Министерство науки и высшего образования Российской Федерации

ФГБОУ ВО «Тверской государственный университет»

Утверждаю:
Руководитель ООП
_____ С.М.Дудаков
«20» 02 2023 г.

Рабочая программа дисциплины (с аннотацией)

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Направление подготовки

01.03.02 Прикладная математика и информатика

Направленность (профиль)

Искусственный интеллект и анализ данных

Для студентов 3 курса

Форма обучения:

очная

Составитель: к.ф.-м.н. доцент Солдатенко И.С.

Тверь, 2023

I. Аннотация

1. Цель и задачи дисциплины

Цель изучения дисциплины является формирование у слушателей компетенций по разработке и применению методов и алгоритмов машинного обучения для решения задач

Задачи дисциплины:

- изучение математических основ методов машинного обучения и соответствующих алгоритмов;
- изучение современных программных сред и библиотек, позволяющих проводить анализ, визуализацию данных, применять современные математические методы машинного обучения;
- развитие практических навыков использования методов машинного обучения в прикладных задачах.

2. Место дисциплины в структуре ООП

Данная дисциплина относится к разделу «Дисциплины профиля подготовки» части, формируемой участниками образовательных отношений Блока 1.

3. Объем дисциплины: 6 зачетных единиц, 216 академических часов, **в том числе:**

контактная аудиторная работа: лекции 62 часов, в т.ч. практическая подготовка 7 часов; практические занятия 62 часов, в т.ч. практическая подготовка 7 часов.

контактная внеаудиторная работа: контроль самостоятельной работы 0, в том числе курсовая работа 0;

самостоятельная работа: 92 часов, в том числе контроль 68.

4. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы

Планируемые результаты освоения образовательной программы (формируемые компетенции)	**Планируемые результаты обучения по дисциплине**
ПК-4 Способен разрабатывать и применять методы машинного обучения для решения задач	ПК-4.1 Проводит анализ требований и определяет необходимые классы задач машинного обучения ПК-4.2 Определяет метрики оценки результатов моделирования и критерии качества построенных моделей ПК-4.3 Принимает участие в оценке, выборе и при необходимости разработке методов машинного обучения

ПК-5 Способен использовать инструментальные средства для решения задач машинного обучения	ПК-5.1 Осуществляет оценку и выбор инструментальных средств для решения поставленной задачи ПК-5.2 Разрабатывает модели машинного обучения для решения задач ПК-5.3 Создает, поддерживает и использует системы искусственного интеллекта, включающие разработанные модели и методы, с применением выбранных инструментов машинного обучения
ПК-7 Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта	ПК-7.1 Осуществляет поиск данных в открытых источниках, специализированных библиотеках и репозиториях ПК-7.2 Выполняет подготовку и разметку структурированных и неструктурированных данных для машинного обучения

5. Форма промежуточной аттестации экзамен (5, 6 семестры).

6. Язык преподавания русский**.**

II. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

1. **Для студентов очной формы обучения**

Учебная программа – наименование разделов и тем	Всего (час.)	Контактная работа (час.)				Самостоятельная работа (час.), в том числе контроль
		Лекции		Практические (лабораторные) работы		
		всего	В т.ч.практическая подготовка	всего	В т.ч.практическая подготовка	
5 семестр						
Терминология	8	2		2		4
Постановка основных задач	8	2		2		4

Математика в машинном обучении: краткий обзор	21	6	2	6	2	9
Оптимизация	14	4	1	4	1	6
Метрические алгоритмы	22	7		7		8
Линейные методы	35	10	1	10	1	15
6 семестр						
Деревья решений	12	4		4		4
Контроль качества и выбор модели	12	4		4		4
Ансамблирование в машинном обучении	17	4	1	4	1	9
Методы, основанные на деревьях: случайный лес, бустинг	14	4	1	4	1	6
Введение в рекомендательные системы	20	6	1	6	1	8
Сложность алгоритмов, переобучение, смещение и разброс	33	9		9		15
ИТОГО	**216**	**62**	**7**	**62**	**7**	**92**

Содержание разделов (тем) дисциплины

№ п/п	Наименование разделов (тем) дисциплины	Содержание разделов (тем) дисциплин
5 семестр		
1.	Терминология	Наука о данных (Data Sciense) Статистика (Statistics) Искусственный интеллект (Artificial Intelligence) Анализ данных (Data Mining) Машинное обучение (Machine learning) Большие данные (Big Data)
2.	Постановка основных задач	Обучение с учителем (с размеченными данными / метками) целевая функция объект

		метка классификация Прогнозирование Пространство объектов признаковое пространство Извлечение признаков Визуализация задач функции ошибки эмпирический риск обучающая выборка Задачи оптимизации в обучении Модель алгоритмов Алгоритм Обучение Обобщающая способность Схема решения задачи машинного обучения Как решаются задачи Обучение без учителя /с неразмеченными данными Обучение с частично размеченными данными трансдуктивное обучение Обучение с подкреплением Структурный вывод Активное обучение Онлайн-обучение Transfer Learning Multitask Learning Feature Learning Проблемы в машинном обучении Примеры модельных задач
3.	Математика в машинном обучении: краткий обзор	Бритва Оккама Теорема о бесплатном сыре Футбольный оракул Сведения из ТВиМС Задание распределений Средние и отклонения Условная плотность, маргинализация и обуславливание Точечное оценивание Оценка максимального правдоподобия Дивергенция Кульбака-Лейблера ковариация и корреляция Оценка плотности гистограммного подхода Парзеновский подход Нормальное распределение Центральная предельная теорема Теория информации

