MSc Projects

University of Glasgow

---

(44) Cluster Analysis of Acute Respiratory Distress Syndrome

Project Description

Acute respiratory distress syndrome (ARDS) is defined as acute hypoxic respiratory failure (PaO2/FiO2<300 mmHg), bilateral chest infiltrates, and the absence of cardiac failure asthe primary diagnosis. Treatment aimed at improving survival of this disease is complicatedby its extreme heterogeneity. A new treatment thought to improve the disease outcome for patients is Extracorporeal membrane oxygenation (ECMO). Of interest is discovering whether groups exist in the biomedical markers data both before and after treatment and whether these clusters connect to the patient's outcome and whether ECMO changes these.

Data

Data are available for 450 patients on biomarkers both before ECMO treatment (markedwith a pretext PreECMO, e.g. PreECMO_RR) and for the first day after ECMO treatment (marked with a pretext Day1ECMO, e.g. Day1ECMO_RR).

(1) PreECMO Data

- Can we find clusters in the PreECMO biomedical markers data?
- Do the clusters found correspond at all to the outcome variables for survival (Hospital_Survival and ECMO_Survival)

Relevant Courses

- Multivariate Methods (main dissertation)

(45) Classification Analysis of Acute Respiratory Distress Syndrome

Project Description

Acute respiratory distress syndrome (ARDS) is defined as acute hypoxic respiratory failure(PaO2/FiO2<300 mmHg), bilateral chest infiltrates, and the absence of cardiac failure asthe primary diagnosis. Treatment aimed at improving survival of this disease is complicatedby its extreme heterogeneity. A new treatment thought to improve the disease outcome for patients is Extracorporeal membrane oxygenation (ECMO). Of interest is discovering whether groups exist in the biomedical markers data both before and after treatment and whether these clusters connect to the patient's outcome and whether ECMO changes these.

Data

Data are available for 450 patients on biomarkers both before ECMO treatment (markedwith a pretext PreECMO, e.g. PreECMO_RR) and for the first day after ECMO treatment (marked with a pretext Day1ECMO, e.g. Day1ECMO_RR).

(1) Day1ECMO Data

- Can we use the PreECMO biomedical markers to accurately predict ECMO survival?
- Do we need all PreECMO variables or just a subset to make accurate predictions?
- What is our expected future performance for these predictions?

Relevant Courses

- Multivariate Methods (main dissertation)

(49) Bayesian Linear Models and Bayesian Lasso

Project Description

Linear models are the most ubiquitous class of statistical models used in practice. In your courses these models were covered extensively using the classical approach. In this project, you will consider the Bayesian approach to inference using linear models, and will consider the problem of variable selection using Lasso regularisation in Bayesian framework.

Data - (1) Communities and Crime Data Set

The data set available from the following address: http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime.The data set contains 1994 records of 128 variables. The response variable isViolentCrimesPerPop, total number of violent crimes per 100K population. Theother variables describe different demographical characteristics of US neighbourhoods. Yourgoal is to build a linear model for predicting the rate of violent crimes from neighbourhood characteristics.First, use a Bayesian formulation of a linear model to infer regression rates. Next, perform Bayesian Lasso for variable selection to decide what are the most informative factors toexplain crime rate

Questions of Interest

- How to formulate a linear model in the Bayesian framework?
- How to perform conjugate and non-conjugate inference for Bayesian linear model?
- How to perform variable selection using Bayesian Lasso?
- What are the most important factors to explain variation in crime rates among neighbourhoods?

Relevant Courses

- Bayesian Statistics
- Advanced Bayesian Methods

(57) Optical Character Recognition

Project Description

Image processing is a difficult task for machines. The relationships linking patterns of pixels to higher concepts are complex and hard to define. For instance, it is easy for a human being to recognize a face or a letter, but defining these patterns in strict rules is difficult.Furthermore, image data are often noisy. There can be many slight variations in how the image was captured depending on the lighting, orientation and positioning of the subject. This project in particular is about optical character recognition (OCR), where the objective is to differentiate among the 26 letters of the English alphabet based on handwritten letters, like shown in the following image:

For the following project, 20,000 handwritten characters were scanned into a computer, converted into pixels and 16 statistical attributes were recorded, following a procedure proposed by Frey and Slate. These attributes measure such characteristics as the horizontal and vertical dimensions of the letter, the proportion of black versus white pixels, and the average horizontal and vertical position of the pixel. The task of this project is to develop and assess a classifier that reads in these attributes and predicts the letter.

Data

The data are available in `fileletterdata.txt`. The first line is a standard line of headings, where the first column (y) indicates the letter, and the following 16 columns (x01 to x16) are 16 integer numbers with the attributes mentioned above.

Questions of Interest

- What classification performance can be obtained with a statistical method, i.e. how close can a machine using a statistical pattern recognition algorithm get to human performance?
- How do linear classification methods compare with non-linear methods, in particular with support vector machines?

