

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«07» марта 2024 г.
Протокол №1

**Рабочая программа дисциплины (модуля)
«Machine Learning (Машинное обучение)»**

Направление подготовки: 02.03.01 Математика и компьютерные науки

Направленность (профиль) подготовки: Математика и компьютерные науки

Квалификация (степень) выпускника: бакалавр

Форма обучения: очная

Срок освоения программы: 4 года

Год набора: 2024

**Москва
2024**

Содержание

1. Краткая характеристика дисциплины (модуля)	3
2. Перечень планируемых результатов обучения	5
3. Тематический план	7
4. Содержание дисциплины (модуля)	7
5. Учебно-методическое обеспечение	8
6. Материально-техническое обеспечение	8
7. Методические и оценочные материалы	10

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Machine Learning (Машинное обучение)» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по специальности 02.03.01 Математика и компьютерные науки, профиль Математика и компьютерные науки, утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) Machine Learning (Машинное обучение) позволяет студентам научиться разрабатывать модели, способные анализировать большие объемы данных и делать предсказания, что находит применение в различных отраслях. Кроме того, знание методов машинного обучения способствует автоматизации процессов и улучшению принятия решений.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки бакалавриата по направлению 02.03.01 Математика и компьютерные науки, профиль Математика и компьютерные науки и входит в обязательную часть Блока 1.

Дисциплина (модуль) изучается на 3 курсе в 5 семестре.

Цель изучения дисциплины (модуля): заключается в формировании у студентов теоретических знаний и практических навыков по основам машинного обучения, овладении студентами инструментарием, моделями и методами машинного обучения, а также приобретении навыков исследователя данных (data scientist) и разработчика математических моделей, методов и алгоритмов анализа данных.

Задачи изучения дисциплины (модуля):

- формирование знаний ключевых терминов и методологии DS, BigData, ML и DL, основные этапы жизненного цикла ML-задачи;
- формирование знаний о видах метрик и функционалов потерь для задач регрессии и классификации, и особенности их использования;
- формирование знаний принципов работы и особенностей метрических алгоритмов (kNN, регрессия Надарая-Ватсона), принципов построения и обучения линейных моделей (линейная, логистическая регрессия, SVM), их плюсы, минусы и ограничения;
- формирование знаний, что такое градиентный спуск, его разновидности и области применения, особенности и принципы работы деревьев решений и ансамблей на их основе;
- формирование знаний сути декомпозиции ошибки на смещение и разброс (bias-variance decomposition), понятий ансамблей (бэггинг, стекинг, blending), алгоритмы Random Forest и бустинг, их преимущества и недостатки, принципов и подходов для uplift-моделирования, способы оценки получаемых моделей;
- формирование знаний о подходах и алгоритмах кластеризации и снижения размерности (k-means, DBSCAN, иерархическая кластеризация, PCA, t-SNE, UMAP, IsoMap), методов детекции аномалий: статистические (z-score, тест Граббса, межквартильный размах), метрические (kNN, LOF, Махаланобис) и ML-(Isolation Forest, One-Class SVM);
- формирование знаний о видах признаков (числовые, категориальные, временные, пространственные, финансовые) и подходов к их генерации и отбору, базовых подходов и инструментов интерпретации моделей машинного обучения (LIME, SHAP, Shapley flow);
- формирование умения проводить базовый анализ данных (EDA) в Python и делать из него выводы, рассчитывать и интерпретировать ключевые метрики качества моделей (регрессия, классификация);

— формирование умения использовать подходы к валидации моделей (hold-out, cross-validation, bootstrap), применять Pipeline, реализовывать и настраивать метрические алгоритмы (kNN, Надарая-Ватсона), применять разные меры расстояний, обучать, настраивать и интерпретировать линейные модели (регрессия, логистическая регрессия, SVM);

— формирование умения реализовывать градиентный спуск (и модификации) и использовать в обучении моделей, обучать и настраивать деревья решений и ансамблевые алгоритмы (бэггинг, стекинг, blending, Random Forest, бустинг), оценивать результаты, строить uplift-модели для оценки причинно-следственных эффектов и оценивать их качество;

— формирование умения реализовывать алгоритмы кластеризации и снижения размерности, интерпретировать результаты, выполнять поиск аномалий с помощью статистических, метрических и ML-алгоритмов, визуализировать их и оценивать качество;

