

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

УТВЕРЖДЕНО

**Директор физтех-школы
электроники, фотоники и
молекулярной физики**

В.В. Иванов

| | |
|----------------------------|---|
| | Рабочая программа дисциплины (модуля) |
| по дисциплине: | Введение в машинное обучение и анализ данных |
| по направлению: | Прикладные математика и физика |
| профиль подготовки: | Физика перспективных технологий: альтернативная энергетика, научное программирование и функциональные материалы Физтех-школа Электроники, Фотоники и Молекулярной Физики кафедра физики высокотемпературных процессов |
| курс: | 2 |
| квалификация: | бакалавр |

Семестр, формы промежуточной аттестации: 4 (весенний) - 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: К.Д. Гольдштейн

Программа обсуждена на заседании кафедры физики высокотемпературных процессов 01.06.2022

Аннотация

Машинное обучение и анализ данных занимают все более важное место в науках, включая физику, также не менее важны эти дисциплины для современной подготовки инженеров. В рамках курса будет дан обзор методов машинного обучения, используемых в реальных научных и инженерных задачах и приведены практические работы для самостоятельного выполнения на языке Python.

1. Цели и задачи

Цель дисциплины

Подготовка будущего физика-исследователя, владеющего современными методами работы с научными данными

Задачи дисциплины

Дисциплина рассматривается как курс, по окончании которого студенты должны овладеть современными методиками анализа данных, необходимыми для работы над прикладными задачами, возникающими перед учеными в области прикладных наук. Курс обобщает и систематизирует представления студентов о программировании; дает новые знания из области прикладной математики и применения методов программирования для решения математических и физических задач.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

| Код и наименование компетенции | Индикаторы достижения компетенции |
|--|--|
| УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач | УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи |
| | УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки |
| | УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи |
| | УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки |
| УК-6 Способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни | УК-6.2 Способен планировать самостоятельную деятельность в решении профессиональных задач; подвергать критическому анализу проделанную работу; находить и творчески использовать имеющийся опыт в соответствии с задачами саморазвития |
| ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности | ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения |
| | ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки |
| | ОПК-1.3 Способен определять границы применимости полученных результатов |

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- Типы моделей машинного обучения, используемых в прикладных физических задачах
- Формулировки наиболее важных математических утверждений из теории анализа данных, способы их применения при помощи Python
- Сферы приложений конкретных моделей машинного обучения

уметь:

- Применять математические и статистические методы для решения прикладных задач.
- Использовать методы программирования и прикладной математики в научных задачах.

владеть:

- Базовыми методами машинного обучения и обработки данных.

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 | Математические основы анализа данных | 4 | | | 2 |
| 2 | Регрессионные модели | 10 | | | 5 |
| 3 | Обучение без учителя | 6 | | | 3 |
| 4 | Глубокое обучение | 10 | | | 5 |
| 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: 4 (Весенний)

1. Математические основы анализа данных

Тема 1.1. Теория вероятностей и математическая статистика в анализе данных

1. Постановка задачи машинного обучения, классификация задач
2. Используемые метрики в анализе данных
3. Метрики качества регрессии, метод максимального правдоподобия

Тема 1.2 Feature Engineering, предобработка данных, валидация результатов

1. Подготовка данных. Embedding. Примеры способов обработки сложных объектов.
2. Методы валидации. Кросс-валидация. Bias-variance дихотомия. Переобучение
3. Методы оптимизации

2. Регрессионные модели

Тема 2.1. Линейная регрессия

1. Аналитическое решение, градиентное решение задачи линейной регрессии.
2. Регуляризация L1 и L2.

Тема 2.2. Логистическая регрессия

Метод максимального правдоподобия.

Минимизация логистической функции потерь. Эквивалентность решений.

Множественные классификации.

Тема 2.3. Регрессия на гауссовских процессах

1. Понятие гауссовского процесса

2. Использование регрессии гауссовских процессов для решения задач анализа данных

3. Обучение без учителя

Тема 3.1. SVD-разложение, метод главных компонент?

1. Формулировка задачи понижения размерности
2. Переход от SVD разложения к методу главных компонент.
3. Связь дисперсии полученной проекции с дисперсией в исходных данных.

