Федеральное государственное бюджетное образовательное
учреждение высшего образования
Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
ФИЛИАЛ МГУ В Г. ДУБНЕ

УТВЕРЖДАЮ
И.о. директора
филиала МГУ в г. Дубне
_____/ Э.Э. Боос /
«24» марта 2024 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наименование дисциплины:

Технологии анализа данных с применением открытых библиотек на языке программирования Python

---

Уровень высшего образования:
магистратура

Направление подготовки / специальность:

01.04.02 "Прикладная математика и информатика" (3++)

Направленность (профиль):

Методы и технологии обработки данных в гетерогенных вычислительных средах

Форма обучения:
очная

Дубна 2024

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по направлению подготовки 01.04.02 "Прикладная математика и информатика" программы магистратуры в редакции приказа МГУ от _____20__ г.

1. Место дисциплины (модуля) в структуре ОПОП ВО:
Дисциплина относится к вариативной части ОПОП ВО.

2. Входные требования для освоения дисциплины (модуля), предварительные условия (если есть):

3. Результаты обучения по дисциплине (модулю):

Планируемые результаты обучения по дисциплине (модулю)		
Содержание и код компетенции.	**Индикатор (показатель) достижения компетенции**	**Планируемые результаты обучения по дисциплине, сопряженные с индикаторами достижения компетенций**
СПК-3. Способность разрабатывать гетерогенные вычислительные среды для решения научных задач крупных проектов, включая проекты класса мегасайенс.	СПК-3.1. Разрабатывает гетерогенные вычислительные среды для решения научных задач крупных проектов, включая проекты класса мегасайенс.	**Знать** Основы библиотек NumPy и pandas. Очистка и предварительная обработка данных. Визуализация данных с помощью Matplotlib и Seaborn: Основные принципы и методы работы нейронных сетей. Их виды и применение в различных областях. **Уметь** Создавать модели для проверки гипотез и построения доверительных интервалов. Проводить корреляционный и регрессионный анализ. Выполнять обучение, тестирование и оценку моделей. **Иметь практический опыт** Анализа данных с применением открытых библиотек на языке программирования Python

4. Объем дисциплины составляет 3 з.е., в том числе 72 академических часов, отведенных на контактную работу обучающихся с преподавателем, 36 академических часов на самостоятельную работу обучающихся.

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

5.1. Структура дисциплины (модуля) по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий (в строгом соответствии с учебным планом)

Наименование разделов и тем дисциплины (модуля), **Форма промежуточной аттестации по дисциплине (модулю)**	**Номинальные трудозатраты обучающегося**			**Всего академических часов**	**Форма текущего контроля успеваемости* (наименование)**
	Контактная работа (работа во взаимодействии с преподавателем) Виды контактной работы, академические часы		**Самостоятельная работа обучающегося, академические часы**		
	Занятия лекционного типа	Занятия семинарского типа			
Раздел 1. Введение в анализ данных с помощью Python	4	4	2	10	опрос
Раздел 2. Эффективная работа с данными в Python	4	4	4	12	опрос
Раздел 3. Применение алгоритмов машинного обучения для прогнозирования	6	6	6	14	опрос
Раздел 4. Обработка временных рядов	4	4	6	14	опрос
Раздел 5. Нейронные сети – основы	4	4	6	14	опрос
Раздел 6. Обработка графических данных	6	6	6	18	опрос
Раздел 7. Основы парсинга веб-страниц с помощью Python.	4	4	4	14	опрос
Раздел 8. Знакомство с инструментами для работы с большими данными.	4	4	2	12	опрос
Другие виды самостоятельной работы (отсутствуют)	—	—			—
Промежуточная аттестация (экзамен)					
Итого	*36*	*36*	*36*	108	—

