UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Proposed projects for the exam

Groups for the Oral Presentation

A google spreadsheet has been created:

- https://docs.google.com/spreadsheets/d/1hJFsOuAkOhTuhfyRZwLhqi-Puz6_hK_801JkQbU-5Ns/edit?usp=sharing
- Please fill the required information

	A	B	C	D	E	F	G
1	Instructions						
2	Oral Exam Dates: choose one of the following:			26-06-2024	17-07-2024	30-08-2024	20-09-2024
3							
4							
5	Name	First Name	UniPD-ID	Group number	Oral Exam Date	Project Name	
6							
7							
8							
9							

Fake News Classifier

Title: Naïve Bayes classifier for Fake News recognition

Fake news are defined by the New York Times as ”a made-up story with an intention to deceive”, with the intent to confuse or deceive people. They are everywhere in our daily life and they come especially from social media platforms and applications in the online world. Being able to distinguish fake contents form real news is today one of the most serious challenges facing the news industry. Naive Bayes classifiers [1] are powerful algorithms that are used for text data analysis and are connected to classification tasks of text in multiple classes. The goal of the project is to implement a Multinomial Naive Bayes classifier in R and test its performances in the classification of social media posts. The suggested data set is available on Kaggle [2].

[1] C. D. Manning, Chapter 13, Text Classification and Naive Bayes, in Introduction to Information Retrieval, Cambridge University Press, 2008.

[2] Fake News Content Detection, KAGGLE data set: https://www.kaggle.com/datasets/ anmolkumar/fake-news-content-detection?select=train.csv

Fake News: build a system to identify unreliable news articles https://www.kaggle.com/ competitions/fake-news/data?select=train.csv

Bayesian Network

Title: Learning the topology of a Bayesian Network from a database of cases using the K2 algorithm

- Given a database of records, it is interesting to construct a probabilistic network which can provide insights into probabilistic dependencies existing among the variables in the database. Such network can be further used to classify future behaviour of the modelled system [2]. Although researchers have made substantial advances in developing the theory and application of belief networks, the actual construction of these networks often remains a difficult, time-consuming task. An efficient method for determining the relative probabilities of different belief-network structures, given a database of cases and a set of explicit assumptions
- The K2 algorithm [3] can be used to learn the topology of a Bayes network [2], i.e. of finding the most probable belief-network structure, given a database

[...]

[1] M. Scutari and J.B. Denis, Bayesian Networks, CRC Press (2022), Taylor and Francis Group

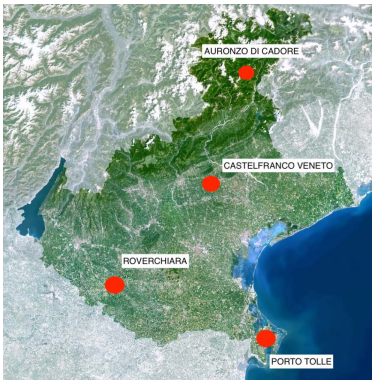
[2] G.F. Cooper and E. Herskovitz, A Bayesian Method for the Induction of Probabilistic Networks from Data, Machine Learning 9 (1992) 309

[3] C. Ruiz, Illustration of the K2 Algorithm for learning Bayes Net Structures, http://web.cs.wpi.edu/~cs539/s11/Projects/k2_algorithm.pdf

ARPAV time series analysis

Title: Bayesian analysis of ARPAV time series on temperature and precipitations

- ARPAV (Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto) is an agency widespread over the territory that collects and analyzes environmental data. Some of the measurement points are quite old and ve a very long time serie (for example in Cavanis, Venice, daily measurements are available since 1900).
- The aim of the project is to analyze the data available in three stations from 1993 to 2021, where the environment is quite different, and study the evolution over time. The stations are located in:
- Auronzo di Cadore (Lat: 46◦33′33′′ N, Long: 12◦25′28′′ E, Alt over sea level: 887 m)
- Castelfranco Veneto (Lat: 45◦40′00′′ N, Long: 11◦55′00′′ E, Alt over sea level: 46 m)
- Porto Tolle (Lat: 44◦56′58′′ N, Long: 12◦19′28′′ E, Alt over sea level: -22 m)
- Roverchiara (Lat: 45◦16′10′′ N, Long: 11◦14′41′′ E, Alt over sea level: 20 m)
- [...]



