Федеральное государственное бюджетное образовательное

учреждение высшего образования

Московский государственный университет имени М.В. Ломоносова

|  |
| --- |
| **УТВЕРЖДАЮ**  **Ректор МГУ имени М.В. Ломоносова**  **В.А. Садовничий**  **«\_\_» \_\_\_\_\_\_\_\_ 2023 г.** |

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)**

**Наименование дисциплины (модуля):**

**Межфакультетский курс**

**«Машинное обучение для решения прикладных задач с помощью Python»**

*наименование дисциплины (модуля)*

**Уровень высшего образования:**

***Бакалавриат, магистратура, специалитет***

*бакалавриат, магистратура, специалитет*

**Направление подготовки / специальность:**

**По всем направлениям подготовки / специальностям**

*(код и название направления/специальности)*

**Направленность (профиль)/специализация ОПОП:**

**По всем направленностям (профилям)/специальностям ОПОП**

*(если дисциплина реализуется в рамках направленности (профиля))*

Форма обучения:

Очная

*очная, очно-заочная*

Рабочая программа рассмотрена и одобрена

*Методической комиссией МГУ, на заседании Ученого совета МГУ*

\_\_.\_\_.2023

Москва 2023

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленными МГУ образовательными стандартами (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования (*программы бакалавриата, магистратуры, реализуемых последовательно по схеме интегрированной подготовки; программы специалитета; программы магистратуры)*

**1.** Место дисциплины (модуля) в структуре ОПОП ВО:

*Межфакультетские учебные курсы относятся к вариативной части и являются дисциплиной по выбору (элективной).*

**2.** Цели и задачи дисциплины (модуля):

Межфакультетский курс посвящен одной из самых передовых областей науки на сегодняшний день - машинному обучению. Слушатели курса познакомятся с классическими методами машинного обучения для решения прикладных задач.

Цель дисциплины — сформировать устойчивые представления о теоретических основах машинного обучения и обучить студентов практическим навыкам работы с алгоритмами искусственного интеллекта в применении к любой области знания. Предполагается, что по окончании курса, студент сможет воспользоваться аппаратом машинного обучения для решения научных и практических задач из любой области.

Задачи дисциплины:

— изучение теоретических основ машинного обучения и алгоритмов;

— изучение базового функционала библиотек Sklearn и XGBoost;

— изучение примеров практического применения алгоритмов машинного обучения в мультидисциплинарном формате, предполагающем демонстрацию применения таких алгоритмов на практических примерах из разных областей науки.

**3.** Входные требования для освоения дисциплины (модуля), предварительные условия:

Изучение данной дисциплины базируется на следующих курсах:

— МФК “Основы программирования и анализа данных на Python” (базовые навыки по программированию на Python упростят освоение программы дисциплины);

— факультатив "Математика для анализа данных" (знание основ линейной алгебры, оптимизации, статистики упростит освоение программы дисциплины).

**4.** Результаты обучения по дисциплине (модулю):

|  |
| --- |
| **Планируемые результаты обучения по дисциплине (модулю)** |
| В результате освоения дисциплины обучающиеся должны:  **знать:**  - теоретические основы и терминологию машинного обучения;  - набор классических алгоритмов, а также области их применения;  - достоинства и недостатки классических алгоритмов машинного обучения;  - возможности применения классических алгоритмов машинного обучения;  - особенности практической реализации алгоритмов машинного обучения;  - особенности практического применения готовых алгоритмов машинного обучения;  - требования к данным в контексте применения каждого из изученных алгоритмов машинного обучения;  - возможности стандартных библиотек по подготовке данных для различных задач машинного обучения;  - как применять алгоритмы машинного обучения для решения практических задач.  **уметь:**  - вызывать классические алгоритмы машинного обучения из стандартных библиотек;  - классифицировать произвольную поставленную задачу с точки зрения терминологии машинного обучения;  - выбирать актуальный алгоритм в зависимости от постановки и контекста поставленной задачи;  - анализировать результат работы алгоритма машинного обучения с учетом контекста поставленной задачи;  - пользоваться документацией стандартных библиотек;  - обрабатывать сырые данные для их анализа при помощи алгоритма машинного обучения;  - выбирать оптимальные алгоритмы машинного обучения в зависимости от ситуации;  - визуализировать и интерпретировать результат работы алгоритма машинного обучения;  **владеть:**  - библиотеками Sklearn и XGBoost;  - навыками работы в классических средах программирования: Jupyter Notebook/Spyder/Google Colab (с учётом индивидуальных предпочтений студента);  - навыками комплексного анализа данных при помощи машинного обучения |

**5.** Объем дисциплины составляет 1 з.е. (36 часов, включая 24 часа на занятия лекционного типа и 12 часов на самостоятельную работу обучающихся).