		Проклятие размерности Сингулярное разложение матрицы (SVD) матричное дифференцирование
4.	Оптимизация	Методы безусловной оптимизации Методы нулевого порядка Методы первого порядка Методы второго порядка Градиентный спуск Наискорейший градиентный спуск Стохастический градиентный спуск Обучение: Пакетное, онлайн, по минибатчам Метод градиентного спуска в машинном обучении Стационарные точки Метод Ньютона Квази-ньютоновские методы Оптимизация с ограничениями
5.	Метрические алгоритмы	Метрические алгоритмы (distance-based) Ближайший центроид (Nearest centroid algorithm) Подход, основанный на близости kNN в задаче классифи-кации kNN в задаче регрессии Обоснование 1NN Ленивые (Lazy) и нетерпеливые (Eager) алгоритмы Весовые обобщения kNN Различные метрики: Минковского, Евклидова, Манхэттенская, Махало-нобиса, Canberra distance, Хэмминга, косинусное, расстояние Джаккарда, DTW, Левенштейна Приложения метрическо-го подхода: нечёткий матчинг таблиц, Ленкор, в DL, классификация текстов Эффективные методы поиска ближайших сосе-дей Регрессия Надарая-Ватсона
6.	Линейные методы	Линейная регрессия Обобщённая линейная регрессия Проблема вырожденно-сти матрицы Регуляризация. Основ-ные виды регуляризации Гребневая регрессия (Ridge Regression) LASSO (Least Absolute Selection and Shrinkage Operator)

		Elastic Net Селекция признаков Ошибка с весами Устойчивая регрессия (Robust Regression) Линейные скоринговые модели в задаче бинарной классификации Логистическая регрессия Probit-регрессия Многоклассовая логистическая регрессия Линейный классифика-тор Персептрон Оценка функции ошибок через гладкую функцию
6 семестр		
7.	Деревья решений.	Деревья решений (CART) Предикаты / ветвления Ответы дерева Критерии расщепления в задачах классификации: Missclassification criteria, энтропийный, Джини Критерии остановки при построении деревьев Проблема переобучения для деревьев Подрезка (post-pruning) Классические алгоритмы построения деревьев ре-шений: ID3, C5.0 Важности признаков Проблема пропусков (Missing Values) Категориальные признаки Сравнение: деревья vs линейные модели
8.	Контроль качества и выбор модели	Проблема контроля качества Выбора модели (Model Selection) в широком смысле Правила разбиения вы-борки Отложенный контроль (held-out data, hold-out set) Скользящий контроль (cross-validation) Бутстреп (bootstrap) Контроль по времени (out-of-time-контроль) Локальный контроль Кривые обучения (Learning Curves) Перебор параметров
9.	Ансамблирование в машинном обучении	Ансамбли алгоритмов: примеры и обоснование комитеты (голосование) / усреднение

		Бэгинг Кодировки / перекоди-ровки ответов, ECOC Стекинг и блендинг Бустинг: AdaBoost, For-ward stagewise additive modeling (FSAM) «Ручные методы» Однородные ансамбли
10.	Методы, основанные на деревьях: случайный лес, бустинг	Случайный лес, его параметры, их настройка Бэггинг и OOB (out of bag) Важность признаков Близость (Proximity) с помощью RF Extreme Random Trees Градиентный бустинг над деревьями, его параметры, современные реализации, Продвинутые методы оптимизации
11.	Введение в рекомендательные системы	Рекомендательные системы Персонализация, онлайн и оффлайн рекомендации Рекомендация по контенту (content based methods) Коллаборативная фильтрация: GroupLens-алгоритм, SVD, SVD++, timeSVD++, адаптация SVD под социальные связи One-class recommendation Факторизационная машина, факторизационная машина с полями (FFM – field-aware factorization machine) Простые методы рекомендаций: FPM – Frequent Pattern Mining Deep Semantic Similarity Model (DSSM) Контекст рекомендации Knowledge-based Recom-mendations Важность объяснений (explanations) Использование дополни-тельной информации Современные тренды в практике построения рекомендательных систем
12.	Сложность алгоритмов, переобучение, смещение и разброс	Проблема обобщения Переобучение Недообучение Сложность алгоритмов Смещение и разброс Способы борьбы с переобучением

III. Образовательные технологии

Учебная программа – наименование разделов и тем *(в строгом соответствии с разделом II РПД)*	Вид занятия	Образовательные технологии
Терминология	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Самостоятельное изучение теоретического материала
Постановка основных задач	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Самостоятельное изучение теоретического материала
Математика в машинном обучении: краткий обзор	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Решение задач 3. Самостоятельное изучение теоретического материала
Оптимизация	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Решение задач 3. Самостоятельное изучение теоретического материала
Метрические алгоритмы	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Решение задач 3. Самостоятельное изучение теоретического материала
Линейные методы	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Решение задач 3. Самостоятельное изучение теоретического материала

Деревья решений.	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Решение задач 3. Самостоятельное изучение теоретического материала
Контроль качества и выбор модели	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Самостоятельное изучение теоретического материала
Ансамблирование в машинном обучении	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Самостоятельное изучение теоретического материала
Методы, основанные на деревьях: случайный лес, бустинг	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Решение задач 3. Самостоятельное изучение теоретического материала
Введение в рекомендательные системы	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Решение задач 3. Самостоятельное изучение теоретического материала
Сложность алгоритмов, переобучение, смещение и разброс	Лекции, практические занятия	1. Изложение теоретического материала (презентация) 2. Самостоятельное изучение теоретического материала

Преподавание учебной дисциплины строится на сочетании лекций и практических занятий и различных форм самостоятельной работы студентов.

