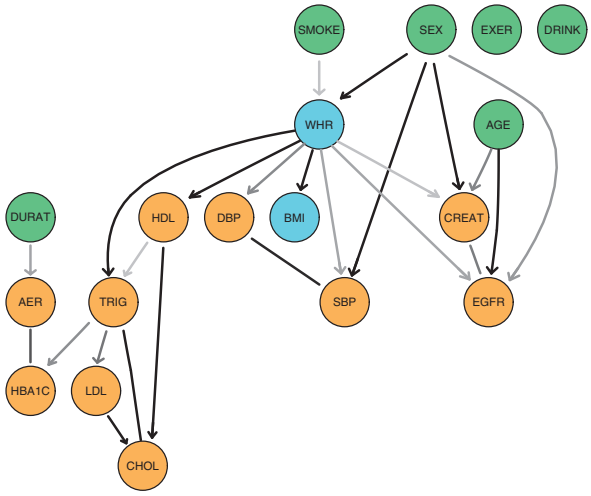
P01 - Learning the topology of a Bayesian Network from a database of cases using the K2 algorithm

A Bayesian belief-network [1] structure is a directed acyclic graph in which nodes represent domain variables and arcs between nodes represent probabilistic dependencies [2]. 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 is described in [2] and [3].



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.

Part 1 After having studied the problem in the suggested literature ([2]-[3]), Implement the algorithm in R and check its performances with the test data set given in [3]: `Ruiz`, `Asia` and `Child` data sets.

Part 2 Implement and test the K2 algorithm with the test data sets ([3]). Compare the results with that obtained with the `bnstruct` R library [4].

Bibliography

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

[2] G. F. Cooper and E. Herskovits, *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`

[4] A. Franzin et al., ***bnstruct**: an R package for Bayesian Network structure learning in the presence of missing data*, Bioinformatics **33(8)** (2017) 1250

[5] F. Sambo and A. Franzin, ***bnstruct**: an R package for Bayesian Network Structure Learning with missing data*, December 12, 2016

P02 - Temporal and spatial analysis of earthquakes in Italy in the last century

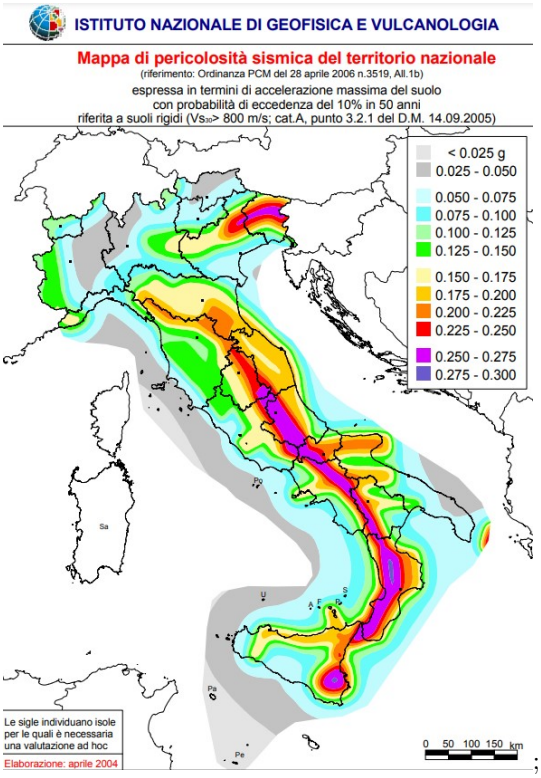
Italy lies at the boundary of the African and Eurasian tectonic plates, and both plates move and smash into each other releasing a lot of energy and making Italy a seismically active zone, especially central Italy (mountain range). The last main earthquake was the MW 6.3 quake that struck L'Aquila (Abruzzo) 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 1923 to 2023.
First of all study the localization of the earthquakes on the italian territory. Study the Gutenberg-Richter [3] law of earthquakes which states that the relationship between the magnitude and total number of earthquakes in any given region and time period of at least that magnitude is given by:

$$N = 10^{a+b\,M}$$

where:

- N is the number of events having a magnitude greater or equal to M
- while a and b are constants.

Estimate the parameter b of the model, which represents the proportion of small and big earthquakes in that region.
Build a sismic map, based in the Gutember-Richter law, similar to that shown in the figure.



