Assignment 3 (ML for TS) - MVA 2022/2023

Matias Etcheverry matias.etcheverry9@gmail.com
Amric Trudel amric.trudel@ens-paris-saclay.fr

April 7, 2023

1 Introduction

Objective. The goal is to implement (i) a signal processing pipeline with a change-point detection method and (ii) wavelets for graph signals.

Warning and advice.

- Use code from the tutorials as well as from other sources. Do not code yourself well-known procedures (e.g. cross validation or k-means), use an existing implementation.
- The associated notebook contains some hints and several helper functions.
- Be concise. Answers are not expected to be longer than a few sentences (omitting calculations).

Instructions.

- Fill in your names and emails at the top of the document.
- Hand in your report (one per pair of students) by Friday 7th April 11:59 PM.
- Rename your report and notebook as follows:
`FirstnameLastname1_FirstnameLastname1.pdf` and
`FirstnameLastname2_FirstnameLastname2.ipynb`.
For instance, `LaurentOudre_CharlesTruong.pdf`.
- Upload your report (PDF file) and notebook (IPYNB file) using this link: https://www.dropbox.com/request/rmETjrLAH9Li3pf8JvOt.

2 Dual-tone multi-frequency signaling (DTMF)

In the last tutorial, we started designing an algorithm to infer from a sound signal the sequence of symbols encoded with DTMF.

Question 1

Finalize this procedure–in particular, find the best hyperparameters. Describe in 5 to 10 lines your methodology and the calibration procedure (give the hyperparameter values).

Answer 1

In this question, we assume we don't know the number of symbols encoded with DTMF. In order to elaborate a working pipeline, we first create a training and a validation dataset made of 100 signals each, with varying number of symbols.

The training dataset is used to create the pipeline, in order to test our ideas. On the contrary the validation dataset is used to find the best hyperparameters. The pipeline is fully described in the notebook. It is mainly divided into 2 steps:

- fit a penalized change point detection to the spectrogram of the signal
- filter and merge detection to produced a consistent list of symbols

However, in order to find the best hyperparameters, we must define a relevant metric. In order to compare sequences of symbol. We use the Levenshtein distance, which historically compares words. This distance is resilient to varying number of symbols between ground-truth sequences and predictions. We finally define an accuracy between a ground-truth list of symbols and its prediction as the normalized Levenshtein distance.

We have the following hyperparameters:

- `nperseg`: Length of each segment in the spectrogram. Set to 512.
- `noverlap`: Number of points to overlap between segments in the spectrogram. Set to 6.
- `min_size`: Minimum length of the segment . Set to 2.
- `min_break_time`: Minimum duration between 2 segment so that they are not merged. Set to 0.2s.
- `min_energy`: Minimum energy of a segment to be declared so. Set to 0.1.

Question 2

What are the two symbolic sequences encoded in the provided signals?

Answer 2

- Sequence 1: `B94B38B#1`
- Sequence 2: `CD11263`

Note: a final 9 is missing in sequence 2. Our best hyperparameters wouldn't find all the symbols at the same time.

3 Wavelet transform for graph signals

Let G be a graph defined a set of n nodes V and a set of edges E. A specific node is denoted by v and a specific edge, by e. The eigenvalues and eigenvectors of the graph Laplacian L are $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $u_1, u_2, \ldots, u_n$ respectively.

For a signal $f \in \mathbb{R}^n$, the Graph Wavelet Transform (GWT) of f is $W_f : \{1, \ldots, M\} \times V \longrightarrow \mathbb{R}$:

$$W_f(m, v) := \sum_{l=1}^{n} \hat{g}_m(\lambda_l) \hat{f}_l u_l(v) \tag{1}$$

where $\hat{f} = [\hat{f}_1, \ldots, \hat{f}_n]$ is the Fourier transform of f and $\hat{g}_m$ are M kernel functions. The number M of scales is a user-defined parameter and is set to $M := 9$ in the following. Several designs are available for the $\hat{g}_m$; here, we use the Spectrum Adapted Graph Wavelets (SAGW). Formally, each kernel $\hat{g}_m$ is such that

$$\hat{g}_m(\lambda) := \hat{g}^U(\lambda - am) \quad (0 \leq \lambda \leq \lambda_n) \tag{2}$$

where $a := \lambda_n / (M + 1 - R)$,

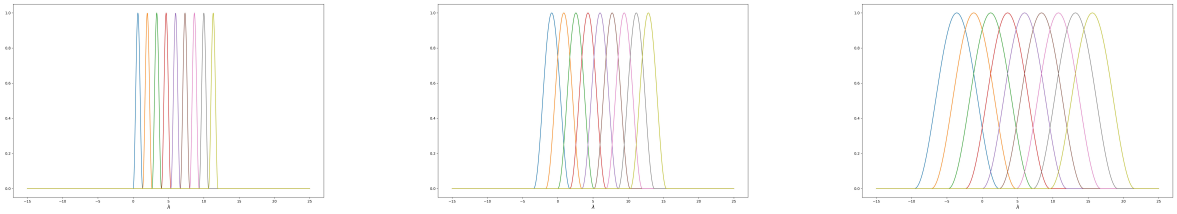
$$\hat{g}^U(\lambda) := \frac{1}{2}\left[1 + \cos\left(2\pi\left(\frac{\lambda}{aR} + \frac{1}{2}\right)\right)\right] \mathbb{1}(-Ra \leq \lambda < 0) \tag{3}$$

and $R > 0$ is defined by the user.

