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

УТВЕРЖДЕНО
и.о. директора физтех-школы физики и исследований им. Ландау
А.А. Воронов

Рабочая программа дисциплины (модуля)

по дисциплине:	Машинное обучение и анализ данных
по направлению:	Прикладные математика и физика
профиль подготовки:	Общая и прикладная физика
	Физтех-школа физики и исследований им. Ландау
	кафедра дискретной математики
курс:	2
квалификация:	магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Экзамен

Аудиторных часов: 60 всего, в том числе:
- лекции: 30 час.
- семинары: 30 час.
- лабораторные занятия: 0 час.

Самостоятельная работа: 45 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Программу составил: Н.А. Волков, ассистент

Программа обсуждена на заседании кафедры дискретной математики 04.06.2020

Аннотация

Курс знакомит слушателей с основными современными методами машинного обучения, включая популярные методы, такие как градиентный бустинг и нейронные сети. Помимо моделей машинного обучения в курсе также освещаются способы оценки качества моделей, методы выбора наилучшей модели. Слушатели учатся также правильно оформлять результаты своих исследований и делать обоснования построенных моделей. В рамках курса предполагается практика на языке Python, включая работу с современными библиотеками машинного обучения. В качестве примера реального приложения машинного обучения, в курсе делается обзор платформ по анализу данных для спортивных соревнований, например, Kaggle.

1. Цели и задачи

Цель дисциплины

Познакомить слушателей с основными задачами машинного обучения и современными методами, включая популярные методы, такие как градиентный бустинг и нейронные сети. Научить работе с широко известными библиотеками машинного обучения. Сформировать навыки решения задач машинного обучения с использованием программного обеспечения и языка программирования Python.

Задачи дисциплины

- изучение методов машинного обучения;
- изучение моделей машинного обучения;
- изучение способов оценки качества моделей и методов выбора наилучшей модели;
- знакомство с нейронными сетями и методами градиентного бустинга;
- приобретение навыков оформления результатов исследования и обоснования построенных моделей;
- приобретение практических навыков программирования на языке Python, включая работу с современными библиотеками машинного обучения;
- знакомство с платформами по анализу данных для спортивных соревнований, в т.ч., Kaggle.

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
УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1 Умеет решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности
	УК-6.2 Оценивает свою деятельность, соотносит цели, способы и средства выполнения деятельности с её результатами
ОПК-3 Способен выбирать и (или) разрабатывать подходы к решению типовых и новых задач в области профессиональной деятельности, учитывая особенности и	ОПК-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)

ограничения различных методов 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-4.1 Способен применять знания и навыки по использованию информационно-коммуникационных технологий для поиска и изучения научной литературы, применения прикладных программных продуктов
	ОПК-4.2 Способен применять знания в области физико-математических наук для решения поставленной задачи, формулирования выводов и оценки полученных результатов
	ОПК-4.3 Способен аргументировано выбирать способ проведения научного исследования
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
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)
	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:

- основные методы машинного обучения;
- основные модели машинного обучения;
- способы оценки качества моделей и методов выбора наилучшей модели;
- основные объекты, процедуры и библиотеки языка Python, необходимые для решения задач машинного обучения.

уметь:

- обосновывать оценку качества модели и метода выбора наилучшей модели;
- оформлять результаты исследования;
- работать с современными библиотеками машинного обучения;
- применять основные объекты и процедуры языка Python, необходимые для решения задач прикладной статистики.

владеть:

- основными методами машинного обучения.
- навыками выбора наилучшей модели для машинного обучения;
- средствами разработки и тестирования программного кода на языке Python, объектами и средствами, предлагаемыми стандартными библиотеками, необходимыми для решения задач машинного обучения.