5.2. Содержание разделов (тем) дисциплины

№ п/п	Наименование разделов (тем) дисциплины	Содержание разделов (тем) дисциплин
1.	Раздел 1. Введение в анализ данных с помощью Python.	Python для анализа данных. Основы библиотек NumPy и pandas. Очистка и предварительная обработка данных. Визуализация данных с помощью Matplotlib и Seaborn:
2.	Раздел 2. Эффективная работа с данными в Python.	Выявление закономерностей и отклонений. Использование описательной статистики для сводного анализа. Создание осмысленных графиков и диаграмм.
3.	Раздел 3. Применение алгоритмов машинного обучения для прогнозирования.	Проверка гипотез и доверительные интервалы. Корреляционный и регрессионный анализ. Обучение, тестирование и оценка моделей.
4.	Раздел 4. Обработка временных рядов.	Обработка данных временных рядов. Визуализация и декомпозиция временных рядов. Прогнозирование с использованием моделей временных рядов.
5.	Раздел 5. Нейронные сети – основы.	Основные принципы и методы работы нейронных сетей. Их виды и применение в различных областях.
6.	Раздел 6. Обработка графических данных.	Базовые понятия задач классификации изображений, детекции и сегментации объектов. Существующие подходы и решения.
7.	Раздел 7. Основы парсинга веб-страниц с помощью Python.	Основы парсинга веб-страниц с помощью Python. Извлечение данных с веб-сайтов. Обработка и очистка данных.
8.	Раздел 8. Знакомство с инструментами для работы с большими данными.	Интеграция грид-технологий, облачных вычислений, технологий Больших данных. Новые решения и перспективы в обработке данных.

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

6.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости

Примеры тем для устного опроса:

1. Какова цель использования NumPy для анализа данных с помощью Python?
2. Объясните разницу между DataFrame и Series в библиотеке pandas.
3. Опишите три распространенных метода обработки выбросов в наборе данных.
4. Какова цель нормализации данных?
5. Сравните Matplotlib и Seaborn для визуализации данных в Python.

6. Объясните концепцию асимметрии в наборе данных и то, как она влияет на распределение данных.
7. Какова цель проверки гипотез при анализе данных?
8. В чем отличия контролируемых и неконтролируемых алгоритмов машинного обучения.
9. Объясните концепцию переобучения в машинном обучении и способы его предотвращения.
10. Что такое гиперпараметры модели, и за что они отвечают?
11. Что такое ансамбль моделей и когда его применять?
12. Что такое алгоритмы кластеризации?
13. Что такое алгоритм K-средних?
14. Анализ временных рядов:
15. Каковы этические соображения при выполнении парсинга веб-страниц?
16. Какова роль PySpark в анализе больших данных с помощью Python.
17. В чем различия между пакетной обработкой и обработкой больших данных в реальном времени.
18. Опишите важность определения четкой постановки задачи перед началом проекта анализа данных.
19. Какие есть преимущества и проблемы параллельных вычислений при анализе данных.
20. Напишите функцию Python для расчета коэффициента корреляции между двумя переменными.
21. Реализуйте простую модель линейной регрессии
22. Что такое машинное обучение и зачем его применяют?
23. Какие задачи можно решать с помощью методов машинного обучения?
24. Что такое нейронная сеть и как она работает?
25. Какие типы алгоритмов машинного обучения существуют?
26. Какие задачи может решать глубокое обучение?
27. Что такое рекуррентные нейронные сети (RNN)?
28. Что такое сверточные нейронные сети (CNN)?
29. Что такое алгоритмы обучения с подкреплением?
30. В чем суть обучения с учителем?

6.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации по дисциплине, критерии и шкалы оценивания

Вопросы к экзамену

1. Роль Python в анализе данных.
2. NumPy и pandas для анализа данных
3. Очистка, предварительная обработка и отображение данных.
4. Концепция нормализации данных и ее значение в машинном обучении.
5. Matplotlib и Seaborn для отображения данных
6. Какова цель исследовательского анализа данных (EDA)?
7. Методы обнаружения выбросов в наборе данных.
8. Влияние асимметрии данных на анализ.
9. Машинное обучение для анализа данных.
10. Контролируемое и неконтролируемое обучение.
11. Этапы построения и оценки модели машинного обучения.
12. Анализ временных рядов.

