Statistical Data Analysis – a course map

Jiří Kléma

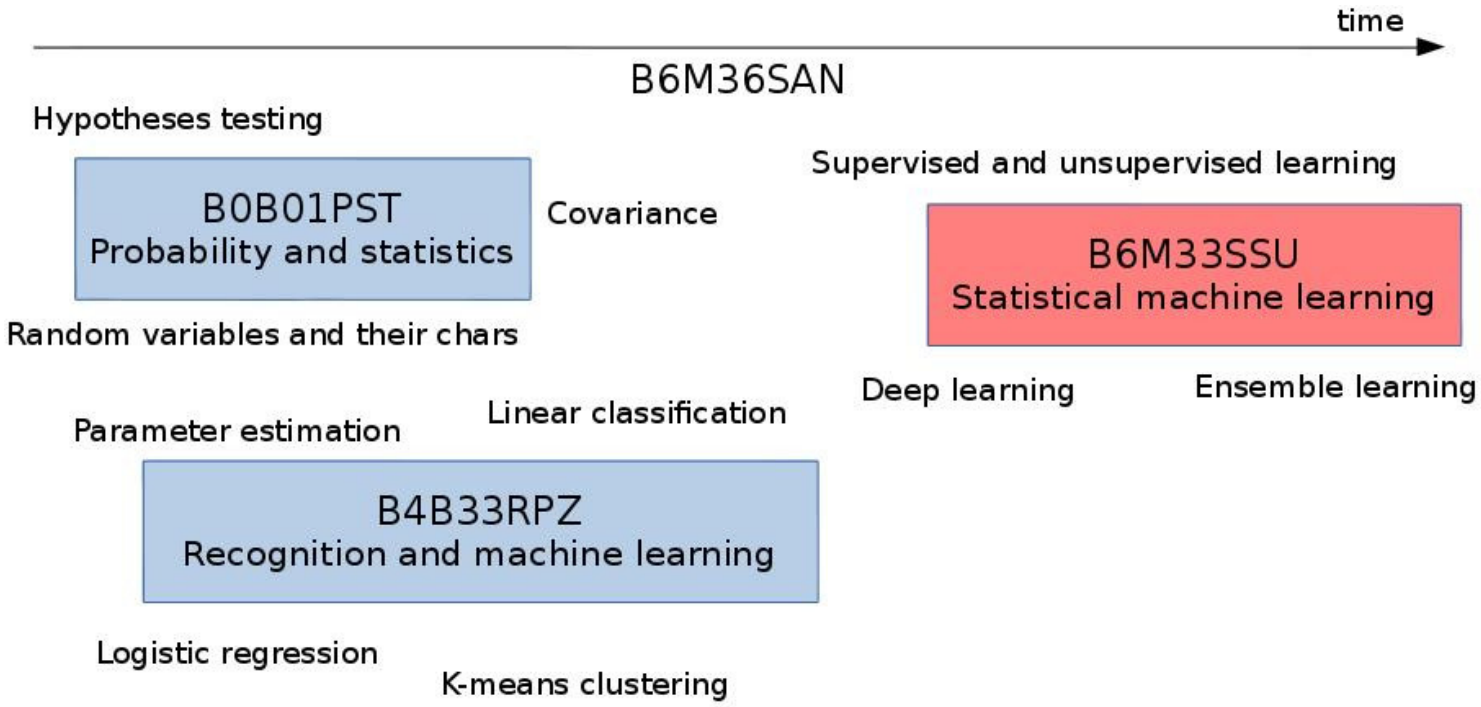
Department of Computer Science,
Czech Technical University in Prague

IDA
Intelligent Data Analysis
RESEARCH GROUP

http://cw.felk.cvut.cz/wiki/courses/b4m36san/start

B4M36SAN

- Purpose
  - This course mainly aims at the statistical methods that help to understand, interpret, visualize and model potentially high-dimensional data. It works with R environment.
- Interactions with other courses



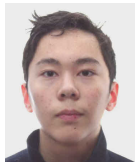
Teachers

Doc. Jiří Kléma (klema@fel.cvut.cz)
CTU, Dept. of Computer Science

Doc. Tomáš Pevný (pevnytom@fel.cvut.cz)
CTU, Dept. of Computer Science, CISCO Technical Leader

Doc. Zdeněk Míkovec (xmikovec@fel.cvut.cz)
CTU, Dept. of Computer Graphics and Interactions

Ing. Jan Blaha (blahaj22@fel.cvut.cz)
CTU, Dept. of Computer Science



Ing. Alikhan Anuarbekov (anuarali@fel.cvut.cz)
CTU, Dept. of Computer Science

IDA/JK Highlights

eBioMedicine
Part of THE LANCET *Discovery Science*

Volume 96, October 2023, 104782

Articles

Novel transcriptomic signatures associated with premature kidney allograft failure