|  |  |
| --- | --- |
| **Вид учебной работы** | **Всего часов** |
| **Аудиторные занятия (всего)** | **24** |
| В том числе: |  |
| Лекции (Л) | 24 |
| Практические занятия (ПЗ) | - |
| Семинары (С) | - |
| Лабораторные работы (ЛР) | - |
| **Самостоятельная работа (всего)** | **12** |
| В том числе: |  |
| Подготовка к семинарским занятиям | - |
| Подготовка домашних заданий | 10 |
| Решение профессиональных задач и т.д. | - |
| Подготовка к аттестации | 2 |
| Вид промежуточной аттестации (зачет, экзамен) | зачет |
| Общая трудоемкость: часы  зачетные единицы | **36** |
| **1** |

**6.** Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий

**Содержательный план дисциплины**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Наименование раздела дисциплины**  **(элемента модуля)** | **Лекции** | **Практические**  **занятия** | **Семинары** | **Самостоятельная работа** | **Всего** |
| 1. | **Тема 1. Введение в ML**  — Постановка задачи ML  — Классификация задач ML  — Классификация алгоритмов ML  — Примеры применения алгоритмов ML в нашей жизни  — Математическая постановка задачи ML  — Модельные примеры различий между разными областями ML | **2** |  |  | **1** | **3** |
| 2. | **Тема 2. Метрические алгоритмы**  — Понятие метрики  — Примеры различных метрик  — Общая идея метрических алгоритмов  — Применение алгоритма ближайшего соседа для решения задачи классификации  — Алгоритм KNN для решения задачи классификации  — Обобщение на задачу регрессии  — Пример реализации алгоритма KNN при помощи библиотеки numpy  — Визуализация метрических методов  — Понятие кросс-валидации  — Стратегии выбора гиперпараметров алгоритма  — Реализация алгоритмов в библиотеке sklearn | **2** |  |  | **1** | **3** |
| 3. | **Тема 3. Линейные модели классификации**  — Задача классификации в многомерном пространстве  — Понятие разделяющей гиперплоскости  — Понятие отступа объекта  — Обобщенный линейный классификатор  — Математическая интерпретация идеи обобщенного метрического классификатора  — Подбор оптимальных параметров линейной модели  — Логистическая регрессия  — Метод опорных векторов  — Практические примеры применения линейных алгоритмов классификации | **2** |  |  | **1** | **3** |
| 4. | **Тема 4. Линейные модели регрессии**  — Понятия признаков объектов и их свойств  — Представление объектов в пространстве признаков-свойств  — Метод наименьших квадратов  — Линейная регрессия как решение метода наименьших квадратов в двумерном случае  — Напоминание матричных операций из линейной алгебры  — Многомерное обобщение линейной регрессии  — Примеры | **2** |  |  | **1** | **3** |
| 5. | **Тема 5. Решающие деревья**  — Интуиция алгоритма построения решающего дерева  — Алгоритм классификации на основе построенных решающих деревьев  — Алгоритм построения решающего дерева – LearnID3  — Достоинства и недостатки решающих деревьев  — Проблема переобучения решающих деревьев  — Визуализация разбиения плоскости на области при помощи решающих деревьев  — Примеры применения решающих деревьев | **2** |  |  | **1** | **3** |
| 6. | **Тема 6. Ансамблевые алгоритмы**  — Интуиция ансамблей  — Разложение ошибки в Bias и Variance  — Ансамбли как метод уменьшения variance-составляющей ошибки  — Простейшие ансамбли  — Ансамбли на основе решающих деревьев  — Случайный лес  — Градиентный бустинг  — Визуализации  — Практические примеры | **2** |  |  | **1** | **3** |
| 7. | **Тема 7. Метрики качества классификации**  — Проблемы наивных методов оценки качества классификации  — Метрика precision  — Метрика recall  — Метрика f1  — Метрики оценки качества мягкой классификации  — Метрика ROC AUC  — Метрика PR AUC  — Многоклассовая классификация  — Подходы One VS All и All VS All  — Метрики качества многоклассовой классификации | **2** |  |  | **1** | **3** |
| 8. | **Тема 8. Задача понижения размерности**  — Концепция обучения без учителя  — Проблемы данных высокой размерности. “Проклятие размерности”  — Введение в матричные разложения  — SVD-разложение  — Снижение размерности при помощи SVD  — Метод главных компонент (PCA)  — Особенности и ограничения работы с PCA  — Алгоритм t-SNE  — Примеры применения всех описанных выше алгоритмов | **2** |  |  | **1** | **3** |
| 9. | **Тема 9. Задача кластеризации**  — Понятие кластеризации данных  — Принципиальные проблемы кластеризации  — Простейшие алгоритмы кластеризации: K-Means  — K-Means как разновидность EM-алгоритма  — Собственноручная имплементация K-Means  — Визуализации  — Достоинства и недостатки K-Means  — Иерархическая кластеризация  — Алгоритм Ланса-Уильямса  — Примеры | **2** |  |  | **1** | **3** |
| 10. | **Тема 10. Рекомендательные системы**  — Историческая ретроспектива  — Постановка задачи  — Общая схема рекомендательных систем  — Типы рекомендательных систем  — Коллаборативная фильтрация  — item2item рекомендации  — Рекомендации user-based и item-based  — Матричные разложения в рекомендательных системах  — Алгоритм ALS  — Оценка качества рекомендаций | **2** |  |  | **1** | **3** |
| 11. | **Тема 11. Обработка временных рядов**  — Компоненты временного ряда  — Тренд  — Стационарные временные ряды  — Критерии стационарности  — Прогнозирование временных рядов | **2** |  |  | **1** | **3** |
| 12. | **Тема 12. Введение в нейронные сети**  — Понятие высокоуровневых признаков  — Мотивация выделения высокоуровневых признаков  — Сложности работы с линейными моделями  — XOR-problem  — Решение XOR-problem при помощи комбинации простых линейных моделей  — История развития идеи перцептрона Розенблатта  — Визуализация  — Идея простейшей полносвязной нейронной сети | **2** |  |  | **1** | **3** |
|  | **Итого по курсу:** | **24** |  |  | **12** | **36** |