— формирование умения создавать и отбирать полезные признаки (временные, географические, финансовые и т.д.) и работать с утечками данных, интерпретировать результаты моделей и визуализировать влияние признаков с помощью SHAP/LIME.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
ОПК-4.	Способен находить, анализировать, реализовывать программно и использовать на практике математические алгоритмы, в том числе с применением современных вычислительных систем	ОПК-4.1.	Знает базовые основы современного математического аппарата, связанного с проектированием, разработкой, реализацией и оценкой качества программных продуктов и программных комплексов в различных областях человеческой деятельности.
		ОПК-4.2.	Умеет использовать этот математический аппарат в профессиональной деятельности.
		ОПК-4.3.	Имеет практический опыт применения современного математического 3 аппарата, связанного с проектированием, разработкой, реализацией и оценкой качества программных продуктов и программных комплексов в различных областях человеческой деятельности.
ОПК-6.	Способен разрабатывать алгоритмы и компьютерные программы, пригодные для практического применения	ОПК-6.1.	Знает алгоритмы разработки, компьютерные программы, а также алгоритмы вычислительной математики в области искусственного интеллекта
		ОПК-6.2.	Умеет разрабатывать математические программные продукты и комплексы с использованием современных технологий программирования в области искусственного интеллекта
		ОПК-6.3.	Имеет практический опыт разработки интеллектуальных информационных систем для визуализации результатов исследований в области искусственного интеллекта
ПК-3.	Способен применять методы математического и алгоритмического моделирования для решения как теоретических, так и практических задач в рамках	ПК-3.1.	Знает основные методы математического и алгоритмического моделирования, а также их применение для решения теоретических и прикладных

	профессиональной деятельности		задач
		ПК-3.2.	Умеет применять методы математического и алгоритмического моделирования для анализа и решения различных задач в области математики и компьютерных наук
		ПК-3.3.	Имеет опыт использования методов математического и алгоритмического моделирования при решении теоретических и прикладных задач в профессиональной деятельности

3. Тематический план

№п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Контактная работа		Контроль	Самостоятельная работа	
Лекции	Практические занятия					
1	Основные термины машинного обучения и смежных областей. Оценка качества моделей	8	20		16	Домашнее задание Коллоквиум
2	Линейные модели	8	22		16	Домашнее задание Проект
3	Деревья решений и модели, основанные на них	8	24		17	Домашнее задание Тест Соревнование
4	Работа с признаками	6	24		17	Домашнее задание Тест
	<i>Экзамен</i>			4		
	Итого:	30	90	4	66	
	Объем дисциплины (модуля) (в ак. ч.)	190				
	Объем дисциплины (модуля) (в зач. ед.)	5				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Основные термины машинного обучения и смежных областей. Оценка качества моделей	Метрики и функционалы потерь. Функционалы потерь и метрики в задачах регрессии и классификации. Контроль качества и выбор модели. Метрические алгоритмы
2	Линейные модели	Линейная регрессия. Логистическая регрессия, SVM
3	Деревья решений и модели, основанные на них	Решающие деревья. Сложность алгоритмов. Ансамбли алгоритмов. Градиентный бустинг. Аплифт-моделирование – принципы и методы
4	Работа с признаками	Подтипы задач и их особенности, кластеризация. Методы понижения размерности. Обнаружение аномалий. Генерация признаков. Отбор признаков. Интерпретация моделей и диагностика сдвига данных

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Платонов, А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. — 2-е изд. — Москва : Издательство Юрайт, 2025. — 89 с. — (Высшее образование). — ISBN 978-5-534-20732-3. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/558662>.

Дополнительная литература:

1. Дьяконов, А.Г. Машинное обучение и анализ данных / А.Г. Дьяконов. — URL: https://github.com/Dyakonov/MLDM_BOOK/blob/main/README.md.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

- столами и стульями;
- компьютерной техникой;
- механическими калькуляторами;
- специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование»	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		
КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное
Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое
Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		

AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Machine Learning (Машинное обучение)» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекция, семинары и домашние задания, тесты, проект, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удастся, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Участие в семинаре (аудиторная работа) – активная работа студента на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре студентам рекомендуется заранее ознакомиться с темой обсуждения, прочитать необходимые материалы и подготовить вопросы. Важно активно слушать и вовлекаться в дискуссию, высказывая свои мнения и аргументируя их. При ответах на вопросы преподавателя стоит быть уверенным, четким и логичным, опираясь на изученный материал. Также полезно поддерживать диалог с однокурсниками, чтобы обогатить обсуждение и расширить свои знания.

