Assignment 1 (ML for TS) - MVA 2021/2022

Eya Ghamgui eya.ghamgui@telecom-paris.fr
Siwar Mhadhbi siwar.mhadhbi@telecom-paris.fr

June 28, 2022

1 Introduction

Objective. The goal is to learn to apply the convolutional dictionary learning procedure and the dynamic time warping distance on the real medical application.

2 General questions

Question 1

Consider the following Lasso regression:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2 \quad + \quad \lambda \|\beta\|_1 \tag{1}$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ the design matrix, $\beta \in \mathbb{R}^p$ the vector of regressors and $\lambda > 0$ the smoothing parameter.

Show that there exists $\lambda_{\max}$ such that the minimizer of (1) is $\mathbf{0}_p$ (a p-dimensional vector of zeros) for any $\lambda > \lambda_{\max}$.

Answer 1

By applying first order condition on (1), we obtain:

$$-X^T(y - X\hat{\beta}) = \lambda\, s \quad (\star)$$

where $s = (s_j)_{j \in [1,..,p]} = \begin{cases} \{+1\} & \text{if} \quad \hat{\beta} > 0 \\ \{-1\} & \text{if} \quad \hat{\beta} < 0 \\ [-1, 1] & \text{if} \quad \hat{\beta} = 0 \end{cases}$

We look for λ_{max} such that the minimizer of (1) is 0_p, *i.e.* $\hat{\beta} = 0_p$.

$$\begin{aligned} \hat{\beta} = 0_p &\iff -X^T y = \lambda\, s \ \ (\text{from}(\star)) \\ &\iff \lambda\, s = -X^T y \ \ (\star\star) \end{aligned}$$

If $\forall j \in \{1, .., p\}, s_j \notin \{-1, 1\}, i.e. |s_j| < 1 \iff ||s||_\infty < 1$, then we have $\hat{\beta} = 0_p$.

$$(\star\star) \iff \lambda||s||_\infty = ||X^T y||_\infty$$

We have,

$$\begin{aligned} ||s||_\infty < 1 &\iff \forall \lambda > 0,\ \lambda||s||_\infty < \lambda \\ &\iff \forall\ \lambda > 0,\ ||X^T y||_\infty < \lambda,\ \text{hence a condition on } \lambda. \end{aligned}$$

Thus, $\forall \lambda > ||X^T y||_\infty$, we have $\hat{\beta} = 0_p$.

Therefore, there exists λ_{max} such that $\hat{\beta} = 0_p$ that we denote as:

$$\lambda_{\max} = ||X^T y||_\infty \tag{2}$$

Question 2

For a univariate signal $\mathbf{x} \in \mathbb{R}^n$ with n samples, the convolutional dictionary learning task amounts to solving the following optimization problem:

$$\min_{(\mathbf{d}_k)_k, (\mathbf{z}_k)_k \| \mathbf{d}_k \|_2^2 \leq 1} \quad \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right\|_2^2 \quad + \quad \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 \tag{3}$$

where $\mathbf{d}_k \in \mathbb{R}^L$ are the K dictionary atoms (patterns), $\mathbf{z}_k \in \mathbb{R}^{N-L+1}$ are activations signals, and $\lambda > 0$ is the smoothing parameter.

Show that

- for a fixed dictionary, the sparse coding problem is a lasso regression (explicit the response vector and the design matrix);
- for a fixed dictionary, there exists $\lambda_{\max}$ (which depends on the dictionary) such that the sparse codes are only 0 for any $\lambda > \lambda_{\max}$.

Answer 2

- We have the following optimization problem:

$$\min_{(\mathbf{d}_k), (\mathbf{z}_k) \| \mathbf{d}_k \|_2^2 \leq 1} \quad \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right\|_2^2 \quad + \quad \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1$$

  First, we want to write the vector $\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k$ as $\mathbf{DZ}$ where $\mathbf{D} \in \mathbb{R}^{N \times KN}$ and $\mathbf{Z} \in \mathbb{R}^{KN}$.