В процессе освоения дисциплины используются следующие образовательные технологии, способы и методы формирования компетенций: лабораторные занятия в компьютерных классах, выполнение индивидуальных заданий в рамках самостоятельной работы.

Самостоятельная работа студентов организуется в форме решения заданий по предложенным тематикам, а также выполнении расчетных или курсовых работ, письменных домашних заданий.

IV. Оценочные материалы для проведения текущей и промежуточной аттестации.

Для проведения текущей и промежуточной аттестации:

ПК-4 Способен разрабатывать и применять методы машинного обучения для решения задач
ПК-5 Способен использовать инструментальные средства для решения задач машинного обучения
ПК-7 Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта

Для всех компетенций один способ аттестации:

Форма аттестации: ответ по темам курса (экзамен)
Способ аттестации: устный
Критерии оценки:

- *ответ целостный, верный, теоретически обоснованный. Ключевые понятия и термины полностью раскрыты. Факты и примеры в полном объеме обосновывают выводы – 30 баллов;*
- *теоретическая аргументация неполная или смысл ключевых понятий не объяснен – 20 баллов;*
- *допущены ошибки, приведшие к искажению смысла. терминологический аппарат раскрыт – 10 баллов;*
- *допущены ошибки, свидетельствующие о непонимании темы. Терминологический аппарат не раскрыт – 0 баллов;*
- *верно решены задачи, иллюстрирующая знание курса – 10 баллов;*
- *при решении задач, допущены арифметические ошибки – 5 баллов;*
- *при решении задач, допущены логические ошибки – 3 балла;*
- *решение задач неверно или отсутствует – 0 баллов.*

Примеры оценочных средств приведены в разделе VI.

V. Учебно-методическое и информационное обеспечение дисциплины

1) Рекомендуемая литература
Основная литература:
1. Ракитский, А. А. Методы машинного обучения: учебно-методическое пособие / А. А. Ракитский. — Новосибирск: Сибирский государственный университет телекоммуникаций и информатики, 2018. — 32 с. — Текст: электронный // Цифровой образовательный ресурс IPR SMART: [сайт]. — URL: https://www.iprbookshop.ru/90591.html

2. Теория и практика машинного обучения: учебное пособие / В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святов. — Ульяновск:

Ульяновский государственный технический университет, 2017. — 291 с. — ISBN 978-5-9795-1712-4. — Текст: электронный // Цифровой образовательный ресурс IPR SMART: [сайт]. — URL: https://www.iprbookshop.ru/106120.html

Дополнительная литература:

1. Сараев, П. В. Методы машинного обучения: методические указания и задания к лабораторным работам по курсу / П. В. Сараев. — Липецк: Липецкий государственный технический университет, ЭБС АСВ, 2017. — 48 с. — Текст: электронный // Цифровой образовательный ресурс IPR SMART: [сайт]. — URL: https://www.iprbookshop.ru/83183.html

2) Программное обеспечение

Компьютерный класс факультета прикладной математики и кибернетики № 4б (170002, Тверская обл., г.Тверь, Садовый переулок, д.35)	
Adobe Acrobat Reader DC - Russian	бесплатно
Apache Tomcat 8.0.27	бесплатно
Cadence SPB/OrCAD 16.6	Государственный контракт на поставку лицензионных программных продуктов 103 - ГК/09 от 15.06.2009
GlassFish Server Open Source Edition 4.1.1	бесплатно
Google Chrome	бесплатно
Java SE Development Kit 8 Update 45 (64-bit)	бесплатно
JetBrains PyCharm Community Edition 4.5.3	бесплатно
JetBrains PyCharm Edu 3.0	бесплатно
Kaspersky Endpoint Security 10 для Windows	Акт на передачу прав ПК545 от 16.12.2022
Lazarus 1.4.0	бесплатно
Mathcad 15 M010	Акт предоставления прав ИС00000027 от 16.09.2011
MATLAB R2012b	Акт предоставления прав № Us000311 от 25.09.2012
Многофункциональный редактор ONLYOFFICE бесплатное ПО	бесплатно
ОС Linux Ubuntu бесплатное ПО	бесплатно
MiKTeX 2.9	бесплатно
MSXML 4.0 SP2 Parser and SDK	бесплатно
NetBeans IDE 8.0.2	бесплатно
NetBeans IDE 8.2	бесплатно
Notepad++	бесплатно
Oracle VM VirtualBox 5.0.2	бесплатно
Origin 8.1 Sr2	договор №13918/M41 от 24.09.2009 с ЗАО «СофтЛайн Трейд»
Python 3.1 pygame-1.9.1	бесплатно
Python 3.4 numpy-1.9.2	бесплатно
Python 3.4.3	бесплатно
Python 3.5.1 (Anaconda3 2.5.0 64-bit)	бесплатно
WCF RIA Services V1.0 SP2	бесплатно
WinDjView 2.1	бесплатно
R Studio	бесплатно

