Assignment 1 (ML for TS) - MVA 2021/2022
refrence : http://www.laurentoudre.fr/ast.html

January 31, 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

First of all, by noting, for a sufficiently large fixed λ, β^* the solution vector of the optimization problem (1), we can notice that:

$$\frac{1}{2} \|y - X\beta^*\|_2^2, \lambda \|\beta^*\|_1 \leq \frac{1}{2} \|y - X\beta^*\|_2^2 + \lambda \|\beta^*\|_1 \leq \frac{1}{2} \|y - X0\|_2^2 + \lambda \|0\|_1 = \frac{1}{2} \|y\|_2^2$$

Thus, the (1) problem can be reformulated as follow:

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

Let $\beta \in \mathbb{R}^p$ such that $\|y - X\beta\|_2 \leq \|y\|_2$ and $\|\beta\|_1 \leq \frac{1}{2\lambda} \|y\|_2^2$, let $\beta^{(i)}$ be the i-th component of β, $\beta^{(-i)}$ the vector β where the i-th component is set to 0 and finally X_i the i-th column of X. We have:

$$\begin{aligned}\Delta_i &:= (\frac{1}{2}\left\|y - X\beta\right\|_2^2 + \lambda\left\|\beta\right\|_1) - (\frac{1}{2}\left\|y - X\beta^{(-i)}\right\|_2^2 + \lambda\left\|\beta^{(-i)}\right\|_1) \\ &= (\frac{1}{2}\left\|y - X\beta\right\|_2^2 - \frac{1}{2}\left\|y - X\beta^{(-i)}\right\|_2^2) + \lambda|\beta^{(i)}| \\ &= (\frac{1}{2}\left\|y - X\beta\right\|_2^2 - \frac{1}{2}\left\|y - X\beta + \beta^{(i)}X_i\right\|_2^2) + \lambda|\beta^{(i)}| \\ &= (-\frac{1}{2}\beta^{(i)2}\left\|X_i\right\|_2^2 - \beta^{(i)} < y - X\beta, X_i >) + \lambda|\beta^{(i)}| \\ &= -\frac{1}{2}|\beta^{(i)}|^2\left\|X_i\right\|_2^2 - \beta^{(i)} < y - X\beta, X_i > +\lambda|\beta^{(i)}|\end{aligned}$$

Using the Cauchy-Schwarz inequality, we get:

$$|-\beta^{(i)} < y - X\beta, X_i > | \leq |\beta^{(i)}|\left\|y - X\beta\right\|_2\left\|X_i\right\|_2 \leq |\beta^{(i)}|\left\|y\right\|_2\left\|X_i\right\|_2$$

Hence:

$$\Delta_i \geq -\frac{1}{2}|\beta^{(i)}|^2\left\|X_i\right\|_2^2 - |\beta^{(i)}|\left\|y\right\|_2\left\|X_i\right\|_2 + \lambda|\beta^{(i)}| = |\beta^{(i)}|((\lambda - \left\|y\right\|_2\left\|X_i\right\|_2) - \frac{1}{2}|\beta^{(i)}|\left\|X_i\right\|_2^2)$$

Thus for $\Delta_i \geq 0$ it suffices that:

$$(\lambda - \left\|y\right\|_2\left\|X_i\right\|_2) - \frac{1}{2}|\beta^{(i)}|\left\|X_i\right\|_2^2 \geq 0 \Leftrightarrow 2\frac{\lambda - \left\|y\right\|_2\left\|X_i\right\|_2}{\left\|X_i\right\|_2^2} \geq |\beta^{(i)}|$$

