

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

УТВЕРЖДЕНО
Проректор по учебной работе

А.А. Воронов

	Рабочая программа дисциплины (модуля)
по дисциплине:	Введение в анализ данных
по направлению:	Биотехнология
профиль подготовки:	Системная и синтетическая биология Физтех-школа Биологической и Медицинской Физики кафедра информатики и вычислительной математики
курс:	3
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 0 час.

семинары: 60 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 75 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 4

Программу составили:

В.К. Хохлов, канд. техн. наук, доцент

М.Н. Герцев, канд. физ.-мат. наук, доцент

М.Д. Юдаев, ассистент

Программа обсуждена на заседании кафедры информатики и вычислительной математики 29.05.2020

Курс посвящен введению в машинное обучение средствами языка Python 3.

1. Цели и задачи

Цель дисциплины

Курс «Введение в анализ данных» рассчитан на студентов, владеющих основами программирования на языке Python 3 и предполагает знание базовых принципов работы компьютера - работы с памятью и дисковой подсистемой. Студенты знакомятся с основами совместной работы над большими проектами с использованием системы контроля версий Git: базовыми модулями языка, использующихся для анализа данных, - numpy, pandas, pytorch и пр.; основам работы с реляционными базами данных при помощи языка SQL; линейной алгеброй и основами теории вероятностей.

Задачи дисциплины

Студенты должны получить практические навыки работы в:

1. Git
2. модулях numpy, pandas, sklearn, pytorch
3. реляционных базами данных
4. языка SQL

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- #. системы контроля версий
- #. модули Python для работы с массивами, таблицами, machine learning
- #. средства визуализации математических данных в Python
- #. основы реляционных баз данных
- #. основы языка SQL
- #. основополагающие принципы и методы машинного обучения

уметь:

- #. осуществлять совместную коллаборацию через git
- #. использовать библиотеки numpy, pandas, matplotlib.pyplot, sklearn, torch
- #. читать, создавать и редактировать реляционные базы данных
- #. создавать простейшие системы машинного обучения

владеть:

- #. распределённой системой управления версиями Git
- #. средствами языка Python для машинного обучения
- #. средой Jupyter

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	Повторение Python 3		4		6
2	Система контроля версий git		4		11
3	Numpy, scipy		8		11
4	Pandas		4		6
5	Визуализация в Python 3		4		4
6	Введение в базы данных		8		6
7	scikit-learn		8		7
8	Введение в машинное обучение		8		9
9	Простейшие задачи машинного обучения		4		9
10	Pytorch		8		6
Итого часов			60		75
Подготовка к экзамену		0 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 6 (Весенний)

1. Повторение Python 3

Краткий обзор средств Python 3, изученных в предыдущих курсах. Среда Jupyter Notebook.

2. Система контроля версий git

Работа основных команд git: init, clone, config, add, commit, push, pull, branch, checkout

3. Numpy, scipy

Основные методы библиотеки numpy для эффективной работы с числовыми массивами

4. Pandas

Приобретение навыков работы с табличными данными средствами Python 3, сравнивая по возможностям с табличными редакторами, как MS Excel

5. Визуализация в Python 3

Средства визуализации различных данных: pyplot, mplot3d

6. Введение в базы данных

Реляционные базы данных и основы языка для работа с ними - SQL.

7. scikit-learn

Библиотека scikit-learn как основная библиотека методов линейной алгебры в Python 3

8. Введение в машинное обучение

Общая постановка задачи обучения по прецедентам. Типы задач Машинного обучения. Обучение с учителем и без учителя. Метрики качества.

9. Простейшие задачи машинного обучения

Задача Классификации. k-NN. Плюсы и минусы метода ближайших соседей. Класс KNeighborsClassifier в Scikit-learn. Дерево решений. Построение дерева. Основные параметры дерева. Плюсы и минусы деревьев решений. Класс DecisionTreeClassifier в Scikit-learn. Выбор параметров модели и кросс-валидация.

10. Pytorch

Навыки использования фреймворка PyTorch для машинного обучения на Python 3. Тензоры PyTorch. Модули autograd и nn.

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

Необходимое оборудование для практических занятий: компьютер и мультимедийное оборудование (проектор, звуковая система).

6. Перечень рекомендуемой литературы

Основная литература

1. Python и анализ данных, Первичная обработка данных с применением pandas, NumPy и IPython / У. Маккини. — Москва, ДМК Пресс, 2020.— URL: <https://e.lanbook.com/book/131721> (дата обращения: 26.01.2021). - Полный текст (Режим доступа : из сети МФТИ / Удаленный доступ)
2. Python для сложных задач: наука о данных и машинное обучение [Текст], [учеб. пособие для вузов] / Дж. Вандер Плас ; [пер. с англ. И. Пальти]. -СПб., Питер, 2018
- 3.

