Федеральное государственное бюджетное образовательное учреждение высшего образования
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М.В.ЛОМОНОСОВА»

ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ

УТВЕРЖДАЮ

Декан факультета ВМК МГУ
Академик РАН

/*И.А. Соколов*/

«20» июня 2022 г.

ПРОГРАММА-МИНИМУМ

кандидатского экзамена по специальности

1.2.1. Искусственный интеллект и машинное обучение

Область науки: 1. Естественные науки

Группа научных специальностей: 1.1. Математика и механика

Отрасль науки: физико-математические науки

Москва 2022

I. Описание программы

Настоящая программа охватывает основополагающие разделы и области знания, в основе данной программы лежат следующие дисциплины: теория вероятностей, математическая статистика, дискретная математика, технологии программирования и машинное обучение.

II. Основные разделы и вопросы к экзамену

1. Математические основы

2. Понятие выборки и генеральной совокупности. Доверительный интервал. Метод максимального правдоподобия. EM-алгоритмы.

3. Статистические критерии проверки гипотез. Критерии значимости. Критерии согласия. Параметрические критерии. t-критерий Стьюдента. Непараметрические критерии.

4. Матричные разложения. Сингулярное разложение.

5. Теория графов. Ориентированные и неориентированные графы и операции над ними. Связь с бинарными отношениями. Критерий связности графов. Неориентированные деревья и их свойства. Теорема Эйлера. Теорема Дирака.

6. Понятие алгоритма и его уточнения: машины Тьюринга, нормальные алгоритмы Маркова, рекурсивные функции. Эквивалентность данных формальных моделей алгоритмов. Понятие об алгоритмической неразрешимости. Примеры алгоритмически неразрешимых проблем.

7. Понятие сложности алгоритмов. Классы P и NP. Полиномиальная сводимость задач. Теорема Кука об NP-полноте задачи выполнимости булевой формулы. Примеры NP-полных задач, подходы к их решению. Точные и приближённые комбинаторные алгоритмы.

8. Примеры эффективных (полиномиальных) алгоритмов: быстрые алгоритмы поиска и сортировки; полиномиальные алгоритмы для задач на графах и сетях (поиск в глубину и ширину, о минимальном остове, о кратчайшем пути, о назначениях).

9. Автоматы. Эксперименты с автоматами. Алгебры регулярных выражений. Теорема Клини о регулярных языках.

10. Алгебра логики. Булевы функции, канонические формы задания булевых функций. Понятие полной системы. Критерий полноты Поста. Минимизация булевых функций в классах нормальных форм.

11. Исчисление предикатов первого порядка. Понятие интерпретации. Выполнимость и общезначимость формулы первого порядка. Понятие модели. Теорема о полноте исчисления предикатов первого порядка.

12. Логические агенты. Представление знаний.

13. Отношения и функции. Отношение эквивалентности и разбиения. Фактор множество. Отношения частичного порядка. Теоретико-множественное и алгебраическое определения решётки, их эквивалентность. Свойства решёток. Булевы решётки. Полные решётки.

14. λ-исчисление, правила редукции, единственность нормальной формы и правила ее достижения, представление рекурсивных функций.

15. Основы комбинаторного анализа. Метод производящих функций, метод включений и исключений. Примеры применения.

16. Алфавитное кодирование. Коды с исправлением ошибок.

17. Основы криптографии. Задачи обеспечения конфиденциальности и целостности информации. Теоретико-информационный и теоретико-сложностной подходы к определению криптографической стойкости. Американский стандарт шифрования AES и стандарт шифрования данных ГОСТ 34.12-2018. Системы шифрования с открытым ключом (RSA и Эль-Гамаля). Цифровая подпись. Методы генерации и распределения ключей.

2. Языки и системы программирования

1. Распределенное программирование. Процессы и их синхронизация. Семафоры, мониторы Хоара. Объектно-ориентированное распределенное программирование. CORBA. Параллельное программирование над общей памятью. Нити. Стандартный интерфейс OpenMP. Распараллеливание последовательных программ. Параллельное программирование над распределенной памятью. Парадигмы SPMD и MIMD. Стандартный интерфейс MPI.

2. Технология разработки и сопровождения программ. Жизненный цикл программы. Этапы разработки, степень и пути их автоматизации. Обратная инженерия. Декомпозиционные и сборочные технологии, механизмы наследования, инкапсуляции, задания типов. Модули, взаимодействие между модулями, иерархические структуры программ.