Anaconda3 2019.07 (Python 3.7.3 64-bit)	бесплатно
Компьютерный класс факультета прикладной математики и кибернетики № 249 (170002, Тверская обл., г.Тверь, Садовый переулок, д.35)	
Cadence SPB/OrCAD 16.6	Государственный контракт на поставку лицензионных программных продуктов 103 - ГК/09 от 15.06.2009
FidesysBundle 1.4.43 x64	Акт приема передачи по договору №02/12-13 от 16.12.2013
Google Chrome	бесплатно
JetBrains PyCharm Community Edition 4.5.3	бесплатно
Kaspersky Endpoint Security 10 для Windows	Акт на передачу прав ПК545 от 16.12.2022
Lazarus 1.4.0	бесплатно
Mathcad 15 M010	Акт предоставления прав ИС00000027 от 16.09.2011
MATLAB R2012b	Акт предоставления прав № Us000311 от 25.09.2012
MiKTeX 2.9	бесплатно
NetBeans IDE 8.0.2	бесплатно
Notepad++	бесплатно
OpenOffice	бесплатно
Origin 8.1 Sr2	договор №13918/M41 от 24.09.2009 с ЗАО «СофтЛайн Трейд»
Python 3.4.3	бесплатно
Python 3.5.1 (Anaconda3 2.5.0 64 bit)	бесплатно
R for Windows 3.3.2	бесплатно
STATGRAPHICS Centurion XVI.II	Акт приема-передачи № Tr024185 от 08.07.2010
Многофункциональный редактор ONLYOFFICE бесплатное ПО	бесплатно
ОС Linux Ubuntu бесплатное ПО	бесплатно

3) Современные профессиональные базы данных и информационные справочные системы

1. ЭБС «ZNANIUM.COM» www.znanium.com;
2. ЭБС «Университетская библиотека онлайн» https://biblioclub.ru/;
3. ЭБС «Лань» http://e.lanbook.com.

4) Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

1. Math-Net.Ru [Электронный ресурс]: общероссийский математический портал / Математический институт им. В. А. Стеклова РАН ; Российская академия наук, Отделение математических наук. - М.: [б. и.], 2010. - Загл. с титул. экрана. - Б. ц.

   URL: http://www.mathnet.ru

2. Университетская библиотека Online [Электронный ресурс]: электронная библиотечная система / ООО "Директ-Медиа". - М. : [б. и.], 2001. - Загл. с титул. экрана. - Б. ц. URL: www.biblioclub.ru

3. Универсальные базы данных EastView [Электронный ресурс]: информационный ресурс / EastViewInformationServices. - М. : [б. и.], 2012. - Загл. с титул. экрана. - Б. ц.

URL: www.ebiblioteka.ru

4. Научная электронная библиотека eLIBRARY.RU [Электронный ресурс]: информационный портал / ООО "РУНЭБ"; Санкт-Петербургский государственный университет. - М. : [б. и.], 2005. - Загл. с титул. экрана. - Б. ц.

URL: www.eLibrary.ru

VI. Методические материалы для обучающихся по освоению дисциплины

Методические указания для обучающихся по подготовке к семинарским занятиям

Важной составляющей данного раздела РПД являются требования к рейтинг-контролю с указанием баллов, распределенных между модулями и видами работы обучающихся.

Максимальная сумма баллов по учебной дисциплине, заканчивающейся экзаменом, по итогам семестра составляет 60 баллов (30 баллов - 1-й модуль и 30 баллов - 2-й модуль).

Обучающемуся, набравшему 40–54 балла, при подведении итогов семестра (на последнем занятии по дисциплине) в рейтинговой ведомости учета успеваемости и зачетной книжке может быть выставлена оценка «удовлетворительно».

Обучающемуся, набравшему 55–57 баллов, при подведении итогов семестра (на последнем занятии по дисциплине) в графе рейтинговой ведомости учета успеваемости «Премиальные баллы» может быть добавлено 15 баллов и выставлена экзаменационная оценка «хорошо».

Обучающемуся, набравшему 58–60 баллов, при подведении итогов семестра (на последнем занятии по дисциплине) в графе рейтинговой ведомости учета успеваемости «Премиальные баллы» может быть добавлено 27 баллов и выставлена экзаменационная оценка «отлично». В каких-либо иных случаях добавление премиальных баллов не допускается.

Обучающийся, набравший до 39 баллов включительно, сдает экзамен.

Распределение баллов по модулям устанавливается преподавателем и может корректироваться.

Для того чтобы семинарские занятия приносили максимальную пользу, необходимо помнить, что упражнение и решение задач проводятся по вычитанному на лекциях материалу и связаны, как правило, с детальным разбором отдельных вопросов лекционного курса. Следует подчеркнуть, что только после усвоения лекционного материала с определенной точки зрения (а именно с той, с которой он излагается на лекциях) он будет закрепляться на семинарских занятиях как в результате обсуждения и анализа лекционного материала, так и с помощью решения проблемных ситуаций, задач.

При этих условиях студент не только хорошо усвоит материал, но и научится применять его на практике, а также получит дополнительный стимул (и это очень важно) для активной проработки лекции.

При самостоятельном решении задач нужно обосновывать каждый этап решения, исходя из теоретических положений курса. Если студент видит несколько путей решения проблемы (задачи), то нужно сравнить их и выбрать самый рациональный. Полезно до начала вычислений составить краткий план решения проблемы (задачи). Решение проблемных задач или примеров следует излагать подробно, вычисления располагать в строгом порядке, отделяя вспомогательные вычисления от основных. Решения при необходимости нужно сопровождать комментариями, схемами, чертежами и рисунками.