Дополнительная литература

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

kaggle.org
machinelearning.ru

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

SQLite3, sqlitebrowser, Python 3, Jupyter Notebook, PyTorch, matplotlib.pyplot

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

Курс предполагает получение обширных практических навыков в программировании на языке Python3.

Широта изучаемых библиотек, не позволяет преподавателю глубоко погружаться в детали каждой, и требует дополнительного самостоятельного изучения студентом. Построение занятий осуществляется в режиме, когда преподаватель объясняет только самые основы нового материала, после чего студентам предоставляется достаточно времени для персональной работы с преподавателем. В связи с этим, для качественного освоения дисциплины студентам необходимо проявлять инициативу и активно работать с преподавателем, задавать множество уточняющих вопросов на интересующие темы. Такой подход студентов позволит существенно сократить время студента в процессе обучения, поскольку самостоятельный поиск методов решения поставленной задачи, как показывает практика, требует несоизмеримо больших затрат со стороны студента и наименее эффективному варианту решения поставленной задачи.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Биотехнология
профиль подготовки: Системная и синтетическая биология
Физтех-школа Биологической и Медицинской Физики
кафедра информатики и вычислительной математики
курс: 3
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Дифференцированный зачет

Разработчики:

В.К. Хохлов, канд. техн. наук, доцент
М.Н. Герцев, канд. физ.-мат. наук, доцент
М.Д. Юдаев, ассистент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников

2. Показатели оценивания компетенций

В результате изучения дисциплины «Введение в анализ данных» обучающийся должен:

знать:

- #. системы контроля версий
- #. модули Python для работы с массивами, таблицами, machine learning
- #. средства визуализации математических данных в Python
- #. основы реляционных баз данных
- #. основы языка SQL
- #. основополагающие принципы и методы машинного обучения

уметь:

- #. осуществлять совместную коллаборацию через git
- #. использовать библиотеки numpy, pandas, matplotlib.pyplot, sklearn, torch
- #. читать, создавать и редактировать реляционные базы данных
- #. создавать простейшие системы машинного обучения

владеть:

- #. распределённой системой управления версиями Git
- #. средствами языка Python для машинного обучения
- #. средой Jupyter

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

Контроль освоения обучающимися учебного материала проводится через реализацию большого совместного проекта по машинному обучению в течении семестра и его устной защитой.

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

Аттестация по дисциплине «Введение в анализ данных» осуществляется в форме дифференцированного зачета. Он проводится в устной форме.

Перечень контрольных вопросов:

1. Понятие предметной области. Понятие модели данных.
2. Гипотеза порождения данных
3. Признаковая матрица и матрица расстояний
4. Типы данных: скалярные и не скалярные, непрерывная и дискретная область значений.
5. Задачи обучения с учителем: классификация и регрессия, функционал качества.
6. Обучение как задача оптимизации. Основные методы и подходы. Регуляризация
7. Обучение без учителя.
8. numpy: представление массивов чисел, основные операции, классы, структуры
9. scipy: методы работы с временными рядами, аналитические преобразования
10. pandas: чтение и запись в файл, компоновка и атрибуция таблиц, фреймы, электронные таблицы
11. Язык SQL: структура запросов на языке SQL.
12. sqlite: основные возможности, интеграция SQL и python
13. Понятие транзакции.
14. sklearn: основные алгоритмы, примеры простых задач обучения

15. Нейросети, их основные типы. Карты Кохонена, RNN, CNN, LSTM, GAN

16. theano, pytorch - основные принципы

17. Глубокое обучение: введение, примеры реализации.

Примеры контрольных заданий:

1. Дан файл с таблицей признаков . неполной и содержащей "мусор". Требуется написать код, корректно считывающий данные из файла и строящий некоторые стандартные проекции (типа главных компонент или карт Кохонена).

2. Дан файл с данными временного ряда (например, запросов на сервер), с пропусками и "мусором". Требуется провести

3. Даны 2 разные базы данных белковых последовательностей. Найти их максимальное пересечение

(наборы, представленные в обеих базах) и оптимальное низкоразмерное представление для них.

Критерии оценивания

Всего работа оценивается по 10-бальной шкале. По конкретным библиотекам языка python есть контесты в системе ejudge, по остальным темам задания в виде jupyter-notebooks.

Итоговая оценка на дифференцированном зачете выставляется с учетом работы в семестре и полноты и качества выполнения курсовой работы.