  We have, for element $i \in [1, N]$:

  - $(\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k)_i = \sum_{k=1}^{K} \sum_{l=1}^{N} \mathbf{z}_k(l) \mathbf{d}_k(i-l) \quad (\star)$

    such that $\mathbf{z}_k(l) = \begin{cases} (\mathbf{z}_k)_l & \text{if } \ 1 \leq l \leq N - L + 1 \\ 0 & \text{otherwise} \end{cases}$

    and $\quad \mathbf{d}_k(i-l) = \begin{cases} (\mathbf{d}_k)_{i-l} & \text{if } \ 1 \leq i - l \leq L \\ 0 & \text{otherwise} \end{cases}$

– $(\mathbf{DZ})_i = \sum_{j=1}^{KN} \mathbf{D}_{ij}\mathbf{Z}_j \quad (\star\star)$

We can write the matrix $\mathbf{D}$ as:

$$\mathbf{D} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 & \dots & \mathbf{H}_K \end{bmatrix} \text{ where } \mathbf{H}_k \text{ is a matrix} \in \mathbb{R}^{N\times N}$$

and the vector $\mathbf{Z}$ as:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_k = \begin{bmatrix} \vdots \\ \mathbf{z}_k(l) \\ \vdots \end{bmatrix} \\ \vdots \\ \mathbf{z}_K \end{bmatrix} \text{ where the } \mathbf{z}_k(l) \text{ are defined previously}$$

Now, using these new definitions, we can write

$$\begin{aligned}(\mathbf{DZ})_i &= (\sum_{k=1}^{K} \mathbf{H}_k\mathbf{Z}_k)_i \\ &= \sum_{k=1}^{K} (\mathbf{H}_k\mathbf{Z}_k)_i \\ &= \sum_{k=1}^{K}\sum_{l=1}^{N} (\mathbf{H}_k)_{il}(\mathbf{z}_k)_l \\ &= \sum_{k=1}^{K}\sum_{l=1}^{N} (\mathbf{z}_k)(l)(\mathbf{H}_k)_{il}\end{aligned}$$

$$\begin{aligned}(\star) = (\star\star) &\iff (\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k)_i = (\mathbf{DZ})_i \\ &\iff \sum_{k=1}^{K}\sum_{l=1}^{N} \mathbf{z}_k(l)\mathbf{d}_k(i-l) = \sum_{k=1}^{K}\sum_{l=1}^{N} \mathbf{z}_k(l)(\mathbf{H}_k)_{il} \\ &\iff (\mathbf{H}_k)_{il} = \mathbf{d}_k(i-l) \text{ where the } \mathbf{d}_k(i-l) \text{ are defined previously}\end{aligned}$$

Finally, we found that $\sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k = \mathbf{DZ}$.

Second, $\sum_{k=1}^{K} \|\mathbf{z}_k\|_1 = \sum_{k=1}^{K}\sum_{l=1}^{N-L+1} |(\mathbf{z_k})_l| = \sum_{k=1}^{K}\sum_{l=1}^{N} |\mathbf{z_k}(l)|$ where $\mathbf{z}_k(l)$ are defined previously

Also, we have: $\sum_{k=1}^{K}\sum_{l=1}^{N} |\mathbf{z_k}(l)| = \sum_{k=1}^{KN} |Z_k| = \|\mathbf{Z}\|_1$. Thus, $\sum_{k=1}^{K} \|\mathbf{z}_k\|_1 = \|\mathbf{Z}\|_1$

In this question, we have shown that the sparse coding problem, for a fixed dictionary, can be written as:

$$\min_{\mathbf{Z}\in\mathbb{R}^{KN}} \quad \|\mathbf{x} - \mathbf{DZ}\|_2^2 \quad + \quad \lambda \|\mathbf{Z}\|_1$$

which is a lasso regression problem.

- For a fixed dictionary, we found that the sparse coding problem is a lasso regression problem. Using question 1, we can say that there exists

$$\lambda_{max} = \left\| \mathbf{D}^T \mathbf{x} \right\|_\infty \quad \text{where } \mathbf{D} \text{ is defined previously}$$

such that the minimizer of the lasso problem is 0_p for any $\lambda > \lambda_{max}$. That is, the sparse codes $\mathbf{Z}$ are only 0_p for any $\lambda > \lambda_{max}$.