Тема 3.2. Решающие деревья.

1. Жадный алгоритм построения дерева. Способы построения деревьев.
2. Использование деревьев при переобучении/недообучении.
3. Ансамбли, бэггинг
4. Лес. Случайные леса.

Тема 3.3. Метод опорных векторов, ядерная регрессия.

1. Постановка задачи для метода опорных векторов.
2. Понятие ядра. Ядерная регрессия.
3. Бустинг. Градиентный бустинг.

4. Глубокое обучение

Раздел 4. Глубокое обучение

Тема 4.1. Нейронные сети.

1. Матричные вычисления.
2. Понятие нейронной сети, полносвязного слоя.
3. Метод обратного распространения ошибки.

Тема 4.2 Функции активации, регуляризация нейронных сетей.

1. Понятие функции активации, их влияние на работу сетей.
2. Функции SoftMax и LogSoftMax.
3. Роль регуляризации при обучении и предсказании.

Тема 4.3 Рекурсивные нейронные сети

1. Применение рекурсивных нейронных сетей.
2. Обратное распространение в случае RNN.
3. Проблема исчезающего градиента.

Тема 4.4 Сверточные нейронные сети

1. Свертка матриц. Переход от полносвязного слоя к сверточному.
2. Обратное распространение и гиперпараметры в случае CNN.

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

учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

6.Перечень рекомендуемой литературы

Основная литература

1. Айвазян С. А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Классификация и снижение размерности. — М. Финансы и статистика. 1989.
2. Айвазян С. А., Енюков И.С., Мешалкин Л.Д. Исследование зависимостей. — М. Финансы и статистика. 1985.
3. Вагин В. Н., Головина Е. Ю., Загорянская А. А, Фомина М. В. Достоверный и правдоподобный вывод в интеллектуальных системах. — М.: Физматлит. 2004.

Дополнительная литература

1. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука. 1979.
2. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. <http://www.ccas.ru/voron>.
3. Головкин В. А. Нейронные сети: обучение, организация и применение. — М.: ИПРЖР. 2001.
4. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.
5. Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. — Новосибирск: Наука, 1985.
6. Ивахненко А. Г., Юрачковский Ю. П. Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1987.
7. Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — 1978. — Т. 33. — С. 5–68.
8. Казанцев В. С. Задачи классификации и их программное обеспечение. — М. Наука. 1990.
9. Лоусон Ч, Хенсон Р. Численное решение задач метода наименьших квадратов. — М. Наука. 1986.
10. Саттон Р.С., Барто Э.Г. Обучение с подкреплением. — БИНОМ, 2011.
11. Хардле В. Прикладная непараметрическая регрессия. — М.: Мир. 1993.

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

не используется

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

Для занятий может потребоваться следующее программное обеспечение:
Интернет-браузер, MS Word, MS Power Point, LaTeX, Adobe Reader, 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 к преподавателю.

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

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

| | |
|---|---|
| по направлению: | Прикладные математика и физика |
| профиль подготовки: | Физика перспективных технологий: альтернативная энергетика, научное программирование и функциональные материалы Физтех-школа Электроники, Фотоники и Молекулярной Физики кафедра физики высокотемпературных процессов |
| курс: | <u>2</u> |
| квалификация: | бакалавр |
| Семестр, формы промежуточной аттестации: 4 (весенний) - Зачет | |
| Разработчик: | К.Д. Гольдштейн |

1. Компетенции, формируемые в процессе изучения дисциплины

| Код и наименование компетенции | Индикаторы достижения компетенции |
|--|--|
| УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач | УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи |
| | УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки |
| | УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи |
| | УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки |
| УК-6 Способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни | УК-6.2 Способен планировать самостоятельную деятельность в решении профессиональных задач; подвергать критическому анализу проделанную работу; находить и творчески использовать имеющийся опыт в соответствии с задачами саморазвития |
| ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности | ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения |
| | ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки |
| | ОПК-1.3 Способен определять границы применимости полученных результатов |

2. Показатели оценивания компетенций

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

знать:

- Типы моделей машинного обучения, используемых в прикладных физических задачах
- Формулировки наиболее важных математических утверждений из теории анализа данных, способы их применения при помощи Python
- Сферы приложений конкретных моделей машинного обучения

уметь:

- Применять математические и статистические методы для решения прикладных задач.
- Использовать методы программирования и прикладной математики в научных задачах.

