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

УТВЕРЖДЕНО
Директор физтех-школы физики и исследований им. Ландау
А.В. Рогачев

Рабочая программа дисциплины (модуля)

по дисциплине: Методы машинного обучения в астрофизике
по направлению: Прикладные математика и физика
профиль подготовки: Общая и прикладная физика
Физтех-школа физики и исследований им. Ландау
кафедра фундаментальных взаимодействий и космологии
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

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

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

Программу составил: О.Е. Калашев, д-р физ.-мат. наук

Программа обсуждена на заседании кафедры фундаментальных взаимодействий и космологии 04.06.2021

Аннотация

Курс посвящен современным методам машинного обучения и алгоритмам анализа данных в парадигме Big Data. На примере актуальных задач физики частиц и астрофизики будут рассмотрены все этапы обработки данных от постановки вопроса до конструирования вычислительного алгоритма. Будут детально рассмотрены принципы построения и механизмы работы нейронных сетей. Курс сопровождается семинарами, проходящими в формате мастер-класса по анализу данных.

1. Цели и задачи

Цель дисциплины

Изучение методов машинного обучения для анализа данных в задачах астрофизики.

Задачи дисциплины

- Знакомство с анализом данных методами машинного обучения, постановка задач и интерпретация результатов
- Освоение методов машинного обучения “с учителем” для задач классификации и регрессии
- Освоение методов машинного обучения “без учителя” для задач структуризации данных и поиска аномалий

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-1.3 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
УК-2 Способен управлять проектом на всех этапах его реализации	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
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-1.3 Понимает междисциплинарные связи в области математики и физики и способен их применять при решении задач профессиональной 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-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, понимает и учитывает на практике границы применимости получаемых решений

ОПК-4 Способен успешно реализовывать решение поставленной задачи, провести анализ результата и представить выводы, применяя знания и навыки в области физико-математических наук и информационно-коммуникационных технологий	ОПК-4.3 Способен аргументировано выбирать способ проведения научного исследования
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

методы машинного обучения «с учителем», и «без учителя» и примеры их применения в астрофизике.

уметь:

применять методы машинного обучения для реальных задач регрессии, классификации и кластеризации в астрофизике.

владеть:

инструментарием для решения задач с помощью программирования на языке python.

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 в машинное обучение.	2			1
2	Введение в анализ данных на Python.	2			1
3	Линейные модели регрессии и классификации.	2			1
4	Деревья решений и метод ближайших соседей.	2			1
5	Пример задачи классификации из астрофизики.	2			1
6	Ансамблевые методы 1.	2			1
7	Ансамблевые методы 2.	2			1
8	Нейронные сети прямого распространения.	2			1
9	Оптимизация нейронных сетей	2			1
10	Сверточные нейронные сети	2			1
11	Применение сверточных нейронных сетей	2			1
12	Модификации сверточных нейронных сетей, используемые в астрофизике.	2			1
13	Рекуррентные нейронные сети	2			1
14	Применение рекуррентных нейронных сетей в астрофизике	2			1

15	Кластеризация.	2			1
Итого часов		30			15
Подготовка к экзамену		0 час.			
Общая трудоёмкость		45 час., 1 зач.ед.			

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

Семестр: 2 (Весенний)

1. Введение в машинное обучение.

Понятие ‘big data’, задачи обработки данных: регрессия, классификация, кластеризация, поиск аномалий, обучение представлениям, уменьшение размерности данных. Пример алгоритма машинного обучения – метод ближайших соседей.

2. Введение в анализ данных на Python.

Основные элементы и конструкции языка Python. Среда разработки jupyter notebook, документирование кода и представление результатов численных расчетов в воспроизводимом виде. Анализ данных и инструменты визуализации. Библиотеки numpy, pandas matplotlib.

3. Линейные модели регрессии и классификации.

Линейная регрессия. Метод наименьших квадратов. Оценка точности модели. Функция цены и метод градиентного спуска. Логистическая регрессия.

4. Деревья решений и метод ближайших соседей.

Как строится дерево решений. Энтропия Шеннона. Пример из библиотеки scikit-learn. Визуализация дерева решений.

5. Пример задачи классификации из астрофизики.

Подготовка данных. Метрики качества классификатора. Интерпретация результатов. Выбор параметров модели и кросс-валидация.

6. Ансамблевые методы 1.

Композиция алгоритмов. Бутстрэп-агрегирование. Случайный лес.

7. Ансамблевые методы 2.

Градиентный бустинг. Постановка задачи. Функциональный градиентный спуск. Алгоритм Фридмана. Пошаговый пример работы.

8. Нейронные сети прямого распространения.

Биологические нейронные сети. Персептрон. Многослойный персептрон. Обучение нейронных сетей. Алгоритм обратного распространения ошибки.

