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

УТВЕРЖДЕНО
Директор физтех-школы биологической и медицинской физики
Д.В. Кузьмин

Рабочая программа дисциплины (модуля)

по дисциплине: Основы биоинформатического анализа генетических данных растений
по направлению: Биотехнология
профиль подготовки: Биомедицинские технологии
Физтех-школа Биологической и Медицинской Физики
центр образовательных программ Физтех-школы биологической и медицинской физики
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен

Аудиторных часов: 30 всего, в том числе:
лекции: 30 час.
семинары: 0 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Подготовка к экзамену: 30 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 1

Программу составил: А.А. Соловьев, д-р биол. наук, ассистент

Программа обсуждена на заседании центра образовательных программ Физтех-школы биологической и медицинской физики 18.03.2022

Аннотация

В курсе рассматриваются ключевые понятия и методы анализа данных. Прежде всего, вводятся базовые термины, понятия и инструменты для работы с данными. После чего рассматриваются различные методы обучения, как с учителем, так и без него. Отдельные занятия посвящены прикладным задачам, таким как прогнозирование временных рядов и обработка изображений. Дается описание устройства рекуррентных и сверточных нейронных сетей.

Курс содержит в себе обсуждение базовых терминов и методов, а также разбор задач, без которых невозможно понимание науки о данных.

Для успешного освоения курса слушателю желательно владеть основами математического анализа, линейной алгебры и теории вероятностей.

1. Цели и задачи

Цель дисциплины

Получение базовых теоретических знаний и практических навыков в области анализа данных и машинного обучения для дальнейшего их использования при изучении дисциплин по соответствующей программе и выполнении НИР в бакалавриате.

Задачи дисциплины

- дать теоретические знания о базовых методах машинного обучения;
- рассказать о цикле задач науки о данных: начиная с поиска и подготовки информации, заканчивая выбором решения и оценкой его качества;
- дать базовые знания и навыки работы с программными инструментами обработки и представления данных в цифровой форме.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
	УК-1.3 Разрабатывает стратегию достижения поставленной цели как последовательность шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области своей профессиональной деятельности, способен на научном языке формулировать профессиональные задачи	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области своей профессиональной деятельности и их практическую значимость
	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- базовые понятия и инструменты науки о данных;
- возможности интернет-ресурсов и программного обеспечения для решения профессиональных задач;
- основные методы решения задач классификации и регрессии, а также кластеризации и понижения размерности;
- классические архитектуры сверточных нейронных сетей.

уметь:

- осуществлять поиск, фильтрацию, сбор и анализ данных, информации и цифрового контента с использованием интернет-браузеров;
- изучать массивы данных с целью поиска в них структуры и находить закономерности;
- строить гипотезы оценки неизвестных параметров систем и проверять их;
- формулировать и решать задачи машинного обучения на размеченных данных;
- понижать размерность данных и кластеризовать их.

владеть:

- навыками усвоения междисциплинарной информации в области математического анализа, теории вероятностей и программирования;
- навыками поиска информации посредством электронных ресурсов;
- базовыми навыками программирования, включая работу в интерактивной вычислительной среде.

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	Введение	2			2
2	Основы Python	2			2
3	Основы математического анализа и линейной алгебры	2			2
4	Библиотеки Python	2			2
5	Градиент и методы оптимизации	2			2
6	Основные понятия теории вероятностей	2			2
7	Машинное обучение	2			2
8	Линейные модели в задачах регрессии	2			2
9	Переобучение и недообучение	2			2
10	Метрики в задачах регрессии и классификации	2			2
11	Качество оценок	2			2
12	Sklearn	4			4
13	Деревья решений	4			4
Итого часов		30			30
Подготовка к экзамену		30 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

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

Семестр: 1 (Осенний)

1. Введение

Python: история и дзен. Anaconda. Jupyter Notebook: терминология, основные элементы интерфейса, горячие клавиши, hello world, язык разметки, магические команды.