13. Этические соображения, связанные с парсингом веб-страниц.
14. Какова роль PySpark в анализе больших данных?
15. Почему важно четко определить постановку задачи в проекте анализа данных?
16. Концепция параллельных вычислений и их роль в анализе данных.
17. Примеры методов сжатия данных в контексте больших данных.
18. Шаги, которые вы предпримете для первоначального исследования данных.
19. Подходы и алгоритмы, используемые в машинном обучении
20. Методы оценки качества моделей
21. Алгоритм K-ближайших соседей
22. Алгоритм решающих деревьев
23. Алгоритм градиентного бустинга
24. Предобучение, недообучение и переобучение
25. Ансамбли моделей и их применение.
26. Основные понятия нейронных сетей.
27. Функции минимизации потерь
28. Обработка естественного языка (NLP)
29. Фреймворки машинного обучения
30. Использование машинного обучения для классификации текстов
31. Машинное обучение для классификации изображений
32. Использование машинного обучения для анализа социальных сетей

Примерные темы семинарских заданий

1. Основы программирования Python, относящиеся к анализу данных. Ключевые библиотеки (NumPy, pandas) и их использование.
2. Matplotlib и Seaborn для визуализации данных
3. Применение статистических методов с использованием Python
4. Введение в библиотеку scikit-learn для машинного обучения на Python.
5. Статистические модели и оценка их точности
6. Методы обработки и анализа данных временных рядов.
7. Применение подхода переноса обучения для решения задачи классификации изображений
8. Обнаружение и отслеживание объектов
9. Парсинг веб-страниц с использованием Python
10. Введение в обработку естественного языка (NLP):
11. Пространственный анализ данных с помощью GeoPandas
12. Поиск аномалий в данных

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине				
Оценка виды оценочных средств	2 (не зачтено)	3 (зачтено)	4 (зачтено)	5 (зачтено)

Знания *(виды оценочных средств: опрос, тесты)*	Отсутствие знаний	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения *(виды оценочных средств: практические задания)*	Отсутствие умений	В целом успешное, но не систематическое умение	В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера)	Успешное и систематическое умение
Навыки (владения, опыт деятельности) ***(виды оценочных средств: выполнение и защита курсовой работы, отчет по практике, отчет по НИР и т.п.)***	Отсутствие навыков (владений, опыта)	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной форме	Сформированные навыки (владения), применяемые при решении задач

7. Ресурсное обеспечение:

7.1. Перечень основной и дополнительной литературы

1. Уэс МакКинни: Python и анализ данных, 2023
2. Бенджио, Гудфеллоу, Курвилль: Глубокое обучение, 2018
3. Криволапов С.Я. Введение в анализ данных. Поиск структуры данных с применением языка Python, 2024
4. Пойнтер Я. Программируем с PyTorch: Создание приложений глубокого обучения, Бестселлеры O'Reilly, 2020

Дополнительная литература

1. Анатолий Постолит. Основы искусственного интеллекта в примерах на Python, 2022
2. Эндрю Траск: Грокаем глубокое обучение, 2019
3. Андрей Бурков: Машинное обучение без лишних слов, 2020

7.2. Перечень лицензионного программного обеспечения, в том числе отечественного производства

При реализации дисциплины может быть использовано следующее программное обеспечение:

1. Операционная система Ubuntu 22.04.
2. Программный продукт Python 3.5.1 (64-bit) Python Software Foundation
3. Операционная система Microsoft Windows 7 корпоративная академическая лицензия
4. Операционная система Microsoft Windows 10 Education академическая лицензия

7.3. Перечень профессиональных баз данных и информационных справочных систем

1. http://www.edu.ru – портал Министерства образования и науки РФ
2. http://www.ict.edu.ru – система федеральных образовательных порталов «ИКТ в образовании»
3. http://www.openet.ru - Российский портал открытого образования
4. http://www.mon.gov.ru - Министерство образования и науки Российской Федерации
5. http://www.fasi.gov.ru - Федеральное агентство по науке и инновациям

7.4. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. Math-Net.Ru [Электронный ресурс] : общероссийский математический портал / Математический институт им. В. А. Стеклова РАН ; Российская академия наук, Отделение математических наук. - М. : [б. и.], 2010. - Загл. с титул. экрана. - Б. ц.
URL: http://www.mathnet.ru

2. Университетская библиотека Online [Электронный ресурс] : электронная библиотечная система / ООО "Директ-Медиа" . - М. : [б. и.], 2001. - Загл. с титул. экрана. - Б. ц. URL: www.biblioclub.ru

3. Универсальные базы данных East View [Электронный ресурс] : информационный ресурс / East View Information Services. - М. : [б. и.], 2012. - Загл. с титул. экрана. - Б. ц.
URL: www.ebiblioteka.ru

4. Научная электронная библиотека eLIBRARY.RU [Электронный ресурс] : информационный портал / ООО "РУНЭБ" ; Санкт-Петербургский государственный университет. - М. : [б. и.], 2005. - Загл. с титул. экрана. - Б. ц.
URL: www.eLibrary.ru

7.5. Описание материально-технического обеспечения.