Соревнование – организованное мероприятие, в рамках которого участники соперничают друг с другом для достижения определенной цели, демонстрируя свои навыки, знания или способности в заданной области.

В процессе подготовки к соревнованию опирайтесь на следующие рекомендации:

1. **Понимание задачи:** внимательно изучите условия соревнования и четко определите задачу, которую необходимо решить. Убедитесь, что вы понимаете, какие метрики будут использоваться для оценки ваших результатов.

2. **Сбор данных:** ознакомьтесь с предоставленным набором данных. Проведите анализ данных, выявите пропущенные значения, выбросы и другие особенности, которые могут повлиять на модель.

3. **Выбор алгоритмов:** исследуйте различные алгоритмы машинного обучения, подходящие для вашей задачи. Начните с простых моделей, затем переходите к более сложным, если это необходимо.

4. **Обучение и валидация:** разделите данные на обучающую и тестовую выборки. Используйте кросс-валидацию для оценки качества модели и избежания переобучения.

5. **Оптимизация гиперпараметров:** Экспериментируйте с настройками алгоритмов, для нахождения оптимальных гиперпараметров.

6. **Документация и презентация:** ведите записи о своих подходах, результатах и выводах. Подготовьте ясную и структурированную презентацию для финального отчета.

7. **Обратная связь и улучшение:** после получения результатов соревнования проанализируйте ошибки и недостатки вашей модели. Используйте этот опыт для улучшения своих навыков в будущем

Проект – исследовательская работа по курсу и презентация результатов.

Для успешной подготовки к проекту: четко определите цели и задачи проекта, распределите роли и обязанности между участниками, а также установите сроки выполнения каждой части работы. Регулярно проводите встречи для обсуждения прогресса и решения возникающих вопросов.

Тест – особая форма проверки знаний. Проводится после освоения одной или нескольких тем и свидетельствует о качестве понимания основных понятий изучаемого материала. Тестовые задания составлены к ключевым понятиям, основным разделам, важным терминологическим категориям изучаемой дисциплины (модуля).

Для подготовки к тесту необходимо знать терминологический аппарат дисциплины (модуля), понимать смысл научных категорий и уметь их использовать в профессиональной лексике. Владение понятийным аппаратом, включённым в тестовые задания, позволяет преподавателю быстро проверить уровень понимания студентами важных методологических категорий.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал. Использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Machine Learning (Машинное обучение)»

Оценивание уровня учебных достижений, обучающихся по дисциплине (модулю), осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме **экзамена**, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом:

Десятибалльная оценка	Оценка за зачет	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
10	Зачтено	Отлично	<p>Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.</p>
9	Зачтено	Отлично	
8	Зачтено	Отлично	
7	Зачтено	Хорошо	<p>Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.</p>
6	Зачтено	Хорошо	
5	Зачтено	Удовлетворительно	<p>Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при</p>
4	Зачтено	Удовлетворительно	

Десятибалльная оценка	Оценка за зачет	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
			самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
3	Не зачтено	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не зачтено	Не сдан	
1	Не зачтено	Не сдан	

Дисциплина (модуль) «Machine Learning (Машинное обучение)» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	15%	13	Набор задач по темам недели
Тест	15%	1	Ответы на вопросы, по изученным темам
Соревнование	25%	1	Мероприятие, в рамках которого участники соперничают друг с другом для достижения определенной цели, демонстрируя свои навыки, знания или способности в заданной области
Проект	25%	1	Исследовательская работа по дисциплине (модулю) и презентация результатов
Экзамен	20%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Формула расчёта итоговой оценки по дисциплине (модулю) «Machine Learning (Машинное обучение)»: $\langle 0,15 \times \text{домашнее задание} + 0,15 \text{ тесты} + 0,25 \times \text{Соревнование} + 0,25 \times \text{Проект} + 0,2 \times \text{Экзамен} \rangle$.

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные задания для теста

Тест 1: Решающие деревья. Сложность алгоритмов. Ансамбли алгоритмов. Градиентный бустинг. Аплифт-моделирование

- Какой из следующих алгоритмов машинного обучения является примером ансамбля?
 - Решающее дерево
 - Градиентный бустинг
 - Линейная регрессия
 - Кластеризация

Ответ: b) Градиентный бустинг

2. Какая из следующих характеристик решающих деревьев является их основным преимуществом?

