

**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:	Методы распознавания образов
по направлению:	Прикладные математика и физика
профиль подготовки:	Компьютерное моделирование физических процессов Физтех-школа Аэрокосмических Технологий кафедра вычислительной физики
курс:	1
квалификация:	магистр

Семестры, формы промежуточной аттестации:

1 (осенний) - Дифференцированный зачет

2 (весенний) - Дифференцированный зачет

Аудиторных часов: 90 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 60 час.

Самостоятельная работа: 90 час.

Всего часов: 180, всего зач. ед.: 4

Программу составил: В.В. Рязанов, канд. физ.-мат. наук

Программа обсуждена на заседании кафедры вычислительной физики 01.06.2020

Аннотация

В рамках курса рассматриваются следующие темы: виды задач машинного обучения, метрики классификации, регрессии, линейные методы и метрические методы, бинарные решающие деревья, нейронная сеть, методы классификации текстов, задача кластеризации, существующие подходы, отбор информативных признаков, создание новых признаков, методы работы с аудио и изображениями, решение задач рекомендаций. Целью дисциплины является научить студента самостоятельно решать сложные прикладные задачи машинного обучения и анализа данных с помощью самостоятельно разработанных программных комплексов. К задачам дисциплины относятся: научить студента решать задачи машинного обучения с помощью готовых программных пакетов (pandas, numpy, Tensorflow, Pytorch, ...), научить студента решать задачи машинного обучения и анализа данных путем самостоятельной реализации методов распознавания, выбора признаков и методов валидации. После освоения курса студент будет знать основные виды задач машинного обучения; основные семейства алгоритмов, используемых для решения задач классификации, регрессии, кластеризации; способы отбора признаков, поиска информативных признаков, создания новых признаков; свойства различных метрик качества, функций потерь и способов валидации. Он будет уметь самостоятельно решать задачи анализа данных с использованием языка Python; пользоваться математическими пакетами, пакетами визуализации, пакетами глубокого обучения в Python; строить нейронные сети с помощью Keras, Tensorflow, Pytorch, а также владеть навыками выбора модели, метрики и способа валидации для конкретной задачи; навыками использования Python и внешних библиотек для решения задач анализа данных.

1. Цели и задачи

Цель дисциплины

научить студента самостоятельно решать сложные прикладные задачи машинного обучения и анализа данных с помощью самостоятельно разработанных программных комплексов.

Задачи дисциплины

1. Научить студента решать задачи машинного обучения с помощью готовых программных пакетов (pandas, numpy, Tensorflow, Pytorch, ...)
2. Научить студента решать задачи машинного обучения и анализа данных путем самостоятельной реализации методов распознавания, выбора признаков и методов валидации.

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-1 Владеет системой фундаментальных научных знаний в области физико-математических наук	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные научные знания в области физико-математических наук
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области математики и физики и способен их применять при решении задач профессиональной деятельности

ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
---	--

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

основные виды задач машинного обучения; основные семейства алгоритмов, используемых для решения задач классификации, регрессии, кластеризации; способы отбора признаков, поиска информативных признаков, создания новых признаков; свойства различных метрик качества, функций потерь и способов валидации.

уметь:

самостоятельно решать задачи анализа данных с использованием языка Python; пользоваться математическими пакетами, пакетами визуализации, пакетами глубокого обучения в Python; строить нейронные сети с помощью Keras, Tensorflow, Pytorch.

владеть:

навыками выбора модели, метрики и способа валидации для конкретной задачи; навыками использования Python и внешних библиотек для решения задач анализа данных.

Материально-техническая база: компьютерный класс.

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	Виды задач машинного обучения	2		4	6
2	Метрики классификации, регрессии	2		4	6
3	Линейные методы и метрические методы	3		6	9
4	Бинарные решающие деревья	4		8	12
5	Нейронная сеть	4		8	12
6	Методы классификации текстов	2		4	6
7	Задача кластеризации, существующие подходы	2		4	6
8	Отбор информативных признаков, создание новых признаков	3		6	9
9	Методы работы с аудио и изображениями	4		8	12
10	Решение задач рекомендаций	4		8	12
Итого часов		30		60	90
Подготовка к экзамену		0 час.			
Общая трудоёмкость		180 час., 4 зач.ед.			

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

Семестр: 1 (Осенний)

1. Виды задач машинного обучения

Задача классификации, регрессии, кластеризации. Разделение выборок. Валидация. Переобучение и data leak. Прикладной подход к решению соревновательных и других задач, этапы работы с данными. Обзор библиотек, языков программирования. Настройка кластера или вычислительной машины.

