Skoltech
Skolkovo Institute of Science and Technology

Course Syllabus

Course Title (in English)	Geometrical Methods of Machine Learning
Course Title (in Russian)	Геометрические методы машинного обучения
Lead Instructor(s)	Bernstein, Alexander

Contact Person	Oleg Kachan
Contact Person's E-mail	oleg.kachan@skoltech.ru

Course Description

Many machine learning problems are fundamentally geometric in nature. The general goal of machine learning is to extract previously unknown information from data, which is reflected in the structure (underlying geometry) of the data. Thus, understanding the shape of the high-dimensional data plays an important role in modern learning theory and data analytics.
Real-world data obtained from natural sources occupy usually only a very small part of the 'observation' space and concentrate non-uniformly along lower dimensional structures, and geometrical methods allow discovering the shape of these structures from given data.
A large part of the course addresses to most popular geometrical model of high-dimensional data called manifold model and introduces modern manifold learning methods. The course includes also Topological Data Analysis that is an emerging trend in exploratory data analysis and data mining and provide a general framework (a set of topological and geometric tools) to analyze high-dimensional, incomplete and noisy data. Necessary short information on differential geometry and topology will be given in the course.
The course will provide examples of the application of geometric and topological methods of data analysis to various applied problems.
The course is useful for MSc and PhD students interested in recent geometrical methods, lying at the interface between Mathematics and Machine learning.
At the end of the course, students will know the basic ideas of a geometric approach to data analysis, possess modern geometric and topological methods of data analysis and be able to apply them to solve basic machine learning problems, such as classification, regression, dimensionality reduction, data presentation and visualization, clustering and others. This knowledge and skills allows them to participate in real-life projects to solve complex applied problems of data analysis.

Course Prerequisites / Recommendations

- Mathematics for Data Science OR
- Numerical Linear Algebra OR
- Optimization Methods OR
- Machine learning

We suppose an attendee be fluent with real analysis (calculus), basics of linear algebra, functional analysis, probability and statistics, graph theory and algorithms.

Аннотация

Многие задачи машинного обучения имеют геометрическую природу. Общая цель машинного обучения - извлечь из данных ранее неизвестную информацию, которая отражается в структуре (геометрии) данных. Поэтому понимание формы многомерных данных играет важную роль в современной теории обучения и аналитике данных.
Данные реального мира, полученные из естественных источников, обычно занимают лишь очень небольшую часть «пространства наблюдений» и концентрируются на структурах более низкой размерности, а геометрические методы позволяют обнаружить форму этих структур по заданным данным.
Значительная часть курса относится к наиболее популярной геометрической модели многомерных данных, называемой «моделью многообразия», в соответствии с которой многомерные данные лежат на или вблизи многообразия меньшей размерности. Курс включает также топологический анализ данных, который активно используется в разведочном и интеллектуальном анализе данных и предоставляет набор топологических и геометрических инструментов для анализа многомерных, неполных и зашумленных данных. Необходимые краткие сведения по дифференциальной геометрии и топологии будут даны в курсе.
В курсе будут приведены примеры применения геометрических и топологических методов анализа данных к различным прикладным задачам.
Курс рассчитан на студентов-магистров и аспирантов, интересующихся новейшими геометрическими и топологическими методами, лежащими на стыке математики и машинного обучения.
По окончанию курса, слушатели будут знать основные идеи геометрического подхода к анализу данных, владеть современными геометрическими и топологическими методами анализа данных и уметь применять их для решения основных задач машинного обучения, таких как классификация, регрессия, снижение размерности, представление и визуализация данных, кластеризация и другие. Эти знания и умения позволяет им участвовать в реальных проектах по решению сложных прикладных задач анализа данных.

Course Academic Level	Master-level course suitable for PhD students
Number of ECTS credits	3

Topic	Summary of Topic	Lectures (# of hours)	Seminars (# of hours)	Labs (# of hours)
Introduction	Geometrical methods in Machine learning: motivation, examples, main tasks, approaches	1		1
Linear models of data	Projection methods in Machine Learning: Principal Component Analysis (PCA), Pursuit projection, random projections	2	2	
Differential geometry (short basics)	Curves, surfaces, tangent spaces, geodesic line, curvature	2		
Intrinsic dimension of nonlinear datasets	Definition and estimation of dataset's intrinsic dimension	2	1	
Heuristic methods of nonlinear data analysis	Multidimensional Scaling, Replicative Neural Networks, Kernel PCA	2	1	
Elements of topology (short basics)	Topological space, manifolds, vector fields on manifold, Riemannian manifold	2		
Manifold model of nonlinear data	Manifold assumption, manifold learning tasks	1		
Dimensionality reduction as Data manifold description	Manifold learning algorithms: Locally Linear Embedding, ISOmetric MAPing (ISOMAP), Laplacian Eigenmaps	2	3	
Data manifold analysis based on subspace learning	Tangent spaces estimation, Local Tangent Space Alignment, Grassman&Stiefel Eigenmaps	1	1	
Statistical problems on Data manifold	Regression on manifolds, density estimation on manifolds	1	0	
Topological Data Analysis	Examples of TDA tasks. Simplicial complexes. Filtration. Persistence diagrams	2	1	

Assignment Type	Assignment Summary
Homework Assignments	Apply learned algorithms to specific datasets
Final Project	Study the proposed articles, choose the appropriate algorithm and apply it to a specific dataset
Final Exam	Prepare a report on a given topic using the knowledge gained in the course. Answer additional questions.

