DOKTORANTŪROS STUDIJŲ DALYKO SANDAS

Dalyko pavadinimas	Mokslo kryptis, kodas	Fakultetas	Katedra
Priešiškas mašininis mokymasis	Informatikos inžinerija (T 007)	Matematikos ir informatikos fakultetas	Duomenų mokslo ir skaitmeninių technologijų institutas

Studijų būdas	Kreditų skaičius ECTS	Studijų būdas	Kreditų skaičius
paskaitos	0 (rudens sem.)	konsultacijos	1
individualus	6	seminarai	0

Dalyko anotacija

Dalyko tikslas: suteikti doktorantui sistemingų žinių apie priešišką mašininį mokymąsi, taip pat apie mašininio mokymosi modelių ir dirbtiniu intelektu (DI) grįstų sistemų saugumo bei privatumo rizikas. Kurso metu taip pat siekiama ugdyti doktoranto gebėjimą kritiškai analizuoti mokslinę literatūrą, formuluoti mokslinio tyrimo problemą, apibrėžti grėsmių modelį, vertinti atakas bei gynybos metodus ir taikyti priešiško mašininio mokymosi koncepcijas įvairiose informatikos inžinerijos srityse. Dalykas grindžiamas vadovaujamu savarankišku mokslinės literatūros studijavimu, individualiomis konsultacijomis ir su disertacijos tema susijusiu individualiu tiriamuoju projektu.
Reikalavimai: Tikimasi, kad doktorantas turės bazinių mašininio mokymosi, giliojo mokymosi ir programavimo „Python“ kalba žinių.
Įvadinė dalis. Priešiškojo mašininio mokymosi apibrėžimas ir istorinė raida, jo vieta DI saugumo, privatumo ir patikimo DI tyrimuose. Mašininio mokymosi modelių ir duomenimis grįstų sprendimų priėmimo sistemų pažeidžiamumai. Priešiškojo mašininio mokymosi vaidmuo šiuolaikiniuose DI saugumo, kibernetinio saugumo ir patikimo DI tyrimuose.

Priešiškas mašininis mokymasis:

- Pagrindinės priešiškojo mašininio mokymosi ir DI saugumo sąvokos.
- Grėsmių modeliai: užpuoliko tikslai, turimos žinios, galimybės ir apribojimai.
- „Baltosios dėžės“, „juodosios dėžės“ ir „pilkosios dėžės“ atakų scenarijai, priklausomai nuo to, kiek informacijos apie modelį turi užpuolikas.
- Tikslinės atakos (angl. *targeted attacks*), kai siekiama priversti modelį priimti konkretų klaidingą sprendimą, ir netikslinės atakos (angl. *untargeted attacks*), kai pakanka bet kokio klaidingo sprendimo.
- Atakos skirtinguose mašininio mokymosi proceso etapuose: duomenų rinkimo, modelio mokymo, validavimo, diegimo, stebėsenos ir modelio atnaujinimo.
- Išvengimo (angl. *evasion*) atakos, kai įvesties duomenys prognozavimo metu pakeičiami taip, kad suklaidintų modelį.
- Priešiški pavyzdžiai ir jų generavimas skirtingų tipų modeliams bei duomenims.
- Duomenų nuodijimo atakos (angl. *poisoning*), kai manipuliuojama mokymo duomenimis, siekiant paveikti apmokytą modelį.
- Paslėpto veikimo atakos (angl. *backdoor*) prieš mašininio ir giliojo mokymosi modelius, kai paslėpta modelio elgsena aktyvuojama tik esant tam tikroms sąlygoms.
- Su privatumu susijusios atakos.
- Atakos ir gynybos metodai taikant skirtingus duomenų tipus: vaizdus, tekstą, lentelinius duomenis, grafus, laiko eilutes, programinės įrangos ir kibernetinio saugumo duomenis.
- Gynybos metodai: priešiškas mokymas (angl. *adversarial training*), atsparus modelio optimizavimas, modelių ansambliai, išankstinis įvesties duomenų apdorojimas, anomalijų aptikimas ir neapibrėžtumo vertinimas.

