Data analysis

Marco Romito
marco.romito@unipi.it
http://people.dm.unipi.it/romito

Katerina Papagiannouli
aikaterini.papagiannouli@unipi.it
https://sites.google.com/view/katerinapapagiannouli

General info

Web page with details
`http://people.dm.unipi.it/romito/Teaching/2025/data`

Recommended prerequisites: elementary probability and statistics, linear algebra, elementary optimization, basic experience of `R` or `python`.

Contents:

- The theory of learning.
- Evaluation of prediction and inference models.
- *Supervised learning*: an exploration of methods for classification (svm, random forests, neural networks, etc.).
- if time allows it, some topics in *Unsupervised learning*.

Timetable: The course will be held in the second term.

Examination methods

Standard: during exam sessions. Required:

- solving 100% of exercises assigned during the course, to be completed individually and handed in by the exam date;
- discussion (oral exam) on the course content and on of one or more of the proposed exercises.

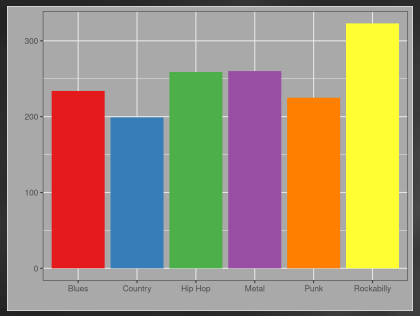
Report: at the end of the course (roughly), recommended to attending students. Required:

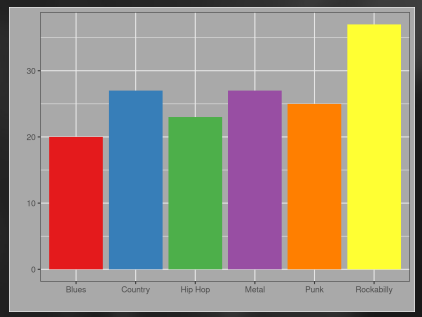
- solving 30% of exercises assigned during the course, possibly completed *in groups*, and handed in by the deadline of the report;
- a report discussing possible solutions, together with their implementation, on a problem that will be assigned by teachers at the end of the course.

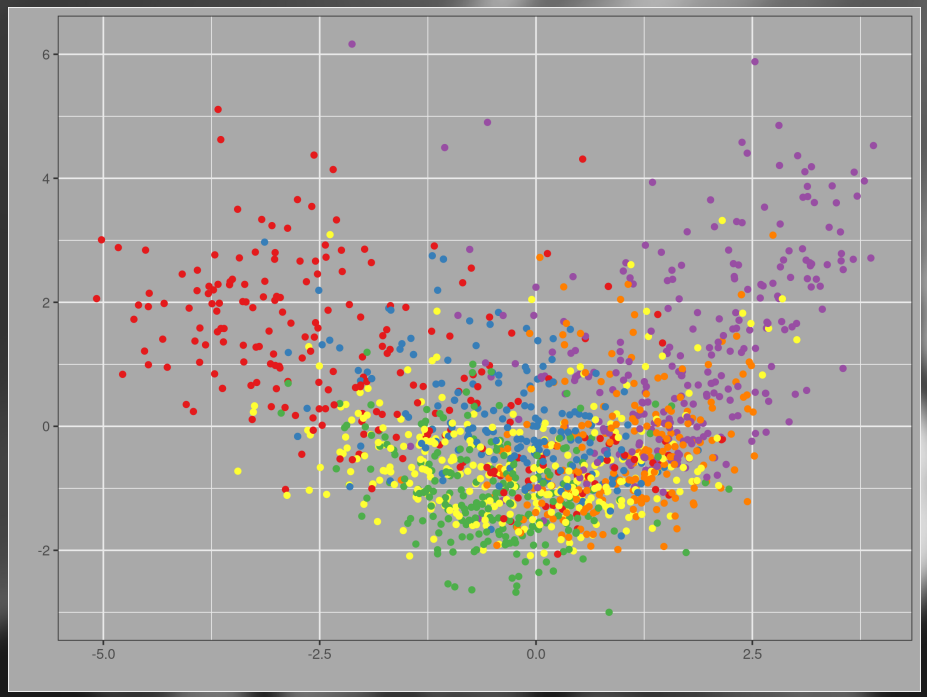
Proposed problem for the report – 2019/20

Development of a model for classification of music tracks genres.
The dataset contains 1500 music tracks of genres:

`Punk` `Rockabilly` `Country` `Metal` `Hip Hop` `Blues`







Spotify API: `duration_ms`, `key`, `mode`, `time_signature`, `acousticness`, `danceability`, `energy`, `instrumentalness`, `liveness`, `loudness`, `speechiness`, `valence`, `tempo`

