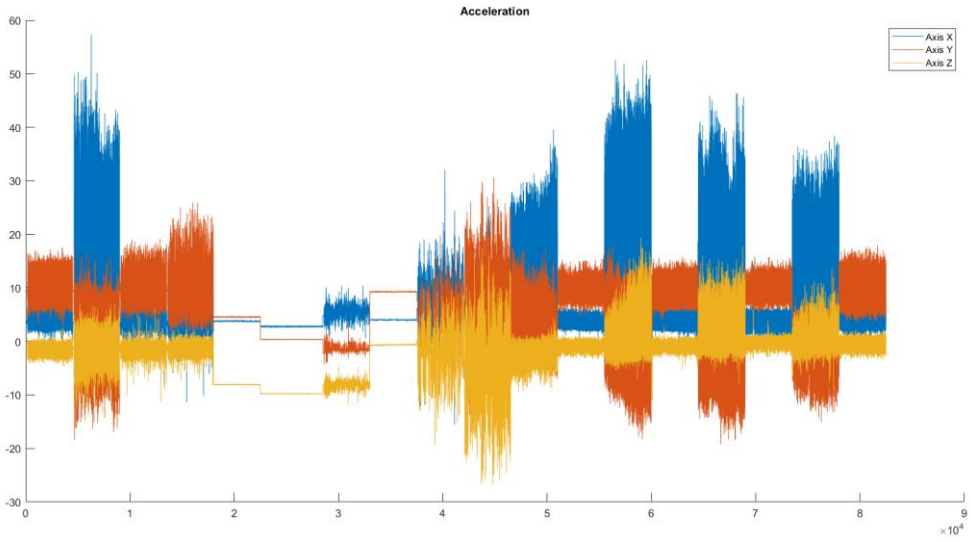
Classification Validation Methods and Decision Tree Classification

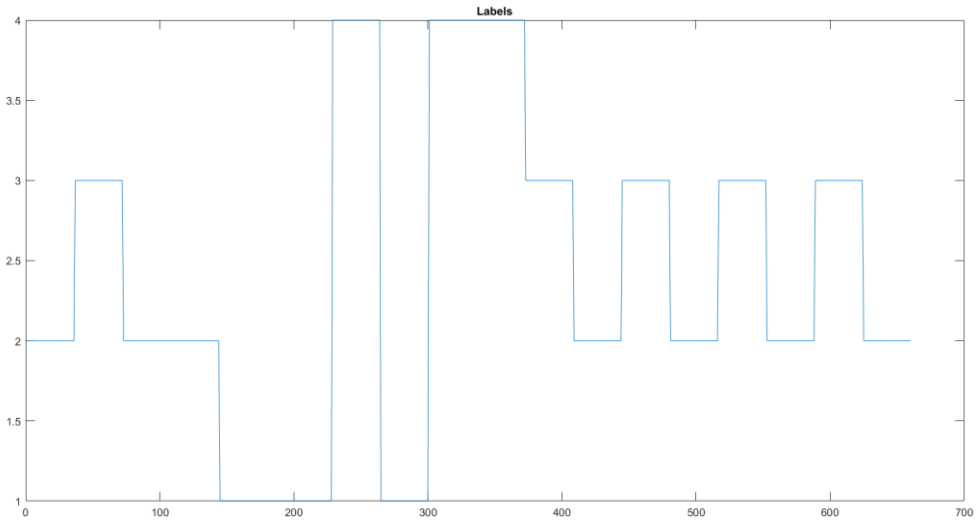
Author: Klára Martinásková

The aim of these lab assignment was practical experience on how to deploy a multi-class classification tree-based algorithm to classify between different physical activities. My assignment was implemented in MATLAB software in two scripts (main script *"task3"* and script for function *"MeanAndStd"*).

After loading training data, I display the signals together in one graph. It is possible to discern from this graph the changing activity during the duration of the signal. If the signal has significant and regular peaks, it is probably a sports activity (running). Conversely, if the signal is without peaks, we can say that it is static activity. Smaller distinct peaks, which are regular, can be described as walking. Peaks that resemble noise are labelled as "*others*".

