Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

УТВЕРЖДЕНО
Проректор по учебной работе и довузовской подготовке
А.А. Воронов

Рабочая программа дисциплины (модуля)

по дисциплине: Анализ данных на Python
по направлению: Прикладные математика и физика
профиль подготовки: Общая и прикладная физика
Физтех-школа физики и исследований им. Ландау
кафедра информатики и вычислительной математики
курс: 2
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 4 (весенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:
лекции: 0 час.
семинары: 0 час.
лабораторные занятия: 60 час.

Самостоятельная работа: 120 час.

Всего часов: 180, всего зач. ед.: 4

Количество контрольных работ, заданий: 4

Программу составил: Т.Ф. Хирьянов

Программа обсуждена на заседании кафедры информатики и вычислительной математики 06.02.2020

Аннотация

Цель дисциплины — научить обучающихся использованию языка программирования Python, включая его уникальные особенности синтаксиса и библиотеки, для решения прикладных и научных задач.

В ходе данного курса обучающийся освоит использование следующих технологий и библиотек:

1. Использование системы контроля версия git.
2. Итерируемые объекты, генераторы и сопроцессы на Python.
3. Декорирование функций и классов.
4. Объектно-ориентированное программирование.
5. Многопоточное программирование.
6. Визуализация данных .
7. Обработка табличных данных.
8. Элементы машинного обучения.

Курс предполагает выполнение серии лабораторных работ, описания которых появляются на сайте курса в течение семестра.

1. Цели и задачи

Цель дисциплины

Освоить инструментарий языка и основных научных библиотек Python для анализа экспериментальных данных.

Задачи дисциплины

– Изучение продвинутых возможностей языка Python 3;

– освоение среды Jupyter;

– освоение инструментария библиотек Pandas, NumPy и других для считывания и обработки данных;

– обучение визуализации данных средствами Matplotlib, Seaborn и других библиотек Python.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
	ОПК-1.3 Способен определять границы применимости полученных результатов
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области

ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований, и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре	ОПК-5.1 Способен решать поставленные задачи в области теоретических и экспериментальных исследований и разработок
	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников
	ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры
ПК-1 Способен планировать и проводить научные эксперименты (в избранной предметной области) и (или) теоретические (аналитические и имитационные) исследования	ПК-1.3 Владеет культурой постановки научной задачи и моделирования естественнонаучных объектов и систем
	ПК-1.4 Умеет строить математические модели для описания и исследования процессов и явлений в соответствующих научных областях
	ПК-1.6 Знает основные правила поведения и работы в современной научной лаборатории
	ПК-1.7 Способен оценивать требуемые ресурсы (материальные и временные) для планирования и проведения научного эксперимента
	ПК-1.8 Владеет навыками работы с современными языками программирования и программными пакетами для научных расчетов
ПК-2 Способен анализировать полученные в ходе научно-исследовательской работы данные и делать научные выводы (заключения)	ПК-2.1 Владеет методами статистической обработки и анализа научных данных
	ПК-2.2 Умеет находить ключевые параметры, определяющие изучаемое явление, и производить численные оценки по порядку величины
ПК-3 Способен выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области	ПК-3.2 Знает области и критерии применимости используемых теоретических подходов и умение оценивать точность приближенных аналитических методов вычислений
	ПК-3.3 Умеет производить оценку точности численных методов, используемых на ЭВМ, вычислительной сложности используемых алгоритмов и объема требуемых вычислительных ресурсов
ПК-4 Способен критически оценивать применимость используемых методик и методов	ПК-4.3 Способен обосновать причинно-следственные отношения используемых понятий и моделей

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- Синтаксические конструкции функционального программирования на Python 3;
- синтаксические основы ООП-программирования на Python 3;
- возможности научных библиотек Python по анализу данных.

уметь:

- Работать в среде Jupyter;
- создавать читабельные программы на языке Python в том числе в формате Jupyter Notebook;
- использовать Pandas, Numpy и другие научные библиотеки для анализа данных;
- визуализировать данные и результаты анализа.

владеть:

Инструментарием языка Python и научных библиотек для анализа данных на практике.

