VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
STUDIJŲ MODULIO KORTELĖ
Informacinių technologijų katedra

A dalis

Modulio pavadinimas

Mašininio mokymosi metodai didelių duomenų apdorojime

Modulio pavadinimas (anglų kalba)

Large Scale Machine Learning

Modulio grupė	**Studijų dalyko**
Modulio blokas	**Mokslo krypties doktorantūros komisijos nustatyti dalykai**
Priklausomybė	**Katedros**

Mokslo krypties ir srities kodas		**Studijos**
T 007	**T 000**	**Doktorantūros**

Modulio kodas

Fakultetas	Katedra	B, A, M, I, D	Modulio Nr.*
F	M I T	D	19001

* modulio registracijos numeris katedroje

Kreditai

Iš viso:	Iš jų: KD, KS, KP
6	0

Atsiskaitymo forma

Į, E1, E2, E, BE, BD, TD, A	KD, KS, KP
E	

Studijų forma		Paskaitoms	Lab. darbams	Pratyboms	Aud. darbui	Sav. darbui	Iš viso
Nuolatinės studijos	F	30	0	0	30	130	160
Ištęstinės studijos	I						

Modulio tikslas

Sukurti pagrindą, taikant šiuolaikinius mašininio mokymosi metodus, sprendžiant praktines didelių duomenų užduotis. Formuoti studentų gebėjimą analizuoti didelius duomenų rinkinius siekiant pasirinkti optimalų mokymosi algoritmą.

Modulio tikslas (anglų kalba)

Creating a base for the application of modern methods of machine learning on big data to solve practical tasks and the formation of students ability to analyze large data sets in order to select optimal learning algorithm.

Suteikiamos žinios ir gebėjimai

Žinos: Darbo su dideliais duomenų rinkiniais ypatumus; Pagrindinius mašininio mokymosi algoritmų kūrimo metodus, kurie atsižvelgia į didelių apdorojamų duomenų masyvų charakteristikas. Gebės: Pasirinkti tinkamą mašininio mokymosi modelį; Pritaikyti mašininio mokymosi technikas sprendžiant praktinius uždavinius.

Suteikiamos žinios ir gebėjimai (anglų kalba)

Know: Features of working with large data arrays; Main approaches to the development of machine learning algorithms that take into account the characteristics of big data. Be able to: Select suitable machine learning models taking into account features of big data; Apply machine learning techniques to solve applied tasks on large data arrays.

Modulio anotacija

Praktikoje dažnai tenka taikyti mašininį mokymosi didelėms duomenims apdoroti. Tai tokie duomenys, kurie dėl įvairių priežasčių negali būti patalpinti vienoje mašinoje. Tokių duomenų atsiranda daugelyje sričių: internetas, finansai, davikliai, NLP. Pagrindiniai sprendžiami uždaviniai šiuo atveju yra:
Kaip paspartinti esamus mašininio mokymosi algoritmus, arba, kaip adaptuoti šiuos algoritmus, kad jie tiktų paskirstytoms sistemoms;
Naujų mašininio mokymosi algoritmų kūrimas, kurie nuo pat pradžių skirti darbui su dideliais duomenimis;
Architektūrų ir skaičiavimo modelių, tinkančių darbui su dideliais duomenimis, kūrimas.

Modulio anotacija (anglų kalba)

In practical tasks, ML often has to work with "big data", i.e. data that is impossible or too long to process on one machine. Such data arise in many areas: the Internet, finance, sensors, NLP. At the junction of ML and CS, the subdiscipline "Large Scale Machine Learning" arose. Research on it is conducted in the following main areas: How to speed up existing ML algorithms or how to transfer them to distributed systems;
Development of new ML algorithms, originally designed for big data;
Development of architectures and computational models suitable for machine learning on big data.

Literatūra (autorius, leidinio pavadinimas, leidykla, metai)

1. Bekkerman R., Bilenko M., Langford J. Scaling up Machine Learning, Cambridge University Press, 2011. - 492 p. - ISBN-13: 978-0521192248
2. Leskovec J., Rajaraman A., Ullman J.D. Mining of Massive Datasets, Cambridge University Press, 2014. - 476 p. - ISBN-13: 978-1107077232.
3. Bishop, C.M. Pattern Recognition and Machine Learning. - Springer, 2006. - 738 p.
4. Cristianini, N. & Shawe-Taylor, J. (2000), An Introduction to Support Vector Machines, Cambridge University Press.
5. Kearns, M. & Vazirani, U. (1994), An Introduction to Computational Learning Theory, MIT Press.
6. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. - Springer, 2009. - 533 p.
7. Shalev-Shwartz Shai, Ben-David Shai. Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014. - 409 p. - ISBN-13: 978-1107057135.