Следует помнить, что решение каждой учебной задачи должно доводиться до окончательного логического ответа, которого требует условие, и по возможности с выводом. Полученный ответ следует проверить способами, вытекающими из существа данной задачи. Полезно также (если возможно) решать несколькими способами и сравнить полученные результаты. Решение задач данного типа нужно продолжать до приобретения твердых навыков в их решении.

При подготовке к семинарским занятиям следует использовать основную литературу из представленного списка, а также руководствоваться приведенными указаниями и рекомендациями. Для наиболее глубокого освоения дисциплины рекомендуется изучать литературу, обозначенную как «дополнительная» в представленном списке.

Методические указания для обучающихся по подготовке к практическим занятиям

Целью практических занятий по данной дисциплине является закрепление теоретических знаний, полученных при изучении дисциплины.

При подготовке к практическому занятию целесообразно выполнить следующие рекомендации: изучить основную литературу; ознакомиться с дополнительной литературой, новыми публикациями в периодических изданиях: журналах, газетах и т. д.; при необходимости доработать конспект лекций. При этом учесть рекомендации преподавателя и требования учебной программы.

При выполнении практических занятий основным методом обучения является самостоятельная работа студента под управлением преподавателя. На них пополняются теоретические знания студентов, их умение творчески мыслить, анализировать, обобщать изученный материал, проверяется отношение студентов к будущей профессиональной деятельности.

Оценка выполненной работы осуществляется преподавателем комплексно: по результатам выполнения заданий, устному сообщению и оформлению работы. После подведения итогов занятия студент обязан устранить недостатки, отмеченные преподавателем при оценке его работы.

Методические указания для самостоятельной работы обучающихся

Приступая к изучению новой учебной дисциплины, студенты должны ознакомиться с учебной программой, учебной, научной и методической литературой, имеющейся в библиотеке университета, встретиться с преподавателем, ведущим дисциплину, получить в библиотеке рекомендованные учебники и учебно-методические пособия, осуществить запись на соответствующий курс в среде электронного обучения университета.

Глубина усвоения дисциплины зависит от активной и систематической работы студента на лекциях и практических занятиях, а также в ходе самостоятельной работы, по изучению рекомендованной литературы.

На лекциях важно сосредоточить внимание на ее содержании. Это поможет лучше воспринимать учебный материал и уяснить взаимосвязь проблем по всей дисциплине. Основное содержание лекции целесообразнее записывать в тетради в виде ключевых фраз, понятий, тезисов, обобщений, схем, опорных выводов. Необходимо обращать внимание на термины, формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации. Желательно оставлять в конспектах поля, на которых делать пометки из рекомендованной литературы, дополняющей материал прослушанной лекции, а также подчеркивающие особую важность тех или иных теоретических положений. С целью уяснения теоретических положений, разрешения спорных ситуаций необходимо задавать преподавателю уточняющие вопросы. Для закрепления содержания лекции в памяти, необходимо во время самостоятельной работы внимательно прочесть свой конспект и дополнить его записями из учебников и рекомендованной литературы. Конспектирование читаемых лекций и их последующая доработка способствует более глубокому усвоению знаний, и поэтому являются важной формой учебной деятельности студентов.

Прочное усвоение и долговременное закрепление учебного материала невозможно без продуманной самостоятельной работы. Такая работа требует от студента значительных усилий, творчества и высокой организованности. В ходе самостоятельной работы студенты выполняют следующие задачи: дорабатывают лекции, изучают рекомендованную литературу, готовятся к практическим занятиям, к коллоквиуму, контрольным работам по отдельным темам дисциплины. При этом эффективность учебной деятельности студента во многом зависит от того, как он распорядился выделенным для самостоятельной работы бюджетом времени.

Результатом самостоятельной работы является прочное усвоение материалов по предмету согласно программы дисциплины. В итоге этой работы формируются профессиональные умения и компетенции, развивается творческий подход к решению возникших в ходе учебной деятельности проблемных задач, появляется самостоятельности мышления.

Решение задач

При самостоятельном решении задач нужно обосновывать каждый этап решения, исходя из теоретических положений курса. Если студент видит

несколько путей решения проблемы (задачи), то нужно сравнить их и выбрать самый рациональный. Полезно до начала вычислений составить краткий план решения проблемы (задачи).

Решение проблемных задач или примеров следует излагать подробно, вычисления располагать в строгом порядке, отделяя вспомогательные вычисления от основных. Решения при необходимости нужно сопровождать комментариями, схемами, чертежами и рисунками.

Следует помнить, что решение каждой учебной задачи должно доводиться до окончательного логического ответа, которого требует условие, и по возможности с выводом.

Полученный ответ следует проверить способами, вытекающими из существа данной задачи. Полезно также (если возможно) решать несколькими способами и сравнить полученные результаты.

Решение задач данного типа нужно продолжать до приобретения твердых навыков в их решении.

Задача — это цель, заданная в определенных условиях, решение задачи — процесс достижения поставленной цели, поиск необходимых для этого средств.

Алгоритм решения задач:

1. Внимательно прочитайте условие задания и уясните основной вопрос, представьте процессы и явления, описанные в условии.

