Statistical Computing

<table>
<thead>
<tr>
<th>Duration</th>
<th>216 hours (6 ESTC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>English level B1 (European Framework of Reference of Communicative Skills)</td>
</tr>
<tr>
<td>Entry requirements</td>
<td>BSc degree in Physics, Math or Computer Sciences. Ehh! Biology grade is also permissible!</td>
</tr>
</tbody>
</table>

About the course

The course provides an overview and introduction to the up-to-date methodology and techniques for non-linear statistical analysis of multidimensional data. Some methods and approaches would be discussed in detail; some computational experiments and lab projects devoted to the real data analysis would support the class-works.

Outline of content

1. Short introductory sub-course in probability theory foundations, and some classical issues of statistics.
2. Bulky data in multidimensional space; an idea of metric space, a brief outline of geometry and topology of metric spaces.
3. Multidimensional data visualization: what one can see and how? What usually remains invisible?
4. Principal component analysis. Factors of divergence.
5. Clustering. What is Clustering Strategy; the curse of dimension.
8. Elastic map technique to visualize multidimensional data and (nonlinear) clustering.

Educator

Michael Sadovsky, Doctor Habitats in Biophysics, Leading Researcher at ICM SB RAS; quarter-Professor, Department of Applied math & computer safety, Siberian Federal University
E-mail: msad@icm.krasn.ru

Course description

Tremendous (faster than exponential) growth of data necessary to be analyzed by experts raises the problem of the development and implementation of the relevant (and adequate) methods and techniques to do it. This demand results in significant progress both in pure and applied mathematics and related disciplines (say, programme design and algorithmic solutions). Hence, the course brings an introduction to the problem of bulky data analysis, and explores (in reasonable scale) the relevant issues in mathematics and related topics.

Course aims

The Aim of the course is present an introduction to the up-to-date ideas, approaches and techniques for multidimensional data analysis. Also, the course is aimed to give students a critical understanding of current technical implementations of some methods and techniques mentioned above, and to train student in the methods application.
Objectives

The objectives of the course are:
1) to give students an understanding of the concept of data analysis and visualization;
2) to provide students with up-to-date knowledge on some methods and techniques of clusterization, data visualization, extraction and retrieval of patterns of interdependence;
3) to provide students with comprehensive understanding of the constraints, advantages and problem points of the methods mentioned above;
4) to make students familiar with some software packages and toolkits used to implement the methods mentioned above into practiced of data analysis.

Learning outcomes

By the end of the course, students will be able
1) to identify and classify main phenomenae and basic peculiarities in multidimensional datasets, in order to select a proper and most efficient methods of the analysis;
2) to apply hierarchial classifications, PCA, $K$-means, mean-shift, and elastic map technique, where necessary;
3) to provide a sketch of interpretation of the results of multidimensional data treatment.

Attendance Policy

Students are expected to attend classes regularly, since the consistent attendance offers the most effective opportunity open to all students to gain command of the concepts and materials of the course. Meanwhile, excuses of various origin are permissible, in case students take a consultation and do the necessary class-work at home (or at their own). Such “hidden extramural” activity must not exceed a quarter of the total course time.

Assessments and Assessment Methods

The course assessment assignments will include (with the draft scheme of student’s grade):
1. Short-response questionnaire $\leqslant 10\%$ (exchangeable with item # 2);
2. Class participation $\leqslant 15\%$ (exchangeable with item # 1);
3. Practically oriented class/home mini-projects 35%;
4. Oral examination (Full course) 40%.

Recommended Reading and Other Handy Skills (optional)


Further reading (yet tentative)

что радует – совершенно официально!)


   Reasonable level in programming is welcomed.

**Special Features**

Statistics is an important tool in various areas of science ranging from biology to sociology. It covers concepts and methods which are able to draw inference based on empirical data, with given level of reliability and confidence.

Being a branch of mathematics, Statistics has strong connections to applications and offers a chance to grow through a number of various specific fields of knowledge and expertise.