Gravitational Waves

Title: Learning the distribution of gravitational wave sources

- Consider a population of sources emitting gravitational waves with n, the density of the population. Let's assume that n is a low density so that the number of sources, even in large volumes, remains relatively small. For the exercise, we ignore cosmological effects (let's consider redshift $z \ll 1$), and we assume that the position of the sources are statistically independent. Let's build a statistical model of the population:

1. given a spherical shell with radius R and thickness ΔR centered on the Sun, what is the probability distribution of the number of sources in our shell ?
2. and, according to that probability distribution, what is the average number and variance of the sources in the shell ?
3. with increasing distance from the Sun, the detection efficiency of the number of sources is decreasing; let's suppose to characterize the amplitude of the gravitational radiation through the maximum strain, h, and we also assume we can neglect the difference on polarization and orientation of the sources. With this assumptions, $h \propto 1/r$.

Energy Resolution of Ge detectors

Title: Energy Resolution of Germanium detectors

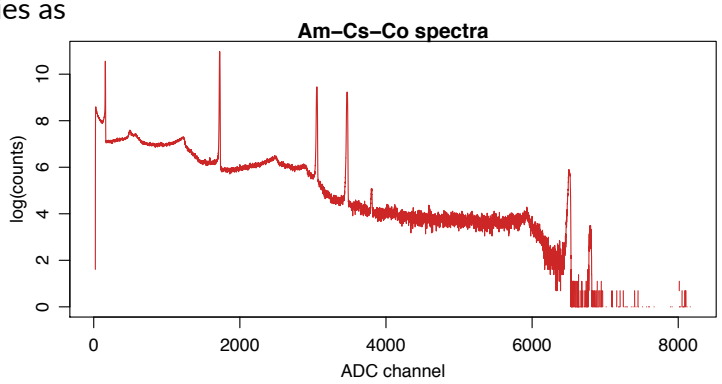
- Germanium detectors have wide fields of application for γ- and X-ray spectrometry thanks to their excellent energy resolution. The energy resolution of these detectors is defined as the width of the detected energy spectra peaks (FWHM); it depends on
  - the statistics of the charge creation process
  - the properties of the detector, and primarily its charge collection efficiency - the electronics noise
- The resolution can be expressed as the squared sum of two terms

$$\mathrm{FWHM} = \sqrt{w_d^2 + w_e^2}$$

- where the first term depends on the detector properties as

$$w_d = 2\sqrt{(2\ln 2)\cdot F\cdot E_\gamma \cdot w}$$

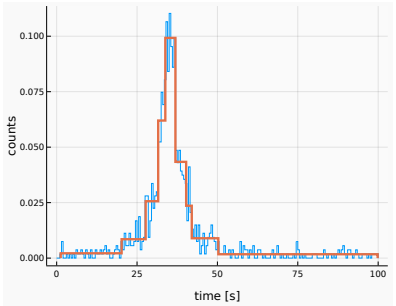
with F the Fano factor, E_γ the energy of the photon deposited energy and w is the electron-hole production energy threshold in germanium (w ~ 3 eV)[1] The other term in eq. 1, w_e is connected with the readout electronics and depends on the detector capacitance, the size of the detector and the bias voltage [...]