2. Повторно прочтите условие для того, чтобы чётко представить основной вопрос, проблему, цель решения, заданные величины, опираясь на которые можно вести поиски решения.

3. Произведите краткую запись условия задания.

4. Если необходимо составьте таблицу, схему, рисунок или чертёж.

5. Определите метод решения задания, составьте план решения.

6. Запишите основные понятия, формулы, описывающие процессы, предложенные заданной системой.

7. Найдите решение в общем виде, выразив искомые величины через заданные.

9. Проверьте правильность решения задания.

10. Произведите оценку реальности полученного решения.

11. Запишите ответ.

Текущий контроль успеваемости осуществляется путем оценки результатов выполнения заданий практических (семинарских) занятий, самостоятельной работы, предусмотренных учебным планом и посещения занятий/активность на занятиях.

В качестве оценочных средств текущего контроля успеваемости предусмотрены:

тестирование

Примеры тестовых заданий

5 семестр

Случайная величина принимает значение из отрезка [0,1], её плотность линейная функция на этом отрезке, в нуле обращается в ноль. Чему равно матожидание с.в.? Совет: здесь и ниже, кроме аналитического решения напишите на Python программу для оценки названных параметров.
0
1/2
2/3
3/4
4/5
1

Чему равна мода этой с.в.?
0
1/2
2/3
3/4
4/5
1
нет правильного ответа

Чему равна медиана этой с.в.?
0
1/2
2/3
3/4
4/5
1
нет правильного ответа

Чему равна дисперсия этой с.в.?

Предположим, что в задаче бинарной классификации с одним признаком объекты класса 1 распределены так, как описано выше, а объекты класса 0 распределены равномерно на отрезке [0,1]. Оба класса равновероятны. Какой оптимальный порог для отнесения объектов к классу 1 (выше него считаем, что они из класса 1), если оба класса равновероятны?
1/3
1/2
2/3
3/4
4/5
Что лучше использовать для определения монотонной зависимости между переменными?
Корреляционный коэффициент Пирсона
Коэффициент корреляции Спирмена
оценку ММП (MLE)
К чему стремится угол между соседними диагоналями n-мерного гиперкуба при увеличении размерности?
0
pi/4
pi/2
нет правильного ответа

Запишите сумму квадратов сингулярных чисел для матрицы [[0,1], [0,1]].

При минимизации функции x^2 методом градиентного спуска с темпом 1.0 и начальной точкой 1.0, какая будет оценка argmin после 4й итерации?

Выберете верные фразы:
в SGD случайный порядок объектов
SGD может использоваться при онлайн-оптимизации (обучении)
SGD может использоваться для минимизации суммы ошибок на объектах моделей классификации / регрессии
SGD – метод оптимизации второго порядка
SGD – это метод условной оптимизации

Какие расстояния численно наибольшие для пары точек (1,1) и (2,2):
Евклидово
Чебышева
Манхэттенское

Пусть даны векторы (1,1,2,2,3,3), (1, 4). Чему равно расстояние DTW?

Решите матричное уравнение Xw=y, W=[[1, 1], [1, 2], [1, 3]], y=[1,2,1] с помощью минимизации невязки. В ответ запишите скалярное произведение вектора w и вектора (3, -1).

С помощью персептронного алгоритма решите систему уравнений a+b>0, 3a-b>0, a-b<0. Начальное приближение (a,b) = (0,0), неравенства просматриваются слева направо. В ответ запишите значение b/a.

6 семестр

Что особенного в деревьях решений вида «oblique decision trees»?
ограничение на глубину
использование предварительной обрезки (pre-pruning)
специальный предикат ветвления
возможность распараллеливания при построении

Почему при построении дерева используют рекурсивную жадную стратегию?
алгоритма оптимизации не существует
задача построения оптимального дерева очень сложна (в одном частном случае это NP-полная проблема)
это, как правило, быстрее градиентного спуска

Рассмотрим 10 объектов, если их упорядочить по первому признаку, то их метки будут чередоваться следующим образом: [0,0,1,0,1,0,1,1,1,0]. Найдите максимальное значение критерия расщепления Missclassification criteria.

Пусть есть категориальный признак со значениями [A, A, B, B, C, C, D, D], с целевыми значениями [2, 0, 2, 8, 3, 5, 0, 4]. Какое будет расщепление со стандартным критерием, использующим дисперсию?
A, B | C, D
A, C | B, D
A, D | B,C
A | C, B, D
A, B, C | D

Что из перечисленного можно использовать для выбора модели (Model Selection):
бутстреп
регуляризацию
разбиение на фолды

В какой модели производится перекодировка целевого признака?
комитеты
ECOC
стекинг
бустинг

В какой модели применяется взятие бустреп-подвыборок?
бэгинг (Bagging)
случайные леса (RF)
стекинг
Feature-Weighted Linear Stacking

Что используется в продвинутых методах реализации градиентного бустинга (как в библиотеке XGBoost)?
принцип минимальной длины (MDL)
вторые производные функции ошибки
автоматический выбор ключевых параметров, например learning_rate

На какие слагаемые раскладывается квадратичная ошибка регрессора (матожидание квадрата разности прогноза и истинного значения)?
шум (noise)
квадрат шума
разброс (variance)
квадрат разброса
смещение (bias)
квадрат смещения

При повышении числа соседей k метода kNN...
увеличивается сложность модели
увеличивается качество на обучении
увеличивается качество на контроле
увеличивается разброс (variance)
увеличивается стабильность

Что из перечисленного приводит к уменьшению переобучения?
аугментация
регуляризация
увеличение объёма выборки

Пример практического задания

Задача машинного обучения с реальными данными, выложенная на https://inclass.kaggle.com/c/dayofweek/

Описание

Для 300000 пользователей дана статистика посещений ресурса за 1099 дней. Необходимо предсказать день недели следующего визита.