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

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Система контроля версий git			2	4
2	Объектно-ориентированное программирование на Python.			6	12
3	Функциональное программирование на Python			6	12
4	Многопоточность в Python			6	12
5	Библиотеки для обработки данных и визуализации			20	40
6	Элементы машинного обучения			20	40
Итого часов				60	120
Подготовка к экзамену		0 час.			
Общая трудоёмкость		180 час., 4 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 4 (Весенний)

1. Система контроля версий git

Создание и настройка репозитория. Клонирование репозитория. Подключение к удаленному репозиторию. Создание коммита, синхронизация с удаленным репозиторием. Работа с ветками: создание веток, слияние веток. Разрешение конфликтов. Организация работы в github: issues, projects.

2. Объектно-ориентированное программирование на Python.

Понятие объекта и класса. Парадигмы ООП. SOLID-принципы. Создание структуры взаимодействующих классов. «Магические» методы классов в Python. Статические и классовые методы. Абстрактные классы. Декомпозиция программы на модули. Менеджер контекста. Обработка исключений.

3. Функциональное программирование на Python

Итерируемые объекты. Генераторы и итераторы. Принцип работы for. Объект range. Ключевое слово yield. Генераторы itertools. Сопроцессы. Работа с файлами.

4. Многопоточность в Python

Поток и процесс. Передача данных между потоками при помощи pipe и общей памяти. GIL. Создание процессов и процессов. Асинхронное выполнение потоков. Библиотеки threading, multiprocessing и asyncio.

5. Библиотеки для обработки данных и визуализации

Построение графиков при помощи matplotlib. Настройки стилей оформления графиков. Трехмерные графики, анимация.

Библиотеки numpy, pytorch, для научных вычислений. Создание тензоров и операции над ними. Соединение тензоров, изменение размеров и порядка координат. Модуль numpy.linalg. Граф вычислений.

Работа с базами данных, библиотека sqlite3. Понятие реляционной базы данных. Язык SQL. Написание запросов к базам данных при помощи библиотеки sqlite3.

6. Элементы машинного обучения

Классификация задач машинного обучения. Алгоритмы решения задач обучения с учителем. Линейная регрессия. Алгоритмы классификации: логистическая регрессия, решающие деревья, kNN. Методы кластеризации: k-means, EM, DBSCAN. Методы понижения размерности: PCA, MDS, SNE. Переобучение и регуляризация. Библиотека scikit-learn.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

- Персональные компьютеры;
- учебная аудитория;
- мультимедииный проектор;
- экран.

6.Перечень рекомендуемой литературы

Основная литература

1. Программирование на Python 3 : Подробное руководство [Текст] = Programming in Python 3 : [учеб. пособие для вузов] / М. Саммерфилд; пер. с англ. А. Киселева .— СПб : Символ-Плюс, 2015 .— 608 с.

Дополнительная литература

1. Язык программирования PYTHON [Текст] : учеб. пособие для вузов / Р. А. Сузи .— 2 изд., испр. — М. : Интернет-Ун-т Информ. Технологий : БИНОМ. Лаб. знаний, 2007 .— 326 с.

2. Язык программирования PYTHON [Текст] : учеб. пособие для вузов / Р. А. Сузи .— М. : Интернет-Ун-т Информ. Технологий : БИНОМ. Лаб. знаний, 2006 .— 326 с.

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

1. cs.mipt.ru/advanced_python
2. https://tatyderb.gitbooks.io/python-express-course/content/
3. python.org
4. github.com
5. pythontutor.ru
6. https://www.coursera.org/learn/programming-in-python
7. https://www.coursera.org/learn/python-osnovy-programmirovaniya
8. https://stepik.org/course/512/syllabus

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На рабочих станциях обучающихся должна быть установлена операционная система GNU\Linux.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Курс предполагает выполнение серии лабораторных работ, описания которых появляются на сайте курса в течение семестра. Присутствие на занятиях обязательно, поскольку суть обучения на данном курсе сводится не к приобретению знаний, но к усвоению навыка программирования и проектирования программ.

Дистанционное взаимодействие обучающихся и преподавателя является важной составляющей частью обучения и происходит через репозиторий системы контроля версий git. Отсутствие коммитов в репозиторий (если оно предполагалось лабораторной работой) или отсутствие реакции на замечания преподавателя к коду является таким же дисциплинарным нарушением, как и непосещение занятий. Сдача выполненных работ на «флешке» или по электронной почте не допускается.

