Кафедра ИБ

Вопросы к государственному экзамену
Магистерская программа
«Искусственный интеллект в кибербезопасности» (гр. 619/2)

1. Задача выполнимости булевых формул (SAT), её сложность. Примеры сведения задач с помощью SMT решателя. Способы решения задачи - решение задачи методом грубой силы, Conflict-Driven Clause Learning алгоритм.
2. Задача выполнимости формулы в теориях (SMT). Примеры сведения задач с помощью SMT решателя. Davis–Putnam–Logemann–Loveland(Theory)-фреймворк. Примеры теорий: теория равенства, теория неинтерпретируемых функций. Их разрешаюшие процедуры.
3. Нетипизированное лямбда исчисление: использование аппликации и абстракции для построения термов, бета-редукция, теорема Чёрча-Россера. Y-комбинатор неподвижной точки. Разрешимость задач обитаемости типа, синтеза терма и проверки типа в нетипизированном лямбда исчислении.
4. Просто типизированное лямбда исчисление. Способы расширения просто типизированного лямбда исчисления в лямбда-кубе: термы, зависимые от термов и типов, типы, зависимые от типов и термов. Соответствие Карри-Ховарда между лямбда исчислениями, теорией множеств и логикой. Типы произведения и суммы, их выражение в логике, а также в языках программирования.
5. Интерпретация нейронных сетей, основные понятия. Классификация методов интерпретации. Методы построения карт значимости и карт активаций классов.
6. Интерпретация нейронных сетей, основные понятия. Методы интерпретации не зависящие от модели (агностические). Метод значений Шейпли. Метод LIME.
7. Интерпретируемые модели машинного обучения.
8. Методы интерпретации не зависящие от модели (агностические). PDP, ICE, ALE, feature importance.
9. Вероятностное пространство. Аксиоматика Колмогорова. Вероятностная мера. Функция распределения вероятностной меры. Понятие случайной величины. Геометрическое понимание случайной величины. Распределение вероятностей и функция распределения.
10. Функция распределения. Плотность распределения. Типы распределения: равномерное, биномиальное, Пуассона, нормальное, экспоненциальное.
11. Условная вероятность и независимость событий. Теорема Байеса. Условные распределения.
12. Математическое ожидание случайной величины. Дисперсия случайной величины. Ковариация случайных величин. Неравенство Коши-Буняковского. Матрица ковариаций вектора случайных величин. Дисперсия суммы независимых случайных величин.
13. Понятие выборки и генеральной совокупности. Доверительный интервал. Метод максимального правдоподобия. EM-алгоритмы.
14. Теория информации. Энтропия по Шеннону. Энтропия объединения. Условная энтропия. Математические свойства энтропии по Шеннону.
15. Нейронные сети. Модель нейрона. MLP. Понятие функции активации. Алгоритм обратного распространения ошибки.
16. Глубокие нейронные сети. Принцип работы слоев: сверточного, полносвязного, пулинг (max pooling, average pooling), нормализации (batch normalization, layer normalization), дропаут.
17. Сверточные нейронные сети. Семейства архитектур: LeNet, AlexNet, VGG, Inception, ResNet, DenseNet, EfficientNet.
18. Рекуррентные нейронные сети. Архитектуры RNN, GRU, LSTM. Затухание градиента, взрыв градиента. Градиентный клиппинг.
19. Механизм внимания. Self-Attention, Multi-head-attention. Маскированное внимание. Архитектура трансформер и использование механизма внимания в ней. Современные языковые модели: двунаправленные энкодеры (BERT), генеративные трансформеры (GPT).
20. Генеративные модели в компьютерном зрении (генеративно-состязательные нейронные сети). Принцип работы генератора и дискриминатора.
21. Нейросетевые модели для работы со звуком. Задача распознавания речи. Задача преобразования речи в текст. Модели Tacatron, Wave2Vec. CTC-loss.
22. Обучение с подкреплением. Основные элементы: среда, агент, функция награды, действия. Монте-Карло, Temporal difference. Проблема исследования и эксплуатации (exporation&expoitation). Алгоритм DQN.
23. Концепция атаки уклонением на нейросетевые модели. Существующие атаки уклонением в разных моделях угроз: белый ящик, черный ящик, атаки в реальном мире.
24. Методы защиты моделей от атак уклонением. Состязательное обучение. Сертификационные методы.

25. Концепция атаки отравлением данных на нейросетевые модели. Существующие атаки отравлением данных и методы защиты моделей от атак данного типа.
26. Концепция атак извлечением. Существующие атаки извлечением и методы защиты моделей от атак данного типа. Методы атак на основе запросов. Дифференциальные атаки.
27. Методы оценки устойчивости моделей машинного обучения к внешним воздействиям.
28. Мониторинг в системах с элементами ТИИ.
29. Публичные облачные провайдеры. Основные концепции и модели. Особенности обеспечения безопасности инфраструктуры, размещенной в облаке.
30. Концептуальная архитектура систем контейнеризации. Механизмы изоляции ядра Linux, необходимые для построения систем контейнеризации.
31. Концептуальная архитектура систем оркестрации контейнеров на примере Kubernetes. Принципы обеспечения безопасности Kubernetes-инфраструктуры.