Метрика качества

Используется простой процент правильных ответов. Например,

```
performance([1,2,2,7], [3,2,2,7]) = 0.75
```

Формат ответа

В загружаемом файле по строкам перечислены идентификаторы пользователей и номера дней их первых визитов по версии вашего алгоритма:

```
id,nextvisit
1, 7
```

Данные

В файле train.csv перечислены даты визитов пользователей. Каждая строка - информация по одному пользователю. Сначала идёт id, потом через пробел номера дней, когда были визиты. Нумерация идёт от некоторого фиксированного момента. Номера могут быть от 1 до 1099 (т.е. статистика охватывает период примерно 3 года). Первый день в нумерации - понедельник.

Необходимо предсказать день недели первого визита после 1099го дня, т.е. для каждого пользователя вычислить

```
0 - нет визита
1 - понедельник
2 - вторник
3 - среда
4 - четверг
5 - пятница
6 - суббота
7 - воскресенье
```

Промежуточная аттестация

5 семестр

Промежуточная аттестация осуществляется в форме экзамена

В качестве средств, используемых на промежуточной аттестации предусматривается: Билеты

6 семестр

Промежуточная аттестация осуществляется в форме экзамена

В качестве средств, используемых на промежуточной аттестации предусматривается: Билеты

Типовые задания для проведения промежуточной аттестации

5 семестр

Вопросы к экзамену

1. Терминология: Наука о данных (Data Sciense), Статистика (Statistics), Искусственный интеллект (Artificial Intelligence), Анализ данных (Data Mining), Машинное обучение (Machine learning), Большие данные (Big Data)
2. Обучение с учителем (с размеченными данными / метками): целевая функция, объект, метка, классификация, прогнозирование
3. Пространство объектов, признаковое пространство, извлечение признаков, визуализация задач
4. Функции ошибки, эмпирический риск, обучающая выборка, задачи оптимизации в обучении, обобщающая способность
5. Модель алгоритмов, алгоритм, обучение, схема решения задачи машинного обучения
6. Обучение без учителя / с неразмеченными данными, обучение с частично размеченными данными, трансдуктивное обучение
7. Обучение с подкреплением, структурный вывод, активное обучение, онлайн-обучение, Transfer Learning, Multitask Learning, Feature Learning
8. Математика в машинном обучении: бритва Оккама, теорема о бесплатном сыре, футбольный оракул, теория информации, проклятие размерности, сингулярное разложение матрицы (SVD), матричное дифференцирование
9. Сведения из ТВиМС: задание распределений, средние и отклонения, условная плотность, маргинализация и обуславливание, точечное оценивание, оценка максимального правдоподобия, дивергенция Кульбака-Лейблера, ковариация и корреляция, нормальное распределение, центральная предельная теорема
10. Оценка плотности: гистограммный подход, Парзеновский подход
11. Оптимизация: методы безусловной оптимизации, нулевого порядка, первого порядка, второго порядка, метод градиентного спуска в машинном обучении, стационарные точки, метод Ньютона, квази-ньютоновские методы, оптимизация с ограничениями
12. Градиентный спуск, наискорейший градиентный спуск, стохастический градиентный спуск, обучение: Пакетное, онлайн, по минибатчам

13. Метрические алгоритмы (distance-based), ближайший центроид (Nearest centroid algorithm), подход, основанный на близости, kNN в задаче классификации / регрессии, обоснование 1NN, ленивые (Lazy) и нетерпеливые (Eager) алгоритмы
14. Весовые обобщения kNN, регрессия Надарая-Ватсона
15. Различные метрики: Минковского, Евклидова, Манхэттенская, Махалонобиса, Canberra distance, Хэмминга, косинусное, расстояние Джаккарда, DTW, Левенштейна, приложения метрического подхода: нечёткий матчинг таблиц, Ленкор, в DL, классификация текстов, эффективные методы поиска ближайших соседей
16. Линейные методы: линейная регрессия, обобщённая линейная регрессия, проблема вырожденности матрицы, регуляризация, основные виды регуляризации, гребневая регрессия (Ridge Regression), LASSO (Least Absolute Selection and Shrinkage Operator), Elastic Net
17. Селекция признаков, ошибка с весами, устойчивая регрессия (Robust Regression)
18. Линейные скоринговые модели в задаче бинарной классификации, логистическая регрессия, Probit-регрессия, многоклассовая логистическая регрессия
19. Линейный классификатор, персептрон, оценка функции ошибок через гладкую функцию