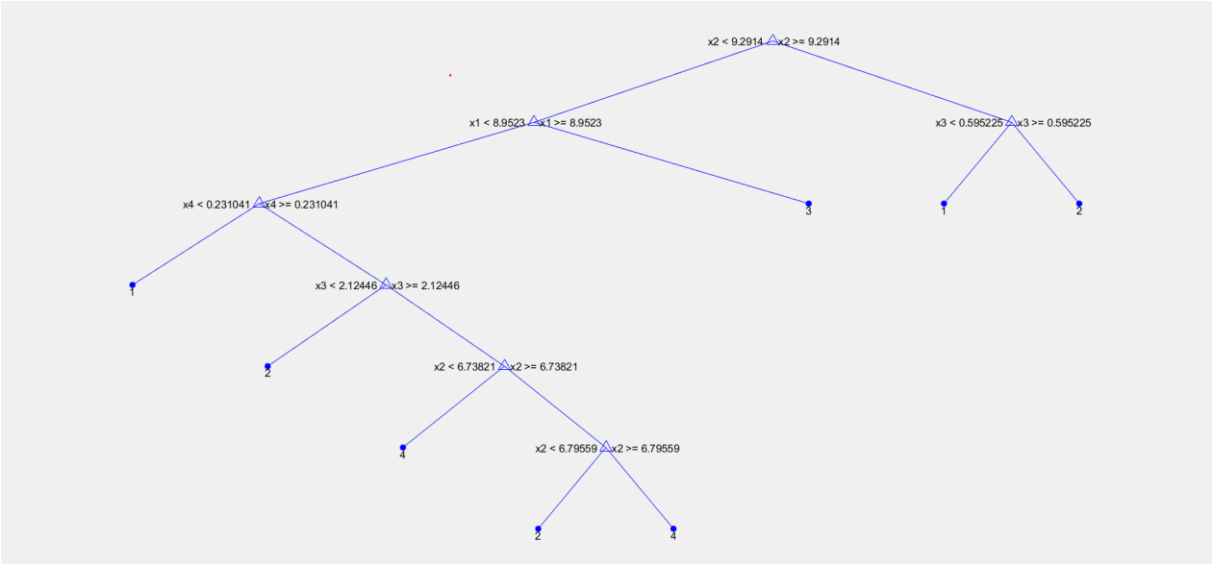


Graph 1: Plot signals



Graph 2: Lables for signal

Next step is extracting features. For each segment the program calculates mean and std for each axis – thanks function "*MeanAndStd*". Based on features we are able to train a decision tree using MATLAB function "*fitctree*". Thanks to the graphical representation we are able to understand very well how the classifier works.



Graph 3: Decision tree

After training the tree, we need to evaluate the classifier to see how well our classifier works. In this assignment, we obtained another data for validation, which were measured in the same way but from another individual. Other solutions for having separate data for validation set could be randomly splitting the dataset into training and validation sets or cross-validation.

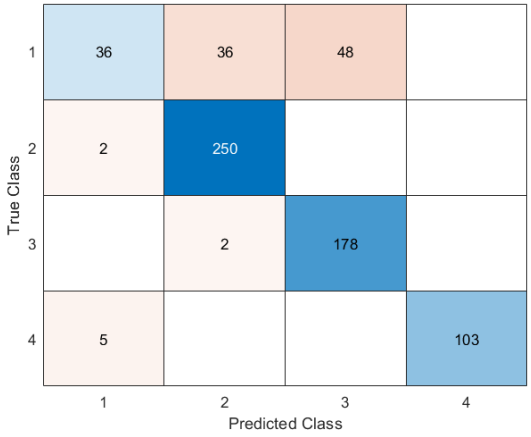
The program finds the same features based on the validation set and classifies the evaluation set features using function "*predict*".

To determine the success of the program the programs finds:

- Confusion Matrix
- True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) of each activity type
- True Positive Rate (sensitivity/recall), True Negative Rate (specifity/selectivity), False Positive Rate (false alarm rate/type I error), False Negative Rate (miss rate/type II error) and Accuracy for each activity type

The results are demonstrated in the following tables.

Confusion Matrix:



TP, TN, FP, FN:

	TP	TN	FP	FN
Label 1	36	533	7	84
Label 2	250	370	2	38
Label 3	178	432	48	2
Label 4	103	552	0	5

TPR, TNR, FPR, FNR, Accuracy:

	TPR	TNR	FPR	FNR	Accuracy
Label 1	0.300	0.987	0.013	0.700	0.862
Label 2	0.992	0.907	0.093	0.008	0.939
Label 3	0.989	0.900	0.100	0.011	0.924
Label 4	0.954	1	0	0.046	0.992

The overall accuracy of the classifier is high with an average accuracy of 92.95 %. The classifier performs the best for Label 4 with a TPR of 0.954 and an accuracy of 99.2 %, which means that it correctly identified 95.4 % of the instances for Label 4. The classifier has the lowest performance for Label 1, with the lowest TPR of 0.3 and an accuracy of 86.2 %, indicating that it struggled to distinguish Label 1 from the other activities. Therefore, the classifier's performance varied across different activity types, with Label 4 being the best-performing and Label 1 being the worst-performing.