- a) Высокая скорость обучения
- b) Простота интерпретации
- c) Высокая точность прогнозирования
- d) Низкая сложность алгоритма

Ответ: b) Простота интерпретации

3. Какой из следующих методов используется для оценки сложности алгоритмов?

- a) Оценка времени обучения
- b) Оценка времени предсказания
- c) Оценка сложности алгоритма по метрике $O(n)$
- d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

4. Какой из следующих алгоритмов машинного обучения является примером градиентного бустинга?

- a) AdaBoost
- b) Gradient Boosting
- c) Random Forest
- d) Support Vector Machine

Ответ: b) Gradient Boosting

5. Какая из следующих характеристик аплифт-моделирования является его основной целью?

- a) Прогнозирование вероятности события
- b) Оценка влияния маркетинговой кампании
- c) Кластеризация клиентов
- d) Обнаружение аномалий

Ответ: b) Оценка влияния маркетинговой кампании

6. Какой из следующих методов используется для построения ансамблей алгоритмов?

- a) Бэггинг
- b) Бустинг
- c) Стэкинг
- d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

7. Какой из следующих алгоритмов машинного обучения является примером решающего дерева?

- a) CART
- b) C4.5
- c) ID3
- d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

8. Какая из следующих характеристик градиентного бустинга является его основным преимуществом?

- a) Высокая скорость обучения
- b) Простота интерпретации
- c) Высокая точность прогнозирования
- d) Низкая сложность алгоритма

Ответ: c) Высокая точность прогнозирования

9. Какой из следующих методов используется для оценки качества аплифт-моделирования?
- a) Оценка времени обучения
 - b) Оценка времени предсказания
 - c) Оценка сложности алгоритма
 - d) Оценка точности прогнозирования

Ответ: d) Оценка точности прогнозирования

10. Какой из следующих алгоритмов машинного обучения является примером ансамбля решающих деревьев?
- a) Random Forest
 - b) Gradient Boosting
 - c) AdaBoost
 - d) Support Vector Machine

Ответ: a) Random Forest

Тест 2: Подтипы задач и их особенности, кластеризация. Методы понижения размерности. Обнаружение аномалий. Генерация признаков. Отбор признаков. Интерпретация моделей и диагностика сдвига данных.

1. Какой из следующих подтипов задач машинного обучения является примером задачи классификации?
- a) Регрессия
 - b) Кластеризация
 - c) Обнаружение аномалий
 - d) Категоризация

Ответ: d) Категоризация

2. Какой из следующих методов используется для понижения размерности данных?
- a) PCA
 - b) t-SNE
 - c) LLE
 - d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

3. Какой из следующих алгоритмов машинного обучения является примером метода обнаружения аномалий?
- a) One-Class SVM
 - b) Local Outlier Factor (LOF)
 - c) Isolation Forest
 - d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

4. Какой из следующих методов используется для генерации новых признаков?
- a) Полиномиальная регрессия
 - b) Логистическая регрессия
 - c) Дерево решений
 - d) Feature Engineering

Ответ: d) Feature Engineering

5. Какой из следующих методов используется для отбора признаков?
- a) Recursive Feature Elimination (RFE)

- b) Correlation-based feature selection
- c) Mutual Information-based feature selection
- d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

6. Какой из следующих методов используется для интерпретации моделей машинного обучения?
- a) SHAP (SHapley Additive exPlanations)
 - b) LIME (Local Interpretable Model-agnostic Explanations)
 - c) TreeExplainer
 - d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

7. Какой из следующих методов используется для диагностики сдвига данных?
- a) Statistical Process Control (SPC)
 - b) Control Charts
 - c) Drift Detection
 - d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

8. Какой из следующих алгоритмов машинного обучения является примером метода кластеризации?
- a) K-Means
 - b) Hierarchical Clustering
 - c) DBSCAN
 - d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

9. Какой из следующих методов используется для понижения размерности данных с использованием линейных преобразований?
- a) PCA
 - b) LLE
 - c) t-SNE
 - d) None of the above

Ответ: a) PCA

10. Какой из следующих методов используется для обнаружения аномалий в данных с использованием машинного обучения?
- a) One-Class SVM
 - b) Local Outlier Factor (LOF)
 - c) Isolation Forest
 - d) Все вышеперечисленные

Ответ: d) Все вышеперечисленные

Примерные задания по соревнованию

Соревнование: "Моделирование Аплифта с помощью Ансамблей и Градиентного Бустинга"

Тема: Разработка эффективных моделей аплифта-моделирования с использованием решающих деревьев, ансамблей алгоритмов и градиентного бустинга.

