COURSE DESCRIPTION CARD - SYLLABUS

Course name

Information Theory Methods in Data Analysis

Course

Field of study

Artificial Intelligence

Year/Semester

1/2

Area of study (specialization)

Profile of study

general academic

Level of study

Second-cycle studies

Course offered in

English

Form of study

full-time

Requirements

elective

Number of hours

Lecture

15

Laboratory classes

15

Other (e.g. online)

Tutorials

Projects/seminars

Number of credit points

3

Lecturers

Responsible for the course/lecturer:

dr hab. inż. Robert Susmaga

e-mail: robert.susmaga@cs.put.poznan.pl

tel.: 616652934

Faculty of Computing and Telecommunications

Piotrowo 2, 60-965 Poznań

Responsible for the course/lecturer:

Prerequisites

Basic knowledge regarding:

a) Calculus (logarithmic function, exponential function),

b) Linear Algebra (vectors, matrices, vector/matrix operations),

c) Probability Theory and Information Theory (probability, entropy).

Basic skills regarding designing, creating and testing computer programs (in a programming language of one's choice) that implement simple processing of static (vectors and matrices) and dynamic (lists, trees) data structures.

(recommended) A fair amount of cognitive curiosity and not less perseverance in pursuing the goals of personal development.

Course objective

The objective of the course is to present a selection of aspects of the Information Theory, one of the most fundamental theories underlying theoretical Computer Science of modern-day. The Information Theory deals with representing, storing and communicating information expressed in the form of symbols. Owing to the fact that many important applications of this theory reach far beyond the core of Computing Science, the presented selection of aspects will be confined to those that are most useful in such domains of the Computer Science as Data Analysis and Data Exploration. These include mainly the different useful measures (in particular: different variants of the Shannon entropy measures), so the course will be focused on both theoretical as well as practical aspects of those particular measures.

Course-related learning outcomes

Knowledge

Student:

- has advanced and in-depth knowledge regarding computing systems related to the information theory and data analysis methods (K2st_W1)

- has theoretically founded general knowledge related to key issues in Information Theory, in particular regarding multidimensional measures, along with their advantages (geometric interpretations) and disadvantages (K2st_W2)

- has advanced detailed knowledge of selected problems occurring within Data Analysis / Exploration,

in particular regarding the characterization of relationships between variables (K2st_W3)

- has some knowledge of development trends and the most important new achievements of computer science, in particular in the field of Information Theory and Data Analysis / Exploration, in which the latest achievements use effective algorithms for multidimensional spaces (K2st_W4)

- knows advanced methods, techniques and tools used in solving complex engineering tasks and conducting research in the field of multivariate data analysis, primarily regarding the use of tools and methods of Information Theory applied to solve Data Analysis and Data Exploration problems (e.g. the possibility of using multidimensional information measures to quantify the levels of dependencies between variables) (K2st_W6).

Skills

Student

– is able to obtain information from literature, databases and other sources (both in Polish and English), integrate them, interpret and critically evaluate them, draw conclusions and formulate and fully justify opinions (K2st_U1)

– is able to plan and carry out experiments, including computer measurements and simulations, interpret the obtained results and draw conclusions and formulate and verify hypotheses related to information theory problems (K2st_U3)

-- is able to use analytical, simulation and experimental methods to formulate and solve engineering tasks and simple research problems, in particular those derived from in the field of Information Theory (K2st_U4)

- is able to - when formulating and solving engineering tasks, in particular regarding machine learning and data mining - integrate knowledge from various areas of mathematics (Information Theory, Data Analysis and Exploration, etc.), also taking into account non-technical aspects (K2st_U5)

- is able to assess the usefulness and the possibility of using new achievements (methods and tools) and new IT products, primarily in the field of Information Theory, and apply them to the analysis and processing of multidimensional data (e.g. allowing to discover dependencies between objects) (K2st_U6)

is able to can make a critical analysis of existing technical solutions (in particular, e.g. in the areas of: Information Theory and data mining - solutions requiring effective characterization of dependencies in data) and propose their improvements (improvements) (K2st_U8)

- is able to assess the usefulness of methods and tools for solving an engineering task in the fields of Information Theory and data mining (K2st_U9)

- is able - using among others conceptually new methods - to solve problems related to Information Theory and data mining (K2st_U10)

-- can - in accordance with the given specification, taking into account non-technical aspects - design a complex IT system and implement this project - at least in part - using appropriate methods, techniques and tools, including existing or designed tools for theoretical and informative analysis data (K2st_U11).

Social competences

Student:

-- understands that in computer science, knowledge and skills very quickly become obsolete (K2st_K1),

-- is able to adequately define priorities for the implementation of a task set by her/himself or others (K2st_K2).

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Formative assessment (laboratory classes): evaluation of the solutions to the assigned programming problems (as they arise).

Final assessment:

-- (laboratory clasess): evaluation of the solutions to the assigned programming problems (final),

-- (lectures): evaluation of the results of a written test (45--60 min) with both multiple choice, short answer and (small) computational questions (mostly: micro-problems to be solved in writing).

Programme content

The course includes, but is not limited to, the following.

The idea of Shannon information (shortly: information). The measure of information content: construction and properties. The idea of the Shannon entropy (shortly: entropy) and its basic mathematical properties, recognition of dimensions in multi-dimensional entropy, graphs and extrema of multi-dimensional entropy, entropy as the measure of information content: construction and properties.

Entropy-related measures in multi-dimensional Data Exploration problems: the idea, properties and interpretations (with focus on multi-dimensional contexts).

Construction/derivation and fundamental properties of: joint entropy, conditional entropy, mutual information, cross-entropy, Kullback-Leibler divergence. Dependencies between joint entropy, conditional entropy, mutual information, cross-entropy and Kullback-Leibler divergence.

Alternative definitions of entropy.

Different aspects of Information Theory tools in popular Data Analysis and Data Exploration areas (mutual information in decision trees, mutual information in document retrieval, cross-entropy in neural networks, etc.).

Exemplary applications of Information Theory aspects in Lossless and Lossy Data Compression.

Teaching methods

Lectures: slide show presentations (theoretical elements, explanations, examples, exercises).

Laboratory classes: designing and creating (in a programming language of one’s choice) programs that solve the assigned problems (which illustrate the ideas and notions presented during the lectures).

Bibliography

Basic

1. D.J.C. MacKay: Information Theory, Inference, and Learning Algorithms, Cambridge University Press, Cambridge, UK, 2003.

2. T.M. Cover, J.A. Thomas, Elements of Information Theory, 2nd Edition, Wiley and Sons, Hoboken, New Jersey, 1991.

Additional

1. K. Sayood (red.): Lossless Compression Handbook, Academic Press, Elsevier Science, San Diego, California, 2003.

2. K. Sayood: Introduction to Data Compression, 3rd Ed., Morgan Kaufmann Publishers, San Francisco, California, 2006.

Breakdown of average student's workload

	Hours	ECTS
Total workload	75	3,0
Classes requiring direct contact with the teacher	30	1,5
Student's own work (literature studies, preparation for laboratory classes, preparation for tests) [1]	45	1,5

[1] delete or add other activities as appropriate