IT resursai:

Savarankiško darbo turinys

Užduoties pavadinimas	Sav. darbo apimtis vienai užduočiai					Užduočių skaičius				Iš viso valandų			
	Rėžis	Priimta											
		NL(S)	I(S)	I(T)	NL(T)	NL(S)	I(S)	I(T)	NL(T)	NL(S)	I(S)	I(T)	NL(T)
Referatas	8-27	22				1				22			
Pasirengimas atsiskaitymui	10-60	20				1				20			
Mokslinis seminaras	20-56	22				4				88			

Savarankiško darbo grafikas

Užduoties tipas	Užduoties pateikimo(*) ir atsiskaitymo(+) savaitė																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Nuolatinės studijos																					
Referatas	*	1																			
	+												1								

Modulio sudarytojai (vardas, pavardė):

Dalius Mažeika
Dmitrij Šešok

Modulio egzaminuotojai (vardas, pavardė):

Dalius Mažeika
Dmitrij Šešok
Profilinės katedros vedėjas

Katedros vedėjas (vardas, pavardė):

Dmitrij Šešok

Doktorantūros komisijos nutarimas

1. Modulis **atestuojamas**
2. Modulis skirtas mokslo krypčiai: **Informatikos inžinerija**
3. Modulio atestacija galioja: nuo 2019-09-01 iki 2024-08-31

Modulį atestavo

Mokslo krypties doktorantūros komisijos pirmininkas (vardas, pavardė)

Data

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
STUDIJŲ MODULIO DARBO PROGRAMA
Informacinių technologijų katedra

B dalis

Modulio pavadinimas

Mašininio mokymosi metodai didelių duomenų apdorojime

Modulio pavadinimas (anglų kalba)

Large Scale Machine Learning

Modulio kodas

Fakultetas		Katedra		B, A, M, I, D	Modulio Nr.*
F	M	I	T	D	19001

* modulio registracijos numeris katedroje

Kreditai

Iš viso:	Iš jų: KD, KS, KP
6	0

Atsiskaitymo forma

Į, E1, E2, E, BE, BD, TD, A	KD, KS, KP
E	

Studijų forma		Paskaitoms	Lab. darbams	Pratyboms	Aud. darbui	Sav. darbui	Iš viso
Nuolatinės studijos	F	30	0	0	30	130	160
Ištęstinės studijos	I						

Paskaitų temų sąrašas

List of the Course lecture topics

Temos (darbo) pavadinimas	Valandų skaičius			
	NL(S)	I(S)	I(T)	NL(T)
1. Įvadas. Kas yra dideli duomenys mašininio mokimosi kontekste. Praktinių uždavinių pavyzdžiai. Kurso apžvalga. 1. Introduction. What is big data in the context of machine learning. Examples of practical problems. Overview of the course program. Estimation / Approximation / Optimization trade-off.	4			
2. Tiesinių modelių Online apmokymas. Tiesiniai modeliai. Praradimų funkcijos. Progressive validation. 2. Online training of linear models. Linear models. Loss functions. Online learning. Progressive validation.	2			
3. Naudingos online apmokymo technikos: objektų svorių apskaita, adaptyvus learning rate, kintamųjų dimensijos įvertinimas. Online Bootstrap. 3. Useful Tricks for Online Learning: Scale Accounting objects, adaptive learning rate, taking into account the dimension of variables. Online Bootstrap.	2			
4. Hiperparametrų optimizavimas. Ameba, lattice, random search, Bayesian optimization. Lygiagretinimo metodai. 4. Optimization of hyperparameters. Ameba, lattice, random search, Bayesian optimization. Parallelization methods.	2			
5. Hashing in machine learning on large data arrays. Hash traits. Hashing in word processing. Hash core Locality-sensitive hashing. 5. Hashing in machine learning on large data arrays. Hash traits. Hashing in word processing. Hash core Locality-sensitive hashing.	4			
6. Methods of working with categorical features. One-hot encoding. Spendimų medžiai. 6. Methods of working with categorical features. One-hot encoding. Decision trees.	4			
7. Factorization machines. Field-aware Factorization Machines. DeepFM. Naive Bayes classifier. Counters. 7. Factorization machines. Field-aware Factorization Machines. DeepFM. Naive Bayes classifier. Counters.	4			
8. Metodas BFGS. BFGS ir Limited-memory BFGS (L-BFGS). Skaičiavimų sudėtingumas. Stochastic BFGS. 8. BFGS method. BFGS and Limited-memory BFGS (L-BFGS). The complexity of the calculations. Stochastic BFGS.	4			
9. Block and coordinate descent, ADMM. Block and coordinate descent for generalized linear models (GLMs). GLMNET algoritmas. Parallelization options (many kernels, distributed systems). 9. Block and coordinate descent, ADMM. Block and coordinate descent for generalized linear models (GLMs). GLMNET Algorithm. Parallelization options (many kernels, distributed systems).	2			
10. ADMM metodas. Consensus and sharing. Implements ADMM for various functions. 10. ADMM method. Consensus and sharing. Implements ADMM for various functions.	2			
Iš viso:	**30**			

Modulio sudarytojai (vardas, pavardė):
Dalius Mažeika
Dmitrij Šešok

Modulio egzaminuotojai (vardas, pavardė):
Dalius Mažeika
Dmitrij Šešok
Profilinės katedros vedėjas

Katedros vedėjas (vardas, pavardė):
Dmitrij Šešok