\mathbf{x}_6 := \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

denotes the sixth scenario where the national average of daily rainfall is between 1.8 $\mathrm{mm/m^2}$ and 2.2$\mathrm{mm/m^2}$; and where the national average daily maximum temperature is above 22°C. It corresponds to a moderate rainfall and high temperature scenario. All scenarios are stored in Table 1 with their associated number of data in the training set.

1.1 What you need to deliver

At the second evaluation, we will evaluate your model for each $k \in \{1, \ldots, 9\}$

$$Z = \begin{bmatrix} Z^{(1)} \\ \vdots \\ Z^{(d_z)} \end{bmatrix}, \mathbf{x}_k = \begin{bmatrix} x_k^{(1)} \\ \vdots \\ x_k^{(9)} \end{bmatrix} \mapsto G_\theta(Z, \mathbf{x}_k)$$

Scenario	C_{R}	C_{T}	n_{data}
1	$R_1 =]-\infty, 1.8]$	$T_1 =]-\infty, 21.2]$	464
2	$R_1 =]-\infty, 1.8]$	$T_2 =]21.2, 22]$	1290
3	$R_1 =]-\infty, 1.8]$	$T_3 =]22, +\infty[$	1678
4	$R_2 =]1.8, 2.2]$	$T_1 =]-\infty, 21.2]$	534
5	$R_2 =]1.8, 2.2]$	$T_2 =]21.2, 22]$	1254
6	$R_2 =]1.8, 2.2]$	$T_3 =]22, +\infty[$	1082
7	$R_3 =]2.2, +\infty[$	$T_1 =]-\infty, 21.2]$	1007
8	$R_3 =]2.2, +\infty[$	$T_2 =]21.2, 22]$	1690
9	$R_3 =]2.2, +\infty[$	$T_3 =]22, +\infty[$	1001

Table 1: Description of scenarios with varying n_{data}.

where G_θ is your generative model parameterized by θ with a mandatory output structure

$$G_\theta\left(\begin{bmatrix} Z^{(1)} \\ \vdots \\ Z^{(d_Z)}, \end{bmatrix}, \begin{bmatrix} x_k^{(1)} \\ \vdots \\ x_k^{(9)} \end{bmatrix}\right) = \begin{bmatrix} \widetilde{Y}^{(1)} \\ \vdots \\ \widetilde{Y}^{(4)} \end{bmatrix},$$

with a common and unknown random vectors Z with $d_z \leq 50$ (**the latent dimension is free as long as it is less than** 50).

Objective For each meteorological scenario $\mathbf{x}_k$, for $k \in \{1, \dots, 9\}$, you have to simulate data $(\widetilde{Y}^{(1)}, \dots, \widetilde{Y}^{(4)})$ similar to the real yields $(Y^{(1)}, \dots, Y^{(4)})$ **given** they were harvested during this k-th scenario.

Information
You can either train a single categorical conditional generator or 9 independent ones.
The problem as it is posed now with categorical conditioning variables (labels) is similar to a categorical conditional generative model. For example, one can train such a model to generate MNIST digits conditionally on the label (*i.e.* the digit).

2 Train-test dataset

The training dataset will be the same, and we will use **different** independent latent variables $Z_1, Z_2, \dots, Z_{n_{\mathrm{eval}}}$ as input noise to generate the appropriate number of testing data in each scenario, where n_{eval} is the number of evaluation points associated to each scenario.

3 Evaluation

The criterion for evaluation is still the *Sliced Wasserstein Distance* that we will compute on each scenario. The evaluation score will be a weighted average of all SWD for each scenario, where the weights are defined with respect to its number of training data, *i.e.*

$$\mathcal{L} = \sum_{k=1}^{9}\left(1 - \frac{n_{\mathrm{data}}^{(k)}}{\sum_{j=1}^{9} n_{\mathrm{data}}^{(j)}}\right)\ell_k,$$

where ℓ_k is the estimated SWD (using the same projection angles as in the first evaluation) computed between the real data and the generated ones for the k-th scenario. We will use $n_{\text{eval}} = 10,000$ **evaluation points** for each scenario.

> **Watch out**
> The smaller the number of training data, the larger the weight in the evaluation score.

3.1 Ranking

Similar as the one in Evaluation 1

4 List of files

The data files of the stations are the same as in Evaluation 1.

- `data/`: folder containing the training data (`station_49.csv`, `station_80.csv`, `station_40.csv`, `station_63.csv`) and an example of a noise file (`noise.npy`). Feel free to use another noise for training your model but keep in mind that
  - d_z must be less or equal to 50,
  - you will be evaluated on a common (to all participants) and unknown standard normal random vector Z.
- `requirements.txt`: text file containing the libraries with their associated versions you used in the `model.py` file. **Do modify ✓**
- `Dockerfile`: docker image in order to create a container. **Do not modify ✗**
- `main.py`: **new** main python file containing the simulation function. **Do not modify ✗**
- `model.py`: **new** python file containing your conditional generative model and for loading the parameters. This file has been modified since your generator must contains now both the latent and the conditional variables. **Do modify ✓**
- `parameters/`: folder where you must put the parameters of your model. **Do modify ✓**
- `run.sh`: bash script to run your simulation. **Do not modify ✗**

Submission. Similar as the one in Evaluation 1

5 Schedule

- Evaluation #1: January 31, 2024 - 11:59 pm Paris time
- Evaluation #2: February 10, 2024 - 11:59 pm Paris time