Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

УТВЕРЖДЕНО:
Директор
А. В. Замятин

Рабочая программа дисциплины

Прикладные аспекты машинного обучения

по направлению подготовки

09.03.03 Прикладная информатика

Направленность (профиль) подготовки:
Искусственный интеллект и большие данные

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
С.П.Сущенко

Председатель УМК
С.П.Сущенко

Томск – 2024

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

ПК-6 Способен разрабатывать и применять методы машинного обучения для решения задач.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-6.1 Проводит анализ требований и определяет необходимые классы задач машинного обучения

ИПК-6.2 Принимает участие в оценке и выборе используемых методов машинного обучения

2. Задачи освоения дисциплины

– Освоить аппарат построения интеллектуальных систем на базе машинного обучения.

– Научиться применять понятийный аппарат интеллектуальных систем с использованием инструментария библиотек Python, R, публичных облачных сервисов, оценивать эффективность их работы и внедрять в приложения для решения практических задач профессиональной деятельности.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к Блоку Факультативных дисциплин.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Седьмой семестр, экзамен

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: «Статистические методы машинного обучения», «Введение в интеллектуальный анализ данных».

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 3 з.е., 180 часов, из которых:
-лекции: 20 ч.
-практические занятия: 40 ч.
Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Разведочный анализ данных. Предварительная обработка данных

Анализ признакового пространства. Описательные статистики. Обработка категориальных данных. Обработка пропущенных значений. Поиск аномалий в данных. Преобразование признаков к единой шкале. Балансировка классов.

Тема 2. Алгоритмы построения классификаторов

Постановка задачи классификации. Метрики оценки качества моделей классификации. Алгоритм К-ближайших соседей. Линейные модели. Наивный Байес. Логистическая регрессия. Метод опорных векторов. Ядерный трюк. Деревья решений.

Тема 3. Алгоритмы построения регрессоров

Постановка задачи регрессии. Метрики оценки качества моделей регрессии. Линейная регрессия. Полиномиальная регрессия. Деревья решений для регрессии. Регрессоры на основе К-ближайших соседей. Метод опорных векторов для регрессии.

Тема 4. Ансамбли моделей и поиск лучшей модели

К-блочная перекрестная проверка. Случайный лес. Бэггинг. Голосование моделей. Бустинг. Стекинг. Поиск гиперпараметров. Автоматизация поиска моделей.

Тема 5. Работа с признаковым пространством. Manifold Learning

Снижение размерности. Backward Elimination. Forward Selection. Метод главных компонент. Факторный анализ. Линейный дискриминантный анализ. Применение алгоритма SVM для преобразования признакового пространства. SVD. Усеченное сингулярное разложение. Многомерное масштабирование. Изометрическое картографирование (Isomap). Локальное линейное владение (LLE). t-распределенное стохастическое вложение соседей (t-SNE). Разномерное приближение и проекция многообразия (UMAP).

Тема 6. Кластеризация данных

Задача кластеризации. Графы и гиперграфы. Жёсткая и мягкая кластеризация. К-средних. Метод локтя. Анализ значений Силуэта и Силуэтных графиков. Статистический подход к кластеризации. EM алгоритм. Определение числа кластеров. Кластеризация, основанная на плотности (DBSCAN). OPTICS. Кластеризация категориальных данных. ROCK. CACTUS. Иерархическая кластеризация. Дивизивные и агломеративные методы. Спектральная кластеризация. Корреляционная и консенсусная кластеризация. Внутренние и внешние метрики оценка качества кластеризации. Относительные методы оценки качества кластеризации.

Тема 7. Анализ временных последовательностей

Временные последовательности. Преобразование временных серий. ARMA модели для стационарных временных последовательностей. ARMA модели для нестационарных временных последовательностей. Определение трендов. Прогноз с помощью ARIMA. Модели ненаблюдаемых компонентов, извлечение сигналов и фильтры. Сезонность и экспоненциальное сглаживание. Волатильность и обобщенные авторегрессионные условные гетероскедастические процессы. Нелинейные случайные процессы. Передаточные функции и авторегрессионное распределенное моделирование задержек. Векторные авторегрессии и причинность по Грейнджеру. Векторные авторегрессии с интегрированными переменными, модели коррекции векторных ошибок и общие тенденции. Композиционный и счетный временной ряд.

Тема 8. Объяснимый искусственный интеллект

Глобальная и локальная интерпретируемость. Графики частичной зависимости. Индивидуальное условное ожидание. График накопленных локальных эффектов (ALE). Важность признаков для моделей, основанных на деревьях. Аддитивные объяснения Шепли (SHAP). Локальные интерпретируемые объяснения независимые от модели (LIME). Профиль Ceteris-paribus. Осцилляции Ceteris-paribus. Локальные диагностические диаграммы. Использование якорей. Понимание о семантической схожести.

Контрфактуальные объяснения. Контрастные объяснения. Количественное тестирование с векторами активации концепций (TCAV).

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится путем контроля посещаемости, проведения контрольных работ, тестов по лекционному материалу, выполнения домашних заданий и фиксируется в форме контрольной точки не менее одного раза в семестр.

Оценочные материалы текущего контроля размещены на сайте ТГУ в разделе «Информация об образовательной программе» - https://www.tsu.ru/sveden/education/eduop/.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Экзамен в седьмом семестре проводится в письменной форме по билетам. Экзаменационный билет состоит из трех частей. Продолжительность экзамена 1,5 часа.