Relevant Courses

- Inference
- Flexible regression
- Generalized linear models
- Mutivariate methods
- Big data analytics
- Introduction to R programming

1 (58) Classifying Baterial Metabolic States with Raman Spectroscopy

Project Description

Raman spectroscopy is a spectroscopic technique used to observe low-frequency modes in a molecular system and is commonly used in chemistry to provide a structural fingerprint by which molecules can be identified. It relies on inelastic scattering of monochromatic light, usually from a laser in the visible, near infrared, or near ultraviolet range. The laser lightinteracts with molecular vibrations, phonons or other excitations in the system, resulting inthe energy of the laser photons being shifted up or down. The shift in energy gives informationabout the vibrational modes in the system. A set of typical Raman spectra is shown in thefigure below.

The objective of the present project is to distinguish between different metabolic states in two unicellular organisms: Chlorella (a single-celled green algae), and Rhodobacter (aproteobacterium). The Raman spectra were obtained in Professor Huabing Yin's group in the School of Engineering, and include 171 strains of Chlorella, and 139 strains of Rhodobacterium.The spectra are discretized, and show the normalized scatter intensities at 498 discrete laser wavelengths. For both unicellular organisms, there are 5 different metabolic states. The objective is to build a statistical classifier to correctly predict the metabolic state from the Raman spectra. To this end, you want to develop and assess a range of classifiers that read in the Raman spectra and predict the metabolic state of the unicellular organism.

Data

The data are available in the `filesdata_chlorella.txt` and `data_Rhodo.txt`. The first line is a standard line of headings, where the first column (y) indicates the metabolic state (aninteger number between a and 5), and the following 498 columns (x001 to x498) show the standardized scatter intensities at 498 laser wavelengths.

Questions of Interest

- How accurately can we predict bacterial metabolic states from Raman spectra?
- How do linear classification methods compare with non-linear methods, in particular with support vector machines?

Relevant Courses

- Inference
- Flexible regression

- Generalized linear models
- Mutivariate methods
- Big data analytics
- Introduction to R programming

(60/61) Finding epigenetic signatures for human ageing (2)

Project Description

Biological ageing of human cells is one of the primary risk factors for the development ofcancer or other lethal diseases. The biology of ageing is a complex process, involving manylayers of interactions among the components of the human cell. Actual chronological age maynot be a good measure for biological age as people may age at different rates, due to genetic, environmental or even lifestyle factors. However, recently developed laboratory experimentsallow for the measurement of various biological factors, from clinical-level measurements toepigenetic ones, such as alterations in the chromosome (histone modifications) and methylationof the DNA, that affect gene activity and function and impact ageing of cells. In this project, you will analyse epigenetic data, on histone modifications and methylation at sites inhuman DNA, in proliferating ("young") and senescent ("old") human cells, to determine a characterization (signature) for biological ageing and estimate the effects of these factors onthe human ageing process

(60) Stratification of ageing-associated modifications in human DNA

Data on several histone modifications and methylation, measured on about 2100 ageing-associated CpG sites in human DNA, from proliferating ("young") and senescent ("old") human cells is available. These sites have been determined, through other biological studies,to be ageing-associated differentially methylated positions (aDMPs) in the DNA. The dataare stored in `aDMPs_Proj1.csv` and contain the following columns.

Questions of Interest

- Can the aDMPs be stratified based on the measured abundances of histone modificationsand methylation observations?
- Does the stratification vary across proliferating cells,senescent cells, or both types of cells taken together?
- What is the effect of each histone modification on the propensity for cell ageing?

Relevant Courses

- Regression modelling, Multivariate methods or Machine Learning (main dissertation).
- Multivariate methods or Machine Learning (advanced chapter)

(61) Determining a epigenetic signature for biological ageing

Data on several histone modifications and methylation, measured on about 285,000 CpG sitesin human DNA, from proliferating ("young") and senescent ("old") human cells is available,measured from an Illumina 450k methylation array. The data are stored in-aDMPs_Proj2.csvand contain the following columns.

Questions of Interest

- Can aDMPs be distinguished from the non-aDMPs based on the histone abundanceand methylation observations (i.e. is there an epigenetic signature for aDMPs)?
- Does the epigenetic signature exist, or vary, across proliferating cells, senescent cells, orboth types of cells taken together?