And since $|\beta^{(i)}| \leq \left\|\beta\right\|_1 \leq \frac{1}{2\lambda}\left\|y\right\|_2^2$ it suffices that: $2\frac{\lambda - \left\|y\right\|_2\left\|X_i\right\|_2}{\left\|X_i\right\|_2^2} \geq \frac{1}{2\lambda}\left\|y\right\|_2^2$, which is equivalent to:

$$\begin{aligned}2\frac{\lambda - \left\|y\right\|_2\left\|X_i\right\|_2}{\left\|X_i\right\|_2^2} \geq \frac{1}{2\lambda}\left\|y\right\|_2^2 &\Leftrightarrow \left\|y\right\|_2^2\left\|X_i\right\|_2^2 \leq 4\lambda(\lambda - \left\|y\right\|_2\left\|X_i\right\|_2) = (2\lambda - \left\|y\right\|_2\left\|X_i\right\|_2)^2 - \left\|y\right\|_2^2\left\|X_i\right\|_2^2 \\ &\Leftrightarrow \sqrt{2}\left\|y\right\|_2\left\|X_i\right\|_2 \leq 2\lambda - \left\|y\right\|_2\left\|X_i\right\|_2 \\ &\Leftrightarrow \frac{1+\sqrt{2}}{2}\left\|y\right\|_2\left\|X_i\right\|_2 \leq \lambda\end{aligned}$$

Hence:

$$\begin{aligned}\frac{1+\sqrt{2}}{2}\left\|y\right\|_2\left\|X_i\right\|_2 \leq \lambda \Rightarrow \Delta_i &\geq 0 \\ &\Rightarrow \frac{1}{2}\left\|y - X\beta^{(-i)}\right\|_2^2 + \lambda\left\|\beta^{(-i)}\right\|_1 \leq \frac{1}{2}\left\|y - X\beta\right\|_2^2 + \lambda\left\|\beta\right\|_1\end{aligned}$$

Therefore, if $\frac{1+\sqrt{2}}{2}\left\|y\right\|_2 \underset{1\leq i\leq p}{max}\left\|X_i\right\|_2 \leq \lambda$ we would have for every component i: $\frac{1}{2}\left\|y - X\beta^{*(-i)}\right\|_2^2 + \lambda\left\|\beta^{*(-i)}\right\|_1 = \frac{1}{2}\left\|y - X\beta^*\right\|_2^2 + \lambda\left\|\beta^*\right\|_1$, and therefore by uniqueness of the minimum value: $\beta^{*(-i)} = \beta^*$, thus $\beta^* = 0$. Therefore:

$$\lambda_{\max} = \frac{1+\sqrt{2}}{2} \|y\|_2 \max_{1 \le i \le p} \|X_i\|_2 \tag{2}$$

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:

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

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

Using the discrete Fourier transform for series of length n, while padding if necessary using zeros, let X be the discrete Fourier transform of x seen as a column (i.e. $X \sim [n, 1]$), Z_k the discrete Fourier transform of z_k seen as a column (i.e. $Z_k \sim [n, 1]$), Z the vertical stacking of the Z_k (i.e. $Z_k \sim [nK, 1]$), D_k the Fourier transform of d_k seen as a column (i.e. $D_k \sim [n, 1]$), and finally D the matrix of size $\sim [n, nK]$ made of $\{D_k\}_{1 \le k \le K}$, where for a given $1 \le i \le n$: $D_{i,(k-1)n+i} = D_k[i]$ for all $1 \le k \le K$, whereas $D_{i,\cdot} = 0$ elsewhere.

By Parsevale's theorem we have:

$$\begin{aligned}
\left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{z}_k * \mathbf{d}_k \right\|_2^2 + \lambda \sum_{k=1}^{K} \|\mathbf{z}_k\|_1 &= \left\| \mathbf{X} - \sum_{k=1}^{K} \mathbf{Z}_k \odot \mathbf{D}_k \right\|_2^2 + \lambda \sum_{k=1}^{K} \|\mathbf{Z}_k\|_1 \\
&= \sum_{i=1}^{n} |\mathbf{X}[i] - \sum_{k=1}^{K} \mathbf{D}_k[i] \mathbf{Z}_k[i]|^2 + \lambda \|\mathbf{Z}\|_1 \\
&= \sum_{i=1}^{n} |\mathbf{X}[i] - \sum_{k=1}^{K} D_{i,(k-1)n+i} Z_{(k-1)n+i}|^2 + \lambda \|\mathbf{Z}\|_1 \\
&= \sum_{i=1}^{n} |\mathbf{X}[i] - \sum_{c=1}^{nK} D_{i,c} Z_c|^2 + \lambda \|\mathbf{Z}\|_1 \\
&= \sum_{i=1}^{n} |\mathbf{X}[i] - \mathbf{D}\mathbf{Z}[i]|^2 + \lambda \|\mathbf{Z}\|_1 \\
&= \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_2^2 + \lambda \|\mathbf{Z}\|_1
\end{aligned}$$