3 Data study

3.1 General information

Context. The study of human gait is a central problem in medical research with far-reaching consequences in the public health domain. This complex mechanism can be altered by a wide range of pathologies (such as Parkinson's disease, arthritis, stroke,...), often resulting in a significant loss of autonomy and an increased risk of fall. Understanding the influence of such medical disorders on a subject's gait would greatly facilitate early detection and prevention of those possibly harmful situations. To address these issues, clinical and bio-mechanical researchers have worked to objectively quantify gait characteristics.

Among the gait features that have proved their relevance in a medical context, several are linked to the notion of step (step duration, variation in step length, etc.), which can be seen as the core atom of the locomotion process. Many algorithms have therefore been developed to automatically (or semi-automatically) detect gait events (such as heel-strikes, heel-off, etc.) from accelerometer and gyrometer signals.

Data. Data are described in the associated notebook.

3.2 Step detection with convolutional dictionary learning

Task. The objective is to perform **step detection**, that is to estimate the start and end times of footsteps contained in accelerometer and gyrometer signals recorded with Inertial Measurement Units (IMUs).

Performance metric. Step detection methods will be evaluated with the **F-score**, based on the following precision/recall definitions. The F-score is first computed per signal then averaged over all instances. Precision and recall rely on the "intersection over union" metric (IoU) that measures the overlap of two intervals $[s_1, e_1]$ and $[s_2, e_2]$:

$$\text{IoU} = \frac{\left|[s_1, e_1] \cap [s_2, e_2]\right|}{\left|[s_1, e_1] \cup [s_2, e_2]\right|}$$

- Precision (or positive predictive value). A detected (or predicted) step is counted as correct if it overlaps (measured by IoU) an annotated step by more than 75%. The precision is the number of correctly predicted steps divided by the total number of predicted steps.
- Recall (or sensitivity). An annotated step is counted as detected if it overlaps (measured by IoU) a predicted step by more than 75%. The recall is the number of detected annotated steps divided by the total number of annotated steps.

The F-score is the geometric mean of the precision and recall:

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Note that an annotated step can only be detected once, and a predicted step can only be used to detect one annotated step. If several predicted steps correspond to the same annotated step, all but one are considered as false. Conversely, if several annotated steps are detected with the same predicted step, all but one are considered undetected.

Example 1.

- Annotation ("ground truth label"): $\big[[80, 100], [150, 250], [260, 290]\big]$ (three steps)
- Prediction: $\big[[80, 98], [105, 120], [256, 295], [298, 310]\big]$ (four steps)

Here, precision is $0.5 = (1 + 0 + 1 + 0)/4$, recall is $0.67 = (1 + 0 + 1)/3$ and the F-score is 0.57.

Example 2.

- Annotation ("ground truth label"): $\big[[80, 120]\big]$ (one step)
- Prediction: $\big[[80, 95]\big]$ (one step)

Here, precision is $0 = 0/1$, recall is $0 = 0/1$ and the F-score is 0.

Question 3

For a single signal, learn a dictionary with manually chosen penalty, number of atoms and length.

Modify Figure 1 to display the original signal and its reconstruction. Modify Figure 2 to display the individual atoms.

Answer 3

To fit the model on the training set, the chosen parameters are $\lambda = 0.1$, $K = 3$ and, $L = 80$.

The reconstruction error (MSE) is equal to 0.0112. The F-score is equal to 0.96.