Question 3

Plot the kernel functions $\hat{g}_m$ for $R = 1$, $R = 3$ and $R = 5$ (take $\lambda_n = 12$) on Figure 1. What is the influence of R?

Answer 3



(a) $R = 1$ (b) $R = 3$ (c) $R = 5$

Figure 1: The SAGW kernels functions

R influences on the smoothness of the signal. Small R will focus on localized patterns while bigger R will highlight large and gradual phenomena over a graph.

We will study the Molene data set (the one we used in the last tutorial). The signal is the temperature.

Question 4

Construct the graph using the distance matrix and exponential smoothing (use the median heuristics for the bandwidth parameter).

- Remove all stations with missing values in the temperature.
- Choose the minimum threshold so that the network is connected and the average degree is at least 3.
- What is the time where the signal is the least smooth?
- What is the time where the signal is the smoothest?

Answer 4

The stations with missing values are Arzal, Batz, Beg Meil, Brest-guipavas, Brignogan, Camaret, Landivisiau, Lannaero, Lanveoc, Ouessant-stiff, Plouay-sa, Ploudalmezeau, Plougonvelin, Quimper, Riec sur belon, Sizun, St nazaire-montoir and Vannes-meucon.

The threshold is equal to 0.8318.

The signal is the least smooth at 2014-01-10 09:00:00.

The signal is the smoothest at 2014-01-24 19:00:00.

Question 5

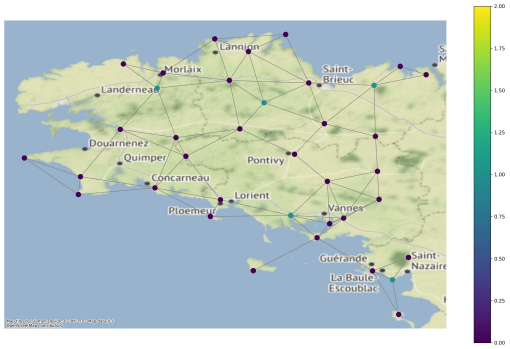
(For the remainder, set $R = 3$ for all wavelet transforms.)

For each node v, the vector $[W_f(1, v), W_f(2, v), \ldots, W_f(M, v)]$ can be used as a vector of features. We can for instance classify nodes into low/medium/high frequency:

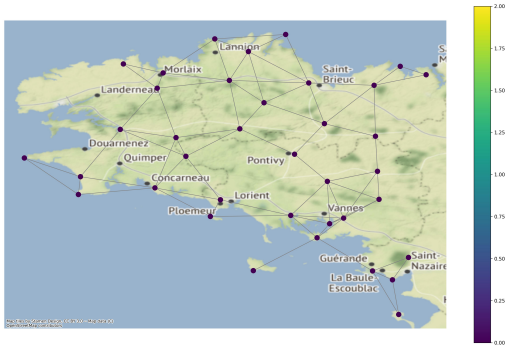
- a node is considered low frequency if the scales $m \in \{1, 2, 3\}$ contain most of the energy,
- a node is considered medium frequency if the scales $m \in \{4, 5, 6\}$ contain most of the energy,
- a node is considered high frequency if the scales $m \in \{6, 7, 9\}$ contain most of the energy.

For both signals from the previous question (smoothest and least smooth) as well as the first available timestamp, apply this procedure and display on the map the result (one colour per class).

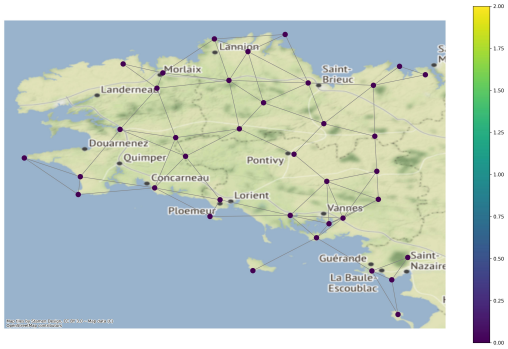
Answer 5



(a) Least smooth signal



(b) Smoothest signal



(c) First available timestamp

Figure 2: Classification of nodes into low/medium/high frequency

Even for the least smooth signal, we mainly have low frequency classes. We suspect an error in the code as we would have expected more medium and high frequency classes.

Question 6

Display the average temperature and for each timestamp, adapt the marker colour to the majority class present in the graph (see notebook for more details).

