ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«24» июня 2025 г.
Протокол №2

Рабочая программа дисциплины (модуля)
«Машинное обучение 1»

Направление подготовки: *38.03.05 Бизнес-информатика*

Направленность (профиль) подготовки: *Бизнес-аналитика*

Квалификация (степень) выпускника: *бакалавр*

Форма обучения: *очная*

Срок освоения программы: *4 года*

Год набора: *2025*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B0C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Москва
2025

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) «Машинное обучение 1» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по направлению 38.03.05 «Бизнес-информатика», профиль «Бизнес-аналитика», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 838 от 29.07.2020 года.

Изучение дисциплины (модуля) «Машинное обучение 1» 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 38.03.05 «Бизнес-информатика», профиль «Бизнес-аналитика» и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений.

Дисциплина (модуль) является выборной и доступна для изучения на 3 курсе в 5 семестре на выбор.

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

Задачи изучения дисциплины (модуля):

— освоение основ теории и алгоритмов машинного обучения;
— развитие умений работы с данными и построения моделей;
— формирование навыков программирования и критического мышления.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

— ключевые понятия и термины (например, обучающая и тестовая выборки, переобучение, регуляризация);
— различные типы машинного обучения (обучение с учителем, обучение без учителя, обучение с подкреплением);
— основные алгоритмы машинного обучения (линейные модели, деревья решений, ансамбли моделей, нейронные сети и др.);
— принципы работы алгоритмов и их применимости к различным типам данных;
— методы оценки производительности моделей (точность, полнота, F-мера, ROC-AUC и др.);
— принципы кросс-валидации и настроек гиперпараметров;
— популярные библиотеки для машинного обучения (например, Scikit-Learn, TensorFlow, Keras, PyTorch);
— основы работы с языком программирования Python и его библиотеками для работы с данными (NumPy, Pandas, Matplotlib);

уметь:

— собирать, очищать и предобрабатывать данные для машинного обучения;
— визуализировать данные для выявления закономерностей и аномалий;
— выбирать и применять подходящие алгоритмы машинного обучения для решения конкретных задач;
— настраивать гиперпараметры моделей и проводить их оценку;

— интерпретировать результаты работы моделей и делать выводы на основе полученных данных;

— выявлять и устранять проблемы, такие как переобучение и недообучение;

владеть навыками:

— использования библиотек для анализа данных и машинного обучения в языке программирования Python;

— критического анализа результатов и обоснования выбора алгоритмов и методов, используемых в проекте;

— самостоятельного и систематического обновления профессиональных знаний в области машинного обучения на основе анализа актуальных исследований и технологических тенденций.

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)
УК-1.	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1.	Знает методы поиска и анализа информации в области аналитики, основные принципы критической оценки источников информации и их релевантности.
		УК-1.2.	Умеет критически оценивать источники информации и синтезировать данные из различных источников для решения задач, применять системный подход к анализу и решению комплексных проблем
		УК-1.3.	Имеет практический опыт работы с современными инструментами и технологиями для обработки информации, формулировании и структурировании задач на основе полученной информации
ОПК-3.	Способен управлять процессами создания и использования продуктов и услуг в сфере информационно-коммуникационных технологий, в том числе разрабатывать алгоритмы и программы для их практической реализации	ОПК-3.1.	Знает принципы управления процессами разработки и внедрения продуктов и услуг в сфере информационно-коммуникационных технологий
		ОПК-3.2.	Умеет разрабатывать алгоритмы и программы, обеспечивающие эффективное создание и использование информационных продуктов
		ОПК-3.3.	Имеет практический опыт в управлении проектами в области информационно-коммуникационных технологий, включая координацию команд и ресурсов для достижения поставленных целей
ОПК-4.	Способен понимать принципы работы информационных технологий; использовать информацию, методы и программные средства ее сбора, обработки и анализа для информационно-аналитической поддержки принятия управленческих решений	ОПК-4.1.	Знает основные принципы работы информационных технологий и их влияние на бизнес-процессы
		ОПК-4.2.	Умеет использовать методы и программные средства для сбора, обработки и анализа информации, обеспечивая качественную информационно-аналитическую поддержку
		ОПК-4.3.	Имеет практический опыт в применении аналитических инструментов для поддержки принятия управленческих решений в организациях