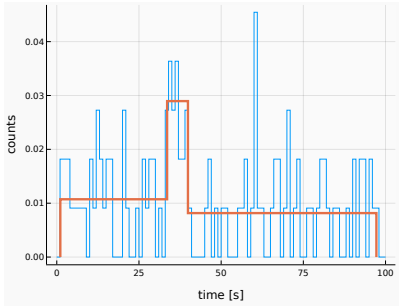


Bayesian Blocks

Title: Bayesian Blocks: a dynamic algorithm for histogram representation

- It is a non-parametric representation of data derived with a bayesian statistical procedure. It has been invented by D. Scargle [1] and applied in the context of astronomical time series analysis. A similar technique available on the market is the kernel density estimation (KDE). As described in [2], it allows to discover local structure in background data, exploiting the full information brought by the data.
- The main idea is based on segmentation of the data interval into variable-sized blocks, each containing consecutive data satisfying some well-defined criteria.
- [1] J. D. Scargle et al., Astrophys. J. 764 (2013) 167
- [2] B. Pollack et al., arXiv:1708.008 10





Earthquakes time series analysis

Title: Temporal and spatial analysis of earthquakes in Italy in the last century

- Italy lies at the boundary of the African and Eurasian tectonic plates, and both plates move and smash into each other releasing a lot of energy and making Italy a seismically active zone, especially central Italy (mountain range). The last main earthquake was the MW 6.3 quake that struck L'Aquila (Abruzzo) in the early morning of April 6, 2009: 297 people were killed, over 1,000 injured, 66,000 made homeless, and many thousands of buildings were destroyed or damaged [1]. In order to better study the relationship between the occurrence of earthquakes and the geological structure, present the temporal and spatial characteristics of earthquakes, explore the temporal and spatial rules of earthquake disasters and determine the seismically active regions in Italy. Investigate and analyze the earthquakes data coming from the USGS Earthquake Hazards Program [2] with the magnitude greater than MW 5.0 that occurred in Italy and surrounding countries during the last hundred years: from 1921 to 2021. Using the R language, study the temporal and spatial characteristics of earthquake data. Apply a time series analysis method (see for instance [3]) to analyze the trend of earthquake magnitude change. At the spatial level, perform a hierarchical cluster analysis to get the range of earthquake active areas.

[1] R. Walters, at al. *The 2009 l'Aquila earthquake (central italy): A source mechanism and implications for seismic hazard*, Geophysical Research Letters, 36 (2009) 17.

[2] USGS Earthquake Hazards Web Site: `https://earthquake.usgs.gov/`

[3] CRAN Task View: Time Series Analysis: `https://cran.r-project.org/web/views/TimeSeries.html`

μ^+ and μ^- lifetime in aluminum

- Title: bayesian analysis of the μ+ and μ− lifetime in aluminum
- The goal of this project is to obtain the lifetimes of positive and negative muons in aluminium. The given dataset contains the time passed between the implantation of the muon and its decay.
- In the first part of the project, a more conventional kind of analysis will be carried out by building an histogram and then fitting it with a curve (in this case two exponentials plus a constant background). The value of the parameters will be then inferred by using bayesian inference on the poissonian likelihood of the fit (binned analysis).
- In the second part the following idea is used: we have three possible kinds of event, decay of either a positive or negative muon or a background event. The first two will follow an exponential distribution, even though having different decay constants, while the background follows a uniform distribution. A weighted average of them will give the distribution of the data. At this point a procedure similar to the one carried out during the lectures will be performed by building the likelihood and posterior. In this way it is possible to avoid the construction of an histogram (unbinned analysis). For both parts JAGS will be used

Vaccine Effectiveness

Title: Inference of Covid-19 vaccines effectiveness and uncertainty using bayesian methods

- Several Covid-19 Vaccines have been authorized by the European Medicines Agency (EMA)
- For the project:

1. collect official data available on the clinical clinical trial performed for each vaccine and compute with JAGS or Stan the efficacy of each Vaccine. Infere the the 95% credibility interval.
2. more recently tests on the efficacy of Vaccine for young people have started. Try to collect available official data from the European medicines Agency (https://www.ema.europa.eu/en) or the U.S. Food and Drug (FDA) (https://www.fda.gov/) and perform a bayesian analysis of the data as a function of the age of the patients

Backup