Using the R language, study the temporal and spatial characteristics of earthquake data. Perform a time series analysis (as a suggestion, use the TTR package [4]). A time series is a series of data points ordered and indexed in time. Most commonly, a time series is a sequence taken at successive equally spaced points in time. It becomes clear that earthquake events do not fit very well with the definition of a time series, since they are extremely irregular in time. Getting insipiration from the work of [5], define two different time series: AEM (Average Earthquakes Magnitude) and MEM (Maximum Earthquakes Magnitude) using a one year time step. Verify if with the two time series it is possible to predict future trends of earthquakes in the same region. Perform a fit of the two time series using the ARIMA model (AutoRegressive Integrated Moving Average).
Finally, perform a hierarchical cluster analysis[1], at the spatial level, to get the range of earthquake active areas.

Bibliography

[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] B. Gutenberg, C. F. Richter, *Frequency of Earthquakes in California*, Bulletin of the Seismological Society of America, 34 (1944) 185

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

[5] H. O. Cekim, et al., *Prediction of the earthquake magnitude by time series methods along the East Anatolian Fault, Turkey*, Earth Science Informatics 14 (2021) 1339

[1] for instance using the `hclust` function

P03a - Study of the energy resolution and uncertainties of germanium detectors using bayesian methods

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

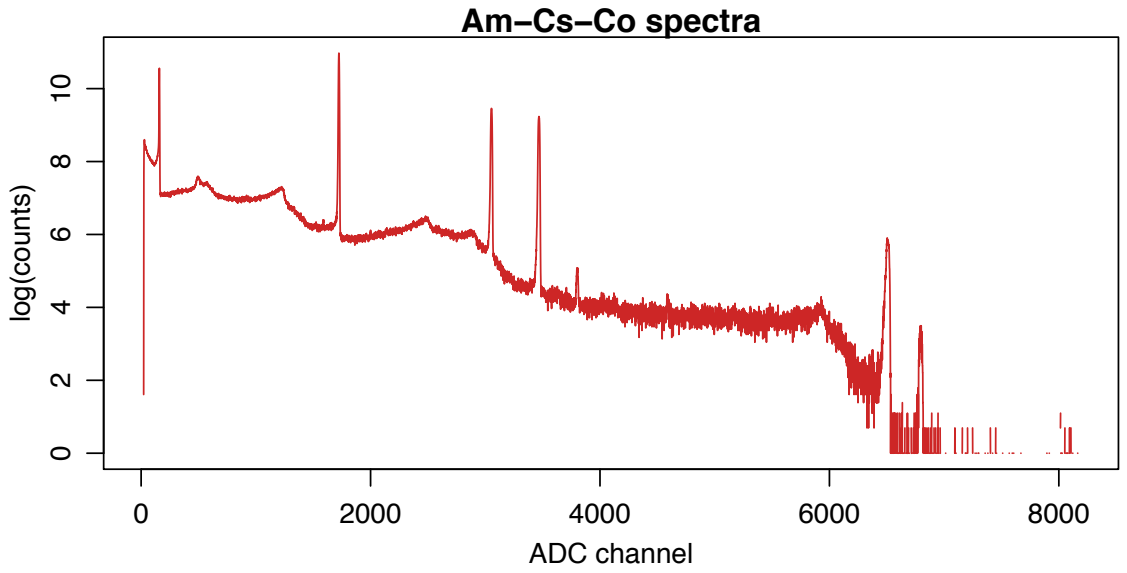
$$\text{FWHM} = \sqrt{w_d^2 + w_e^2} \ , \tag{1}$$

where the first term depends on the detector properties as

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

with F the Fano factor[1], E_γ the energy of the photon deposited energy and w is the electron-hole production energy threshold in germanium (w $\sim$ 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.

The following plot shows an uncalibrated energy spectrum collected with a Germanium detector irradiated by a combination of three sources: ^{241}Am, ^{60}Co and ^{137}Cs[2].



According to [2], the source nuclides emit the following photons:

Nuclide	^{241}Am	^{137}Cs	^{60}Co	
Photon energy (keV)	59.54	661.66	1173.24	1332.51

and these are the first four peaks (starting from the left side) visible in the figure. Similar spectra have been collected with other gamma sources (i.e. Th-228).

1. using statistical methods similar to that presented during the course, infere the FWHM of each γ peak for all available γ sources

[1] The Fano factor is an inherent property of the material.

2. assuming a linear response of the detector, as a function of energy, perform a calibration of the detector, associating the centroid of each peak to the nominal value of the detected γ full energy peak

3. using a MCMC method (with either JAGS or stan), study the behavious of the energy resolution as a function of the photon energy and infere the parameters of eq.1 and 2.