2. Основы Python

Типы данных: числовые, последовательностей, наборов, сопоставления, NoneType. Операторы: if, for, while, break, continue, pass. Анонимные функции. Работа с файлами: способы чтения и записи.

3. Основы математического анализа и линейной алгебры

Множество. Функция. Предел. Производная и ее геометрический смысл. Дифференциал и его геометрический смысл. Экстремумы. Выпуклость. Линейное пространство. Матрицы.

4. Библиотеки Python

Numpy: способы создания массива, базовые и унарные операции, манипуляции с формой, линейная алгебра. SciPy: optimize и методы оптимизации функций, linalg и отличие от numpy. MatPlotLib: pyplot для построения графиков. Pandas: основные компоненты, работа с csv.

5. Градиент и методы оптимизации

Частная производная и градиент. Функция потерь. Градиентный спуск. Методы случайного поиска. Метод имитации отжига. Эволюционные алгоритмы. Метод Нелдера-Мида.

6. Основные понятия теории вероятностей

Главные свойства вероятности. Условная и полная вероятности. Распределения: нормальное, равномерное на отрезке, Бернулли, биномиальное, Пуассона, дискретное. Гистограммы. Характеристики распределений: матожидание, медиана, мода.

7. Машинное обучение

Введение. Задачи обучения с учителем. Задачи обучения без учителя. Типы признаков.

8. Линейные модели в задачах регрессии

Линейная регрессия. Обучение модели линейной регрессии. Стохастический градиентный спуск. Линейная классификация. Функции потерь в задачах классификации.

9. Переобучение и недообучение

Основные причины низкой производительности алгоритмов машинного обучения. Методы выявления. Кросс-валидация. Регуляризация. Геометрический смысл регуляризации.

10. Метрики в задачах регрессии и классификации

Среднеквадратичная ошибка. Средняя абсолютная ошибка. Коэффициент детерминации. Квантильная ошибка. Матрица ошибок. Точность. Полнота. F-мера.

11. Качество оценок

Кривая точности-полноты. Рабочая характеристика приемника. Площадь под кривой. Дисбаланс классов.

12. Sklearn

Sklearn: datasets и наборы данных, model_selection и разбиение данных и итераторы перекрестной проверки, linear_model и линейные модели, metrics и оценки качества.

13. Деревья решений

Деревья решений в задачах регрессии и классификации. Критерии информативности. «Стрижка» деревьев. Композиции деревьев. Случайный лес. Градиентный бустинг.

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

Компьютерный класс для проведения занятий, оснащенный мультимедийным оборудованием (проектором).

6.Перечень рекомендуемой литературы

Основная литература

1. Интеллектуальные системы , учебник / Л. Н. Ясницкий. — Москва, Лаборатория знаний, 2020.— URL: http://books.mipt.ru/book/301409 (дата обращения: 10.03.2021). - Полный текст (Режим доступа : из сети МФТИ / Удаленный доступ)

Дополнительная литература

Литература кафедры:

1. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. – 1999.

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

1. https://drive.google.com – сервис хранения, редактирования и синхронизации файлов.
2. https://docs.scipy.org/doc/scipy/tutorial/index.html#user-guide – руководство пользователя библиотеки SciPy, предназначенной для выполнения научных и инженерных расчётов.

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

Программное обеспечение: Python, Jupyter Notebook (Anaconda).

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

Успешное освоение дисциплины требует большого объема самостоятельной работы студента. Самостоятельная работа включает в себя поиск информации, проработку учебного материала, подготовку к сдаче экзамена.

Студентам доступна возможность консультаций с преподавателем, как индивидуальных, так и групповых, посредством сервисов для мгновенного обмена сообщениями или, при необходимости, сервисов организации и проведения видеоконференций.

Руководство и контроль за самостоятельной работой студента осуществляется преподавателем в форме индивидуальных консультаций, выборочных опросов на занятиях и контрольных вопросов.