Thus for a fixed dictionary $\mathbf{D}$, the problem of learning a sparse coding is a Lasso regression problem:

$$\underset{Z \in \mathbb{R}^{nK}}{min} \|\mathbf{X} - \mathbf{D}\mathbf{Z}\|_2^2 + \lambda \|\mathbf{Z}\|_1$$

For such a problem, the matrix $\mathbf{D}$ defined above is the design matrix, while $\mathbf{X}$ is the response vector.

Using the result of the previous question, the penalty λ from which the zero vector is the solution to the previous problem is at least:

$$\lambda_{\max} = \frac{1+\sqrt{2}}{2} \|\mathbf{X}\|_2 \underset{1 \leq k \leq K}{max} \|\mathbf{D}_k\|_2 = \frac{1+\sqrt{2}}{2} \|\mathbf{x}\|_2 \underset{1 \leq k \leq K}{max} \|\mathbf{d}_k\|_2 \leq \frac{1+\sqrt{2}}{2} \|\mathbf{x}\|_2 \tag{4}$$

where we have used the fact that the L_2 norms of the columns of D are the same as the L_2 norms of the discrete Fourier transforms D_k.

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 $[s_2, e_2]$:

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

- 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

For this fitting, the chosen parameters are $\lambda = 0.5$, $K = 5$, $L = 90$.

The reconstruction error (MSE) is equal to $4.52 \cdot 10^{-3}$. The F-score is equal to 0.37.

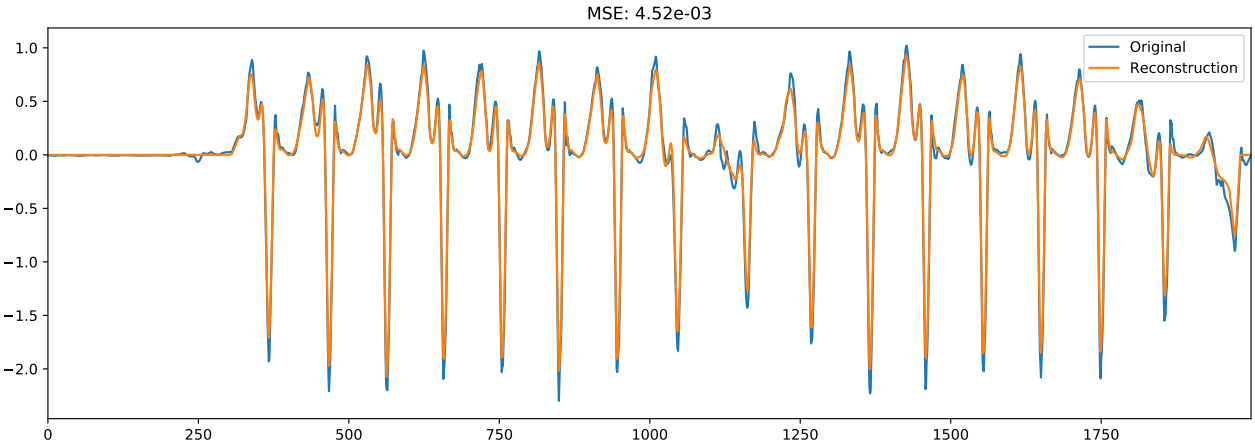


Figure 1: Original signal and its reconstruction (see Question 3).