Bibliography

[1] K. Debertin and R. G. Helmer, *Gamma- and X-ray spectrometry with semiconductor detectors*, North-Holland, 1988

[2] Laboratoire national Henri Becquerel, tables of eveluated data on radioactive nuclides, `http://www.nucleide.org/DDEP_WG/DDEPdata.htm`

P03b - Study of the energy resolution and uncertainties of Germanium and NaI detectors using bayesian methods

NaI and Germanium detectors have wide fields of application for γ- and X-ray spectrometry thanks to their good 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

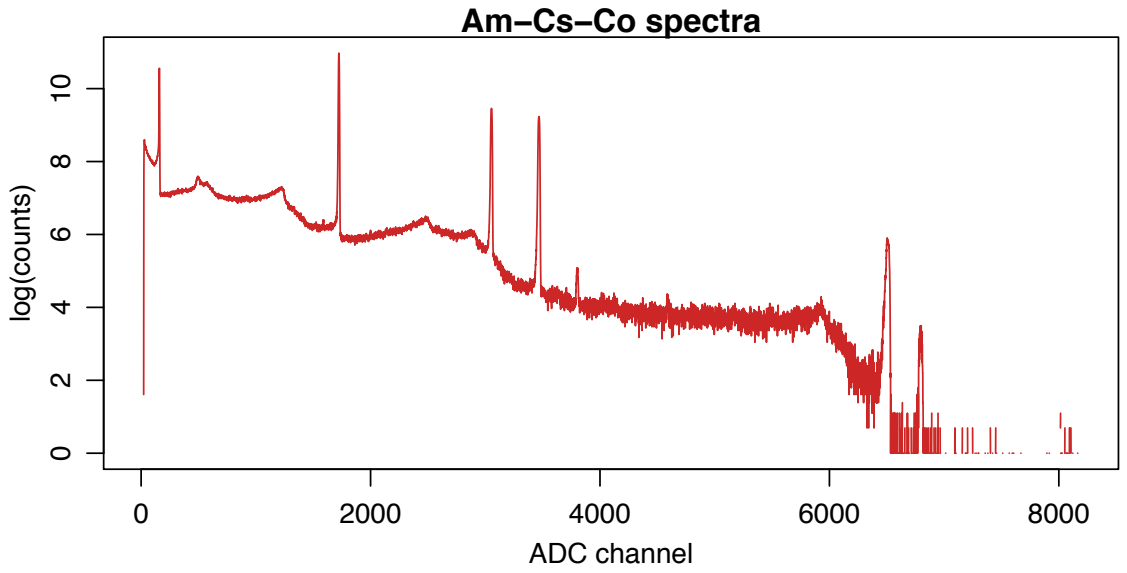
$$\text{FWHM} = \sqrt{w_d^2 + w_e^2} \ , \tag{1}$$

where the first term depends on the detector properties as

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

with F the Fano factor[1], E_γ the energy of the photon deposited energy and w is the electron-hole production energy threshold in germanium (w $\sim$ 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.

The following plot shows an uncalibrated energy spectrum collected with a Germanium detector irradiated by a combination of three sources: ^{241}Am, ^{60}Co and ^{137}Cs[2].



According to [2], the source nuclides emit the following photons:

Nuclide	^{241}Am	^{137}Cs	^{60}Co	
Photon energy (keV)	59.54	661.66	1173.24	1332.51

and these are the first four peaks (starting from the left side) visible in the figure. Similar spectra have been collected with other gamma sources (i.e. Th-228).

1. using statistical methods similar to that presented during the course, infere the FWHM of each γ peak for all available γ sources

[1] The Fano factor is an inherent property of the material.

2. assuming a linear response of the detector, as a function of energy, perform a calibration of the detector, associating the centroid of each peak to the nominal value of the detected γ full energy peak
3. using a MCMC method (with either JAGS or stan), study the behavious of the energy resolution as a function of the photon energy and infere the parameters of eq.1 and 2.

Perform a study on the energy resolution for a NaI detector. Data are given for the following gamma sources:

- ^{60}Co
- ^{22}Na
- ^{137}Cs

Verify in Ref. [2] the energy of the calibration photons for the given sources and produce the calibration curves as those obtained for the Germanium detector.