Задания:

1. **Задача 1:** Разработайте модель апlifта-моделирования, используя решающие деревья, для прогнозирования выгоды от маркетинговой кампании на основе набора признаков, включающего демографические и поведенческие данные клиентов.
2. **Задача 2:** Создайте ансамбль алгоритмов, включающий градиентный бустинг, для улучшения точности прогнозирования апlifта-моделирования по сравнению с моделью из задачи 1.
3. **Задача 3:** Проведите анализ сложности алгоритмов, используемых в задачах 1 и 2, и оцените их эффективность в зависимости от размера обучающей выборки.

Критерии оценки:

1. **Точность прогнозирования:** Оценка точности моделей апlifта-моделирования по метрикам MAE и RMSE.
2. **Сложность алгоритмов:** Оценка сложности алгоритмов, используемых в задачах 1 и 2, по метрикам времени обучения и времени предсказания.
3. **Эффективность ансамблей:** Оценка эффективности ансамблей алгоритмов, используемых в задаче 2, по метрикам улучшения точности прогнозирования по сравнению с моделью из задачи 1.
4. **Качество презентации:** Оценка качества презентации результатов, включая ясность и структурированность отчета, а также визуализацию данных.

Примерное задание для проекта

Задание для проекта по теме: Линейные модели. Линейная регрессия. Логистическая регрессия. SVM

Студентам предлагается провести анализ данных с использованием линейной регрессии, логистической регрессии и метода опорных векторов (SVM). Для этого необходимо выбрать набор данных из открытых источников (например, Kaggle, UCI Machine Learning Repository) и выполнить следующие шаги:

1. Выбор и описание данных:

- Выберите набор данных, который подходит для анализа с использованием линейной и логистической регрессии, а также SVM.
- Опишите данные: какие переменные присутствуют, каковы их типы, и какую задачу вы собираетесь решить (регрессия или классификация).

2. Предобработка данных:

- Проведите анализ данных на наличие пропусков и аномалий.
- Выполните необходимые преобразования, такие как нормализация, кодирование категориальных переменных и удаление выбросов.

3. Моделирование:

- Постройте модель линейной регрессии и оцените её качество, используя такие метрики, как MSE (среднеквадратичная ошибка) или R^2 .
- Постройте модель логистической регрессии и оцените её качество, используя такие метрики, как accuracy, precision, recall и F1-score.
- Постройте модель SVM и сравните её качество с предыдущими моделями.

4. Сравнительный анализ:

- Сравните результаты всех трех моделей. Обсудите, какая модель показала лучшие результаты и почему.
- Проанализируйте влияние различных факторов на качество моделей.

5. Выводы:

- Сделайте выводы о применимости различных моделей для решения поставленной задачи и предложите рекомендации по выбору модели в зависимости от специфики данных.

Критерии оформления проекта:

Структура:

- Введение (описание задачи и целей проекта).
- Описание данных (выбор набора данных и его характеристика).
- Предобработка данных (методы и результаты).
- Моделирование (описание каждой модели, результаты и метрики).
- Сравнительный анализ (сравнение моделей и выводы).
- Заключение (общие выводы и рекомендации).

Оформление:

- Проект должен быть оформлен в виде документа (например, PDF или Word).
- Используйте графики и таблицы для визуализации результатов.
- Все коды должны быть оформлены и комментированы (можно использовать Jupyter Notebook или аналогичные инструменты).

Объем:

Минимальный объем проекта — 10 страниц (без учета графиков и таблиц).

Сдача проекта:

- Проект необходимо представить в электронном виде до установленного срока (указать дату).
- Презентация проекта (10-15 минут) должна быть подготовлена для защиты: краткое изложение целей, методов, результатов и выводов.
- Оценка проекта будет проводиться на основе качества анализа, глубины понимания темы, ясности изложения и оригинальности подхода.