6 семестр

Вопросы к экзамену

1. Деревья решений (CART), предикаты / ветвления, птветы дерева, критерии расщепления в задачах классификации: Missclassification criteria, энтропийный, Джини, критерии остановки при построении деревьев, проблема переобучения для деревьев, подрезка (post-pruning), классические алгоритмы построения деревьев решений: ID3, C5.0
2. Важности признаков, проблема пропусков (Missing Values), категориальные признаки, сравнение: деревья vs линейные модели
3. Проблема контроля качества, выбора модели (Model Selection) в широком смысле, правила разбиения выборки, кривые обучения (Learning Curves)
4. перебор параметров
5. Отложенный контроль (held-out data, hold-out set), скользящий контроль (cross-validation), бутстреп (bootstrap), контроль по времени (out-of-time-контроль), локальный контроль
6. Ансамбли алгоритмов: примеры и обоснование, комитеты (голосование) / усреднение, бэгинг, кодировки / перекодировки ответов, ECOC

7. Стекинг и блендинг, бустинг: AdaBoost, Forward stagewise additive modeling (FSAM), «Ручные методы», однородные ансамбли
8. Случайный лес, его параметры, их настройка, бэггинг и OOB (out of bag), важность признаков, близость (Proximity) с помощью RF, Extreme Random Trees
9. Градиентный бустинг над деревьями, его параметры, современные реализации, продвинутые методы оптимизации
10. Рекомендательные системы, персонализация, онлайн и оффлайн рекомендации, рекомендация по контенту (content based methods), One-class recommendation, использование дополнительной информации, современные тренды в практике построения рекомендательных систем
11. Коллаборативная фильтрация: GroupLens-алгоритм, SVD, SVD++, timeSVD++, адаптация SVD под социальные связи
12. Факторизационная машина, факторизационная машина с полями (FFM – field-aware factorization machine)
13. Простые методы рекомендаций: FPM – Frequent Pattern Mining, Deep Semantic Similarity Model (DSSM), контекст рекомендации, Knowledge-based Recommendations, важность объяснений (explanations)
14. Сложность алгоритмов, переобучение, смещение и разброс: проблема обобщения, переобучение, недообучение, сложность алгоритмов, смещение и разброс, способы борьбы с переобучением

Задачи к экзамену

В предыдущей задаче пусть указанные распределения – распределения классов 0 и 1 в задаче бинарной классификации. Оба класса равновероятны. Какая вероятность, что объект x=1 принадлежит классу 0?

1/2	2/3	3/4	1

При минимизации функции x^2 методом градиентного спуска с темпом 0.5 и начальной точкой 1.0, какая будет оценка argmin после 1й итерации?

– 0.5	0	0.5	1

Выберите верные фразы

Для селекции признаков обычно используют L2-регуляризацию	Логистическая регрессия – ленивый алгоритм
Евклидово расстояние – частный случай расстояния Махалонобиса	С помощью персептронного алгоритма можно решать системы линейных уравнений

Чему рано максимальное значение MC (Missclassification criteria)?

0	0.5	e	1

Пусть дана выборка целевых значений: 1, 2, 3 (упорядочено по времени получения меток). Используется модель константных алгоритмов (ответ равен среднему по всем меткам обучения). Функция ошибки – MAE (средний модуль отклонения). Чему равна средняя ошибка при контроле LOOCV (контроля по одному)?

0.5	2/3	1	3/2

В каком ансамбле следует использовать неустойчивые модели?

бэгинг	случайные леса	бустинг	ECOC

Что происходит при увеличении числа деревьев в градиентном бустинге (отметьте все варианты)?

ошибка на обучении падает	ошибка на контроле падает	ошибка на обучении возрастает	ошибка на контроле возрастает

Выберите верные фразы:

Критерий gini используется для построения деревьев в задаче регресии	В экстремальных лесах (Extreme Random Trees) используется вычисление градиента ошибки
Контроль по фолдам используется для отбора модели	Аугментация – способ увеличения обучающей выборки

Пример экзаменационного билета

1. Ансамбли алгоритмов: примеры и обоснование, комитеты (голосование) / усреднение, бэгинг, кодировки / перекодировки ответов, ECOC
2. Стекинг и блендинг, бустинг: AdaBoost, Forward stagewise additive modeling (FSAM), «Ручные методы», однородные ансамбли
3. Пусть дана выборка целевых значений: 1, 2, 3 (упорядочено по времени получения меток). Используется модель константных алгоритмов (ответ равен среднему по всем меткам обучения). Функция ошибки – MAE (средний модуль отклонения). Чему равна средняя ошибка при контроле LOOCV (контроля по одному)?

0.5	2/3	1	3/2

VII. Материально-техническое обеспечение

Учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, Учебная аудитория № 310 (170002, Тверская область, г.Тверь, пер. Садовый, д.35)	Набор учебной мебели, меловая доска
Учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, Учебная аудитория № 205 (170002, Тверская область, г.Тверь, пер. Садовый, д.35)	Набор учебной мебели, экран, проектор.
Учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, Учебная аудитория № 318 (170002, Тверская область, г.Тверь, пер. Садовый, д.35)	Набор учебной мебели, экран, проектор.
Учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, Учебная аудитория № 3л (170002, Тверская область, г.Тверь, пер. Садовый, д.35)	Набор учебной мебели, экран, компьютер, проектор, МФУ.

VIII. Сведения об обновлении рабочей программы дисциплины

№ п.п.	Обновленный раздел рабочей программы дисциплины	Описание внесенных изменений	Реквизиты документа, утвердившего изменения
1	11. 2) Программное обеспечение	Внесены изменения в список ПО	От 24.08.2023 года, протокол № 1 ученого совета факультета
2	V. 1) Рекомендуемая литература	Обновление ссылок на литературу	От 24.08.2023 года, протокол № 1 ученого совета

			факультета