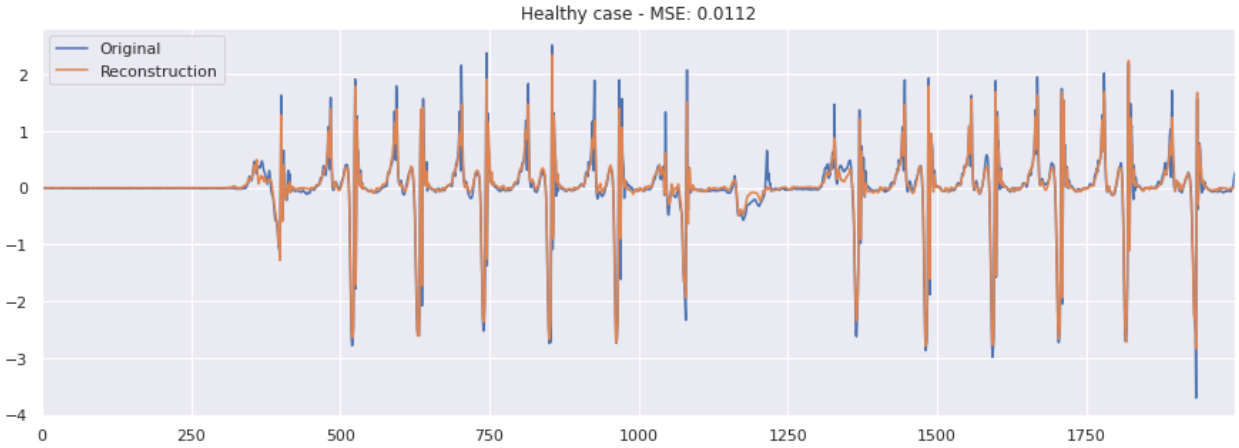


Figure 1: Original signal and its reconstruction (see Question 3).

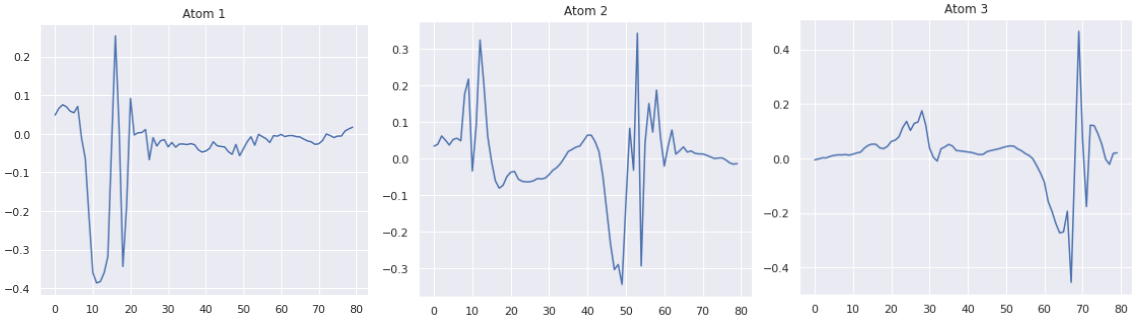


Figure 2: Individual atoms (see Question 3).

Comment:

By changing the values of the hyperparameters manually, we can notice some variance of the F-score as shown in the attached notebook where F-scores go from 0.53 to 0.96 depending also on the underlying signal. Thus, we need to perform a grid search to choose the best combination of parameters that maximizes the F-score value. Moreover, for the same combination of parameters, running the code several times gives different scores, which means that the calculation of the steps to compute the score is too sensitive to small variations.

Question 4

Using only the training set, find with a 5-fold cross-validation among the candidates values (see notebook) the best combination of (λ, K, L) for the step detection task.

Provide the optimal values of (λ, K, L) and the associated average F-score and MSE.

Answer 4

We performed a 5-fold cross-validation on the training set in order to find the best combination of (λ, K, L) for the step detection task.

As the cross-validation process takes too much time in our problem, we performed on intuitively-chosen values among the provided candidates :

- penalty : $\lambda \in [0.1, 0.2, 0.5, 0.8]$
- number of atoms : $K \in [2, 3, 4, 5, 6, 7]$
- length of atoms: $L \in [50, 80, 100, 150]$

After around four hours of training, the best combination we obtained is : $(\lambda, K, L) = (0.1, 7, 80)$ with average F-score equal to 0.87 and MSE around 0.0075.

Comment:

In questions 3 and 4, in order to compute the steps signals, we chose to select the best atom that gives the best F-score. Indeed, this atom is chosen to describe the beginning of the step pattern, which leads us to the best starting point of the step, and thus to a correct detection of the steps signals.