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Биотехнология
профиль подготовки: Биомедицинские технологии
Физтех-школа Биологической и Медицинской Физики
центр образовательных программ Физтех-школы биологической и медицинской физики
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен

Разработчик: А.А. Соловьев, д-р биол. наук, ассистент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
	УК-1.3 Разрабатывает стратегию достижения поставленной цели как последовательность шагов, предвидя результат каждого из них и оценивая их влияние на внешнее окружение планируемой деятельности и на взаимоотношения участников этой деятельности
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области своей профессиональной деятельности, способен на научном языке формулировать профессиональные задачи	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области своей профессиональной деятельности и их практическую значимость
	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации

2. Показатели оценивания компетенций

В результате изучения дисциплины «Основы биоинформатического анализа генетических данных растений» обучающийся должен:

знать:

- базовые понятия и инструменты науки о данных;
- возможности интернет-ресурсов и программного обеспечения для решения профессиональных задач;
- основные методы решения задач классификации и регрессии, а также кластеризации и понижения размерности;
- классические архитектуры сверточных нейронных сетей.

уметь:

- осуществлять поиск, фильтрацию, сбор и анализ данных, информации и цифрового контента с использованием интернет-браузеров;
- изучать массивы данных с целью поиска в них структуры и находить закономерности;
- строить гипотезы оценки неизвестных параметров систем и проверять их;
- формулировать и решать задачи машинного обучения на размеченных данных;
- понижать размерность данных и кластеризовать их.

владеть:

- навыками усвоения междисциплинарной информации в области математического анализа, теории вероятностей и программирования;
- навыками поиска информации посредством электронных ресурсов;
- базовыми навыками программирования, включая работу в интерактивной вычислительной среде.

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

Текущий контроль осуществляется в форме индивидуального опроса. По результатам опроса выставляется промежуточная оценка. К экзамену допускаются студенты, получившие положительную оценку.

Типовые вопросы:

1. Оценка распределения выборочного среднего случайной величины при разных объёмах выборок.

- В чем заключается разница между полученными распределениями при различных значениях n?
- Как меняется точность аппроксимации распределения выборочных средних нормальным с ростом n?

2. Линейная регрессия с использованием синтетических данных, смоделированных на основе исследования.

- Как отобразить попарные зависимости признаков с помощью библиотеки Pandas?
- Для чего используется метод pandas.DataFrame.boxplot?
- Какая функция по двум параметрам w_0 и w_1 вычисляет квадратичную ошибку приближения зависимости y от x прямой линией $y = w_0+w_1 \cdot x$?
- С помощью какого метода из модуля scipy.optimize можно найти минимум функции?

3. Логистическая регрессия в задаче бинарной классификации.

- Опишите способы обработки отсутствующих в наборе данных вещественных значений.
- Опишите работу sklearn.feature_extraction.DictVectorizer, используемого для преобразования категориальных признаков.
- Зачем используются стратификация и масштабирование вещественных признаков?
- Зачем используется полиномиальное преобразование признаков?

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

Типовые вопросы для подготовки к экзамену:

1. Градиент в задачах оптимизации.
2. Методы оптимизации гиперпараметров в машинном обучении.
3. Центральная предельная теорема.
4. Линейные модели в задачах регрессии.
5. Метод градиентного спуска в задаче линейной регрессии.
6. Стохастический градиентный спуск.
7. Линейная классификация (классификаторы, функции потерь).
8. Недообучение и переобучение. Методы борьбы с ними.
9. Виды обучения и признаков.
10. Напишите функцию, которая приближает функцию на отрезке с помощью многочленов, а также отображает на графике исходную функцию и получившуюся.
11. Напишите функцию, которая минимизирует функцию на отрезке.
12. Напишите функцию, реализующую шаг стохастического градиентного спуска.
13. Напишите функцию, реализующую стохастический градиентный спуск, предполагая, что шаг градиентного спуска уже реализован (на каждой итерации в вектор (список) должно записываться текущее значение среднеквадратичной ошибки).
14. Классифицируйте набор данных digits с использованием линейной регрессии. «Отложите» 10% и отобразите эффективность прогнозирования на этих данных.
15. Покажите, как на результат работы классификатора влияют изменения в обучающих данных и насколько отличаются друг от друга разбиения, генерируемые K-кратной перекрестной проверкой (K-fold cross-validation). Для этого отобразите кривые ROC различных наборов данных, созданных в результате K-кратной перекрестной проверки. Используйте любой набор и классификатор.
16. Отобразите первый признак набора данных diabetes точками на двумерном графике. На нем же изобразите линией предсказание линейной регрессии. Также рассчитайте коэффициенты, среднеквадратическую ошибку и коэффициент детерминации (r2 score).
17. Продемонстрируйте влияние коллинеарности на коэффициенты оценки, используя Ridge-регрессию. Постройте двумерный график зависимости вектора коэффициентов от параметра регуляризации. В качестве обучающих данных используйте матрицу Гильберта.