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	Основные задачи машинного обучения	2	2		3
2	Регуляризация в линейной регрессии	4	4		6
3	Логистическая регрессия	2	2		3
4	Метод главных компонент	2	2		3
5	Решающие деревья	2	2		3
6	Случайные леса	2	2		3
7	Бустинг	2	2		3
8	Продвинутые методы построения композиций	2	2		3
9	Работа с признаками	2	2		3
10	SVM (Метод опорных векторов)	2	2		3
11	Кластеризация	2	2		3
12	Нейронные сети	4	4		6
13	Свёрточные нейронные сети	2	2		3
Итого часов		30	30		45
Подготовка к экзамену		30 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 3 (Осенний)

1. Основные задачи машинного обучения

Регрессия, классификация, кластеризация.

2. Регуляризация в линейной регрессии

Мультиколлинеарность, ее признаки. Регуляризация в линейной регрессии: ридж и лассо регрессии, свойства моделей. Формула оценки коэффициентов в ридж-регрессии. Эквивалентные задачи условной оптимизации и их интерпретация.

3. Логистическая регрессия

Свойства логистической функции. Постановка задачи логистической регрессии, формула градиентного спуска, стохастический градиентный спуск. Переобучение модели.

4. Метод главных компонент

Причины избыточности информации в данных, типы методов снижения размерности. Метод главных компонент (PCA) как выбор направлений с максимальной дисперсией, формулы перехода в сжатое пространство и обратно. Дисперсии образа, выбор размерности сжатого пространства на основе доли необъясненной дисперсии.

5. Решающие деревья

Решающее дерево, процесс его построения и выбора разбиения в вершине. Критерии информативности (gini, энтропийный, MSE). Критерии останова, выбор ответа в листе.

6. Случайные леса

Bias-variance разложение. Беггинг, b-v разложение для него.
Случайный лес, анализ случайного леса при помощи b-v разложения. Out-of-Bag ошибка.

7. Бустинг

Бустинг, его построение и построение новой базовой модели. Формулы для сдвигов для разных функций потерь. Аналогия с градиентным спуском. Сокращение шага. Стохастический градиентный спуск. Анализ бустинга с помощью b-v разложения.
Градиентный бустинг над деревьями, перенастройка ответов в листьях.

8. Продвинутые методы построения композиций

XGBoost, CatBoost, LightGBM, стеккинг, варианты его обучения. Блендинг, StackNet.

9. Работа с признаками

Работа с числовыми признаками (трансформации, квантование). Работа с порядковыми и категориальными признаками. Mean encoding, регуляризация для него.

10. SVM (Метод опорных векторов)

Оптимальная разделяющая гиперплоскость для линейно разделимых классов, ее ширина полосы. Постановка задачи SVM в линейно разделимом случае и в общем случае. Двойственная, классификация объектов на три типа. Ядерный трюк в задаче SVM, свойства ядра, примеры. История SVM.

11. Кластеризация

Задача кластеризации. Метрика качества задачи кластеризации. K-means, оптимизируемый функционал, начальные приближения. Mini-batch K-means, K-means++. EM-алгоритм. DBSCAN.

12. Нейронные сети

Модель нейрона, однослойная сеть. Метод обратного распространения ошибки для двухслойной нейросети. Методы оптимизации для нейронных сетей. Функции активации. Проблема затухания и взрыва градиента. Dropout. Batch Normalization. Автоенкодеры.

13. Свёрточные нейронные сети

Свёртка, padding, stride. Receptive field, интерпретация нейронов с помощью receptive field. Pooling.

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

Компьютерный класс, оснащенный компьютерами с установленным ПО среды программирования Python. Или стандартная учебная аудитория и ноутбуки с установленным ПО среды программирования Python.

6.Перечень рекомендуемой литературы

Основная литература

1. Машинное обучение [Текст]/Х. Бринк, Дж. Ричардс, М. Феверолф, Real-World Machine Learning, -СПб., Питер, 2017

2. Python и машинное обучение [Текст] = Python Machine Learning : крайне необходимое издание по новейшей предсказательной аналитике для более глубокого понимания методологии машинного обучения / С. Рашка; пер. с англ. А. В. Логунова .— М. : ДМК Пресс, 2017 .— 418 с.: ил. - Предм. указ.: с. 408-417. - 200 экз. - ISBN 978-5-97060-409-0 (в пер.) .— Полный текст (Доступ из сети МФТИ / Удаленный доступ).