Question 5

Display on Figure 3 the atoms learned for Question 4.

Answer 5

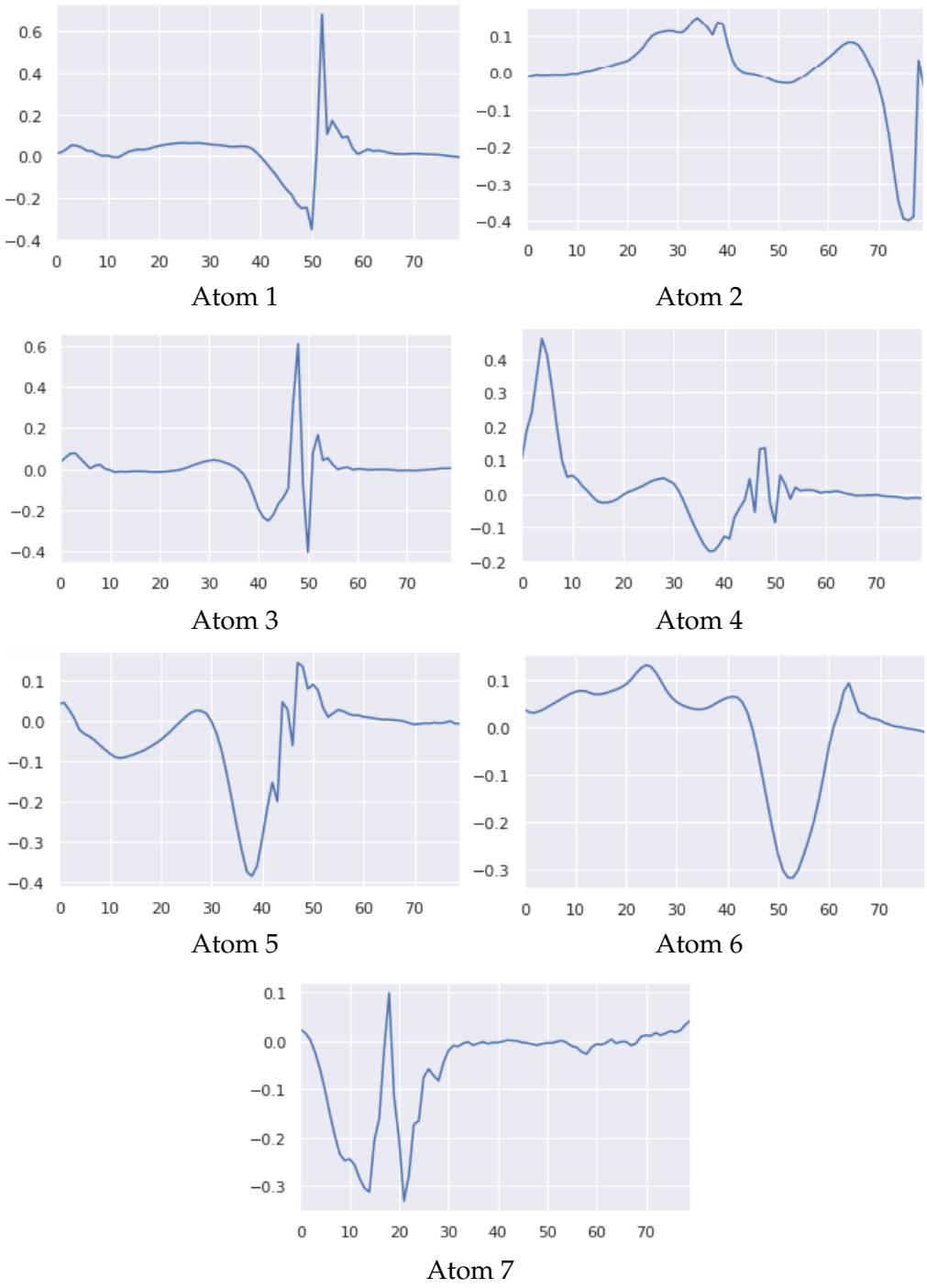


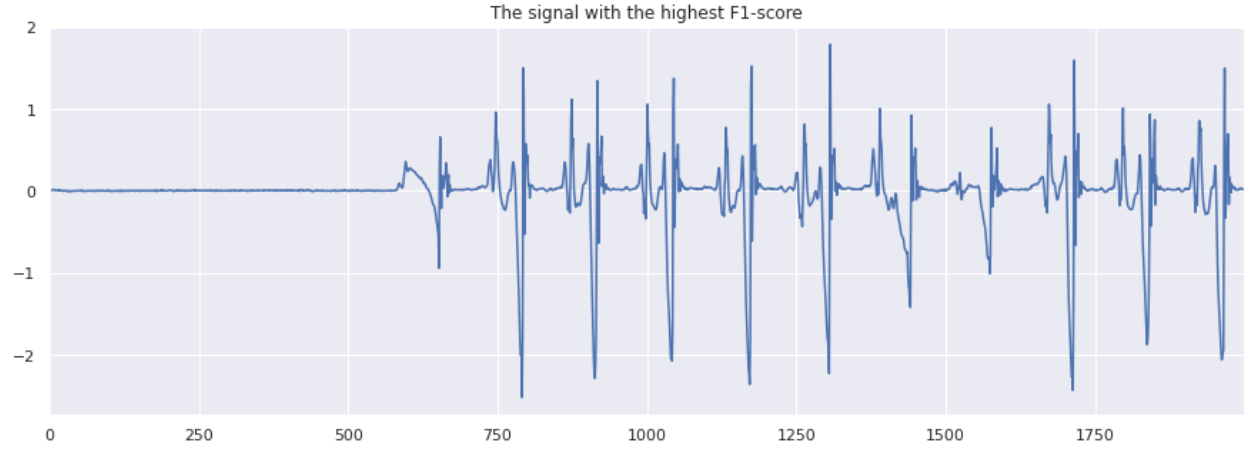
Figure 3: Individual atoms (see Question 5).

Question 6

Display on Figure 4 the signals from the test set with highest and lowest F-score. Comment briefly.

Answer 6

Applying the model with the best hyperparameters on the test set, we obtained 0.78 as F-score. We display in Figure 4 the signals with the best and the worst F-score.



Test signal with the best F-score

Test signal with the worst F-score

Figure 4: Best and worst scores (see Question 6).

Comment:

The signal with the highest F-score represents regular periodic patterns. These patterns are smooth and thus the steps detection task becomes easier. Whereas the signal with the lowest F-score contains noisy patterns with strong perturbations that mislead the steps detection. This may be the reason why the learned atoms are more similar to the patterns of the first signal, hence the discrepancy between the corresponding F-score values.

3.3 Step classification with the dynamic time warping (DTW) distance

Task. The objective is to classify footsteps then walk signals between healthy and non-healthy.

Performance metric. The performance of this binary classification task is measured by the F-score.

Question 7

Combine the DTW and a k-neighbors classifier to classify each step. Find the optimal number of neighbors with 5-fold cross-validation and report the optimal number of neighbors and the associated F-score. Comment briefly.

Answer 7

The goal of this part is to find the best $KNeighborsClassifier$. To do so, we started by dividing the data into training and test sets, with 25% of the data for the test set.

Applying 5-fold $GridSearchCV$ on $KNeighborsClassifier$ on the train set with DTW distance as metric and $f1$-score for scoring, we obtained 5 as best number of neighbors with $f1$-score equal to 0.86. We tried various numbers from 3 to 9 with a step of 2 to obtain odd numbers for our binary classification task. We discarded the value 1 to avoid the case of assigning to each step signal the class of only the closest one which is a trivial case and does not lead in general to good classification unlike assigning the majoritarian class of the nearest ones.

Evaluating the best trained estimator gives $f1$-score equal to 0.82 on the test set.