Оценочные материалы для проведения промежуточной аттестации размещены на сайте ТГУ в разделе «Информация об образовательной программе» - https://www.tsu.ru/sveden/education/eduop/.

11. Учебно-методическое обеспечение

а) Электронный учебный курс по дисциплине в электронном университете «LMS IDO»

б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.

12. Перечень учебной литературы и ресурсов сети Интернет

а) основная литература:

– Eklas Hossain. Machine Learning Crash Course for Engineers. Springer. – 2024. ISBN 978-3-031-46989-3. https://doi.org/10.1007/978-3-031-46990-9. – 453 p.

– Gerald Friedland. Information-Driven Machine Learning. Data Science as an Engineering Springer. – 2024. ISBN 978-3-031-39476-8. https://doi.org/10.1007/978-3-031-39477-5. – 267 p.

– Tong Zhang. Mathematical Analysis of Machine Learning Algorithms. Manning Publications Co. – 2023. ISBN: 9781633439023. DOI: 10.1017/9781009093057. – 453 p.

– Дайзенрот Марк Питер, Альдо Фейзал А., Чен Сунь Он. Математика в машинном обучении. — СПб.: Питер, 2024. — 512 с.: ил. — (Серия «Для профессионалов»). ISBN 978-5-4461-1788-8

– Лакшманан, В. Машинное обучение. Патгерны проектирования: Пер. с англ./ В. Лакшманан, С. Робинсон, М. Мунн. – СПб.: БХВ-Петербург, 2022. – 448 с.: ил. ISBN 978-5-9775-6797-8.

– Амейзен Эммануэль. Создание приложений машинного обучения: от идеи к продукту. — СПб.: Питер, 2023. — 256 с.: ил. — (Серия «Библиотека программиста»). ISBN 978-5-4461-1773-4

б) дополнительная литература:

– Simon Thompson. Managing Machine Learning Projects. From Design to Deployment. Douglas J. Santry. Demystifying Deep Learning. An Introduction to the Mathematics of Neural Networks. The Institute of Electrical and Electronics Engineers, Inc. IEEE Press Wiley – 2024. Hardback ISBN: 9781394205608 – 247 p.

– Paul Fieguth. An Introduction to Pattern Recognition and Machine Learning. Springer. – 2022. ISBN 978-3-030-95993-7. https://doi.org/10.1007/978-3-030-95995-1. – 471 p.

– Michael Munn, David Pitman. Explainable AI for Practitioners. Designing and Implementing Explainable ML Solutions. O'Reilly Media, Inc. – 2023. ISBN 978-1-098-11913-3. – 259 p.

– S. Sumathi, Suresh V. Rajappa, L. Ashok Kumar, Surekha Paneerselvam Machine Learning for Decision Sciences with Case Studies in Python CRC Press. – 2022. ISBN: 978-1-032-19356-4. DOI: 10.1201/9781003258803. – 454 p.

– T.V. Geetha, S. Sendhilkumar. Machine Learning. Concepts, Techniques and Applications. CRC Press. – 2023. ISBN: 978-1-032-26828-6. – 455 p.

– Бурков А. Инженерия машинного обучения/ пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2022. – 306 с.: ил. ISBN 978-5-93700-125-2.

в) ресурсы сети Интернет:

– The AI community building the future. The platform where the machine learning community collaborates on models, datasets, and applications. – https://huggingface.co/

– OpenAI. – https://openai.com/

– Tensorflow. An end-to-end platform for machine learning. – https://www.tensorflow.org/

– PyTorch documentation. – https://pytorch.org/

– IBM. What is deep learning? – https://www.ibm.com/topics/deep-learning

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

– Microsoft Office Standart 2013 Russian: пакет программ. Включает приложения: MS Office Word, MS Office Excel, MS Office PowerPoint, MS Office On-eNote, MS Office Publisher, MS Outlook, MS Office Web Apps (Word Excel MS PowerPoint Outlook);

– публично доступные облачные технологии (Google Docs, Google Colab, Яндекс диск).

- Пакет Anaconda
- Средства языков программирования и анализа данных R и Python
- Библиотеки для машинного и глубокого обучения: Scikit-learn, NumPy, Matplotlib.pyplot, Seaborn, PyTorch, Keras/TensorFlow, OpenAI Gym.

б) информационные справочные системы:

– Электронный каталог Научной библиотеки ТГУ – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system

– Электронная библиотека (репозиторий) ТГУ – http://vital.lib.tsu.ru/vital/access/manager/Index

– ЭБС Лань – http://e.lanbook.com/

– ЭБС Консультант студента – http://www.studentlibrary.ru/

– Образовательная платформа Юрайт – https://urait.ru/

– ЭБС ZNANIUM.com – https://znanium.com/

– ЭБС IPRbooks – http://www.iprbookshop.ru/

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

Аудитории для проведения занятий лекционного и семинарского типа индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации в смешенном формате («Актру»).

15. Информация о разработчиках

Аксёнов Сергей Владимирович, к.т.н., кафедра теоретических основ информатики (ТОИ) Института прикладной математики и компьютерных наук (ИПМКН) Национальный исследовательский Томский государственный университет (НИ ТГУ), доцент каф. ТОИ