Type of Assessment Graded

Grade Structure

Activity Type	Activity weight, %
Homework Assignments	30
Final Project	35
Final Exam	35

A: 80

B:	70
C:	60
D:	50
E:	50
F:	0
Attendance Requirements	Mandatory with Exceptions
Course Term (in context of Academic Year)	Term 4
Course Delivery Frequency	Every year

Students of Which Programs do You Recommend to Consider this Course as an Elective?

Masters Programs	PhD Programs
Data Science	

Required Textbooks	ISBN-13 (or ISBN-10)
Wasserman, L.: All of Nonparametric Statistics. Springer Texts in Statistics, Berlin (2007)	978-0-387-30623-0

Recommended Textbooks	ISBN-13 (or ISBN-10)
Burges Christopher J.C. Dimension Reduction: A Guided Tour. Foundations and Trends in Machine Learning, 2010, 2(4): 275 – 365. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/FnT_dimensionReduction.pdf	9781601983787
Cox T.F., Cox M.A.A. Multidimensional Scaling. Chapman and Hall, 2001.	9781584880943
Jollie T. Principal Component Analysis. New-York, Springer, 2002. http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principal%20Component%20Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf	9780387954424
Ma Y, Fu Y, eds. Manifold Learning Theory and Applications. London: CRC Press, 2011. http://www.gbv.de/dms/goettingen/689943164.pdf	9781439871096
Gorban AN, Kegl B, Wunsch D, Zinovyev AY. Principal Manifolds for Data Visualisation and Dimension Reduction. Springer, Berlin - Heidelberg - New York, 2008. http://pca.narod.ru/contentsgkwz.htm	9783540737490
Jost, J. Riemannian Geometry and Geometric analysis, 6th edn. Springer-Verlag, Berlin, Heidelberg (2011) http://cds.cern.ch/record/1666885/files/9783540773405_TOC.pdf	9783319618593
Lee, J.M. Manifolds and Differential Geometry. Graduate Studies in Mathematics, 107. American Mathematical Society, Providence (2009)	9780821848159

Papers	DOI or URL
A. V. Bernstein, Manifold learning in statistical tasks. Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki, 2018, Volume 160, Book 2, 229–242	http://www.mathnet.ru/links/af4197b9f7c9feecbad3b6a40c191c50/uzku1447.pdf

Software
Python

Knowledge
1. Know and understand an essence of main models and methods of geometrical data analysis 2. Be fluent with theoretical foundations and algorithmic techniques of modern geometrical data analysis

Skill
1. Identify geometrical nature of various machine learning problems and approaches to their solutions 2. Choose appropriate methods and algorithms when using geometrical approaches in solving the machine learning tasks

Experience
1. Ability to work with research literature on using the geometrical data analysis approaches and methods for solving the machine learning tasks 2. Propose and implement the studied methods for solving the machine learning tasks

Select Assignment 1 Type	Homework Assignments
Input Example(s) of Assignment 1 (preferable)	Apply several studied algorithms for estimating the intrinsic dimensionality of a given dataset and compare their results
Assessment Criteria for Assignment 1	The student expected to demonstrate understanding of how to choose appropriate methods and techniques, how to use it correctly in the homework task, and how to compare the obtained solutions. Maximum level:4
Select Assignment 2 Type	Final Project
Input Example(s) of Assignment 2 (preferable)	Topological Data Analysis for COVID-19 Diagnosis (apply the appropriate method for extracting the TDA features from CT images with COVID 19 or Effusion and apply the Deep Learning technique for classification task)
Assessment Criteria for Assignment 2	The student expected to present the project presentation, structured perfectly with clear explanation of all essential parts of the project. The presentation should be delivered in a confident manner with respect to the audience. Student’s participation in discussion and Q&A session is active and relevant the project goal and results. Maximum level:4
Select Assignment 3 Type	Final Exam
Input Example(s) of Assignment 3 (preferable)	Explain the essence of the Locally Linear Embedding, ISOMAP and Laplacian Eigenmaps methods and explain the difference in approaches underlying these methods.
Assessment Criteria for Assignment 3	The student expected to present the answers to the given questions with clear explanation of all essential parts of the question topics. The answers should be delivered in a confident manner to demonstrate understanding by Student of the essence of the questions and knowledge of the main results in this area. Maximum level:4

Upload a File (if needs to be)	https://ucarecdn.com/af086ecc-430b-4f34-84be-404aa44e88b9/