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

Дифференцированный зачет проводится с учетом текущей успеваемости и результатов сдачи курсовой работы. При необходимости, в процессе собеседования со студентом проводится выборочный опрос на знание контрольных вопросов, предлагаются типовые задачи.

3. Перечень типовых контрольных заданий, используемых для оценки знаний, умений, навыков

Аттестация по дисциплине «Базы данных» осуществляется в форме дифференцированного зачета. Дифференцированный зачет проводится в устной форме.

Перечень контрольных вопросов:

1. Понятие предметной области. Понятие модели предметной. Понятие модели данных.
2. Сетевая и иерархическая модели.
3. Основные понятия реляционной модели: домен, отношение, кортеж, атрибут, операции над отношениями.
4. Типы данных: скалярные и нескаларные.
5. Понятие базы данных и реляционной базы данных.
6. Понятие переменной отношения (реляционной переменной).
7. Понятие кортежной переменной.
8. Переменная, определенная на домене.
9. Операция соединения: внутреннее, внешнее (левое, правое), полное
10. Операция деления.
11. Эквивалентность реляционной алгебры и реляционного исчисления на кортежах
12. Эквивалентность реляционной алгебры и реляционного исчисления на доменах.
13. Отношение, тип, объект, домен, кортеж: взаимосвязь понятий.
14. Язык SQL: структура запросов на языке SQL.
15. Язык SQL: связь с реляционной алгеброй.
16. Язык SQL: связь с реляционным исчислением на кортежах.
17. Язык SQL. Работа с отсутствующими значениями (NULL).
18. Вложенные запросы в языке SQL.
19. Структура хранимой процедуры/функции в языке SQL.
20. Понятие языка определения данных. Определение пользовательского типа данных.
21. Понятие языка определения данных. Создание таблицы.
22. Понятие языка определения данных. Определение ограничений.
23. Триггеры.
24. Особенности хранимых процедур и функций в СУБД MS SQL Server.
25. Особенности хранимых процедур и функций в СУБД Oracle. Пакеты.
26. Общая архитектура СУБД.
27. Понятие транзакции.
28. Организация хранения данных на жестком диске.
29. Виды (представления, views). Материализованные представления.
30. Индексы: назначение и организация.

Примеры контрольных заданий:

Типовое задание предполагает написание типового запроса на языках реляционной алгебры, реляционного исчисления и SQL:

1. Запросы на извлечение данных из одного отношения с помощью одной-двух реляционных операций (ограничение, проекция)
2. Запросы на извлечение данных из двух-трех отношений с помощью двух-трех реляционных операций (включая, как минимум, одно соединение)
3. Запросы на извлечение данных из двух-трех отношений с помощью двух-трех реляционных операций (включая, как минимум, одно деление)
4. Запросы на извлечение данных из нескольких отношений с помощью нескольких реляционных операций (включая соединение или/и деление и теоретико-множественные операции)

5. Запросы, содержащие вложенные запросы
6. Комплексные запросы, включающие сложные вложенные запросы, с множественными операциями соединения/деления.

Примеры задач:

1. Найти производители, детали которых (хотя бы одна) продаются во всех городах (в которых вообще есть магазины).
2. Найти магазины (в Москве), в которых продаются детали весом не более 15 кг, которые производятся в Париже или Берлине.
3. Найти магазины, у которых ассортимент продаваемых деталей такой же, как у данного магазина X.

Типовой вариант контрольно-тестовой работы (КТР) включает 4 теоретических задания в форме закрытых тестовых вопросов и 3 задание на составление запроса. Примеры вариантов – в приложении. Проведение КТР предусматривается как в рамках полусеместрового контроля, так и, при необходимости, в рамках дифференцированного зачета (в рамках дифференцированного зачета КТР может использоваться как вместо, так и совместно с устным собеседованием). Если КТР проводится на дифференцированном зачете, то задачи нужно решать в трех вариантах: на языках реляционной алгебры, реляционного исчисления и SQL.

4. Критерии оценивания

КТР, проводимые в рамках полусеместрового контроля, в варианте с 4мя теоретическими вопросами и 3мя задачами, оцениваются по следующей схеме: каждый теоретический вопрос оценивается в 1 балл, каждая задача в 2 балла – по 1 баллу за каждую версию запроса (на языке алгебры и на языке исчисления). Всего, таким образом, работа оценивается по 10-бальной шкале.

Итоговая оценка на зачете выставляется с учетом работы в семестре и полноты и качества выполнения курсовой работы.

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

Дифференцированный зачет проводится с учетом текущей успеваемости и результатов сдачи курсовой работы. При необходимости, в процессе собеседования со студентом проводится выборочный опрос на знание контрольных вопросов, предлагаются типовые задачи.