владеть:

- Базовыми методами машинного обучения и обработки данных.

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

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- проработку учебного материала (по конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к практическим занятиям, выполнение двух индивидуальных домашних заданий.

Примеры вопросов, заданий, тем для подготовки к текущему контролю

- Что такое кросс-валидация? Что такое переобучение и недообучение?
- Чем гиперпараметры отличаются от параметров?
- Какие существуют алгоритмы кластеризации?
- Как выглядит функционал логистической регрессии?

- Опишите метод опорных векторов.
- Опишите принцип работы сверточного слоя.
- Опишите принцип работы рекуррентного слоя.

Примеры тем проектов:

1. Предсказание ширины запрещенной зоны полупроводника
2. Моделирование потока жидкости с помощью нейросети
3. Кластеризация двумерных материалов с дефектами
4. Описание нелинейного осциллятора при помощи нейросети.

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

Вопросы к зачету:

1. Постановка задач машинного обучения, их классификация, примеры.
 2. Метрики качества классификации: accuracy, balanced accuracy, precision, recall, ROC-AUC.
 3. Метрики качества регрессии: MSE, MAE, R2, другие варианты.
 4. Метод максимального правдоподобия.
 5. Метод ближайших соседей.
 6. Постановка задачи линейной регрессии. Аналитическое решение МНК. Градиентное решение задачи линейной регрессии.
 7. Регуляризация: L1 и L2. Свойства, вероятностная интерпретация.
 8. Логистическая регрессия. Эквивалентность решений, полученных методом максимального правдоподобия и минимизации логистической функции потерь.
 9. Мультиклассовые классификации, их свойства.
 10. Метод опорных векторов, ядерная регрессия.
 12. Алгоритм PCA. Связь с SVD-разложением, дисперсией.
 13. Этапы построения модели: тренировка, валидация, тестирование. Роль каждого этапа.
- Переобучение.
14. Способы валидации. Кросс-валидация. Bias-variance tradeoff.
 15. Понятие информации, информационной энтропии. Критерии информативности: энтропийный, Джини.
 16. Жадный алгоритм построения дерева. Стандарты для построения деревьев.
 17. Ансамблирование. Бэггинг. Метод случайных подпространств.
 18. Случайный лес, другие леса.
 19. Бустинг. Градиентный бустинг.
 20. Матричные вычисления. Матричное дифференцирование.
 21. Понятие нейронной сети. Полносвязный слой. Логистическая регрессия как простейшая нейронная сеть. Метод обратного распространения ошибки, правило цепочек.
 22. Логистическая ошибка, кросс-энтропия. Функции активации, их влияние на работу сети, вычислительная сложность. Softmax и LogSoftmax функции
 23. Методы оптимизации для обучения нейронных сетей. Градиентный спуск. Стохастический градиентный спуск.
 24. Регуляризация нейронных сетей. Разница в поведении при обучении и предсказании.
 25. Рекурсивные нейронные сети. Обратное распространение через слой RNN.
- Проблема исчезающего градиента и ее решения.
26. Свертка матриц. Сверточный слой, обратное распространение через него. Гиперпараметры в случае CNN.

Критерии оценивания

Оценка "зачтено" 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.

Оценка "Не зачтено" выставляется студенту в случае большого количества недочетов и неправильных ответов, а также пассивной работе в ходе занятий, многие учебные задания не выполнены.

По усмотрению лектора допускается выставление оценки по результатам выполнения группового семестрового проекта. Проект считается выполненным, если показывает полное освоение материала, слаженную групповую работу, соответствующие теоретическим оценкам и экспериментальным данным результаты

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

Оценка выставляется по результату опроса в конце семестра. По усмотрению лектора допускается выставление оценки по результатам выполнения группового семестрового проекта. Проект считается выполненным, если показывает полное освоение материала, слаженную групповую работу, соответствующие теоретическим оценкам и экспериментальным данным результаты.