Дополнительная литература

1. Математические основы машинного обучения и прогнозирования [Текст] / В. В. Вьюгин ; Моск. физ.-техн. ин-т (гос. ун-т), Лаб. структурных методов анализа данных в предсказательном моделировании (ПреМоЛаб), Ин-т проблем передачи информации им. А. А. Харкевича РАН - М.МЦНМО,2013

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 занятиях используются мультимедийные технологии, включая демонстрацию презентаций. Также занятия могут проходить в дистанционном виде посредством видеоконференции.

Среда программирования: Python.

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.

Самостоятельные задания могут выдаваться и оцениваться дистанционно.

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 дискретной математики
курс: 2
квалификация: магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Экзамен

Разработчик: Н.А. Волков, 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
УК-6 Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1 Умеет решать задачи собственного личностного и профессионального развития, определять и реализовывать приоритеты совершенствования собственной деятельности
	УК-6.2 Оценивает свою деятельность, соотносит цели, способы и средства выполнения деятельности с её результатами
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-3.1 Способен анализировать задачу, планировать пути решения, предлагать и комбинировать способы решения
	ОПК-3.2 Способен использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки (техники)
	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-4.1 Способен применять знания и навыки по использованию информационно-коммуникационных технологий для поиска и изучения научной литературы, применения прикладных программных продуктов
	ОПК-4.2 Способен применять знания в области физико-математических наук для решения поставленной задачи, формулирования выводов и оценки полученных результатов
	ОПК-4.3 Способен аргументировано выбирать способ проведения научного исследования
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
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-3 Способен профессионально работать с исследовательским и испытательным оборудованием (приборами и установками, специализированными пакетами прикладных программ) в избранной предметной области	ПК-3.1 Понимает принципы работы используемого оборудования (специализированных пакетов прикладных программ)
	ПК-3.2 Способен проводить эксперимент (моделирование) с использованием исследовательского оборудования (пакетов прикладных программ)

программ) в избранной предметной области	ПК-3.3 Способен оценивать точность полученных экспериментальных (численных) результатов

2. Показатели оценивания компетенций

В результате изучения дисциплины «Машинное обучение и анализ данных» обучающийся должен:

знать:

- основные методы машинного обучения;
- основные модели машинного обучения;
- способы оценки качества моделей и методов выбора наилучшей модели;
- основные объекты, процедуры и библиотеки языка Python, необходимые для решения задач машинного обучения.

уметь:

- обосновывать оценку качества модели и метода выбора наилучшей модели;
- оформлять результаты исследования;
- работать с современными библиотеками машинного обучения;
- применять основные объекты и процедуры языка Python, необходимые для решения задач прикладной статистики.

владеть:

- основными методами машинного обучения.
- навыками выбора наилучшей модели для машинного обучения;
- средствами разработки и тестирования программного кода на языке Python, объектами и средствами, предлагаемыми стандартными библиотеками, необходимыми для решения задач машинного обучения.

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

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия или в конце занятия по пройденной теме.

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

Список вопросов:

1. Чем отличаются задачи обучения с учителем от обучения без учителя?
2. На какие виды делится обучение с учителем?
3. Что такое главные компоненты в методе главных компонент?
4. Как при построении случайного леса получаются разные решающие деревья?
5. Имеет ли смысл проводить беггинг над линейной регрессией?
6. Какой глубины стоит брать решающие деревья для построения бустинга? А для случайного леса?
7. В каком пространстве происходит градинентный спуск при построении градиентного бустинга?
8. В чем отличие XGBoost от простого градиентного бустинга?
9. Можно ли с помощью SVM получить нелинейный классификатор?
10. Что такое нейрон?
11. Из чего состоит слой в нейронной сети?
12. В чем проблема затухания градиента в нейронной сети?

Примеры контрольных заданий:

1. Для каждого X i из обучающей выборки определим множество X i , которое состоит из всех объектов x, для которых X i ближайший к x объект из обучающей выборки. Доказать, что X i является выпуклым многогранником.
2. Приведите пример задачи обучения с учителем, для которой метод Add для отбора признаков дает неоптимальный набор, то есть существует набор из меньшего количества признаков, дающий такое же или лучшее качество.

