

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

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики**

А.М. Райгородский

	Рабочая программа дисциплины (модуля)
по дисциплине:	Машинное обучение на больших объемах данных
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра алгоритмов и технологий программирования
курс:	2
квалификация:	магистр

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

Аудиторных часов: 45 всего, в том числе:

лекции: 30 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 60 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 2

Программу составили:

А.А. Драль, старший преподаватель

О.Н. Ивченко, старший преподаватель

И.Е. Трофимов, ассистент

А.В. Зухба, ассистент

Н.Н. Притыковская, ассистент

Программа обсуждена на заседании кафедры алгоритмов и технологий программирования 13.05.2024

Аннотация

Целью данного курса является изучение применения машинного обучения при обработке больших данных. На практических занятиях будут разбираться методы машинного обучения в применении к большим объемам текста.

1. Цели и задачи

Цель дисциплины

- обучить студентов особенностям работы с алгоритмами машинного обучения в условиях современных объёмов данных.

Задачи дисциплины

- приобретение студентами навыков адаптации известных алгоритмов к большим объемам данных.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен управлять проектом на всех этапах его жизненного цикла	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
	УК-2.3 Способен организовать и координировать работу участников проекта, обеспечивать работу команды необходимыми ресурсами
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области информатики и вычислительной техники, 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 информатики и вычислительной техники 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
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- современные алгоритмы машинного обучения и их адаптации к анализу больших объемов данных.

уметь:

- применять алгоритмы и подходы машинного обучения в условиях современного строения и объёмов данных.

владеть:

- фреймворками и алгоритмами распределенного машинного обучения в условиях больших 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	Методы оптимизации и линейные модели	8	3		15
2	Алгоритмы работы с графами большого размера	7	4		15
3	Информационный поиск	7	4		15
4	Рекомендательные системы	8	4		15
Итого часов		30	15		60
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. Методы оптимизации и линейные модели

Машинное обучение с учителем на больших данных. Закон Ципфа. Тематическое моделирование. Метод стохастического градиента. Постановка задачи. Оптимизации обучения на больших данных: градиентный спуск, стохастический градиент. Признаки. Пространства признаков, веса признаков, нормализация признаков. Генерация и хеширование признаков. Онлайн обучение линейных моделей. Метод стохастического градиента: выбор функции потерь. Оценка качества метода стохастического градиента. Алгоритм Бутстрап. Хеширование, чувствительное к расстоянию (LSH). Меры сходства: расстояние Жаккара, Хемминга, косинусное расстояние, Евклидово расстояние. Оптимизация и тестирование гиперпараметров. Симплекс-метод.

2. Алгоритмы работы с графами большого размера

Графы, их виды. Стохастический граф. Представление графа: матрицы смежности, инцидентности, достижимости. Списки смежности. Алгоритмы перевода из одного представления в другое. Социальный граф. Задача поиска общих друзей в социальном графе. Язык DSL. Граф пользовательских предпочтений. Использование подхода BigData в анализе графов.

3. Информационный поиск

Постановка ранжирования. Основные подходы к решению задачи ранжирования. Метрики измерения точности ранжирования. Кликовые модели. Тематическое моделирование и его связь с ранжированием. Проблемы тематического моделирования при больших данных. AD-LDA, его недостатки, Y!LDA, Mr. LDA. ARTM. Архитектура библиотеки BigARTM. Online LDA и его применение в Vowpal Wabbit.

4. Рекомендательные системы

Рекомендательные системы, постановка задачи предсказания / рекомендации. Классификация рекомендательных систем. Неперсонализированные рекомендательные системы, content-based рекомендательные системы. Задачи коллаборативной фильтрации, транзакционные данные и матрица субъекты—объекты. Корреляционные методы, методы сходства (neighbourhood) - user-based, item-based. Латентные методы на основе матричных разложений. Методы ALS и iALS.

Современные рекомендательные системы: рекомендательные системы, основанные на учете контекста (context aware); аспектные рекомендательные системы (aspect-aware), рекомендательные системы на основе тензорных разложений.

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

Для лекционных занятий:

Учебная аудитория, оснащенная мультимедиа-проектором и экраном.

Для практических занятий:

Компьютерный класс. Каждый компьютер должен иметь выход в интернет и ПО для подключения к удалённым серверам.

Удалённый Hadoop-кластер.

6. Перечень рекомендуемой литературы

Основная литература

1. Методы и средства вычислений с объектами. Аппликативные вычислительные системы [Текст] : [учеб. пособие для вузов] / В. Э. Вольфенгаген ; Ин-т актуального образования <ЮрИнфоР-МГУ>, Каф. перспективных компьютерных исследований и информационных технологий .— М. : JurInfoR, 2004 .— 789 с. — (Компьютерные науки и информационные технологии). - 2000 экз. - ISBN 5-89158-100-0 (в пер.) .
2. Комбинаторная логика в программировании. Вычисления с объектами в примерах и задачах [Текст] : [учеб. пособие для вузов] / В. Э. Вольфенгаген ; НОУ Ин-т Актуального образования "ЮрИнфоР-МГУ, Каф. перспективных компьт. исслед. и информ. технологий .— 3-е изд., доп. и перераб. — М. : Ин-т "ЮрИнфоР-МГУ, 2008 .— 384 с.
3. Нейронные сети [Текст] : полный курс / С. Хайкин ; пер. с англ. Н. Н. Куссуль, А. Ю. Шелестова ; под ред. Н. Н. Куссуль .— 2-е изд., испр. — М. : Вильямс, 2006 .— 1103 с.