Answer 6

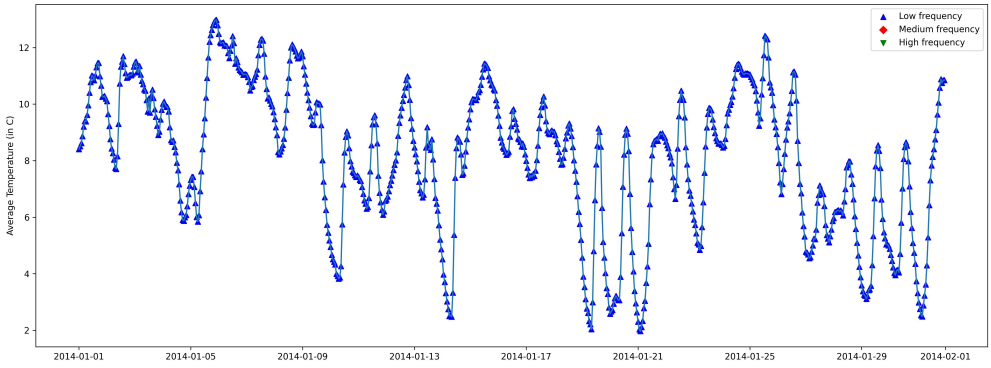


Figure 3: Average temperature. Markers' colours depend on the majority class.

The appearance of this curve follows the previous remark. We expected more medium and high frequency majority classes, especially when the signal is moving abruptly, like on the 21st.

Question 7

The previous graph G only uses spatial information. To take into account the temporal dynamic, we construct a larger graph H as follows: a node is now *a station at a particular time* and is connected to neighbouring stations (with respect to G) and to itself at the previous timestamp and the following timestamp. Notice that the new spatio-temporal graph H is the Cartesian product of the spatial graph G and the temporal graph G' (which is simply a line graph, without loop).

- Express the Laplacian of H using the Laplacian of G and G' (use Kronecker products).
- Express the eigenvalues and eigenvectors of the Laplacian of H using the eigenvalues and eigenvectors of the Laplacian of G and G'.
- Compute the wavelet transform of the temperature signal.
- Classify nodes into low/medium/high frequency and display the same figure as in the previous question.

Answer 7

- Let $\otimes$ be the Kronecker product and $H = G \otimes G'$. The Laplacian of H is given by:

$$L(H) = L(G) \otimes I_{|V(G)|} + I_{|V(G')|} \otimes L(G')$$

- Let n and m be the number of vertices of G and G' respectively. Thus, the Laplacian of H has $n + m$ eigenvectors. Let a and b be 2 eigenvectors of G and G' respectively, with eigenvalues λ and μ. Then:

$$\begin{aligned} L(H)(a \otimes b) &= L(G) \otimes I_n + I_m \otimes L(G')(a \otimes b) \\ &= (L(G) \otimes I_n)(a \otimes b) + (I_m \otimes L(G'))(a \otimes b) \\ &= (L(G)a) \otimes (I_n b) + (I_m a) \otimes (L(G')b) \\ &= (\lambda a) \otimes b + a \otimes (\mu b) \\ &= (\lambda + \mu)(a \otimes b) \end{aligned}$$

  This proves that $a \otimes b$ is a eigenvector of H with eigenvalue $\lambda + \mu$. In the end, the eigenvectors of H are all the combinations $a \otimes b$ where a and b are eigenvectors of G and G' with eigenvalues $\lambda + \mu$.

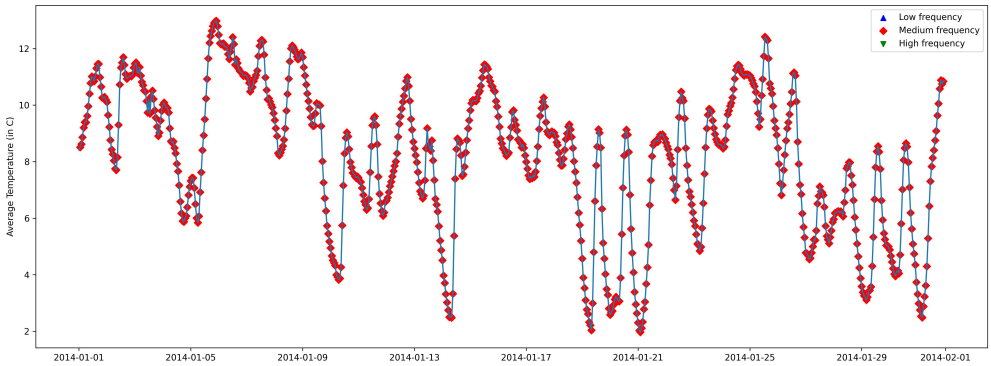


Figure 4: Average temperature. Markers' colours depend on the majority class.

When considering temporal information, we expect smoother marker color changes than on question 6. The curve above is not accurate.