Список рекомендованной литературы

[1] Daniel Kroening, Ofer Strichman. Decision Procedures An Algorithmic Point of View. Springer, 2016.
[2] Rob Nederpelt, Herman Geuvers “Type theory and formal proof”, Cambridge university press, 2014.
[3] Samuel Mimram “Program = proof”, Independently published, 2020.
[4] Benjamin C. Pierce “Types and programming languages”, the MIT press, 2002
[5] Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/
[6] Kamath U., Liu J. Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning. – Springer, 2021. – С. 1-310.
[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.
[8] Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. — СПб.: Питер, 2018. — 480 с.
[9] Б.В. Гнеденко. Курс теории вероятностей. Изд. 6-е, перераб. и доп. — М.: Наука. Гл. ред. физ.-мат. лит., 1988.
[10] В.Е. Гмурман. Теория вероятностей и математическая статистика. 9-е издание, стереотипное – М.: Высш. шк., 2003.
[11] Stephen Boyd, Lieven Vandenberghe. Introduction to Applied Linear Algebra. Cambridge University Press 2018.
[12] Garrett Thomas. Mathematics for Machine Learning. Berkeley, 2018.
[13] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
[14] 8. Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
[15] 9. Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." 2017 ieee symposium on security and privacy (sp). IEEE, 2017.
[16] 10. Brendel, Wieland, et al. "Accurate, reliable and fast robustness evaluation." Advances in neural information processing systems 32 (2019).
[17] 11. Hirano, Hokuto, and Kazuhiro Takemoto. "Simple iterative method for generating targeted universal adversarial perturbations." Algorithms 13.11 (2020): 268.
[18] 12. Brown, Tom B., et al. "Adversarial patch." arXiv preprint arXiv:1712.09665 (2017).
[19] 13. Papernot, Nicolas, Patrick McDaniel, and Ian Goodfellow. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." arXiv preprint arXiv:1605.07277 (2016).
[20] 14. Papernot, Nicolas, et al. "The limitations of deep learning in adversarial settings." 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016.
[21] 15. Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[22] 16. Jang, Uyeong, Xi Wu, and Somesh Jha. "Objective metrics and gradient descent algorithms for adversarial examples in machine learning." Proceedings of the 33rd Annual Computer Security Applications Conference. 2017.

[23] 17. Chen, Shang-Tse, et al. "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2018.
[24] 18. Chen, Pin-Yu, et al. "Ead: elastic-net attacks to deep neural networks via adversarial examples." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
[25] 19. Chen, Jianbo, Michael I. Jordan, and Martin J. Wainwright. "Hopskipjumpattack: A query-efficient decision-based attack." 2020 ieee symposium on security and privacy (sp). IEEE, 2020.
[26] 20. Kotyan, Shashank, and Danilo Vasconcellos Vargas. "Adversarial Robustness Assessment: Why both and Attacks Are Necessary." arXiv preprint arXiv:1906.06026 (2019).
[27] 21. Guo, Chuan, et al. "Simple black-box adversarial attacks." International Conference on Machine Learning. PMLR, 2019.
[28] 22. Engstrom, Logan, et al. "Exploring the landscape of spatial robustness." International Conference on Machine Learning. PMLR, 2019.
[29] 23. Chen, Pin-Yu, et al. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models." Proceedings of the 10th ACM workshop on artificial intelligence and security. 2017.
[30] 24. Brendel, Wieland, Jonas Rauber, and Matthias Bethge. "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models." arXiv preprint arXiv:1712.04248 (2017).
[31] 25. Rahmati, Ali, et al. "Geoda: a geometric framework for black-box adversarial attacks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[32] 26. Carlini, Nicholas, et al. "Extracting training data from large language models." 30th USENIX Security Symposium (USENIX Security 21). 2021.
[33] 27. Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. 2015.
[34] 28. Shokri, Reza. "Bypassing backdoor detection algorithms in deep learning." 2020 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2020.
[35] 29. Shafahi, Ali, et al. "Poison frogs! targeted clean-label poisoning attacks on neural networks." Advances in neural information processing systems 31 (2018).
[36] 30. Turner, Alexander, Dimitris Tsipras, and Aleksander Madry. "Clean-label backdoor attacks." (2018).
[37] 31. Biggio, Battista, Blaine Nelson, and Pavel Laskov. "Poisoning attacks against support vector machines." arXiv preprint arXiv:1206.6389 (2012).
[38] 32. Aghakhani, Hojjat, et al. "Bullseye polytope: A scalable clean-label poisoning attack with improved transferability." 2021 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2021.
[39] 33. Tramèr, Florian, et al. "Stealing Machine Learning Models via Prediction {APIs}." 25th USENIX security symposium (USENIX Security 16). 2016.
[40] 34. Wang, Binghui, and Neil Zhenqiang Gong. "Stealing hyperparameters in machine learning." 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 2018.
[41] 35. Orekondy, Tribhuvanesh, Bernt Schiele, and Mario Fritz. "Knockoff nets: Stealing functionality of black-box models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[42] 36. Marko Luksa, Kubernetes in Action, Manning, 624 pages, 2018
[43] 37. Liz Rice, Container Security, O'Reilly Media, 2020
[44] 38. Eyal Estrin, CloudSecurity Handbook, Packt Publishing, 2022
[45] 39. Chris Dotson, Practical Cloud Security, O'Reilly Media, 2019