Дополнительная литература

1. Прикладная статистика. Принципы и примеры [Текст] : [учеб. пособие для вузов] / Д. Кокс, Э. Снелл ; пер. с англ. Е. В. Чепурина ; под ред. Ю. К. Беляева .— М. : Мир, 1984 .— 200 с.
2. Параллельное программирование многопоточных систем с разделяемой памятью [Текст] : учеб. пособие для вузов / А. Г. Тормасов .— М : Физматкнига, 2014 .— 208 с.

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

<https://www.coursera.org/specializations/big-data-engineering> - специализация из 5 курсов, посвящённая тематике обработки больших данных.

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

Microsoft Azure Machine Learning

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

Успешное освоение курса требует большой 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 4 индивидуальных домашних заданий.

Промежуточный контроль знаний проводится в виде письменных опросов (мини-тестов) по теории.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра алгоритмов и технологий программирования
курс:	<u>2</u>
квалификация:	магистр

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

Разработчики:

А.А. Драль, старший преподаватель
О.Н. Ивченко, старший преподаватель
И.Е. Трофимов, ассистент
А.В. Зухба, ассистент
Н.Н. Притыковская, ассистент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен управлять проектом на всех этапах его жизненного цикла	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
	УК-2.3 Способен организовать и координировать работу участников проекта, обеспечивать работу команды необходимыми ресурсами
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области информатики и вычислительной техники, 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 информатики и вычислительной техники 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
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками

2. Показатели оценивания компетенций

В результате изучения дисциплины «Машинное обучение на больших объемах данных» обучающийся должен:

знать:

- современные алгоритмы машинного обучения и их адаптации к анализу больших объемов данных.

уметь:

- применять алгоритмы и подходы машинного обучения в условиях современного строения и объёмов данных.

владеть:

- фреймворками и алгоритмами распределенного машинного обучения в условиях больших данных.

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

1. Какие способы построения рекомендательных систем Вы знаете? Кратко опишите суть каждого из них.
2. Обработка естественного языка: примеры задач, этапы решения задач естественного языка.
3. Обработка естественного языка: применение регулярных, контекстно-свободных и контекстно-зависимых грамматик.

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

1) Методы оптимизации и линейные модели.

1. Машинное обучение с учителем на больших данных. Закон Ципфа. Тематическое моделирование.
2. Метод стохастического градиента. Постановка задачи. Оптимизации обучения на больших данных: градиентный спуск, стохастический градиент.
3. Признаки. Пространства признаков, веса признаков, нормализация признаков. Генерация и хеширование признаков.
4. Онлайн обучение линейных моделей. Метод стохастического градиента: выбор функции потерь. Оценка качества метода стохастического градиента. Алгоритм Бутстрап.
5. Хеширование, чувствительное к расстоянию (LSH). Меры сходства: расстояние Жаккара, Хемминга, косинусное расстояние, Евклидово расстояние.
6. Оптимизация и тестирование гиперпараметров. Симплекс-метод.

2) Алгоритмы работы с графами большого размера.

1. Графы, их виды. Стохастический граф.
2. Представление графа: матрицы смежности, инцидентности, достижимости. Списки смежности. Алгоритмы перевода из одного представления в другое.
3. Социальный граф. Задача поиска общих друзей в социальном графе. Язык DSL.
4. Граф пользовательских предпочтений.
5. Использование подхода BigData в анализе графов.

3) Информационный поиск.

1. Постановка ранжирования. Основные подходы к решению задачи ранжирования.
2. Метрики измерения точности ранжирования. Кликовые модели.
3. Тематическое моделирование и его связь с ранжированием.
4. Проблемы тематического моделирования при больших данных. AD-LDA, его недостатки, Y!LDA, Mr. LDA. ARTM. Архитектура библиотеки BigARTM. Online LDA и его применение в Vowpal Wabbit.

4) Рекомендательные системы.

1. Постановка задачи рекомендаций. Неперсонализированные рекомендательные системы. Content-Based системы. Использование подхода BigData в рекомендательных системах.
2. Коллаборативная фильтрация и матричная факторизация.
3. Проблема “холодного старта”. Методы понижения размерности.

Типовой пример экзаменационного билета:

1. Content-based рекомендательная система: построение и применение; сложности при построении на Spark.
2. Каковы основные сложности обучения тематических моделей на больших массивах данных? Какие возможны методы их решения?
3. Какие архитектуры параллельного обучения тематических моделей вы можете предложить для одной машины? Для MPI-кластера? Для Hadoop MapReduce?
4. Обработка естественного языка: применение регулярных, контекстно-свободных и контекстно-зависимых грамматик.

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 недочётами.

Оценка "хорошо" (7 баллов) выставляется студенту, если он твёрдо 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 твёрдо 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

При проведении экзамена обучающемуся предоставляется 60 минут на подготовку. Опрос обучающегося по билету не должен превышать двух астрономических часов.

Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины и своими конспектами.

Билет состоит из 4 вопросов, каждый из которых оценивается в 0,5 балла. Каждый вопрос относится к одному из 4 модулей.

Итоговая оценка по курсу складывается из оценки за выполненные в ходе семестра практические задания и оценки за ответы на теоретические вопросы на экзамене.

Накопленные баллы за работу в семестре:

Форма контроля	Макс. балл
Домашнее задание по предсказанию вероятности клика	2,5
Домашнее задание по анализу социального графа	2,5
Домашнее задание по информационному поиску	2,5
Домашнее задание по рекомендательным системам	2,5

Всего

10

Примечание. 3 из 4 заданий проверяются с помощью системы поддержки соревнований Kaggle. За призовые места (1-е - 3-е на потоке) есть возможности получить бонус в размере не больше 2 баллов за задание.