2. Метрики классификации, регрессии

Log-loss. ROC-AUC. Precision/Recall. Выбор правильной метрики исходя из условий задачи. Наилучшие константы для метрик классификации/регрессии. Визуализация данных. Поиск аномалий.

3. Линейные методы и метрические методы

Метод KNN. Линейная регрессия. Логистическая регрессия. L1/L2 регуляризация. Метод опорных векторов.

4. Бинарные решающие деревья

Обучение бинарных деревьев. Информационные критерии бинарных деревьев. Случайный лес. Градиентный бустинг. Прикладная настройка алгоритмов и выбор параметров.

5. Нейронная сеть

Нейронные сети для бинарной классификации, многоклассовой классификации и регрессии. Функции активации. Работа с пакетами Tensorflow, Keras, Pytorch. Сверточные нейронные сети, обзор различных видов внутренних слоев. Визуализация и интерпретация работы сверточной сети. Прикладные применения сверточных нейронных сетей. Рекуррентные сети.

Семестр: 2 (Весенний)

6. Методы классификации текстов

Получение признаков из текста. Этапы предобработки текста. TF-IDF. Получение эмбедингов с помощью word2vec, doc2vec. Skipgram и CBOW архитектуры. Классификация с помощью нейронных сетей. Bert, механизмы Attention.

7. Задача кластеризации, существующие подходы

Метрики качества кластеризации. EM-алгоритм. K-Means. Иерархическая кластеризация. Анализ результатов кластеризации.

8. Отбор информативных признаков, создание новых признаков

Понижение размерности выборки. PCA алгоритм. T-SNE алгоритм.

9. Методы работы с аудио и изображениями

Представление аудио в цифровом виде. Waveform и спектрограмма. Классификация аудио с помощью сверточных нейронных сетей. Классификация изображений с помощью нейронных сетей. Style transfer.

10. Решение задач рекомендаций

Коллаборативная фильтрация, content-based рекомендации.

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

Компьютер и мультимедийное оборудование (проектор, звуковая система).

6.Перечень рекомендуемой литературы

Основная литература

Фонд литературы базовой кафедры

1. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации. В сборнике "Проблемы кибернетики", М.: "Наука", 1978, стр. 6-68.
2. Журавлев Ю.И. Избранные научные труды. М.: "Магистр", 1998, 420 с.
3. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М., "Наука", 1974.
4. Ежегодник "Распознавание, классификация, прогноз (математические методы и их применение)", М.: "Наука", Вып.1 (1988), 2 (1989), 3 (1990).
5. Дуда Р.,Харт П. Распознавание образов и анализ сцен, М.: "Мир", 1976, 511 с.
6. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск, Институт математики им. С.Л.Соболева СО РАН, 1999, 268 с.

Дополнительная литература

Фонд литературы базовой кафедры

1. Хардле В. Прикладная непараметрическая регрессия. — М.: Мир. 1993.
2. Шурыгин А. М. Прикладная стохастика: робастность, оценивание, прогноз. — М. Финансы и статистика. 2000.

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

Не используются

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

На лекционных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

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

Методические рекомендации позволяют студенту оптимальным образом организовать процесс обучения. В рабочей программе приведено примерное распределение часов аудиторной и внеаудиторной нагрузки по различным темам данной дисциплины.

Для успешного освоения данной дисциплины студенту необходимо посещать лекции и семинарские занятия,

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Прикладные математика и физика
профиль подготовки:	Компьютерное моделирование физических процессов Физтех-школа Аэрокосмических Технологий кафедра вычислительной физики
курс:	<u>1</u>
квалификация:	магистр

Семестры, формы промежуточной аттестации:

- 1 (осенний) - Дифференцированный зачет
- 2 (весенний) - Дифференцированный зачет

Разработчик: В.В. Рязанов, канд. физ.-мат. наук

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-1 Владеет системой фундаментальных научных знаний в области физико-математических наук	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные научные знания в области физико-математических наук
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области математики и физики и способен их применять при решении задач профессиональной деятельности
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности

2. Показатели оценивания компетенций

В результате изучения дисциплины «Методы распознавания образов» обучающийся должен:

знать:

основные виды задач машинного обучения; основные семейства алгоритмов, используемых для решения задач классификации, регрессии, кластеризации; способы отбора признаков, поиска информативных признаков, создания новых признаков; свойства различных метрик качества, функций потерь и способов валидации.