3. Отладка, тестирование, верификация и оценивание сложности программ. Генерация тестов. Системы генерации тестов. Срезы программ (slice, chop) и их применение при отладке программ и для генерации тестов.

4. Методы спецификации программ. Методы проверки спецификации. Схемное, структурное, визуальное программирование. Разработка пользовательского интерфейса, стандарт CUA, мультимедийные среды интерфейсного взаимодействия.

3. Искусственный интеллект и машинное обучение

1. Типы задач: классификация, регрессия, прогнозирование, ранжирование, обнаружение аномалий. Методы обучения: обучение с учителем, обучение без учителя, обучение с подкреплением.

2. Функции потерь для задач машинного обучения. Градиентный спуск.

3. Функционалы качества для задач классификации Accuracy, Precision, Recall, F1, ROC AUC. Функционалы качества для задач регрессии: MSE, RMSE, MAE, MAPE.

4. Линейные модели. Логистическая регрессия. Концепция переобучения и недообучения. Методы валидации качества алгоритма. Регуляризация. L1/L2 регуляризация, множитель Лагранжа.

5. Решающее дерево, бинаризация признаков, алгоритм построения. Решающие деревья, случайный лес, градиентный бустинг. Методы их обучения, критерий информативности, критерий остановки. Ансамблирование моделей: мажорантное голосование, блендинг, бустинг, бэгинг.

6. Метрики оценки качества для задач классификации. Метрики оценки качества для задач регрессии

7. Метод опорных векторов (SVM). Разделяющая гиперплоскость. Функция ядра. Трюк с ядром (KernelTrick)

8. Байесовский подход в машинном обучении. Наивный байесовский классификатор.

9. Кластеризация. Алгоритмы кластеризации. Метод k-средних. DBSCAN. Иерархическая кластеризация.

10. Методы снижения размерности. Метод главных компонент. SNE, t-SNE. UMAP.

11. Типы признаков и их обработка. Нормализация данных, масштабирование, обработка категориальных признаков. Векторизация текстовых данных с помощью bag-of-words, tf-idf.

12. Методы оптимизации. Градиентный спуск, SGD, AdaGrad, Adam, RMSProp, момент Нестерова.

13. Методы работы с временными рядами. Модель ARIMA.

14. Статистические методы интерпретации моделей машинного обучения. SHAP, LIME.

4. Нейронные сети

1. Нейронные сети. Модель нейрона. MLP. Понятие функции активации. Алгоритм обратного распространения ошибки.

2. Глубокие нейронные сети. Принцип работы слоев: сверточного, полносвязного, пулинг (maxpooling, averagepooling), нормализации (batchnormalization, layernormalization), дропаут

3. Сверточные нейронные сети. Семейства архитектур: LeNet, AlexNet, VGG, Inception, ResNet, DenseNet, EfficientNet.

4. Способы искусственного расширения набора данных для увеличения обобщающей способности. Аугментация, генерация, симуляция. Перенос обучения (Transfer learning).

5. Рекуррентные нейронные сети. Архитектуры RNN, GRU, LSTM. Затухание градиента, взрыв градиента. Градиентный клиппинг.

6. 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).

7. Контрастивное обучение. Примеры из компьютерного зрения и языковых задач. Сиамские сети. Функции потерь: constrastiveloss, tripletloss.

8. Генеративные модели в компьютерном зрении (генеративно-состязательные нейронные сети). Принцип работы генератора и дискриминатора.

9. Autoencoder. Variational autoencoder. Примеры прикладных задач. Концепция сжатия информации. KL-дивергенция. Трюк с репараметризацией.

10. Задачи компьютерного зрения. Задачи детекции. Описание принципов работы R-CNN, YOLO. Задача сегментации. Семантическая сегментация. Паноптическая сегментация. Описание принципа работы U-net. Задача распознавания лиц. Подходы для метрического обучения.

11. Трансформеры в компьютерном зрении. Описание принципа работы архитектуры ViT и его разновидностей.

12. Нейросетевые модели для работы со звуком. Задача распознавания речи. Задача преобразования речи в текст. Модели Tacatron, Wave2Vec. CTC-loss.