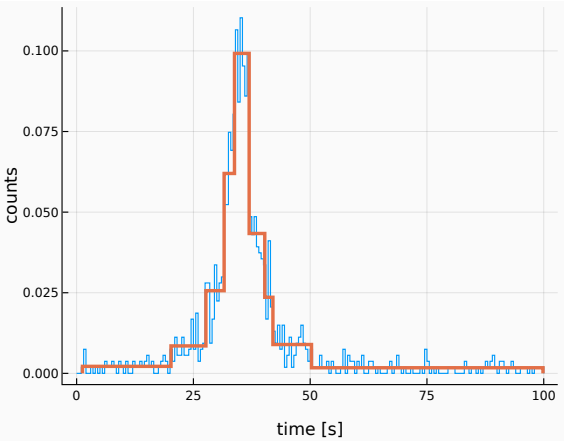
Bibliography

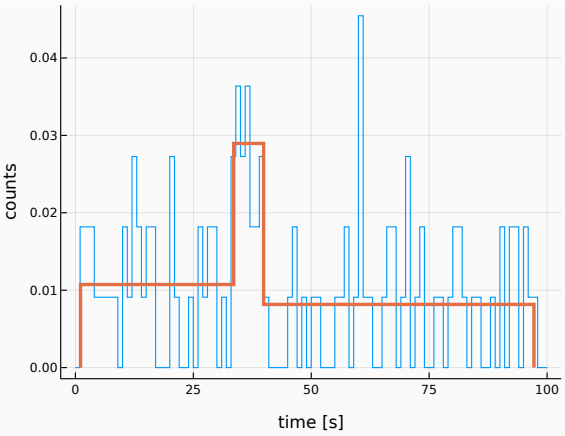
[1] K. Debertin and R. G. Helmer, *Gamma- and X-ray spectrometry with semiconductor detectors*, North-Holland, 1988

[2] Laboratoire national Henri Becquerel, tables of eveluated data on radioactive nuclides, `http://www.lnhb.fr/home/nuclear-data/`

P04 - 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 struture 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.





Write an algorithm in R and then test its performances with different sets of data.

Try to think other possible application of the method.

Bibliography

[1] J. D. Scargle *et al.*, Astrophys. J. **764** (2013) 167

[2] B. Pollack *et al.*, arXiv:1708.008 10

[3] J. D. Scargle *et al.*, Astrophys. J. **504** (1998) 405

P05 - Inference of Covid-19 vaccines effectiveness and uncertainty using bayesian methods

The European Medicines Agency (EMA) has authorized several different brand of Covid-19 Vaccines:

- Comirnaty [1];
- Nuvaxovid [2];
- Bimervax [3];
- Ronapreve [4];
- Xevudy [5];
- Spikevax [6].

For the project:

1. collect official data available in [1]-[6], on the 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.

Bibliography

[1] `https://www.ema.europa.eu/en/medicines/human/EPAR/comirnaty`

[2] `https://www.ema.europa.eu/en/medicines/human/EPAR/nuvaxovid`

[3] `https://www.ema.europa.eu/en/medicines/human/EPAR/bimervax`

[4] `https://www.ema.europa.eu/en/medicines/human/EPAR/ronapreve`

[5] `https://www.ema.europa.eu/en/medicines/human/EPAR/xevudy`

[6] `https://www.ema.europa.eu/en/medicines/human/EPAR/spikevax-previously-covid-19-vaccine-moderna`

P06 - Determination of the muon magnetic moment

Muons are long-lived particle, produced in the decays of pions and kaons originating from the interactions of primary cosmic rays with the Earth's atmosphere. Muons decay via weak interactions and, according to [1], their lifetime is

$$2.1969811 \pm 0.0000022 \ \mu\text{s}$$

Moreover, parity violation is also present in the decay which proceeds as follow:

$$\mu^+ \to e^+ + \nu_e + \overline{\nu}_\mu \quad \text{and} \quad \mu^- \to e^- + \overline{\nu}_e + \nu_\mu$$

According to [2] and [3], simple experiments can be performed measuring muons that decay in a thick absorber. If the absorber is immersed in a constant magnetic field, the muon spin, before the decay, proceeds with a frequency

$$\omega = g_\mu \frac{eB}{2m_\mu c}$$

The decay proceeds mainly along the direction of the spin of the muon and therefore, if the muon is (partly) polarized, the detected signal varies with time with ω, spin precession angular frequency.

Analyzing the data collected without and with the magnetic field, setup a Markov Chain Monte Carlo that allows to extract the muon lifetime τ_μ (B off) and the muon precession frequency ω (B on).
The magnetic field is realized with a solenoid and its intensity, at the center, is approximately 5.6 mT.

Further details on the apparatus and measurement principle can be found in [2] and [3].

Bibliography

[1] S. Navas *et al.*, (Particle Data Group), Phys. Rev. D 110 (2024) 03001
`https://pdglive.lbl.gov/Particle.action?node=S004&init=0`

[2] C. Amsler, *The determination of the muon magnetic moment from cosmic rays*, American Journal of Physics, 42 (1974) 12.