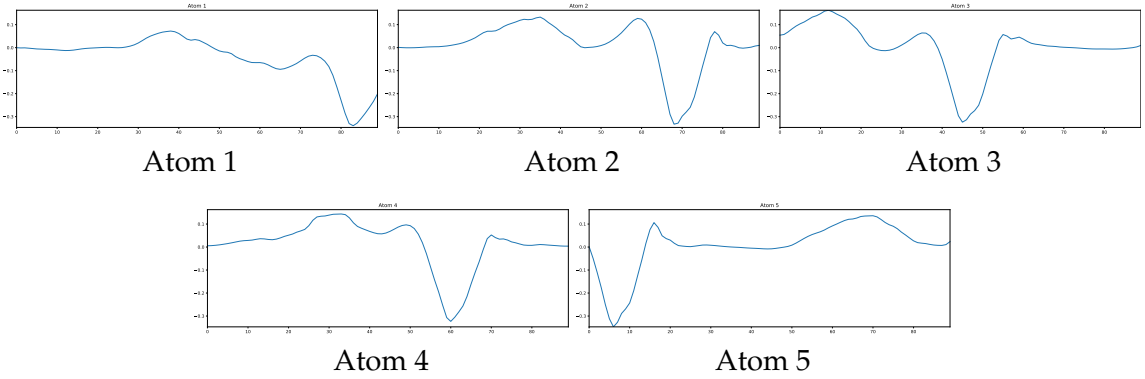


Figure 2: Individual atoms (see Question 3).

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

The optimal parameters found at the end of the grid-search, with a cross-validation procedure using 5-folds, are:

$$\lambda = 0.9, K = 4, L = 80$$

The average associated F_1 score is:

$$F_1 = 0.595(+/\text{-}0.592)$$

The root mean square error associated to such model is:

$$MSE = 7.47 \cdot 10^{-2}$$

Question 5

Display on Figure 3 the atoms learned for Question 4.

Answer 5

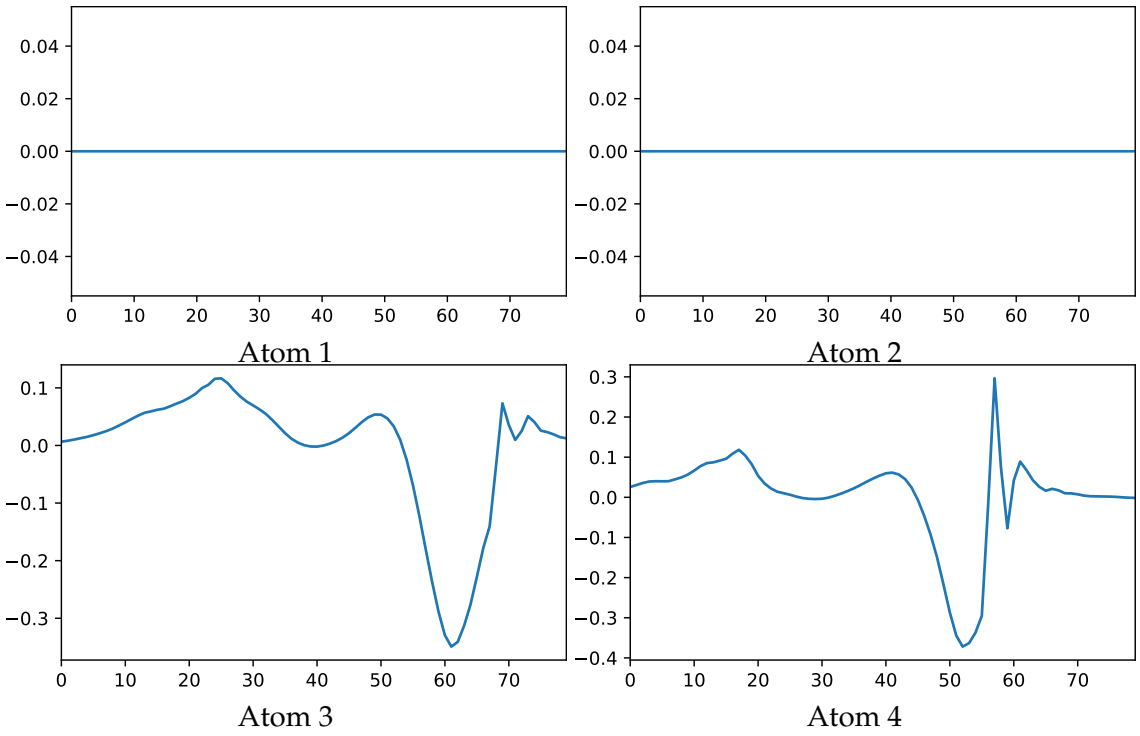
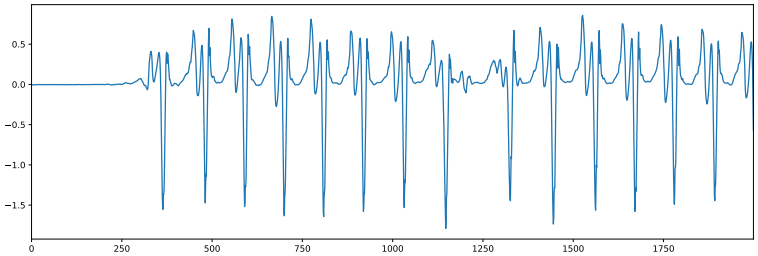