13. Обучение с подкреплением. Основные элементы: среда, агент, функция награды, действия. Монте-Карло, Temporal difference. Проблема исследования и эксплуатации (exporation&expoitation). Алгоритм DQN.

14. Вопросы практической реализации нейронных сетей в условиях ограничения вычислительных ресурсов. Дистилляция. Прунинг. Квантизация.

15. Методы интерпретации нейронных сетей. Градиентные методы: GradCAM, Integratedgradients, Noise Tunnel. Методы на основе механизма внимания: матрица внимания, CLEAR, SCOUTER. ИНСна основе графов.

5. Робастность и анализ методов глубокого обучения

1. Концепция атаки уклонением на нейросетевые модели. Существующие атаки уклонением и методы защиты моделей от атак данного типа. Принципы работы FGSM, PGD, семейство атак Карлинии Вагнера, атакаBrendel&Bethge, Universal Adversarial Perturbations, Adversarial Patch, Decision Tree Attack, Jacobian Saliency Map, DeepFool, NewtonFool, ShapeShifter, Elastic Net, HopSkipJump Attack, Threshold Attack, Pixel Attack, SimBA, Spatial Transformation, ZOO

(Zeroth Order Optimization), Decision-based/Boundary Attack, Geometric Decision-based Attack (GeoDA).

2. Концепция атак извлечением данных. Существующие атаки извлечением данных на модели и методы защиты моделей от атак данного типа. Подходы к извлечению текстовых данных из лингвистических моделей. Подходы к извлечению данных из моделей, работающих с изображениями.

3. Концепция атаки отравлением данных на нейросетевые модели. Существующие атаки отравлением данных и методы защиты моделей от атак данного типа. Принципы работы Adversarial Backdoor Embedding, Clean Label Feature Collision Attack, Clean-Label Backdoor Attack, Poisoning Attack on Support Vector Machines, Bullseye Polytope.

4. Концепция инверсионных атак. Существующие инверсионные атаки и методы защиты моделей от атак данного типа. Методы атак на основе запросов. Дифференциальные атаки. Атаки по побочным каналам.

5. Методы формальной верификации моделей машинного обучения. Верификация на основе ограничений. Абстрактная верификация.

6. Методы оценки устойчивости моделей машинного обучения к внешним воздействиям.

III. Основная литература

1. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org.
2. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. — СПб.: Питер, 2018. — 480 с.
3. Б.В. Гнеденко. Курс теории вероятностей. Изд. 6-е, перераб. и доп. — М.: Наука. Гл. ред. физ.-мат. лит., 1988.
4. В.Е. Гмурман. Теория вероятностей и математическая статистика. 9-е издание, стереотипное – М.:Высш. шк., 2003.
5. Stephen Boyd, Lieven Vandenberghe. Introduction to Applied Linear Algebra. Cambridge University Press 2018.
6. Aws Albarghouthi. Introduction to Neural Network Verification https://arxiv.org/abs/2109.10317
7. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
8. Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).

9. Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." 2017 ieee symposium on security and privacy (sp). IEEE, 2017.
10. Brendel, Wieland, et al. "Accurate, reliable and fast robustness evaluation." Advances in neural information processing systems 32 (2019).
11. Hirano, Hokuto, and Kazuhiro Takemoto. "Simple iterative method for generating targeted universal adversarial perturbations." Algorithms 13.11 (2020): 268.
12. Brown, Tom B., et al. "Adversarial patch." arXiv preprint arXiv:1712.09665 (2017).
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).
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.
15. Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deep fool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
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.
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.
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.
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.
20. Kotyan, Shashank, and Danilo Vasconcellos Vargas. "Adversarial Robustness Assessment: Why both L_0 and L_∞ Attacks Are Necessary." arXiv preprint arXiv:1906.06026 (2019).
21. Guo, Chuan, et al. "Simple black-box adversarial attacks." International Conference on Machine Learning. PMLR, 2019.
22. Engstrom, Logan, et al. "Exploring the landscape of spatial robustness." International Conference on Machine Learning. PMLR, 2019.
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.