Примерные домашние задания

Домашнее задание по теме «Основные термины машинного обучения и смежных областей. Оценка качества моделей»

Задание 1: Метрики и функционалы потерь

Найдите примеры функционалов потерь, используемых в задачах регрессии и классификации. Опишите их особенности и применения. Кроме того, объясните понятие средней гипотезы и ее связь с функционалами потерь.

Задание 2: Контроль качества и выбор модели

Опишите методы контроля качества моделей машинного обучения, такие как валидация и кросс-валидация. Explain понятие ошибки внутри и вне выборки и ошибки обобщения. Кроме того, объясните неравенство Хёфдинга и его применение в контроле качества моделей.

Задание 3: Метрические алгоритмы

Найдите примеры метрических алгоритмов, используемых в задачах классификации и регрессии. Опишите их особенности и применения. Кроме того, объясните понятие размерности Вапника-Червоненкиса и ее связь с метрическими алгоритмами.

Домашнее задание по теме «Линейные модели»

Задание 1: Линейная регрессия

Дан набор данных, содержащий информацию о цене на жилье и площади квартир в городе. Используя линейную регрессию, построить модель, которая позволяет предсказать цену на жилье на основе площади квартиры.

- Найдите коэффициенты линейной регрессии (β_0 , β_1) используя метод наименьших квадратов.
- Оцените качество модели с помощью коэффициента детерминации (R^2).
- Дайте интерпретацию результатов и объясните, как можно использовать эту модель для предсказания цены на жилье.

Задание 2: Логистическая регрессия и SVM

Дан набор данных, содержащий информацию о клиентах банка и их кредитной истории. Используя логистическую регрессию и SVM, построить модели, которые позволяют предсказать вероятность того, что клиент вернет кредит.

- Найдите коэффициенты логистической регрессии (β_0 , β_1) используя метод максимального правдоподобия.
- Построить модель SVM с линейным ядром и оценить ее качество с помощью коэффициента точности.
- Дайте интерпретацию результатов и объясните, как можно использовать эти модели для предсказания вероятности возврата кредита.

Задание 3: Линейные модели

Дан набор данных, содержащий информацию о влиянии различных факторов на результаты студентов в университете. Используя линейные модели, построить модель, которая позволяет предсказать результаты студентов на основе следующих факторов: возраст, пол, количество посещенных занятий и средний балл по предметам.

- Найдите коэффициенты линейной регрессии (β_0 , β_1 , ..., β_n) используя метод наименьших квадратов.
- Оцените качество модели с помощью коэффициента детерминации (R^2).
- Дайте интерпретацию результатов и объясните, как можно использовать эту модель для предсказания результатов студентов.

Домашнее задание по теме: «Решающие деревья. Сложность алгоритмов. Ансамбли алгоритмов. Градиентный бустинг. Аплифт-моделирование – принципы и методы»

Задача 1. Реализуйте алгоритмы построения дерева с критерием информационного выигрыша и критерием Джини и определению класса по мажоритарному классу в листе. Найдите оптимальную глубину дерева в обоих случаях (в отрезке 2-10).

Задача 2. Примените метод SVM (например, из библиотеки sklearn) для датасета blobs2. Визуализируйте результат (разбиение плоскости и опорные вектора) при разных вариантах ядер (линейное; полиномиальное степеней 2,3,5; RBF).

Задача 3. Найдите максимум функции с помощью алгоритма кросс-энтропийного поиска, изображая распределение на каждом шаге.

Примерный перечень вопросов для подготовки к экзамену

1. Каковы основные типы функционалов потерь в задачах регрессии и классификации?
2. Что такое метрика точности и как она используется в задачах классификации?
3. Каковы основные недостатки функционала потерь MSE в задачах регрессии?
4. Каковы основные типы функционалов потерь в задачах классификации и как они различаются?
5. Каковы основные недостатки функционала потерь Cross-Entropy в задачах классификации?
6. Каковы основные типы метрик в задачах регрессии и как они используются?
7. Каковы основные типы метрик в задачах классификации и как они используются?
8. Каковы основные недостатки функционала потерь MSE в задачах классификации?
9. Каковы основные недостатки функционала потерь Cross-Entropy в задачах регрессии?

