Bayesian Statistics Course 2021/2022 Outline

Jure Demšar and Erik Štrumbelj

Grading

The final grade is composed of two parts, 50% from the homework and the other 50% either from an exam or from a practical project. The homework part is mandatory for all students, if all goes to plan there will be 11 exercises in total. Unless otherwise noted, all homework is due in 1 week and should be submitted as a 1 page report in the pdf format. General feedback to your homework will be provided on the group level via the course's forum. If some of you want more detailed, individual feedback about your work, we will happily provide it on request. Each homework will be graded on a scale from 1 to 5:

- 5 – the best submission (or multiple best submissions within the margin of error),
- 4 – submissions of an above average quality,
- 3 – average submissions,
- 2 – below average, but acceptable,
- 0 – unacceptable.

To get a positive grade, you need to gather at least 25 points from homework. Only 10 exercises out of 11 will count towards your grade, meaning that your worst submission will be discarded when calculating the final homework grade.

The traditional way to to finish the course will be via a written exam. Only students that gathered at least 25 points from the homework are allowed to take the exam. The exam lasts for 120 minutes, during the exam you are allowed to use all literature (books, computers, the internet ...). Note though that you are strictly prohibited to use tools for communication (e.g. messaging, Discord, Facebook ...). If we see you using such a tool, you will receive a negative grade (no questions asked!) and will be prohibited to take then next scheduled exam. In the exam you will be asked to provide a constructive discussion or a critique of actual real-world data analyses. For example, you will be provided an excerpt from a paper that used an inappropriate statistical model, your task will be to identify and expose this weakness in a constructive fashion and also provide a solution that fixes the problem.

The alternative to the exam is the project. In the last month of the course, we will invite a couple of best students (based on homework) to form groups and collaborate on real-world data analysis problems. We plan to submit the results of theses projects to scientific journals, so your goal here will be to execute a data analysis of the highest standard. Obviously, you will be listed as co-authors on the submitted paper. Note here that by taking this way you will not be able to complete the course in January or February 2022. Projects will span at least over a couple of months, do not worry though we will assure that you will be able to finish the course by the end of the school year (June 2022 or September 2022 at the latest).

Lectures

Lectures will span over 14 weeks. Below is a more detailed outline of this year's edition of the course. Note here that what is listed below is a only a plan of execution and might change slightly during the course of the year.

Introduction to Statistical Enquiry [5. 10. 2021]

The main goal of this lecture is to provide an introduction to the process of statistical enquiry. What is statistical enquiry, why is it important, and how do we approach it in a systematic way? We base our introduction on Wild and Pfannkuchs 4 dimensions of statistical enquiry, with special focus on the 1st dimension, the PPDAC investigative cycle (Problem, Plan, Data, Analysis, Conclusions).

Probabilistic Thinking [12. 10. 2021]

The main goal of this lecture is to illustrate how uncertainty is part of our everyday lives but in order to deal with it in a systematic way we require the rigor of probability theory. The language of probability theory allows us not only to express uncertainty but also to provide probabilistic interpretations of processes of interest, which then allows us to infer their properties from the data they generate. That is, probability theory is the very foundation of applied statistics.

Probabilistic Programming [19. 10. 2021]

Statistical (probabilistic) models are used for describing how certain data we are interested in was generated. Probabilistic programming languages offer both a methodological way for

specifying statistical models and tools for performing automated inferences on these models.

Stan is a widespread probabilistic programming language, it offers an intuitive framework for specifying statistical models along with algorithms for full Bayesian inference for continuous variable models through Markov chain Monte Carlo methods such as the No-U-Turn sampler, an adaptive form of Hamiltonian Monte Carlo sampling.

The Generalized Linear Model [26. 10. 2021]

The intercept and slope parameters in simple linear regression are easy to interpret and can as such provide key insight into our data generating process. When modeling the relationship between independent and dependent variables, simple linear regression often gives suboptimal results. We need a more powerful tool, especially so in cases where the dependent variable is not metric. Generalized linear model (GLM) is a powerful tool capable of working with dependent variables of various scale types (metric, ordinal, nominal and count) and error distributions other than normal. The simple linear model is actually the simplest of all GLMs. In the GLM, beta coefficients are usually not as easy to interpret as in the case of simple linear regression. Fortunately, we only need some relatively easy math to make them understandable!