Petra Hruba [a], Jiri Klema [b], Anh Vu Le [b], Eva Girmanova [a], Petra Mrazova [a], Annick Massart [c], Dita Maixnerova [d], Ludek Voska [e], Gian Benedetto Piredda [f], Luigi Biancone [g], Ana Ramirez Puga [h]

Molecular Oncology

Research Article | Open Access

Expression of circular RNAs in myelodysplastic neoplasms and their association with mutations in the splicing factor gene *SF3B1*

Iva Trsova, Andrea Hrustincova, Zdenek Krejcik, David Kundrat, Aleš Holoubek, Karolina Staflova, Lucie Janstova, Sarka Vanikova, Katarina Szikszai, Jiri Klema, Petr Rysavy **... See all authors**

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE | Full Access

Improved recovery and annotation of genes in metagenomes through the prediction of fungal introns

Anh Vu Le, Tomáš Větrovský, Denis Barucic, Joao Pedro Saraiva, Priscila Thiago Dobbler, Petr Kohout, Martin Pospíšek, Ulisses Nunes da Rocha, Jiří Kléma, Petr Baldrian

First published: 10 August 2023 | https://doi.org/10.1111/1755-0998.13852

BMC Bioinformatics

Research | Open Access | Published: 27 September 2022

circGPA: circRNA functional annotation based on probability-generating functions

Petr Ryšavý, Jiří Kléma & Michaela Dostálová Merkerová

BMC Bioinformatics **23**, Article number: 392 (2022) | Cite this article

Leukemia

Article | Open Access | Published: 03 May 2022

MYELODYSPLASTIC NEOPLASM

***RUNX1* mutations contribute to the progression of MDS due to disruption of antitumor cellular defense: a study on patients with lower-risk MDS**

Monika Kaisrlikova, Jitka Vesela, David Kundrat, Hana Votavova, Michaela Dostalova Merkerova, Zdenek Krejcik, Vladimir Divoky, Marek Jedlicka, Jan Fric, Jiri Klema, Dana Mikulenkova, Marketa Stastna Markova, Marie Lauermannova, Iolana Mertova, Jacqueline Soukupova Maaloufova, Anna Ionasova

PLOS ONE

OPEN ACCESS | PEER-REVIEWED

RESEARCH ARTICLE

On transformative adaptive activation functions in neural networks for gene expression inference

Vladimír Kunc, Jiří Kléma

Published: January 14, 2021 • https://doi.org/10.1371/journal.pone.0243915

See the preprint

0 Save	7 Citation
1,576 View	0 Share

The key terms

- Multivariate statistical analysis
  - concerned with data that consists of **sets** of measurements on a number of individuals,
  - statistical approach based on **stochastic data models**
    - a certain model is assumed (a class of models),
    - its parameters are learned based on data,
  - more than independent testing of the individual variables (i.e., univariate tests known from introductory statistical courses),
  - intertwined variables, **examined simultaneously**,
  - not only the extensions of univariate and bivariate procedures,
  - examples: multivariate analysis of variance,
    multivariate discriminant analysis.

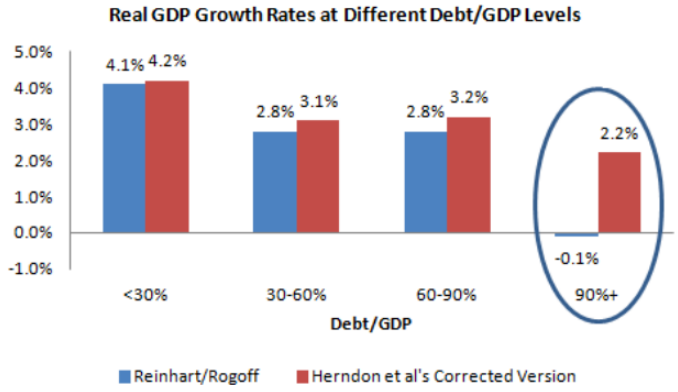
The key terms

- Applied statistics
  - in general, rather a branch of study than a course,
  - in here, the course could be understood as an opportunity to bring the (previously learned) methods to practice,
  - in labs, stress on applications and their implementation in R.
- Statistical inference/learning
  - close interaction with (statistical) machine learning,
  - sometimes it is difficult to distinguished these two fields
    - as their goals are interchangeable,
  - the most striking distinctions
    - different schools – statistics is a subfield of mathematics, machine learning is a subfield of computer science,
    - different eras – for centuries versus modern,
    - different degree of assumptions – larger versus smaller.