10. Каковы основные типы функционалов потерь в задачах регрессии и классификации и как они различаются?
11. Каковы основные типы методов контроля качества в задачах машинного обучения?
12. Каковы основные недостатки метода перекрестной проверки в задачах машинного обучения?
13. Каковы основные типы критериев выбора модели в задачах машинного обучения?
14. Каковы основные недостатки критерия выбора модели по минимальной ошибке в задачах машинного обучения?
15. Каковы основные типы методов выбора модели в задачах машинного обучения?
16. Каковы основные типы метрических алгоритмов в задачах машинного обучения?
17. Каковы основные недостатки алгоритма k-NN в задачах машинного обучения?
18. Каковы основные типы алгоритмов кластеризации в задачах машинного обучения?
19. Каковы основные недостатки алгоритма k-means в задачах машинного обучения?
20. Каковы основные типы алгоритмов понижения размерности в задачах машинного обучения?
21. Каковы основные типы линейных моделей в задачах машинного обучения?
22. Каковы основные недостатки линейной регрессии в задачах машинного обучения?
23. Каковы основные типы логических моделей в задачах машинного обучения?
24. Каковы основные недостатки логистической регрессии в задачах машинного обучения?
25. Каковы основные типы методов оптимизации в задачах машинного обучения?
26. Каковы основные типы решающих деревьев в задачах машинного обучения?
27. Каковы основные недостатки решающих деревьев в задачах машинного обучения?
28. Каковы основные типы ансамблей в задачах машинного обучения?
29. Каковы основные недостатки ансамблей в задачах машинного обучения?
30. Каковы основные типы методов градиентного бустинга в задачах машинного обучения?

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Каков основной тип функционала потерь в задачах классификации? A) MSE (Mean Squared Error) B) Cross-Entropy C) Accuracy D) Precision	В	ОПК-6
2.	Каков основной недостаток метода перекрестной проверки в задачах машинного обучения? A) Высокая вычислительная сложность B) Низкая точность оценки качества модели C) Требуется большое количество данных D) Не может быть использован для оценки качества модели	В	ОПК-6
3.	Каков основной тип линейной модели в задачах машинного обучения? A) Логистическая регрессия B) Линейная регрессия C) SVM (Support Vector Machine)	В	ОПК-6

4.	Каков основной тип ансамбля в задачах машинного обучения? A) Градиентный бустинг B) RandomForest C) AdaBoost D) все вышеперечисленное	D	ПК-3
5.	Каков основной метод понижения размерности в задачах машинного обучения? A) PCA (Principal Component Analysis) B) t-SNE (t-Distributed Stochastic Neighbor Embedding) C) LLE (Local Linear Embedding) D) все вышеперечисленное	D	ОПК-6
6.	Каков основной тип проблемы, которая может возникнуть при использовании решающих деревьев в задачах машинного обучения? A) Overfitting B) Underfitting C) Error D) все вышеперечисленное	A	ОПК-4
7.	Каков основной тип метода оптимизации в задачах машинного обучения? A) Градиентный спуск B) Stochastic Gradient Descent (SGD) C) Adam D) все вышеперечисленное	D	ПК-3
8.	Как называется метрика, используемая для оценки качества классификации, которая измеряет долю правильно классифицированных объектов?	Точность/точность	ОПК-4
9.	Какой функционал потерь чаще всего используется в задачах регрессии для измерения разницы между предсказанными и фактическими значениями?	MSE	ОПК-4
10.	Как называется метод, который объединяет несколько слабых моделей для создания более сильной модели?	Бустинг	ОПК-4
11.	Какой алгоритм кластеризации основан на разделении данных на k кластеров, минимизируя внутрикластерное расстояние?	K-средних	ОПК-6
12.	Какой метод в градиентном бустинге используется для минимизации ошибки на каждом шаге?	Градиентный спуск	ОПК-6
13.	Как называется процесс, при котором модель обучается на основе уже размеченных данных?	Обучение с учителем	ОПК-6
14.	Какой критерий используется для оценки качества кластеризации, основанный на расстоянии между кластерами и внутри кластеров?	Силуэт	ПК-3
15.	Как называется метрика, которая измеряет, насколько хорошо модель разделяет классы в задачах бинарной классификации?	ROC-AUC	ПК-3
16.	Какой алгоритм используется для уменьшения размерности данных, сохраняя их важные характеристики?	PCA /метод главных компонент	ПК-3
17.	Какой подход в машинном обучении включает использование ансамблей моделей для повышения общей точности?	Ансамблирование	ПК-3