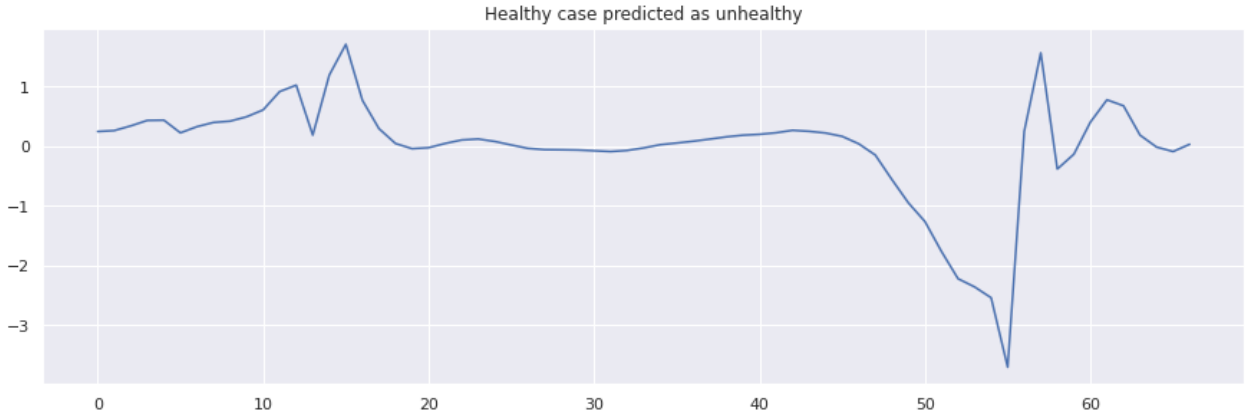
Comment:

Our best estimator classifies a given step signal as "healthy" or "non-healthy" based on its 5 nearest step signals determined using DTW distance and labels it as "healthy" if at least 3 of the nearest signals are "healthy" or "non-healthy" otherwise.

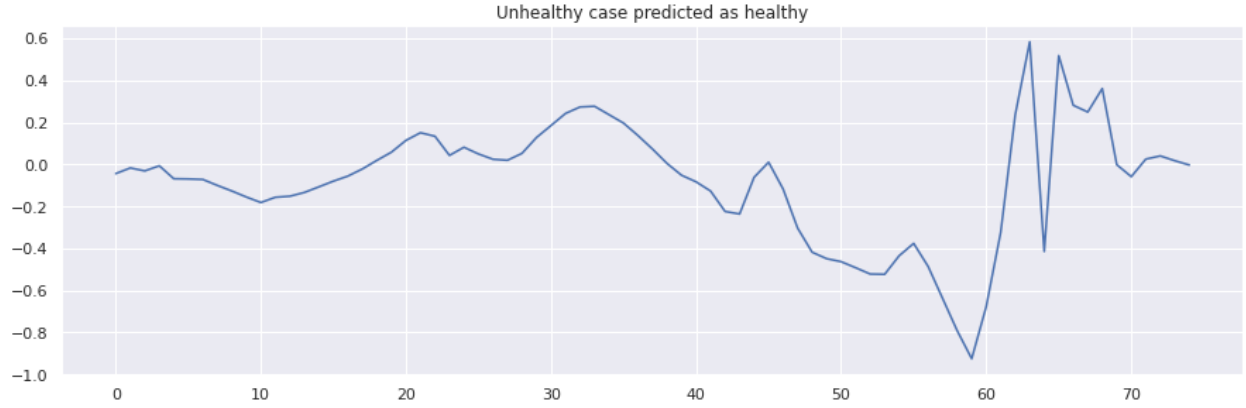
Question 8

Display on Figure 5 a badly classified step from each class (healthy/non-healthy).

Answer 8



Badly classified healthy step



Badly classified non-healthy step

Figure 5: Examples of badly classified steps (see Question 8).

Comment:

From the previous plots, we can clearly differentiate the two signals as "healthy" or "non-healthy". In fact, the healthy signal is smoother while the non-healthy signal has more perturbations. However, the general forms of the two steps signals are close, which may explain the misclassification. Indeed, for each of these signals, the closest 3 or more signals are of the opposite class.