Describe the Process [2. 11. 2021]

When developing statistical models we often fall into the habit of merely fitting some distributions onto our data. In this lecture we will show why this is bad an suboptimal. We should do our best to try and understand the actual data generating process and model it!

Questionnaires [9. 11. 2021]

Measuring things accurately and precisely is difficult to start with, but even more so when we try to measure peoples' psychological characteristics, opinions, preferences, etc. We will discuss the questionnaire as a measuring device. How do we design, test, and validate one. How do we select the type of question and scale. And what are the most common mistakes.

Predictive Checking [16. 11. 2021]

Bayesian modeling is an iterative process, we start by picking an initial model and settings its priors. To investigate whether our priors have undesirable effects that contradict domain knowledge we start by performing prior predictive checks. Once we are happy with our priors, we fit the model and diagnose the fitting process (traceplot, convergence, diagnostics ...). If all is good, we then execute posterior predictive checks to evaluate whether our model is suitable for answering the questions we are asking.

Model Selection with Cross-Validation [23. 11. 2021]

Cross-validation is a widely spread technique for estimating how well our models generalize (how they fare when encountering unknown data). Since in Bayesian statistics we are always working with probability distributions we can use log-score as our go-to model evaluation measure. This allows us to resort to information theory for calculating good cross-validation approximations without actually performing the actual (often very time consuming) cross-validation. An important point to emphasize here is that cross-validation and regularization are not mutually exclusive, they go hand in hand and traditionally we use both at the same time.

Hierarchical Models [30. 11. 2021]

In Bayesian modeling hierarchical models (also called multilevel models) are an extremely powerful tool. As already emphasized a couple of times now, when modeling we should try to describe the data generating process (and not fit some distributions to some data). Since data generating processes often have a hierarchical structure (e.g. groups of students, multiple repetitions of an experiment ...), hierarchical models enable us to efficiently describe such data generating processes.

Sampling [7. 12. 2021]

In this lecture we focus on the basics of sampling in the context of survey sampling. We cover three of the most common probability sampling approaches: simple random sampling, stratified sampling, and cluster sampling. We also briefly discuss non-probability sampling methods: convenience sampling, judgment sampling, quota sampling, and snowball sampling.

Priors [14. 12. 2021]

Defining priors is an integral part in Bayesian statistics. The main purpose of priors is to introduce prior expert knowledge about the domain into our models. As we know by now, this is not the only usage of priors, they are also useful for regularization and can be of help during the sampling process which can lead to stabilization of inferences in certain models. Based on the amount of information priors provide they are commonly categorized into three groups: non-informative priors, weakly informative priors and informative priors. This document briefly explains the differences between these groups and provides some guidance about when to use certain priors. Since one of the main advantages of Bayesian statistics is its ability to facilitate existing knowledge to empower our models the second part of this document talks about approaches for eliciting relevant information from experts.

Modelling time sensitive data [21. 12. 2021]

We are often faced with data that is time sensitive (time-series). Since the time component is usually of paramount importance in such data we have to be extra careful when handling and modelling our data. For this purpose, we will take a look at specialized Bayesian models for modelling time sensitive data.

Holiday [28. 12. 2021]

Happy holidays!

Spatial models [4. 1. 2022]

In several scientific fields (e.g. weather, astronomy, geography) the data almost always has a spatial component to it. Simply discarding this component will most likely lead to suboptimal or even erroneous models. In this lecture we will learn how we can empower our Bayesian models with spatial information.

Putting it all together [11. 1. 2022]

Before the final lecture you will self-study a few parts from different research papers. We will use this lecture to discuss your comments and criticism of these papers. We will try to utilize all that we have learned and it will serve as final preparation for the exam. We will also discuss your opinion on this year's edition of the Bayesian statistics course.

Lab practice

The lab practice slot will be used for more hands-on oriented work. Since the work here will be more dynamic we do not have a clear schedule. Some of the activities that await you are:

- designing a questionnaire,
- communicating with a non-statistics expert,
- practically oriented guest lectures.

During weeks that will have no organized programme, lab practice slot will be used for consultations. You can use this opportunity to ask anything course or homework related. Consultations will be held online on https://uni-lj-si.zoom.us/my/juredemsar.