Machine Learning

Nijmegen, Netherlands

Instructor: George Miloshevich

Problem 3

§ Heat waves §

You will use the tools from machine learning to classify heat waves over the region of interest with data generated by a climate simulation of intermediate complexity: PlaSim. This project is based on active area of research (e.g. Freddy Bouchet group at ENS de Lyon), where investigators are attempting to see how much skill can be extracted from data to predict extreme events using neural networks. Extreme events are rare by definition, thus require extensive statistics in order to be describable. This implies great computational costs, especially when it comes to running highly complex realistic models. Thus one of the hopes of the community is that models will eventually merge with machine learning tools to simplify the analysis.

In this project, we will simply attempt to predict heat waves over a large area ($1000km \times 1000km$). The precise definition will be based on the 95-th percentile of the following time series:

$$A(t) := \frac{1}{T}\int_t^{t+T} \frac{1}{|\mathcal{D}|}\int_D \left(T_{2m} - \mathbb{E}\left(T_{2m}\right)\right)(\vec{r}, u)\mathrm{d}\vec{r}\,\mathrm{d}u, \tag{0.1}$$

where $T = 5$ days corresponds to the duration of the event of interest, T_{2m} a two-meter temperature field, referred to as tas in the files we provide and $\mathcal{D}$ a domain of interest. We have tested the case of “France”, defined as the area enclosed by 5 W, 8 E longitudes and 42 N, 52 N latitudes. In principle, for excess mortality yearly maxima could be more important to look at, but since the project is not so applied, we work with anomalies. This way we achieve richer statistics as the whole summer is taken as the time domain of heat waves.

It is known that heat waves may be correlated with soil moisture deficits as well as stationary anticyclonic geopotential height anomalies. These variables are expected to play a role because the former correlates with latent heat evaporation, whereas the latter with advection. Thus we provide you with the relevant fields, above 30 N latitude. In principle deep learning is highly demanding when it comes to the amount of data. To achieve a balance between usability and reduce the memory footprint we have provided you with only 500 years of the simulation, where climatology has been subtracted. However this amount of data is much larger than what one has in the observational record. Another thing to consider is that training is most efficient when performed on powerful GPUs. If you are running the jupyter notebook ew2.ipynb locally it is recommended to have a good machine, or resort to Google Colab, or perform manual data reduction. The instructions are written directly inside the notebook, and in fact the first problem requires you to execute provided scripts. Since you will be working in groups it is a good idea to plan how you can divide labor. We have also included assignments labelled “OPTIONAL”, which require extra work. You should skip them when first working on the notebook. Only do them if you have extra time.

$$H = -\mathbb{E}_X \sum_{k=0}^{K-1} Y_k \log\left[\hat{p}_k(x)\right], \text{for K classes} \tag{0.2}$$

We note that when performing classification neural networks typically minimize the cross-entropy (see equation (0.2)). This achieves the goal of maximizing the probability that the model corresponds to the reality given the data (provided the latter is representative). This way the loss corresponds to the logarithmic probabilistic score, which penalizes heavily the labels that are mis-classified. The algorithm computes the gradients with respect to the weights and biases in each layer using back-propagation. The successive layers allow us to retain the information that is hierarchical, so theoretically one expects that the data processing capability improves with deeper networks. However, as in conventional machine learning (simple classifiers) no algorithm can perform well if not provided enough data, especially if the models are particularly complex (have too many weights). In image recognition the revolution in the last decade happened precisely because of availability of data, where highly over-parametrized deep models became very successful. When studying extreme events, one encounters the technical difficulties of imbalanced datasets, which do play an important role in data science, such as banking fraud prevention etc. When predicting heat waves however, we have an additional aspect: the fact that a priori we do not expect to the patterns seen in geopotential, soil moisture etc to carry the full information for deterministic classification. Even if we had a complete state of the system because of chaos the deterministic prediction is known to be impossible.

Please see the relevant files on https://github.com/AlessandroLovo/EW2-heatwaves

Problem 1: Basic data operations, plotting

(1) Fetch the data we uploaded to Google Drive/owncloud

(2) Visualize the data, describe the typical spatial and time scales.

(3) Compute 2-m temperature anomaly $A(t)$ for $T = 5$ and choose the appropriate area, e.g. France

(4) Assign labels based on the 95th percentile of $A(t)$

Problem 2: Prepare the pipelines

(1) Stack the fields (temperature, geopotential, soil moisture) and possibly remove the irrelevant information

(2) Split data into training and validation and perform data normalization or rescaling (0-1)

(3) OPTIONAL: *Implement k-fold cross-validation* to allow measurements of the skill error

(4) Construct the neural network architecture: either convolutional or fully-connected

Problem 3: Train Neural network

(1) Perform the fit choosing the optimizer and compiling with the appropriate learning rate

(2) Evaluate the predictive capability as a function of training epoch

(3) OPTIONAL: *Implement early stopping callback*

(4) OPTIONAL: *How does your success scale with the amount of data you use to train your network?*

Problem 4: Analysis

(1) Feed different inputs inside the architecture and describe what matters for successful prediction

(2) Generate images (possibly also composite maps) of True/False Positives and Negatives

(3) OPTIONAL: Predict heat waves 15 to 30 days in advance. What changes?

(4) OPTIONAL: Perform hyper-parameter optimization via our optuna tutorial