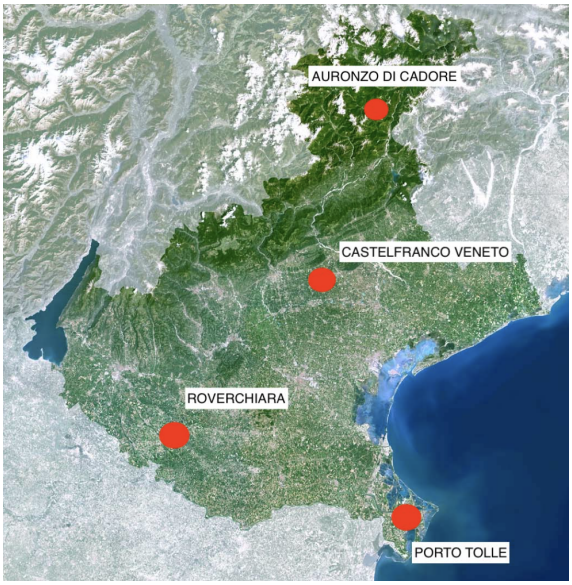
[3] D. Bosnar emphet al., *A simple setup for the determination of the cosmic muon magnetic moment*, American Journal of Physics, 90 (2022) 8

P07 - Bayesian Analysis of ARPAV time series on temperatures and precipitations

ARPAV (Agenzia Regionale per la iPrevenzione 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).



Part 0

Study the evolution over time of the min, max and average temperatures computed over one month.

Part 1

Analysis of the evolution of the annual average of the min, max and daily averate temperature over time (1993 - 2021):

- study the trend of the annual averages and compare them with a constant value or a rising trend (for instance linear or quadratic)
- perform an analysis using a Bayesian linear regression with JAGS or STAN
- perform the same analysis uwing a constant regression
- perform a Bayesian hypothesis test comparing the two results
- do you see correlations between data measured at different stations ?

Part 2
Analysis of the evolution of the annual difference of the min, max and daily averate temperature over time (1993 - 2021):

- assuming you found an increasing trend in the temperature, compute it considering 4-years intervals and compare the results with those shown in SNPA (Sistema Nazionale per la Protezione Ambiente) in [1].
- perform an analysis using a Bayesian linear regression with JAGS or STAN

Part 3
Using the `forecast` R package [2], which provides methods and tools for displaying and analysing univariate time series, analyze the data. The library contains also the ARIMA (Autoregressive Integrated Moving Average), which allows to perform the equivalent of a linear regression in time series, where data is not stationary. Analyze your data and try to predict the evolution in the next years (average the data over multiple years, if needed).

Bibliography

[1] `https://www.snpambiente.it/wp-content/uploads/2021/06/Rapporto-SNPA-21_2021.pdf`

[2] `https://cran.r-project.org/web/packages/forecast/index.html`

P08 - Naive 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 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]. Possible suggested lables for classifying the text are the following:

- True - 5
- Not-Known - 4
- Mostly-True - 3
- Half-True - 2
- False - 1
- Barely-True - 0

The Kaggle dataset [2] consists of a training set wth 10,240 instances and a test set wth 1,267 instances.

- divide the dataset into a training, validation and testing set;
- tokenize each word in the data set (convert uppercase to lowercase) and split into tokens;
- clean the collection of words from stop words;
- perform token normalization: create equivalence classes so that similar tokens are mapped in the same class
- build the vocabulary and perform feature selection
- show the results

Apply the developed methods and technique to a new dataset [3] which is characterized by only two labels: `1` $\rightarrow$ `unreliable` and `0` $\rightarrow$ `reliable`.

Draw your conclusions on the results obtained on the two data sets.

Bibliography

[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`

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

P09 - Learning the distribution of gravitational waves 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 pupulation:

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 neglet the difference on polarization and orientation of the sources. With this assumptions, $h \propto 1/r$.

4. let's assume that the detection efficiency is a sigmoid function that can be approximated with the following Gaussian integral

$$\epsilon(h) = \int_{-\infty}^{h} \frac{\exp\left[-(h' - h_\circ)^2/(2w^2)\right]}{\sqrt{2\pi w^2}} dh'$$

   with $h_\circ$ the strain produced by a source located at distance $r_\circ$ corresponding to a detection efficiency $\epsilon(h_\circ) = 0.5$.

Perform an analysis of the detection efficiency and on its uncertainty as a function of the source density n.