Statistical Learning

Phd Programme in Economics and Statistics (ECOSTAT), DEMS, University of Milano-Bicocca

Instructor

Rajen Shah is a Reader in Statistics at the University of Cambridge, where he previously completed his PhD under Richard Samworth. He was awarded the Royal Statistical Society Research Prize in 2017. His research interests include high-dimensional data analysis, causal inference and aspects of computational statistics. Currently he is Associate Editor for Annals of Statistics, JRSS-B and Biometrika.
`http://www.statslab.cam.ac.uk/~rds37/`

Course objectives

Over the last 25 years, the sorts of datasets that statisticians have been challenged to study have changed greatly. Where in the past, we were used to datasets with many observations with a few carefully chosen variables, we are now seeing datasets where the number of variables can run into the thousands and greatly exceed the number of observations. For example, in genomics settings we may have gene expression values measured for several thousands of genes, but only for a few hundred tissue samples. The classical statistical methods are often simply not applicable in these "high-dimensional" situations.

We will begin the *first module* by introducing ridge regression, a simple generalisation of ordinary least squares. Our study of this will lead us to some beautiful connections with functional analysis and ultimately one of the most successful and flexible classes of learning algorithms: kernel machines. We will then introduce the Lasso, a remarkable method that has been at the centre of much of the developments that have occurred in high-dimensional statistics, and will allow us to perform regression in the seemingly hopeless situation when the number of parameters we are trying to estimate is larger than the number of observations.

The *second module* will begin with a more in depth study of some theoretical properties of the Lasso. We will cover a few of the many extensions of the Lasso and describe an algorithm for efficient computation. Next, we will study graphical modelling. Whereas the earlier material concerns methods for relating a particular response to a large collection of (explanatory) variables, graphical modelling will give us a way of understanding relationships between the variables themselves. Ultimately we would like to infer causal relationships between variables based on (observational) data. This may seem like a fundamentally impossible task, yet we will show how by developing the graphical modelling framework further, we can begin to answer such causal questions. Statistics is not only about developing methods that can predict well in the presence of noise, but also about assessing the uncertainty in our predictions and estimates. In the final part of the course, we will tackle the problem of quantifying uncertainty in high-dimensional settings by studying the recently developed debiased Lasso method, which is currently the subject of a great deal of research activity.

Schedule

Module I: Kernel machines and an introduction to the Lasso
Module II: The Lasso and friends, Graphical modelling, causality and high-dimensional inference

Module I

- Monday 5 October 10-13
- Tuesday 6 October 11-13
- Wednesday 7 October 11-13
- Friday 9 October 11.30-13
- Monday 12 October 11-13
- Tuesday 13 October 11-13
- Friday 16 October 11.30-13

Module II

- Monday 19 October 10-13
- Tuesday 20 October 11-13
- Wednesday 21 October 11-13
- Friday 23 October 11.30-13
- Monday 26 October 11-13
- Tuesday 27 October 11-13
- Friday 30 October 11.30-13

We will host the lectures as a Zoom webinar. If you are enrolled to the course, you will receive an email with the link before the webinar begins. We offer the possibility to enrol in single modules of 14 hours but we strongly recommend to take the full course because its structure is intended as a single block of 28 hours.

Prerequisites
The overall flavour of the course overall will be theoretical and methodological. The material should be understandable for anyone with a thorough knowledge of undergraduate Statistics (particularly linear regression), probability (no measure theory required, but familiarity with e.g. multivariate Gaussians and conditional probabilities is essential), real analysis and linear algebra.

Enrollment
The course is addressed to Phd students, researchers, academics, professionals and practitioners with strong quantitative background (see Prerequisites) and motivation.
Application period: September 14, 2020 - September 25, 2020.
Applications should be addressed via e-mail to the ECOSTAT Administration Office (Dr. Silvia Locatelli, e-mail: phdecostat@unimib.it) and should include the following documents as attachments:

i) filled-out application form;

ii) updated curriculum vitae.

Admission is conditional on place availability. No fees are requested. The ECOSTAT Administration Office will contact non-admitted candidates only. The application form can be downloaded from the following website (section Events): `https://www.dems.unimib.it/en/research/phd-programme`

Contacts
For more information:
Prof. Matteo Manera, Coordinator of ECOSTAT, e-mail: matteo.manera@unimib.it
Prof. Aldo Solari, ECOSTAT faculty member, e-mail: aldo.solari@unimib.it

For administrative issues:
Dr. Silvia Locatelli, ECOSTAT Administration Office, e-mail: phdecostat@unimib.it