DOKTORANTŪROS STUDIJŲ DALYKO SANDAS

Dalyko pavadinimas	Mokslo kryptis (šaka) kodas	Fakultetas	Katedra
Netiesiniai statistikos modeliai masinių duomenų analizėje (7 ECTS, 190 val)	Informatika (09 P), Matematika (01P), Informatikos inžinerija (07T)	Matematikos ir informatikos fakultetas	Duomenų mokslo ir skaitmeninių technologijų institutas

Studijų būdas	Kreditų skaičius ECTS	Studijų būdas	Kreditų skaičius
paskaitos (pavasario sem.)	1 ECTS (30 val.)	konsultacijos	1 ECTS (30 val.)
individualus	4 ECTS (100 val.)	seminarai (pavasario sem.)	1 ECTS (30 val.)

Dalyko anotacija

Dalyko tikslas – papildyti studentų turimas žinias apie automatinį (mašininį) mokymąsi netiesinio statistinio modeliavimo žiniomis akcentuojant kritinį statistinį mąstymą.

Kurso kontekste didieji duomenys (angl. *Big data*) suprantami, kaip masiškai be konkretaus tikslo surinkti duomenys ir yra vadinami masiniais duomenimis. Pastarieji duomenys, kaip taisyklė, yra heterogeniški nuo paprasčiausių mažų tekstinių įrašų iki akcijų informacijos kas minutę ar viso genomo duomenų. Šiuos duomenis galima analizuoti traktuojant juos kaip juodąją dėžę (angl. *Black box*) ir naudoti automatinio (mašininio) mokymosi metodus. O galima taikyti statistinį modeliavimą ir, atsižvelgiant į duomenų generavimo procesą, daryti prielaidas apie duomenų skirstinį ir jas testuoti. Kuo konkretesnės (griežtesnės) prielaidos dera su duomenimis, tuo turiningesnė ir subtilesnė gautų rezultatų interpretacija ir išvados. Todėl automatinio (mašininio) mokymosi metodus prasminga derinti su netiesinio (parametriniais ir neparametriniais) statistinio modeliavimo metodais.

Pagrindinės temos:

- Klasikiniai ir robastiniai tiesiniai metodai (Agresti)
- Apibendrintieji tiesiniai (AT) modeliai, modelio pasirinkimas ir statistinės išvados (Agresti)
- Pakartotiniai stebiniai, atsitiktiniai veiksniai ir ilgalaikių stebėjimų duomenys (Faraway)
- Mišriųjų veiksnių modeliai, nenormalusis atsako kintamasis (Faraway)
- Bajeso daugialygis (hierarchinis) modeliavimas, Markovo grandinių Monte Karlo metodai (Madigan)
- AT modelio reguliarizavimas (Agresti)
- Neparametrinė regresija (Faraway), apibendrintieji adityvūs modeliai ir glodinimo metodai (Agresti, Faraway)

Studentų gebėjimai: Kurso metu studentai, išanalizavę skirtą literatūrą, pasirinktiems duomenims realizuoja apibendrintuosius tiesinius ir neparametrinius regresijos modelius. Esant poreikiui, masinių duomenų analizei moka pritaikyti reguliarizacijos metodus, sudėtingų struktūrų tyrime -- Bajeso metodus. Kursą išklausęs studentas suvokia ir geba įvertinti duomenų ir jų analizės rezultatų neapibrėžtumą ir patikimumą. Metodai realizuojami R kalba arba kita studentui ir dėstytojui abipusei priimtina programine įranga.

Atsiskaitymas

Atsiskaitymą sudaro dvi dalys – uždavinių sprendimas ir mokslinis referatas. Abi dalys yra vertinamos 10 balų skalėje. Uždavinių sprendimas sudaro 40 proc. galutinio vertinimo.

Naudojant sukauptas žinias studentas išsprendžia šiuos uždavinius iš Wasserman knygos:

- 15.6 skyrius nuo 4 uždavinio;
- 21.7 skyriaus 7 uždavinys;

- 22.13 skyriaus 3 ir 6 uždaviniai
- 24.7 skyriaus 5 uždavinys.

Gavus teigiamą vertinimą iš uždavinių sprendimo yra rašomas referatas. Referatą turėtų sudaryti šios dalys:

1. Trumpas ruošiamos disertacijos pristatymas (kodėl netiesiniai modeliai svarbūs jūsų disertacijoje?).
2. Mokslinių straipsnių analizė iš disertacijos srities, kuriuose taikomi netiesiniai modeliai (kokie modeliai ir kokiems uždaviniams spręsti pritaikyti?)
3. Empirinė dalis. Netiesinių modelių parinkimas ir realizavimas pasirinktam duomenų masyvui (duomenų aprašymas, žvalgomoji analizė, modelių parinkimas ir realizavimas, gautų rezultatų aprašymas ir diskusija). Pasirinktas duomenų masyvas ir modelius realizuojantis programinis kodas turi būti prieinamas kurso dėstytojams.

Pagrindinė literatūra

Agresti, A., 2015. Foundations of linear and generalized linear models. John Wiley & Sons. http://bayanbox.ir/view/7443147326514856944/Foundations-of-Linear-and-Generalized-Linear-Models-Wiley-Series-in-Probability-and-Statistics-Alan-Agresti-2015.pdf

Faraway, J.J., 2016. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models (Vol. 124). CRC press.

Madigan, D., Ridgeway G., 2002. Bayesian Data Analysis for Data Mining. https://pdfs.semanticscholar.org/58be/87292ce8c4a0eefca6dd5430368f4af4e177.pdf

Wasserman, L., 2004. All of statistics: a concise course in statistical inference (Vol. 26). New York: Springer.

Papildoma literatūra

Alpaydin, E., 2010. Introduction to Machine Learning. The MIT Press. ISBN-10: 0-262-01243-X, ISBN-13: 978-0-262-01243-0. http://cs.du.edu/~mitchell/mario_books/Introduction_to_Machine_Learning_-_2e_-_Ethem_Alpaydin.pdf

Han, J., Kamber, M., 2006. Data Mining Concepts and Techniques. 2nd ed. CA: Morgan Kaufmann Publishers is an imprint of Elsevier. http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf
https://pdfs.semanticscholar.org/02e0/bc77460469aefec5bd794ee6c4efc15e6adb.pdf

Konsultuojančiųjų dėstytojų vardas, pavardė	Mokslo laipsnis	Svarbiausieji darbai mokslo kryptyje (šakoje) paskelbti per pastaruosius 5 metus
Audronė Jakaitienė	Dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Audronė+Jakaitienė
Marijus Radavičius	Dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Marijus+Radavičius
Olga Kurasova	Dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Olga+Kurasova