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Прикладные математика и физика
профиль подготовки: Общая и прикладная физика
Физтех-школа физики и исследований им. Ландау
кафедра информатики и вычислительной математики
курс: 2
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 4 (весенний) - Дифференцированный зачет

Разработчик: Т.Ф. Хирьянов

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
	ОПК-1.3 Способен определять границы применимости полученных результатов
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований, и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре	ОПК-5.1 Способен решать поставленные задачи в области теоретических и экспериментальных исследований и разработок
	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников
	ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры
ПК-1 Способен планировать и проводить научные эксперименты (в избранной предметной области) и (или) теоретические (аналитические и имитационные) исследования	ПК-1.3 Владеет культурой постановки научной задачи и моделирования естественнонаучных объектов и систем
	ПК-1.4 Умеет строить математические модели для описания и исследования процессов и явлений в соответствующих научных областях
	ПК-1.6 Знает основные правила поведения и работы в современной научной лаборатории
	ПК-1.7 Способен оценивать требуемые ресурсы (материальные и временные) для планирования и проведения научного эксперимента
	ПК-1.8 Владеет навыками работы с современными языками программирования и программными пакетами для научных расчетов

ПК-2 Способен анализировать полученные в ходе научно-исследовательской работы данные и делать научные выводы (заключения)	ПК-2.1 Владеет методами статистической обработки и анализа научных данных
	ПК-2.2 Умеет находить ключевые параметры, определяющие изучаемое явление, и производить численные оценки по порядку величины
ПК-3 Способен выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области	ПК-3.2 Знает области и критерии применимости используемых теоретических подходов и умение оценивать точность приближенных аналитических методов вычислений
	ПК-3.3 Умеет производить оценку точности численных методов, используемых на ЭВМ, вычислительной сложности используемых алгоритмов и объема требуемых вычислительных ресурсов
ПК-4 Способен критически оценивать применимость используемых методик и методов	ПК-4.3 Способен обосновать причинно-следственные отношения используемых понятий и моделей

2. Показатели оценивания компетенций

В результате изучения дисциплины «Анализ данных на Python (ЛФИ)» обучающийся должен:

знать:

- Синтаксические конструкции функционального программирования на Python 3;
- синтаксические основы ООП-программирования на Python 3;
- возможности научных библиотек Python по анализу данных.

уметь:

- Работать в среде Jupyter;
- создавать читабельные программы на языке Python в том числе в формате Jupyter Notebook;
- использовать Pandas, Numpy и другие научные библиотеки для анализа данных;
- визуализировать данные и результаты анализа.

владеть:

Инструментарием языка Python и научных библиотек для анализа данных на практике.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Пример лабораторной работы по анализу данных, используемой для текущего контроля:

Загрузите датасет titanic.csv и, используя описанные выше способы работы с данными, найдите ответы на вопросы

1. Какое количество мужчин и женщин ехало на корабле? В качестве ответа приведите два числа через пробел.
2. Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров. Ответ приведите в процентах (число в интервале от 0 до 100, знак процента не нужен), округлив до двух знаков.
3. Какую долю пассажиры первого класса составляли среди всех пассажиров? Ответ приведите в процентах (число в интервале от 0 до 100, знак процента не нужен), округлив до двух знаков.
4. Какого возраста были пассажиры? Посчитайте среднее и медиану возраста пассажиров. В качестве ответа приведите два числа через пробел.
5. Коррелируют ли число братьев/сестер/супругов с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch.
6. Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name).

Это задание — типичный пример того, с чем сталкивается специалист по анализу данных. Данные очень разнородные и шумные, но из них требуется извлечь необходимую информацию. Попробуйте вручную разобрать несколько значений столбца Name и выработать правило для извлечения имен, а также разделения их на женские и мужские.

Если ответом является нецелое число, то целую и дробную часть необходимо разграничивать точкой, например, 0.42. При необходимости округляйте дробную часть до двух знаков.