ПК-1.	Способен использовать основные методы естественнонаучных, экономических и ИТ-дисциплин в профессиональной деятельности для теоретического и экспериментального исследования	ПК-1.1.	Знает ключевые методы естественнонаучных, экономических и ИТ-дисциплин, применяемые в профессиональной деятельности
		ПК-1.2.	Умеет интегрировать различные методологические подходы для проведения теоретических и экспериментальных исследований
		ПК-1.3.	Имеет практический опыт применения методов в реальных проектах для достижения научных и практических результатов
ПК-2.	Способен использовать соответствующий математический аппарат и инструментальные средства для обработки, анализа и систематизации информации по теме исследования для решения задач профессиональной деятельности	ПК-2.1.	Знает основные математические методы и инструментальные средства, применяемые для обработки и анализа информации
		ПК-2.2.	Умеет эффективно использовать математический аппарат для систематизации данных и решения профессиональных задач
		ПК-2.3.	Имеет практический опыт работы с инструментами анализа информации в рамках исследовательских проектов
ПК-8.	Способен под руководством специалиста более высокой категории осуществлять планирование и организацию проектной деятельности на основе стандартов управления проектами	ПК-8.1.	Знает принципы и стандарты управления проектами
		ПК-8.2.	Умеет разрабатывать планы и организовывать проектную деятельность в соответствии с установленными стандартами
		ПК-8.3.	Имеет практический опыт участия в проектной работе, включая планирование и координацию задач

3. Тематический план

№п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары			
1	Основные термины машинного обучения и смежных областей. Оценка качества моделей	8	8		8	Домашнее задание
2	Линейные модели	12	12		12	Домашнее задание
3	Деревья решений и модели, основанные на них	20	20	4	18	Домашнее задание Коллоквиум Проект
4	Обучение без учителя, работа с признаками	20	20	6	18	Домашнее задание Коллоквиум
	Экзамен			4		
	Итого:	***60***	***60***	***14***	***56***	
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 в задачах регрессии и классификации. Контроль качества и выбор модели. Метрические алгоритмы
2	Линейные модели	Линейная регрессия. Логистическая регрессия, метод опорных векторов
3	Деревья решений и модели, основанные на них	Решающие деревья. Сложность алгоритмов. Ансамбли алгоритмов Случайные леса. Градиентный бустинг. Аплифт-моделирование
4	Обучение без учителя, работа с признаками	Обучение без учителя. Кластеризация и методы понижения размерности. Обнаружение аномалий. Генерация и отбор признаков. Интерпретация моделей и диагностика сдвига данных: Частые вопросы на собеседованиях по темам машинного обучения

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. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных : учебник / П. Флах. — 2-е изд. — Москва : ДМК Пресс, 2023. — 401 с. — ISBN 978-5-89818-300-4. — URL: https://znanium.ru/catalog/product/2085038.

2. Мэрфи, К. П. Вероятностное машинное обучение: введение : практическое руководство / К. П. Мэрфи ; пер. с англ. А. А. Слинкина. — Москва : ДМК Пресс, 2023. — 990 с. — ISBN 978-5-93700-119-1. — URL: https://znanium.ru/catalog/product/2109489.

3. Шолле, Ф. Глубокое обучение на Python : практическое руководство / Ф. Шолле. - Санкт-Петербург : Питер, 2023. - 576 с. - (Библиотека программиста). - ISBN 978-5-4461-1909-7. - Текст : электронный. - URL: https://znanium.com/catalog/product/2123389.

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.

2. 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.

3. Введение в статистическое обучение с примерами на языке R / Джеймс Г. [и др.]. — 2-е изд., испр. — Москва : ДМК Пресс, 2018. — 455 с. : ил. — Библиогр. : с. 450—455.

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) «Машинное обучение 1» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекция, семинары и домашние задания, проект, коллоквиумы, 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 удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Семинар – это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

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.

Коллоквиум – устные ответы на вопросы, список которых известен студенту заранее.

В процессе подготовки к коллоквиуму необходимо проанализировать учебные материалы, ознакомившись с лекциями, учебниками и дополнительными источниками, акцентируя внимание на ключевых темах. Рекомендуется создать структурированные конспекты, выделяя основные идеи, термины и формулы.

