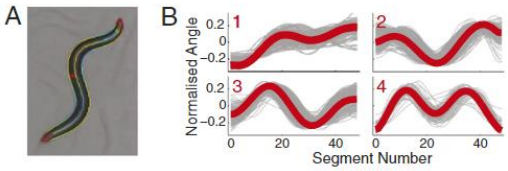
Ciências ULisboa | Informática

Advanced Machine Learning

Project 2 – Temporal Data

Introduction

Caenorhabditis elegans is a roundworm commonly used as a model organism in the study of genetics. The movement of these worms is known to be a helpful indicator for understanding behavioral genetics. Brown et al. (2013) describe a system for recording the motion of worms on an agar plate and measuring a range of human-defined features. It has been shown that the space of shapes *Caenorhabditis elegans* adopts on an agar plate can be represented by combinations of four base shapes, or eigenworms. Once the worm outline is extracted, each frame of worm motion can be captured by four scalars representing the amplitudes along each dimension when the shape is projected onto the four eigenworms.



Dataset

The data relates to 258 traces of worms converted into four "eigenworm" series. The eigenworm data are lengths from 17984 to 100674 (sampled at 30 Hz, so from 10 minutes to 1 hour) and in four dimensions (eigenworm 1 to 4). There are five classes: N2, goa-1, unc-1, unc-38 and un63. N2 is wildtype (i.e., normal) the other 4 are mutant strains. These datasets are the first dimension only (first eigenworm) averaged down so that all series are lengths 900 (the single hour-long series is discarded). This smoothing is likely to discard discriminatory information.

We address the problem of classifying individual worms as wild-type or mutant based on the time series of the first eigenworm, down-sampled to second-long intervals. We have 258 cases, which we split 70%/30% into a train and test set. Each series has 900 observations, and each worm is classified as either wild-type (the N2 reference strain – 109 cases; class 1) or one of four mutant types (149 cases; class 2): goa-1 (44 cases); unc-1 (35 cases); unc-38 (45 cases) and unc-63 (25 cases). The data were extracted from the *C. elegans* behavioral database WormWeb.

The following files are provided:

- "worms_trainset.csv" – file with the training dataset, where the first column has the class label (181x901 matrix)
- "worms_testset.csv" – file with the test dataset, where the first column has the class label (77x901 matrix)

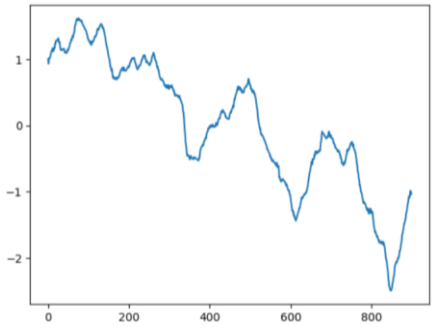
Work plan

This project has two main research questions (or two main parts):

1. Can we classify the type of worm using the information provided by the eigenworm series?
2. For a specific worm, how can we model its motion, i.e., the eigenworm?

To answer those research questions, we need to perform the following steps:

- Prepare the data in the correct formats to be used by the methods and pre-process the data, if required.
- Identify the best classifier model and/or representation method for our dataset; this might include hyperparameter selection in some cases.
- Perform time series analysis to model the movement of one single worm. In this case, we will consider the worm in the train set indexed by 5 (see figure below).



Submission

Each group should submit the report written in Jupyter Notebook with the name **AAA2324_P2_xx_yy.ipynb**, where xx and yy should be replaced by each of the student numbers of the group. The report should include (but not limited to):

- The identification of the members of the group
- Exploratory data analysis and preprocessing steps required for the project
- Implementation of the models considered to answer each research question
- Discussion of the models and results obtained.

The report should also include your **explanations and justifications for your decisions**, as well as the **code and corresponding outputs** obtained.

Beware that **I will not run the code of all groups, but I might run some groups selected at random**, so I need to have explicit outputs in the report to confirm your conclusions.

Deadline

The deadline for this project is **November 26th at 23:59 in Moodle**. The groups without access to Moodle to submit the project should send the project by email.

References

Brown, A. E., Yemini, E. I., Grundy, L. J., Jucikas, T., & Schafer, W. R. (2013). "A dictionary of behavioral motifs reveals clusters of genes affecting *Caenorhabditis elegans* locomotion." Proceedings of the National Academy of Sciences, 110(2), 791-796.