Пример дополнительных вопросов при сдаче лабораторной работы:

1. Выберите верные утверждения:
- Объекты описываются с помощью признаков
- Одна из задач машинного обучения — научиться делать прогнозы для объектов
- Одна из задач машинного обучения — научиться делать прогнозы для признаков
- Признаки описываются с помощью объектов

2. Что из этого — корректные названия типов признаков?
- Устойчивые признаки
- Нетривиальные признаки
- Номинальные (категориальные) признаки
- Числовые (количественные) признаки
- Бинарные признаки

3. Какие из этих задач являются задачами классификации?
- Прогноз оценки студента по пятибалльной шкале на экзамене по машинному обучению в следующей сессии
- Поиск групп похожих пользователей интернет-магазина
- Разделение книг, хранящихся в электронной библиотеке, на научные и художественные
- Прогноз температуры на следующий день

4. Какая из этих фраз наиболее точно описывает переобучение?
- Переобучение — это ситуация, в которой алгоритм выдает недетерминированные ответы на новых данных (то есть при разных запусках на одном и том же объекте можно получить разные предсказания)
- Переобучение — это ситуация, в которой алгоритм показывает одинаково плохое качество и на обучающей выборке, и на новых данных
- Переобучение — это ситуация, в которой алгоритм часто отказывается от построения прогноза на новых данных.
- Переобучение — это ситуация, в которой алгоритм показывает хорошее качество на обучающей выборке, но при этом плохо работает на новых данных

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Перечень типовых заданий на дифференцированном зачете:

1. Реализуйте свой класс Complex для комплексных чисел
- Добавьте конструктор класса
- Реализуйте операции проверки на равенство, сложения, вычитания, произведения и деления комплексных чисел (eq, add, sub, mul, truediv)
- Реализуйте операцию модуля (abs)
- Класс должен давать осмысленный вывод как при print, так и просто при вызове в ячейке ноутбука

2. Вам нужно узнать доступность набора IP адресов. Неэффективный вариант представлен ниже. Реализуйте то же самое, но используя threading.

3. Напишите генератор, выводящий первые n чисел Фибоначчи.

4. Реализуйте класс BinTree двоичного дерева, и его итератор, который обходит дерево в порядке обхода в глубину.

5. Решите без использования циклов средствами NumPy (каждый пункт решается в 1-2 строчки)
- Создайте вектор с элементами от 12 до 42
- Создайте вектор из нулей длины 12, но его пятый элемент должен быть равен 1
- Создайте матрицу (3, 3), заполненную от 0 до 8
- Найдите все положительные числа в np.array([1,2,0,0,4,0])
- Умножьте матрицу размерности (5, 3) на (3, 2)
- Создайте матрицу (10, 10) так, чтобы на границе были 0, а внутри 1
- Создайте рандомный вектор и отсортируйте его

6. Реализуйте запрос FULL OUTER JOIN в sqlite3.
7. Создайте таблицы с указанными столбцами и заполните их произвольными данными.

- Books (id, author, title, publish_year)
- Readers (id, name)
- Records (reader_id, book_id, taking_date, returning_date)

Постройте к таблицам указанные SELECT запросы:

- Запрос возвращает id и названия книг, находящихся в данный момент на руках у читателей.
- Запрос возвращает имена читателей и названия книг, которые они когда-либо брали.
- Запрос возвращает количество книг для каждого автора.

Критерии оценивания

10 - Обучающийся ответил на все вопросы, но не с первой попытки.

9 - Обучающийся допустил не более одной ошибки или воспользовался помощью преподавателя.

8 - Обучающийся, работая самостоятельно, допустил не более двух численных ошибок в лабораторной работе.

7 - Обучающийся если он ответил на подавляющее большинство вопросов в лабораторной работе, может быть с помощью преподавателя или товарищей.

6 - Обучающийся ответил на подавляющее большинство вопросов в лабораторной работе, может быть с помощью преподавателя или товарищей.

5 - Обучающийся ответил на основные вопросы в лабораторной работе, может быть с помощью преподавателя или товарищей.

4 - Обучающийся ответил на основные вопросы в лабораторной работе

3 - Обучающийся ответил на некоторые вопросы в лабораторной работе.

2 - Обучающийся не справился с работой.

1 - Обучающийся демонстрирует полное отсутствие знаний по предмету или пытался выдать чужую работу за свою.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Итоговая аттестация по дисциплине «Анализ данных на Python» осуществляется в форме дифференцированного зачета.

Оценка за зачёт выставляется с учётом оценок лабораторных работ, выполняемых в течение семестра.