уметь:

самостоятельно решать задачи анализа данных с использованием языка Python; пользоваться математическими пакетами, пакетами визуализации, пакетами глубокого обучения в Python; строить нейронные сети с помощью Keras, Tensorflow, Pytorch.

владеть:

навыками выбора модели, метрики и способа валидации для конкретной задачи; навыками использования Python и внешних библиотек для решения задач анализа данных.

Материально-техническая база: компьютерный класс.

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

Типовые контрольные задания

1. Виды задач машинного обучения
2. Метрики классификации, регрессии
3. Линейные методы и метрические методы
4. Бинарные решающие деревья
5. Нейронная сеть
6. Методы классификации текстов

7. Задача кластеризации, существующие подходы
8. Отбор информативных признаков, создание новых признаков
9. Методы работы с аудио и изображениями
10. Решение задач рекомендаций

Критерии оценивания:

отлично (10) - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

отлично (9) - выставляется студенту, показавшему свободное оперирование знаниями учебной программы дисциплины, выполнение заданий творческого характера.

отлично (8) - выставляется студенту, показавшему владение программным учебным материалом с наличием несущественных ошибок в действиях, самостоятельно исправляемых учащимся.

хорошо (7) - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускается в ответе или в решении задач некоторые неточности.

хорошо (6) - выставляется студенту если он осознает воспроизведение программного учебного материала, в том числе и различной степени сложности, с несущественными ошибками, затруднения в применении отдельных навыков.

хорошо (5) - выставляется студенту если теоретическое содержание освоено не полностью, некоторые практические навыки сформированы недостаточно, в некоторых случаях были допущены ошибки.

удовлетворительно (4) - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации.

удовлетворительно (3) - выставляется студенту в случае большого количества недочетов и неправильных ответов, а также пассивной работе в ходе занятий, многие учебные задания не выполнены.

неудовлетворительно (2) - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

неудовлетворительно (1) - выставляется студенту, который не освоил теоретическое и практическое содержание курса, все выполненные учебные задания содержат грубые ошибки.

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

Примеры типовых вопросов

9ый семестр:

1. Обосновать выбор схемы валидации или метрики на примере практической задачи (2 балла)
2. Написать программный комплекс с загрузкой данных, разделением выборки, обучением модели, предсказанием для новых объектов (4 балла)
3. Объяснить работу и математическую модель одного из методов классификации или регрессии (случайный лес, логистическая регрессия, нейронная сеть и т.д.) (4 балла)

10ый семестр:

Написание нейронной сети для классификации аудиосообщений

1. Сбор аудио-данных, загрузка, разделение объектов на классы.
 2. Написание сверточной нейронной сети, разделение выборок на обучающую, валидационную, тестовую. Обоснование функции потерь нейронной сети.
 3. Развитие архитектуры нейронной сети, аугментация объектов, обработка объектов с целью улучшения качества распознавания или скорости работы/обучения.
1. Выбор задачи (соревнование Kaggle, собственный проект и т.п.), сбор и подготовка данных (3 балла)
 2. Выбор модели и ее настройка, выбор схемы валидации, метрики (3 балла)

3. Отбор признаков, составление новых признаков для улучшения показателей метрики. (4 балла)

Критерии оценивания

отлично (10) - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

отлично (9) - выставляется студенту, показавшему свободное оперирование знаниями учебной программы дисциплины, выполнение заданий творческого характера.

отлично (8) - выставляется студенту, показавшему владение программным учебным материалом с наличием несущественных ошибок в действиях, самостоятельно исправляемых учащимся.

хорошо (7) - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускается в ответе или в решении задач некоторые неточности.

хорошо (6) - выставляется студенту если он осознает воспроизведение программного учебного материала, в том числе и различной степени сложности, с несущественными ошибками, затруднения в применении отдельных навыков.

хорошо (5) - выставляется студенту если теоретическое содержание освоено не полностью, некоторые практические навыки сформированы недостаточно, в некоторых случаях были допущены ошибки.

удовлетворительно (4) - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации.

удовлетворительно (3) - выставляется студенту в случае большого количества недочетов и неправильных ответов, а также пассивной работе в ходе занятий, многие учебные задания не выполнены.

неудовлетворительно (2) - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

неудовлетворительно (1) - выставляется студенту, который не освоил теоретическое и практическое содержание курса, все выполненные учебные задания содержат грубые ошибки.

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

Промежуточная аттестация по дисциплине «Методы распознавания образов» осуществляется в форме дифференцированного зачета.

Оценка выставляется по сумме баллов, набранных за выполнение заданий в течение семестра.