

DATA MINING 2 – Project Guidelines

Francesco Spinnato
francesco.spinnato@sns.it

a.a. 2022/2023



UNIVERSITÀ DI PISA

Project Guidelines

There are 2 datasets for this project:

1. A tabular dataset which is needed for Module 1 and 2
2. A time series dataset which is needed for Module 3

Module 1

Data Understanding and Preparation

Explore and prepare the dataset. You are allowed to take inspiration from existing notebooks you can find online and figure out your personal research perspective (from choosing a subset of variables to the class to predict...). You are welcome in creating new variables and performing all the pre-processing steps the dataset needs.

Module 1

Outliers

- Identify the top 1% outliers: adopt at least three different methods from different families (e.g., density-based, angle-based...) and compare the results.
- Visualize the outliers in a 2 or 3d scatter plot using at least one dimensionality reduction technique.
- Deal with the outliers in a way you see fit, e.g. by removing them from the dataset or by treating the anomalous variables as missing values and employing replacement techniques. In this second case, you should check that the outliers are not outliers anymore. Justify your choices in every step.

Module 1

Imbalanced Learning

- Define one simple unbalanced classification tasks and solve it with Decision Tree or KNN.
- If the dataset is already unbalanced leave it as it is, otherwise turns the dataset into an imbalanced version (e.g., 96% - 4%, for binary classification).
- Then solve the classification task using the Decision Tree or KNN by adopting at least 2 techniques of imbalanced learning (Undersampling, Oversampling).

Module 2

Advanced Classification

- Solve the classification task defined in Module 1 (or define new ones) with the other classification methods analyzed during the course: Logistic Regression, Support Vector Machines, Neural Networks, Ensemble Methods, Gradient Boosting Machines.
- Always perform hyper-parameter tuning phases and justify your choices (which are the best parameters? which parameters did you test and why?).
- Evaluate each classifier with the techniques presented in DM1: accuracy (or precision, recall, F1-score etc), ROC curve (or lift, precision-recall etc).
- Besides the numerical evaluation draw your conclusions about the various classifiers (e.g. for Neural Networks: what are the parameter sets or the convergence criteria which avoid overfitting? For Ensemble classifiers how the number of base models impact the classification performance? What is revealing the feature importance of Random Forests?)

Module 2

Advanced Regression

- Define a multivariate regression task and solve it using 2 advanced regression approaches (not linear).
- Compare and evaluate the approaches using appropriate metrics.

Module 3

Data Understanding and Preparation

- Explore and prepare the time series dataset.
- Preprocess the dataset in order to be able to run time series clustering; motif/anomaly discovery and classification. If the dataset is too big for these tasks, you can use approximations (e.g. SAX, PAA etc)

Module 3

Motifs/Discords

- Analyze the dataset for finding motifs and/or anomalies. Visualize and discuss them and their relationship with shapelets.

Clustering

- Use at least two clustering algorithm on time series using an appropriate distance.
- Analyze the clusters and highlight similarities and differences and visualize the clusters using at least 2 dimensionality reduction techniques.

Module 3

Classification

Define one (or more) classification task and solve it using:

- KNN with at least two distances
 - Euclidean/Manhattan
 - DTW
- Shapelets
 - Analyze the shapelets retrieved
- At least one other method (rocket, muse, cnn, rnn etc)

Module 4

Explainability

- Try to use one or more explanation methods (e.g., TREPAN, LIME, LORE, SHAP, Counterfactual Explainers, etc.) to illustrate the reasons for the classification in one of the steps of the previous tasks.