18. Отобразите Ridge-коэффициенты в зависимости от L2-регуляризации. Изобразите различными цветами на двумерном графике вектор коэффициентов как функцию параметра регуляризации. Также постройте второй график зависимости среднеквадратичной ошибки между коэффициентами, найденными моделью, и вектором коэффициентов базовой линейной модели от параметра регуляризации.

19. Решите задачу 17, используя L1-регуляризатор.

20. Решите задачу 18, используя L1-регуляризатор.

Примеры билетов

Билет 1

1. Линейная классификация (классификаторы, функции потерь).
2. Недообучение и переобучение. Методы борьбы с ними.

Билет 2

1. Градиент в задачах оптимизации.
2. Методы оптимизации гиперпараметров в машинном обучении.

Критерии оценивания

оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины при ответе на контрольные вопросы и имеющему отличные результаты по контрольной работе и лабораторным работам,

оценка «отлично (9)» выставляется студенту, показавшему систематизированные, глубокие знания учебной программы дисциплины при ответе на контрольные вопросы и имеющему отличные результаты по контрольной работе и лабораторным работам,

оценка «отлично (8)» выставляется студенту, показавшему глубокие знания учебной программы дисциплины при ответе на контрольные вопросы и имеющему отличные результаты по контрольной работе и лабораторным работам,

оценка «хорошо (7)» выставляется студенту, продемонстрировавшему твердые, систематизированные знания материала при ответе на контрольные вопросы и имеющему отличные или хорошие результаты по контрольной работе и лабораторным работам,

оценка «хорошо (6)» выставляется студенту, продемонстрировавшему хорошие знания материала при ответе на контрольные вопросы и имеющему хорошие результаты по контрольной работе и лабораторным работам,

оценка «хорошо (5)» выставляется студенту продемонстрировавшему, хорошие (с минимальным количеством неверных ответов) знания материала при ответе на контрольные вопросы и имеющему хорошие результаты по контрольной работе и лабораторным работам,

оценка «удовлетворительно (4)» выставляется, если во время ответа студент показывает нетвердое знание базовых положений, связанных с материалом контрольных вопросов, но имеет хороший результат по контрольной работе и лабораторным работам,

оценка «удовлетворительно (3)» выставляется, если во время ответа студент показывает разрозненный характер знаний, нечеткие, но без грубых ошибок, формулировки базовых положений, связанных с материалом контрольных вопросов и имеет удовлетворительный результат по контрольной работе и лабораторным работам.

оценка «неудовлетворительно (2-1)» выставляется, если во время ответа на контрольные вопросы, студент показывает, что не знает большей части основного содержания курса.

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

При проведении экзамена студенту из перечня в письменной форме задаются 2-3 вопроса (не более 5), также учитываются активность студента на занятиях.

Экзамен по дисциплине является заключительным этапом изучения курса и имеет целью проверку знаний студентов по теории, а также навыков самостоятельной работы с рекомендованной литературой и интернет-ресурсами.

Экзамен проводится в устной форме.