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).
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.
26. Carlini, Nicholas, et al. "Extracting training data from large language models." 30th USENIX Security Symposium (USENIX Security 21). 2021.
27. Fredrikson, Matt, SomeshJha, 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.
28. Shokri, Reza. "Bypassing backdoor detection algorithms in deep learning." 2020 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2020.
29. Shafahi, Ali, et al. "Poison frogs! targeted clean-label poisoning attacks on neural networks." Advances in neural information processing systems 31 (2018).
30. Turner, Alexander, Dimitris Tsipras, and Aleksander Madry. "Clean-label backdoor attacks." (2018).
31. Biggio, Battista, Blaine Nelson, and Pavel Laskov. "Poisoning attacks against support vector machines." arXiv preprint arXiv:1206.6389 (2012).
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.
33. Tramèr, Florian, et al. "Stealing Machine Learning Models via Prediction {APIs}." 25th USENIX security symposium (USENIX Security 16). 2016.
34. Wang, Binghui, and Neil Zhenqiang Gong. "Stealing hyperparameters in machine learning." 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 2018.
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.
36. Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." Proceedings of the 2017 ACM on Asia conference on computer and communications security. 2017.
37. Oh, SeongJoon, Bernt Schiele, and Mario Fritz. "Towards reverse-engineering black-box neural networks." Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer, Cham, 2019. 121-144.
38. Krishna, Kalpesh, et al. "Thieves on sesame street! model extraction of bert-based apis." arXiv preprint arXiv:1910.12366 (2019).

39. He, Xuanli, et al. "Model extraction and adversarial transferability, your bert is vulnerable!." arXiv preprint arXiv:2103.10013 (2021).
40. Sha, Zeyang, et al. "Can't Steal? Cont-Steal! Contrastive Stealing Attacks Against Image Encoders." arXiv preprint arXiv:2201.07513 (2022).
41. Szyller, Sebastian, et al. "Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks." arXiv preprint arXiv:2104.12623 (2021).
42. Chen, Kangjie, et al. "Stealing deep reinforcement learning models for fun and profit." Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. 2021.
43. Wallace, Eric, Mitchell Stern, and Dawn Song. "Imitation attacks and defenses for black-box machine translation systems." arXiv preprint arXiv:2004.15015 (2020).
44. Milli, Smitha, et al. "Model reconstruction from model explanations." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.
45. Jagielski, Matthew, et al. "High accuracy and high fidelity extraction of neural networks." 29th USENIX Security Symposium (USENIX Security 20). 2020.
46. Carlini, Nicholas, Matthew Jagielski, and Ilya Mironov. "Cryptanalytic extraction of neural network models." Annual International Cryptology Conference. Springer, Cham, 2020.
47. W. Hua, Z. Zhang and G. E. Suh, "Reverse Engineering Convolutional Neural Networks Through Side-channel Information Leaks," 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC), 2018, pp. 1-6, doi: 10.1109/DAC.2018.8465773.
48. Duddu, Vasisht, et al. "Stealing neural networks via timing side channels." arXiv preprint arXiv:1812.11720 (2018).
49. Ахо, Сети Р., Ульман Дж. Компиляторы: Принципы, техника реализации и инструменты. М. 2016.
50. Введение в криптографию. Под ред. В.В. Ященко. Санкт-Петербург: МЦНМО. 2012.
51. Кнут Д. Искусство программирования, т. 1 – 3. ИД «Вильямс» М., СПб.,2017.
52. Харари Ф. Теория графов. М. 1973.

IV. Дополнительная литература

1. Garrett Thomas. Mathematics for Machine Learning. Berkeley, 2018.
2. Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. NewYork: springer, 2009.
3. Е.Е. Тыртышников. Матричный анализ и линейная алгебра. Москва, 2005.

4. Davidson-Pilon, Cameron. Bayesian methods for hackers: probabilistic programming and Bayesian inference. Addison-Wesley Professional, 2015.
5. Cormen, Thomas H., et al. Introduction to algorithms. MIT press, 2009.

V. Автор программы

— к.ф.-м.н. Чижов И.В.

VI. Критерии оценивания

Критерии и показатели оценивания ответа на экзамене			
2	3	4	5
Неудовлетворительно	**Удовлетворительно**	**Хорошо**	**Отлично**
Фрагментарные знания актуальных проблем и тенденций в развитии компьютерных наук и прикладной математики	Неполные знания актуальных проблем и тенденций в развитии компьютерных наук и прикладной математики	Сформированные, но содержащие отдельные пробелы знания актуальных проблем и тенденций в развитии компьютерных наук и прикладной математики	Сформированные и систематические знания актуальных проблем и тенденций в развитии компьютерных наук и прикладной математики