9. Оптимизация нейронных сетей

Проблема переобучения и методы регуляризации. Глубокие нейронные сети, проблема обнуления градиентов и способы борьбы с ней.

10. Сверточные нейронные сети

Анализ изображений. Операции свертки и масштабирования. Архитектура сверточных нейронных сетей.

11. Применение сверточных нейронных сетей

Примеры задач регрессии и классификации изображений из астрономии и астрофизики частиц.

12. Модификации сверточных нейронных сетей, используемые в астрофизике.

Операции свертки на непрямоугольных решетках. Сверточные нейронные сети на сфере. Анализ направлений прихода космических лучей сверхвысоких энергий.

13. Рекуррентные нейронные сети

Регрессия на временных рядах. Рекуррентные нейронные сети; архитектура Long Short-Term Memory (LSTM), парадигма seq2seq

14. Применение рекуррентных нейронных сетей в астрофизике

Детектор гравитационных волн LIGO. Использование LSTM для прогнозирования фона и выделения сигнала о грав. волн

15. Кластеризация.

Постановка задачи. Методы кластеризации. Задача поиска аномалий.

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

Учебная аудитория, доска, медиапроектор, экран.

6.Перечень рекомендуемой литературы

Основная литература

1. Открытый курс машинного обучения. Open Data Science на habr
2. Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015
3. К. В. Воронцов, Лекции по алгоритмам кластеризации и многомерного шкалирования.

Дополнительная литература

1. D. Ivanov et al, “Using deep learning to enhance event geometry reconstruction for the telescope array surface detector”, Machine Learning: Science and Technology 2020, http://iopscience.iop.org/10.1088/2632-2153/abae74
2. O. Kalashev et al, “Identifying nearby sources of ultra-high-energy cosmic rays with deep learning”, JCAP 2020, e-Print: 1912.00625
3. Tri Nguyen et al. “Extending the reach of gravitational-wave detectors with machine learning”, LIGO Document P1800129-v1
4. N. Krachmalnicoff et al. “Convolutional neural networks on the HEALPix sphere: a pixel-based algorithm and its application to CMB data analysis”, A&A Vol 628, August 2019 e-Print: 1902.04083
5. Idan Shilon et al., “Application of Deep Learning methods to analysis of Imaging Atmospheric Cherenkov Telescopes data.”, Astropart.Phys. 105 (2019) 44-53, e-Print: 1803.10698

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

1. Курс лекций. Машинное обучение и анализ данных. МФТИ, Яндекс. https://www.coursera.org/specializations/machine-learning-data-analysis
2. Курс лекций. Python для анализа данных, МФТИ, ФРОО, Mail.ru Group https://www.coursera.org/learn/python-for-data-science
3. Курс лекций. Нейронные сети. Институт биоинформатики. https://stepik.org/course/401/promo
4. Базовый курс neurohive.io https://neurohive.io/ru/osnovy-data-science/
5. Machine Learning Yearning, Andrew Ng https://www.deeplearning.ai/

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

- Python
- jupyter notebook
- google colab

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

Студент, изучающий дисциплину, должен, с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

В результате изучения дисциплины студент должен знать основные определения и понятия, уметь применять полученные знания для решения различных задач.

Успешное освоение курса требует:

– посещения всех занятий, предусмотренных учебным планом по дисциплине;

– ведения конспекта занятий;

– напряжённой самостоятельной работы студента.

Самостоятельная работа включает в себя:

– чтение рекомендованной литературы;

– проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;

– решение задач, предлагаемых студентам на занятиях;

– подготовку к выполнению заданий текущей и промежуточной аттестации.

Показателем владения материалом служит умение без конспекта отвечать на вопросы по темам дисциплины.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями преподавателю.

Возможен промежуточный контроль знаний студентов в виде решения задач в соответствии с тематикой занятий.

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 фундаментальных взаимодействий и космологии
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

Разработчик: О.Е. Калашев, д-р 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-1.3 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
УК-2 Способен управлять проектом на всех этапах его реализации	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
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-1.3 Понимает междисциплинарные связи в области математики и физики и способен их применять при решении задач профессиональной 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-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и ограничения различных методов решения	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.3 Владеет аналитическими и вычислительными методами решения, понимает и учитывает на практике границы применимости получаемых решений
ОПК-4 Способен успешно реализовывать решение поставленной задачи, провести анализ результата и представить выводы, применяя знания и навыки в области физико-математических наук и информационно-коммуникационных технологий	ОПК-4.3 Способен аргументировано выбирать способ проведения научного исследования
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

2. Показатели оценивания компетенций

В результате изучения дисциплины «Методы машинного обучения в астрофизике» обучающийся должен:

знать:

методы машинного обучения «с учителем», и «без учителя» и примеры их применения в астрофизике.