- Modelio atsparumo vertinimas: atakos sėkmės rodiklis, atsparus tikslumas (angl. *robust accuracy*), užklausų skaičiaus apribojimas, klaidingai teigiamų ir klaidingai neigiamų sprendimų rodikliai, įvesties modifikavimo ribos ir modelio užklausų skaičius.
- Eksperimentų atkuriamumas ir patikimas eksperimentinio tyrimo planavimas priešiškojo mašininio mokymosi tyrimuose.

Kibernetinio saugumo taikymai:

- Mašininio mokymosi taikymas kenkėjiškos programinės įrangos aptikimui, įsibrovimų aptikimui, naudotojų autentifikavimui, elgsenos biometrijai ir anomalijų aptikimui.
- Specifiniai kibernetinio saugumo duomenų iššūkiai: klasių disbalansas, triukšmas, didelis dimensijų skaičius, duomenų pasiskirstymo kaita laike (angl. *concept drift*), priešiška elgsena ir besikeičiantys atakų modeliai.
- Statinė ir dinaminė kenkėjiškų programų analizė iš mašininio mokymosi perspektyvos.
- Požymių manipuliavimas, priešiškas obfuskavimas, gerybinių požymių įterpimas ir kenkėjiškų programų klasifikatorių apėjimas.
- Kibernetinio saugumo sistemose taikomų mašininio mokymosi modelių vertinimas priešiškomis sąlygomis.
- Saugus mašininio mokymosi modelių diegimas, stebėsena ir atnaujinimas kibernetinio saugumo taikymuose
- Atsakingas priešiško mašininio mokymosi metodų naudojimas kibernetinio saugumo tyrimuose.

Paaiškinamasis ir patikimas DI:

- Paaiškinamojo DI metodai modelio analizei, klaidų analizei ir sprendimų palaikymui.
- Lokalūs paaiškinimai, paaiškinantys atskiras prognozes, ir globalūs paaiškinimai, apibūdinantys bendrą modelio elgseną.
- Paaiškinamojo DI taikymas kibernetiniame saugume, kenkėjiškų programų aptikime, autentifikavime ir anomalijų aptikime.
- Interpretuojamumo metodai: LIME, SHAP ir reikšmingumo žemėlapiai (angl. *saliency maps*).
- Galimas manipuliavimas paaiškinimais ir klaidinančio interpretuojamumo rizika.
- Priešiškos atakos prieš paaiškinimus ir klaidinančio interpretuojamumo rizikos.

Praktinė užduotis: Parengti individualų tiriamąjį darbą, susijusį su doktoranto disertacijos tema. Darbas turėtų apimti mokslinės literatūros apžvalgą, grėsmių modelio ar tyrimo klausimo suformulavimą, atitinkamų priešiško mašininio mokymosi metodų analizę ir eksperimentinį, lyginamąjį arba koncepcinį tyrimą. Galutinis rezultatas gali būti parengtas kaip mokslinio tyrimo ataskaita.

Atsiskaitymas / Vertinimas: individualaus projekto pristatymas ir žodinis egzaminas.

Pagrindinė literatūra
Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2019). Adversarial Machine Learning. Cambridge University Press.
Vorobeychik, Y., & Kantarcioglu, M. (2018). Adversarial Machine Learning. Morgan & Claypool.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. http://www.deeplearningbook.org
Molnar, C. (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (3nd ed.). https://christophm.github.io/interpretable-ml-book/
Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.
Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion, 107.
MITRE. ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. https://atlas.mitre.org/
Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Konsultuojančiųjų dėstytojų vardas, pavardė	Mokslo laipsnis	Svarbiausieji darbai mokslo kryptyje (šakoje) paskelbti per pastaruosius 5 metus
Viktor Medvedev	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Viktor+Medvedev
Olga Kurasova	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Olga+Kurasova
Rokas Gipiškis	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Rokas+Gipiškis