3. Пусть (X 1 , Y 1), ..., (X m , Y m) — выборка, X i — некоторый объект, а Y∉ {0, 1} — метка класса. В этой выборке P объектов класса 1 и N объектов класса 0, m = P +N . Предикат φ выделяет p объектов класса 1 и n объектов класса 0. Покажите эквивалентность статистического и энтропийного критериев при m → +∞, то есть IStat(p, n)~ IGain(p, n). Вероятности в определении энтропийного критерия оцените частотным способом.

4. Рассмотрим модель градиентного бустинга. Выписать формулы антиградиента, по которым обучается новая компонента в модели, а так же шаг, с которым она добавляется в модель для квадратичная и гауссовской функций потерь.

5. Покажите, что логистическая регрессия является линейным методом классификации, то есть поверхность, разделяющая все пространство на два класса при классифкации, линейна.

6. Предложите способ, как можно применить метод SVM для многоклассовой классификации.

Пример билета экзамена:

1. Приведите пример задачи обучения с учителем, для которой метод Add для отбора признаков дает неоптимальный набор, то есть существует набор из меньшего количества признаков, дающий такое же или лучшее качество.

2. Покажите, что логистическая регрессия является линейным методом классификации, то есть поверхность, разделяющая все пространство на два класса при классифкации, линейна.

Критерии оценивания

Оценка «отлично (10)» выставляется обучающемуся, если он показал всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенноприменять их на практике при решении конкретных задач, свободное и правильноеобоснование принятых решений;

оценка «отлично (9)» выставляется обучающемуся, если он показал всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенноприменять их на практике при решении конкретных задач, свободное и правильноеобоснование принятых решений, но при этом были допущены небольшие неточности, которыебыли самостоятельно обнаружены и исправлены;

оценка «отлично (8)» выставляется обучающемуся, если он показал всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенноприменять их на практике при решении конкретных задач, свободное и правильноеобоснование принятых решений, но при этом были допущены небольшие неточности, которые после указания экзаменатора были самостоятельно исправлены;

оценка «хорошо (7)» выставляется обучающемуся, если он твердо знает материал, грамотно ипо существу излагает его, умеет применять полученные знания на практике, но допускаетнеточности в ответе или делает несущественные ошибки при решении задач;

оценка «хорошо (6)» выставляется обучающемуся, если он твердо знает материал, грамотно ипо существу излагает его, умеет применять полученные знания на практике, но допускаетнебольшие ошибки в ответе и (или) при решении задач;

оценка «хорошо (5)» выставляется обучающемуся, если он твердо знает материал, грамотно ипо существу излагает его, умеет применять полученные знания на практике, но отвечаетнеуверенно и (или) допускает ошибки при решении задач;

оценка «удовлетворительно (4)» выставляется обучающемуся, показавшему фрагментарный, разрозненный характер знаний, неточные формулировки базовых понятий, нарушениялогической последовательности в изложении программного материала, если при этом онвладеет основными разделами учебной программы, необходимыми для дальнейшего обученияи может применять полученные знания по образцу в стандартной ситуации;

оценка «удовлетворительно (3)» выставляется обучающемуся, показавшему фрагментарный, разрозненный характер знаний, неточные формулировки базовых понятий, нарушениялогической последовательности в изложении программного материала, не владеющемунекоторыми разделами учебной программы, но умеющему применять полученные знания пообразцу в стандартной ситуации;

оценка «неудовлетворительно (2)» выставляется обучающемуся, который не знает большейчасти основного содержания учебной программы дисциплины, допускает грубые ошибки вформулировках основных понятий дисциплины и не умеет использовать полученные знанияпри решении типовых практических задач;

оценка «неудовлетворительно (1)» выставляется обучающемуся, показавшему полное незнание учебной программы дисциплины.

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

Итоговая аттестация по дисциплине «Машинное обучение и анализ данных» осуществляется в форме экзамена. Экзамен проводится в устной форме по билетам. При проведении экзамена обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.