Figure 3: Individual atoms (see Question 5).

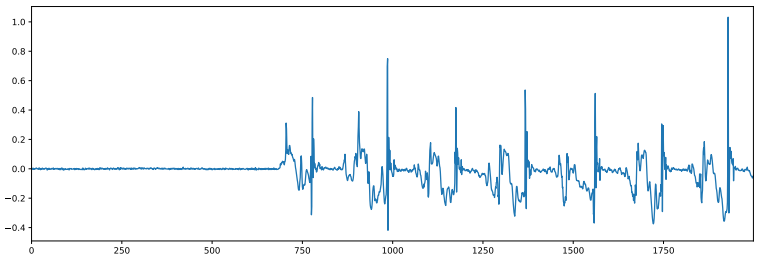
Question 6

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

Answer 6



Test signal with the best F-score



Test signal with the worst F-score

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

We can notice that the signal with the best-detected footsteps is a very regular signal with a large amplitude, while the signal with the worst-detected footsteps is a signal whose amplitude is - relatively - lower, with a very low level of regularity , to the point that footsteps can even be mistaken for noise!

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 search for the optimal number of neighbors was performed through a cross-validation procedure involving the range of values between 1 and 20.

The optimal number of neighbors found by such procedure is:

$$K = 5$$

The associated model has a score F_1 equal to:

$$F_1 = 0.770(+/\text{-}0.134)$$

We can then observe that the increase in the number of neighbors used in a KNN model does not necessarily imply an increase in the performance of such a model.

Question 8

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

Answer 8

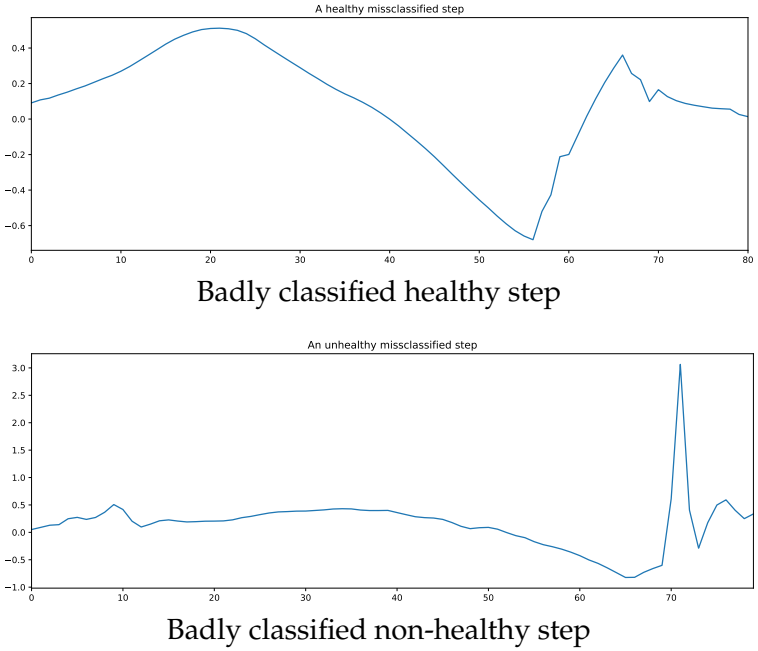


Figure 5: Examples of badly classified steps (see Question 8).