B4M36SAN – stories and jokes

- The Reinhart-Rogoff error – or how not to Excel at economics

Reinhart, C. M. and Rogoff, K. S. (2010a). Growth in a Time of Debt. *American Economic Review: Papers & Proceedings*, 100.



	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

- What is the difference between statistics, ML, AI and data mining?

Changes in this and previous year

- Mainly as a reaction to feedback from students,
- changes in lectures
  - (generalized) linear models as an universal multivariate data analysis tool,
  - 1 less lectures due to the dean's day on 20th November.
- practical changes in labs
  - more stress on understanding of concepts,
  - only then programming,
  - submissions in R as well as Python,
  - the course evaluation takes activity into account more significantly
    - 5 times 1 activity points can be obtained on the lab day for submissions,
    - other 5 bonus activity points for interaction during labs.

Syllabus

#	Lect	Content
1.	JK	Introduction, course map, review of the basic stat terms/methods.
2.	JK	Multivariate regression (continuous, linear regression, p-vals).
3.	JK	Multivariate regression (overfitting, model shrinkage).
4.	JK	Multivariate regression (non-linear, polynomial and local regression).
5.	JK	Discriminant analysis (categorical, LDA, logistic regression).
6.	JK	Generalized linear models, special cases.
7.	JK	Dimension reduction (PCA and kernel PCA).
8.	JK	Dimension reduction (other non-linear methods).
9.	TP	Anomaly detection.
10.	TP	Robust statistics.
11.	ZM	Empirical studies, their design and evaluation. Power analysis.
12.	JK	Clustering (basic methods).
13.	JK	Clustering (advanced methods, spectral clustering).

R package

- **R – the platform selected for labs**
  - the leading tool for statistics,
  - one of the main tools in data analysis and machine learning,
  - it is free, open-source and platform independent,
  - a large community of developers and users
    $\rightarrow$ a great variety of libraries, tutorials, mailing lists,
  - easy to integrate with other languages (C, Java, Python),
  - we actually use it,
  - bottlenecks in memory management, speed, and efficiency,
- alternatives
  - **Python** with its data analysis libraries (more general use),
  - **Matlab** (popular at FEL for its forte in control, Simulink etc.),
  - **Julia** a compiled language, modern features (GPU, parallel computing), simple to learn.

The key prerequisities – a brief review

- probability, independence, conditional probability, Bayes theorem,
- random variables, random vector,
- their description, distribution function, quantile function,
- categorical and continuous random variables,
- characteristics of random variables,
- the most common probability distributions,
- random vector characteristics, covariance, correlation, central limit theorem,
- measures of central tendency and dispersion, sample mean and variance,
- point and interval estimates of population mean and variance,
- maximum likelihood estimation, EM algorithm,
- statistical hypotheses testing,
- parametric and non-parametric tests,
- multiple comparisons problem, family wise error rate and false discovery rate.

Exam – the prerequisities make a part of it

- Sample questions (see the course web page for a larger list)
  - Explain in your own words the meaning of *p-value*. Assume that a p-value of a test is 0.028. What is the probability that its H_0 does not hold? Does it have any connection with the level of significance α?

Exam – the prerequisities make a part of it

- Sample questions (see the course web page for a larger list)
  - Explain in your own words the meaning of *p-value*. Assume that a p-value of a test is 0.028. What is the probability that its H_0 does not hold? Does it have any connection with the level of significance α?
  - $p = P(observation\_like\_this\_or\_more\_extreme|null) = P(o|H_0)$
  - $P(H_0|o) = \frac{P(o|H_0)P(H_0)}{P(o)} = \frac{P(o|H_0)P(H_0)}{P(o|H_0)P(H_0)+P(o|H_a)(1-P(H_0))}$
  - H_0 probability decreases with decreasing p-value of a correct statistical test, however, it is also a function of unexpectedness of the alternative hypothesis and the effect size (both can be hidden variables),
- an illustrative example: Did the sun just explode?
  - https://xkcd.com/1132/
  - H_0: the sun did not change, H_a: the sun has gone nova,
  - $P(H_0) = .999$, $P(o|H_0) = .028$, $P(o|H_a) = 0.972 \ldots P(H_0|o) = .97$.

The main references

:: Resources (slides, scripts, tasks) and reading

- G. James, D. Witten, T. Hastie and R. Tibshirani: **An Introduction to Statistical Learning with Applications in R.** Springer, 2014.
- T. Hastie, R. Tibshirani and J. Friedman: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** Springer, 2009.
- A. C. Rencher, W. F. Christensen: **Methods of Multivariate Analysis.** 3rd Edition, Wiley, 2012.
- research papers referenced in the individual lectures ...