Филиал МГУ в г. Дубне, ответственный за реализацию данной Программы, располагает соответствующей материально-технической базой, включая современную вычислительную технику, объединенную в локальную вычислительную сеть, имеющую выход в Интернет. Используются специализированные компьютерные классы, оснащенные современным оборудованием.

Материальная база подразделения соответствует действующим санитарно-техническим нормам и обеспечивает проведение всех видов занятий (лабораторной, практической, дисциплинарной и междисциплинарной подготовки) и научно-исследовательской работы обучающихся, предусмотренных учебным планом.

8. Методические рекомендации по организации изучения дисциплины

8.1. Формы и методы преподавания дисциплины

(перечисляются в соответствии с таблицей 5.1.)

Используемые формы и методы обучения:
лекции
семинарские занятия

самостоятельная работа студентов.

В процессе преподавания дисциплины преподаватель использует как классические формы и методы обучения (лекции и семинарские занятия), так и активные методы обучения.

При проведении лекционных занятий преподаватель использует аудиовизуальные, компьютерные и мультимедийные средства обучения, а также демонстрационные и наглядно-иллюстрационные (в том числе раздаточные) материалы.

Семинарские занятия проводятся в форме проблемной ситуации, когда некоторый аспект рассмотренной темы излагается преподавателем более подробно. Часть информации конспектируется. Большая часть времени выделена на работу с использованием компьютерной техники и программного обеспечения.

В рамках курса используются активные и интерактивные методы обучения в процессе проведения занятий. Основными особенностями интерактивных занятий являются интерактивные практические упражнения и задания, которые выполняются обучающимися не только и не столько на закрепление изученного материала, но и на самостоятельное изучение нового.

8.2. Методические рекомендации преподавателю

Перед началом изучения дисциплины преподаватель должен ознакомить студентов с видами учебной и самостоятельной работы, перечнем литературы и интернет-ресурсов, формами текущей и промежуточной аттестации, с критериями оценки качества знаний для итоговой оценки по дисциплине.

При проведении лекций, преподаватель:

1) формулирует тему и цель занятия;

2) излагает основные теоретические положения;

3) с помощью мультимедийного оборудования и/или под запись дает определения основных понятий, расчетных формул;

4) проводит примеры из отечественного и зарубежного опыта, дает текущие статистические данные для наглядного и образного представления изучаемого материала;

5) в конце занятия дает вопросы для самостоятельного изучения.

Для семинарских занятий

Подготовка к проведению занятий проводится регулярно. Организация преподавателем семинарских занятий должна удовлетворять следующим требования: количество занятий должно соответствовать учебному плану программы, содержание планов должно соответствовать программе, план занятий должен содержать перечень рассматриваемых вопросов.

Во время семинарских занятий используются словесные методы обучения, как беседа и дискуссия, что позволяет вовлекать в учебный процесс всех слушателей и стимулирует творческий потенциал обучающихся.

При подготовке семинарскому занятию преподавателю необходимо знать план его проведения, продумать формулировки и содержание учебных вопросов, выносимых на обсуждение.

В начале занятия преподаватель должен раскрыть теоретическую и практическую значимость темы занятия, определить порядок его проведения, время на обсуждение каждого учебного вопроса. В ходе занятия следует дать возможность выступить всем желающим и предложить выступить тем слушателям, которые проявляют пассивность.

Целесообразно, в ходе обсуждения учебных вопросов, задавать выступающим и аудитории дополнительные и уточняющие вопросы с целью выяснения их позиций по существу обсуждаемых проблем, а также поощрять выступление с места в виде кратких дополнений. На занятиях проводится отработка практических умений под контролем преподавателя

8.3. Методические рекомендации студентам по организации самостоятельной работы.