Proposed problem for the report – 2020/21

Development of a model aimed at finding the most suitable MSC code of a scientific paper, based on title and abstract. Two dataset have been assigned: a *training* dataset containing title and abstract of 2837 papers in probability, and a *validation* dataset of 1419 papers. For instance:

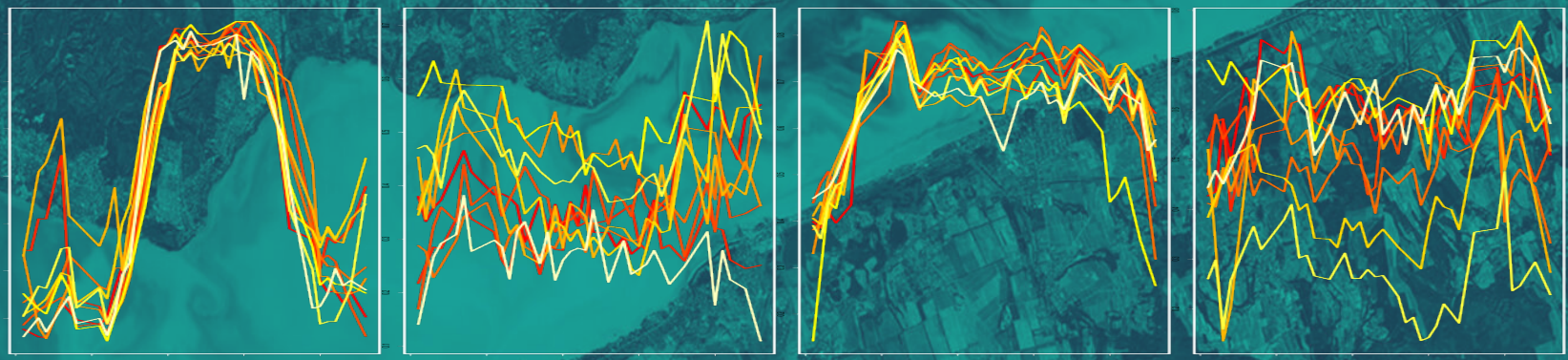
Solving the KPZ equation
Martin Hairer
We introduce a new concept of solution to the KPZ equation which is shown to extend the classical Cole-Hopf solution. This notion provides a factorisation of the Cole-Hopf solution map into a "universal" measurable map from the probability space into an explicitly described auxiliary metric space, composed with a new solution map that has very good continuity properties. The advantage of such a formulation is that it essentially provides a pathwise notion of a solution, together with a very detailed approximation theory. In particular, our construction completely bypasses the Cole-Hopf transform, thus laying the groundwork for proving that the KPZ equation describes the fluctuations of systems in the KPZ universality class.
Comments: 102 pages, many figures
Subjects: Probability (math.PR); Mathematical Physics (math-ph); Analysis of PDEs (math.AP)

MSC classes: 60H15, 35Q82, 60K35

Proposed problem for the report – 2021/22

Development and implementation of a classification method for predicting the land use, based on short time series of geographical areas.

The time series have been acquired by open data provided by the *Sentinel-2* mission of ESA, dell'ESA, designed to monitor green areas and provide support in the event of natural disasters.



https://sentinel.esa.int/web/sentinel/missions/sentinel-2

Proposed problem for the report – 2022/23

The preliminary aim is to analyse and develop extensively classification methods to produce an accurate prediction of the histological stage of biliary cirrhosis of a series of patients, based on of tabular (numerical) data or possibly images concerning their state of health.

Once satisfactory methods of classification have been developed, the main task is to explore possible methods to produce measures or estimates of uncertainty in classification.

Indeed, a major challenge for the implementation of automated systems in critical applications (such as medical diagnostics, self-driving vehicles, etc.) is to what extent the system can fail a prediction. After all, the evaluation of uncertainty is embedded into our behaviour. For example, a human driver slows down, or a doctor requires further investigations in the case of significant uncertainties.

`https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset`
`https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset`

Proposed problem for the report - 2023/24

Development of methods of anomaly detection in classification problems, with emphasis on the fact that results should be as much informative as possible and able to identify not only *which* data points are outliers, but also *why* data points are outliers.

Anomaly detection broadly refers to the process of identifying outliers in data. Essentially outliers are the data points that stand out among all other data points, and whose values are somewhat unexpected and deviate significantly from the standard patterns. Recent applications in anomaly detection are for instance developed in predictive maintenance in manufacturing facilities, to detect any deviations from the equipment's normal performance long before the machine fails.

Three possible datasets: the classical MNIST digits dataset, electrical power demand time series dataset, network intrusions in LAN traffic dataset.