уметь:

применять методы машинного обучения для реальных задач регрессии, классификации и кластеризации в астрофизике.

владеть:

инструментарием для решения задач с помощью программирования на языке python.

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

1. С помощью линейной регрессии аппроксимировать функцию sin(x) + 0.1ξ , где ξ - случайный шум с нормальным распределением, полиномом пятой степени в интервале от 0 до pi/2. . Визуализировать результат.
2. С помощью дерева решений с глубиной от 1 до 10 аппроксимировать функцию sin(x) + 0.1ξ , где ξ - случайный шум с нормальным распределением, в интервале от 0 до pi/2. Визуализировать результат. Сравнить устойчивость деревьев разной глубины к случайным выбросам в данных.

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

Перечень контрольных вопросов.

1. Что такое переобучение. Как с ним бороться на примере нейронных сетей и деревьев решений.
2. При оптимизации метапараметров алгоритмов классификации, например глубины дерева или числа нейронов в промежуточном слое сети используется отложенная выборка, по которой не происходит обучение. Что произойдет, если в примерах отложенной выборки будет дисбаланс классов. Каким образом можно с этим бороться.
3. В тестовой выборке образцов для задачи бинарной классификации имеется N примеров класса A и M примеров класса B. Обученный классификатор имеет точность (accuracy) 90%. Как понять много это или мало?
4. Можно ли любую сверточную сеть представить как многослойный персептрон?
5. Можно ли многослойный персептрон представить как сверточную сеть?
6. В чем основное преимущество блоков Long Short-Term Memory (LSTM) перед простейшими рекуррентными нейронными сетями.

Примеры контрольных заданий

1. В табличном файле приведены наблюдаемые признаки и тип (активные ядра галактик или пульсары) объектов, наблюдаемых телескопом FERMI LAT. Обучить классификатор, основанный на деревьях решений и оценить его точность. Определить типы объектов признаки которых приведены во втором файле. Какие из признаков – самые важные?

2. В табличном файле приведены наблюдаемые признаки и тип (активные ядра галактик или пульсары) объектов, наблюдаемых телескопом FERMI LAT. Обучить классификатор, основанный на многослойном персептроне и оценить его точность. Определить типы объектов признаки которых приведены во втором файле.

3. На сайте Королевской обсерватории Бельгии http://www.sidc.be/silso/datafiles размещены исторические данные

о числе пятен на Солнце с 1 января 1818 года по настоящее время. Известно, что солнечная активность циклична с периодом около 11 лет. Кроме того, в солнечной активности существуют закономерности на больших и меньших временных масштабах. Таким образом предсказание числа пятен на следующий день может опираться на данные за 22 прошедших года, то есть на более, чем 8000 известных значений. Построить нейронную сеть, решающую данную задачу.

4. Эксперимент Telescope Array регистрирует широкие атмосферные ливни, вызванные космическими лучами ультравысоких энергий. Предполагается, что первичные частицы - протоны или ядра химических элементов. Для каждого события в результате реконструкции определяются 16 наблюдаемых параметров (описание физического смысла параметров см. в работе https://arxiv.org/abs/1808.03680. По адресу ftp://cluster.inr.ac.ru/pub/ML/TASD/ размещены результаты Монте-Карло моделирования эксперимента для ШАЛ, вызванных первичными протонами (p.dat), ядрами гелия (he.dat), азота (n.dat) и железа (fe.dat). Кроме того, размещены три неизвестных смеси (unknown1.dat), (unknown2.dat), (unknown3.dat). Известно, что в первом неизвестном наборе присутствуют только протоны и железо.

Определить состав первичных частиц в каждом из трех неизвестных наборов.

5. В эксперименте Telescope Array проводится поиск гамма-квантов ультравысоких энергий E>100 ЭэВ. Для каждого события в результате реконструкции определяются 16 наблюдаемых параметров (описание физического смысла параметров см. в работе http://arxiv.org/abs/arXiv:1811.03920. По адресу ftp://cluster.inr.ac.ru/pub/ML/TASD/ размещены результаты Монте-Карло моделирования эксперимента для ШАЛ, вызванных первичными протонами, task5_p.dat и гамма-квантами, task5_gamma.dat. Кроме того, размещен неизвестный набор task5_unknown.dat, в котором присутствует несколько событий, вызванных гамма-квантами. Оценить число гамма-квантов в неизвестном наборе.

Критерии оценивания

Оценка отлично 10 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 9 баллов - 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.

Оценка отлично 8 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо 7 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо 6 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

Оценка хорошо 5 баллов - выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно 4 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка удовлетворительно 3 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

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

Оценка неудовлетворительно 1 балл - выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

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

Дифференцированный зачёт проводятся в устной форме. При проведении дифференцированного зачёта обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.