Приступая к изучению новой учебной дисциплины, студенты должны ознакомиться с учебной программой, учебной, научной и методической литературой, имеющейся в библиотеке университета, встретиться с преподавателем, ведущим дисциплину, получить в библиотеке рекомендованные учебники и учебно-методические пособия, осуществить запись на соответствующий курс в среде электронного обучения университета.

Глубина усвоения дисциплины зависит от активной и систематической работы студента на лекциях и практических занятиях, а также в ходе самостоятельной работы, по изучению рекомендованной литературы.

На лекциях важно сосредоточить внимание на ее содержании. Это поможет лучше воспринимать учебный материал и уяснить взаимосвязь проблем по всей дисциплине. Основное содержание лекции целесообразнее записывать в тетради в виде ключевых фраз, понятий, тезисов, обобщений, схем, опорных выводов. Необходимо обращать внимание на термины, формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации. Желательно оставлять в конспектах поля, на которых делать пометки из рекомендованной литературы, дополняющей материал прослушанной лекции, а также подчеркивающие особую важность тех или иных теоретических положений. С целью уяснения теоретических положений, разрешения спорных ситуаций необходимо задавать преподавателю уточняющие вопросы. Для закрепления содержания лекции в памяти, необходимо во время самостоятельной работы внимательно прочесть свой конспект и дополнить его записями из учебников и рекомендованной литературы. Конспектирование читаемых лекций и их последующая доработка способствует более глубокому усвоению знаний, и поэтому являются важной формой учебной деятельности студентов.

Методические указания для самостоятельной работы обучающихся

Прочное усвоение и долговременное закрепление учебного материала невозможно без продуманной самостоятельной работы. Такая работа требует от студента значительных усилий, творчества и высокой организованности. В ходе самостоятельной работы студенты выполняют следующие задачи: дорабатывают лекции, изучают рекомендованную литературу, готовятся к практическим занятиям, к коллоквиуму, контрольным работам по отдельным темам дисциплины. При этом эффективность учебной деятельности студента во многом зависит от того, как он распорядился выделенным для самостоятельной работы бюджетом времени.

Результатом самостоятельной работы является прочное усвоение материалов по предмету согласно программы дисциплины. В итоге этой работы формируются профессиональные умения и компетенции, развивается творческий подход к решению возникших в ходе учебной деятельности проблемных задач, появляется самостоятельности мышления.

Решение задач

При самостоятельном решении задач нужно обосновывать каждый этап решения, исходя из теоретических положений курса. Если студент видит несколько путей решения проблемы (задачи), то нужно сравнить их и выбрать самый рациональный. Полезно до начала вычислений составить краткий план решения проблемы (задачи).

Решение проблемных задач или примеров следует излагать подробно, вычисления располагать в строгом порядке, отделяя вспомогательные вычисления от основных. Решения при необходимости нужно сопровождать комментариями, схемами, чертежами и рисунками.

Следует помнить, что решение каждой учебной задачи должно доводиться до окончательного логического ответа, которого требует условие, и по возможности с выводом.

Полученный ответ следует проверить способами, вытекающими из существа данной задачи. Полезно также (если возможно) решать несколькими способами и сравнить полученные результаты.

Решение задач данного типа нужно продолжать до приобретения твердых навыков в их решении.

Задача — это цель, заданная в определенных условиях, решение задачи — процесс достижения поставленной цели, поиск необходимых для этого средств.

Алгоритм решения задач:

1. Внимательно прочитайте условие задания и уясните основной вопрос, представьте процессы и явления, описанные в условии.

2. Повторно прочтите условие для того, чтобы чётко представить основной вопрос, проблему, цель решения, заданные величины, опираясь на которые можно вести поиски решения.

3. Произведите краткую запись условия задания.

4. Если необходимо составьте таблицу, схему, рисунок или чертёж.

5. Определите метод решения задания, составьте план решения.

6. Запишите основные понятия, формулы, описывающие процессы, предложенные заданной системой.

7. Найдите решение в общем виде, выразив искомые величины через заданные.

9. Проверьте правильность решения задания.

10. Произведите оценку реальности полученного решения.

11. Запишите ответ.

9. Разработчик (разработчики) программы.

к.т.н. Ужинский А.В.