Relevant Courses

- Regression modelling, Multivariate methods or Machine Learning (main dissertation).
- Multivariate methods or Machine Learning (advanced chapter)

(62/63) Determining Genetic Variation Associated with Heart Disease (2)

Project Description

It has long been known to scientists and clinicians that heart disease is a complex set of conditions that are caused in part by environmental or lifestyle factors, but also has a significant connection to the underlying genetics of an individual. The genetic signature of every human being is unique, encoded in their DNA, which can be represented as a long (of length about 3 billion) string of nucleotides, A, C, G and T, in some specific order. Many common health conditions are caused due to variations from the "normal" DNA ata few specific positions on the genome. Genetic variation in individuals often occurs assingle alterations (mutations) in different positions of the genome, termed "single nucleotidepolymorphisms" or SNPs. Genome-wide association studies (GWAS) are a popular method for studying and determining the locations of these SNPs. Using experimental plates that contain millions of SNPs from hundreds or thousands of people, the goal of GWAS is to detect which SNPs are associated with a particular disease outcome. Much recent evidence indicates that two or more SNPs often work in combination to produce a genetic effect, which suggests that multiple regression methods with variable selection may be a potential way to determine causal SNPs. In this project, you will study genetic and clinical data collected at a Glasgow medical centreand try to determine which factors play a part in the development of heart disease. The genotype, or genetic composition at each location of the genome is typically given by one of 3 possibilities, aa, ab, or bb, where a and b take values from the set {A,C,G,T}. These three possibilities are usually encoded numerically by 0, 1, and 2 for purposes of statistical analysis.In a typical genetic experiment (called a genome-wide association study) to study the effectof genetic variation on some characteristic (phenotype) or disease, data is collected from thousands of individuals with varying levels of the phenotype (or disease), and their DNA sequenced for about 500,000-1,000,000 locations on their genomes. It is still an extremely challenging problem to detect which SNPs are associated with the phenotype of interest, compounded by high volumes of data, high levels of missingness, and high correlations amongSNPs that are located in certain neighborhoods in the genome. In this project, you will study a simplified version of this problem in which a small set ofcandidate SNPs (that have been selected by other means, and may have an impact on the phenotype of interest) are given to you, along with a number of measurements on certain clinical covariates, and measurements on some features representing aspects of heart disease.

(62) Determining genetic factors associated with high blood pressure

Data is provided in two files. The first `filebloodpressure.csv`, contains information onsystolic and diastolic blood pressure for patients at the clinic, along with a number of clinical measurements. The file contains the following columns.

Questions of Interest

- Are any of the measured clinical covariates associated with high blood pressure?
- Do one or more of the candidate SNPs appear to be associated with high blood pressure?
- How much of blood pressure variation can be explained by clinical/lifestyle factors,genetic factors, or both?

Relevant Courses

- Regression modelling, Data Analysis, Multivariate methods or Machine Learning (main dissertation).
- Big Data Analytics, or Bayesian Statistics/Advanced Bayesian methods (advanced chapter).

(63) Determining genetic and lifestyle factors underlying blood sugarand cholesterol levels

Data on several clinical measurements and candidate SNPs are available for this project, stored in the `filegwasHDLglu.csv`, containing the following columns:

Questions of Interest

- How do levels of fasting glucose and HDL vary among different segments of the clinicalpopulation? • How well can the variation in fasting glucose be explained by lifestyle factors?
- Can fasting glucose level prediction be improved by accounting for genetic variation inspecific SNPs?
- Are HDL levels associated with lifestyle factors, genetic factors, or both?

Relevant Courses

- Regression modelling, Data Analysis, Multivariate methods or Machine Learning (maindissertation).
- Big Data Analytics, or Bayesian Statistics/Advanced Bayesian Methods (advancedchapter).

(79) How well can you establish the geographical origin of a DNA sequence?

Project Description

One morning, at a large international statistics conference, a body is found slumped over the lectern. From the murder scene, a sample of blood, which does not match the victim andhence is presumed to be from the perpetrator, is recovered. Mitochondrial DNA (mtDNA) is successfully extracted from the blood sample. The question to be investigated is: can any inference be made about where in the world the perpetrator came from? DNA sequences differ between individuals and the different sequences occur at different frequencies in different populations. Databases of samples of sequences from around the world are available. If the perpetrator's sequence is common only in a restricted part of the world, the legal system could be on to a winner. For example, it could potentially be useful to the police in refining their pool of suspects. DNA sequences can be thought of as a sort of high-dimensional multivariate data. In this project, you will investigate how well short mitochondrial DNA sequences allow the assignment of sequences to their population of origin. In principle, you can use any classification approach that you think appropriate. The datahas many variables so dimension-reduction techniques (such as principal components analysis, PCA) might be applied at the outset. The first task will be to investigate whether a broad continental assignment is possible

Data

The data consist of short mitochondrial DNA sequences from 1394 subjects from different human populations. You are not given the raw sequences (strings of the letters A, G, C, T representing the chemical constituents, called nucleotides, of DNA: adenine, guanine, cytosineand thymine). Rather, for every position in the sequence in the sequence where there isvariability between individuals in the sample, you are given information on which individualsshare the same letter, as described below. The data table has 1394 rows (subjects) and 206 columns (variables) as follows.

Questions of Interest

- Do there appear to be systematic genetic difference between the three continental groups (after reducing the number of variables)?
- How well in general can individual sequences be assigned to continents?
- Is there an optimal degree of dimension reduction that makes classification as good aspossible?
- Do there appear to be systematic genetic difference between the 11 population groups (after reducing the number of variables)
- How well in general can individual sequences be assigned to populations?

Relevant Courses

- Multivariate Methods (main dissertation)