**7.** Фонд оценочных средств (ФОС, оценочные и методические материалы) для оценивания результатов обучения по дисциплине.

**7.1.** Критерии оценивания

Поскольку межфакультетский курс является лекционным, а аттестация предусмотрена в форме зачета без оценки, аттестация слушателей будет проводиться исходя из выполнения домашних заданий. Домашнее задание будет считаться выполненным, если за него будет набрано не менее 70% баллов от максимально возможного количества баллов. Домашние задания будут представлены в форме задач с автоматической проверкой.

**7.2.** Список вопросов к зачету:

1. Метрические методы

- Что такое косинусная схожесть?

- Верно ли, что функция расстояния в метрических алгоритмах классификации и регрессии всегда удовлетворяет неравенству треугольника?

(доп. вопросы: что такое неравенство треугольника? Можете ли привести пример такой функции расстояния?)

- Манхэттенское расстояние между двумя точками.

- Идея алгоритма k-NN. В каких задачах применяется. Эффективность. Влияние значения k на время работы и и качество получаемого решения. Недостатки k-NN.

2. Линейные классификаторы

- Почему классификаторы называются линейными? Какое предположение о зависимости в данных делается в этом случае?

- Support Vector Machine, или метод опорных векторов.

- Какую задачу решает логистическая регрессия? Что возвращается в результате применения логистической регрессии?

- Расшифруйте TF-IDF (TF(t, d) - term frequency, IDF(t, D) - inverse document frequency, tf-idf(t, d, D) = tf(t, d) \* idf(t, D))

- Градиентный спуск. Градиент.

- Отличия SVM от Логистической регрессии

- Недостатки SVM

3. Линейная регрессия

- Какие метрики можно применять для оценки качества построенной линейной модели?

- Для чего еще, кроме решения самой задачи регрессии, довольно часто используют алгоритм построения линейной регрессии?

- Одномерный случай лин. регрессии. Сколько параметров у модели?

4. Решающие деревья

- Это алгоритм без учителя, с учителем?

- Недостатки решающих деревьев.

- Достоинства решающих деревьев.

5. Ансамбли

- Что такое ассамблирование нескольких моделей?

- Мотивация для использования ансамблей.

- Какими должны быть базовые алгоритмы в ансамбле?

- В чем отличие градиентного бустинга от случайного леса?

6. Метрики качества классификации

- Когда метрика accuracy может сбить с толку?

- Что такое F1 мера?

- PR-кривая.

- ROC-кривая.

- Многоклассовая классификация. One vs All, All vs All. Качество многоклассовой классификации.

7. Понижение размерности

- Проклятие размерности. Проблемы.

- SVD разложение. Применение.

- Что такое эмбеддинг?

- PCA, или метод главных компонент.

- t-SNE. Что это такое?

8. Кластеризация

- К какому типу задач относится кластеризация?

- Проблемы постановки задачи кластеризации?

- K-Means. Сходимость. Недостатки.

- Иерархическая кластеризация.

9. Рекомендательные системы

- Коллаборативная фильтрация.

- Knowledge-based и User-based подходы.

- Матрица User-Item, разреженность матрицы.

- ALS разложение

- SVD vs ALS разложение.

10. Временные ряды

- Что такое временной ряд?

- Компоненты временного ряда?

- Стационарный временной ряд?

- Автокорреляция.

- Приведите примеры алгоритмов прогнозирования временных рядов.

11. Нейронные сети

- Перцептрон Розенблатта

- Приведите пример функций активации. Почему они имеют такую форму?

- Многослойный перцептрон.

- XOR проблема и ее решение.

- Применение нейронных сетей.

**8.** Ресурсное обеспечение.

Основная литература

1. Кристофер Бишоп. Pattern Recognition and Machine Learning. Издательство: Springer, 2006.

Дополнительная литература

1. Андрей Михайлович Миронов. Машинное обучение: учебное пособие. Москва : МАКС Пресс, 2019.
2. Константин Вячеславович Воронцов. Конспекты курса лекций “Машинное обучение”: http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5\_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5\_(%D0%BA%D1%83%D1%80%D1%81\_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9,\_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2).

**9.** Преподаватели дисциплины: Мигачёва О.А., Горохов О.Е.