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) «Машинное обучение 1»

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, если иное не указано в силлабусе дисциплины:

Десятибалльная оценка	**Пятибалльная оценка**	**Общая характеристика результата обучения по дисциплине (модулю)**
10	Отлично	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.
9	Отлично	
8	Отлично	
7	Хорошо	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы,
6	Хорошо	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		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.
5	Удовлетворительно	Студент обладает базовыми знаниями по дисциплине, но испытывает трудности при 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.
4	Удовлетворительно	
3	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 сдан	
1	Не сдан	

Дисциплина (модуль) «Машинное обучение 1» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	30%	10	Набор задач по темам недели
Коллоквиум	30%	3	Устные ответы на вопросы, список которых известен студенту заранее
Проект	20%	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) «Машинное обучение 1»: «0,3 х среднее за домашние задания + 0,3 х среднее за коллоквиумы + 0,2 х Проект + 0,2 х Экзамен».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные вопросы к коллоквиуму

Коллоквиум 1: Линейные модели

Вопросы:

1. Что такое линейная регрессия и в каких задачах она применяется?
2. Как интерпретировать коэффициенты в линейной модели?
3. Что такое функция потерь в линейной регрессии? Какая функция используется чаще всего?
4. В чём разница между линейной и логистической регрессией?

5. Как логистическая регрессия преобразует линейную комбинацию признаков в вероятность?

6. Что такое S-образная (сигмоидная) функция и зачем она нужна?

7. Объясните, как работает метод опорных векторов (SVM) в задаче классификации.

8. Что такое зазор (margin) и почему SVM старается его максимизировать?

9. Какое влияние оказывает параметр C в SVM? Что происходит при слишком малом или слишком большом значении?

10. Зачем нужна регуляризация в линейных моделях? Чем отличаются L1 и L2 регуляризация?

Критерии оценивания:

— Точность и корректность определений и объяснений.

— Понимание математической и логической основы моделей.

— Умение объяснить, как и зачем применяются модели на практике.

— Способность сравнивать подходы и аргументировать выбор метода.

— Чёткость, логичность и структурированность ответа.

Коллоквиум 2: Деревья решений и ансамбли

Вопросы:

1. Как строится решающее дерево? Какие критерии используются для разбиения узлов?

2. Что такое прирост информации (information gain) и неоднородность Джини?

3. Почему решающие деревья склонны к переобучению? Какие способы борьбы существуют?

4. Как работает алгоритм случайного леса (Random Forest)?

5. Почему в случайном лесе используются подвыборки признаков при каждом разбиении?

6. Как устроен градиентный бустинг (например, XGBoost)? Чем он отличается от случайного леса?

7. Что такое остатки (residuals) и как они используются в бустинге?

8. Как ансамбли моделей повышают качество предсказаний по сравнению с отдельными деревьями?

9. Что такое аплифт-моделирование? В каких задачах оно применяется?

10. Приведите пример бизнес-задачи, где аплифт-модель полезнее обычной модели отклика.

Критерии оценивания:

— Понимание принципов построения деревьев и ансамблей.

— Умение объяснить разницу между Random Forest и Gradient Boosting.

— Способность описать, как избежать переобучения и оценить важность признаков.

— Понимание специфических подходов, таких как аплифт-моделирование.

— Ясность, логичность и правильное использование терминологии.

Коллоквиум 3: Обучение без учителя и работа с признаками

Вопросы:

1. Что такое обучение без учителя? Приведите примеры задач.

2. Как работает алгоритм k-means? Как он обновляет центры кластеров?

3. Какие проблемы есть у k-means? В каких случаях он работает плохо?

4. Как определить оптимальное число кластеров (например, метод локтя или силуэт)?

5. Что такое метод главных компонент (PCA)? Как он помогает при работе с данными?

6. Как интерпретировать главные компоненты и что такое доля объяснённой дисперсии?

7. Какие методы обнаружения аномалий вы знаете? Опишите один из них.

8. В чём идея изоляционного леса (Isolation Forest)? Почему он эффективен для аномалий?

9. Какие способы генерации и отбора признаков существуют?

10. Что такое сдвиг данных (data drift)? Почему его важно отслеживать в реальных системах?

Критерии оценивания:

— Понимание алгоритмов кластеризации, понижения размерности и обнаружения аномалий.

— Умение объяснить, зачем нужны методы работы с признаками.

— Осознание важности мониторинга моделей в продакшене (например, data drift).

— Способность привести практические примеры и объяснить логику работы методов.

— Чёткость, последовательность и грамотное использование терминов.

Примерное задание для проекта

Задание для проекта по теме: Линейные модели

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 ($\beta 0$, $\beta 1$) 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 (β0, β1) 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 (β0, β1, ..., βn) 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: «Деревья решений и модели, основанные на них»

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.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1	Назовите обобщающее название основных типов задач машинного обучения.	Задачи обучения (с учителем, без учителя, с подкреплением)	УК-1
2	Укажите интерпретацию коэффициента детерминации R^2 для идеальной модели линейной регрессии.	Идеальная подгонка ($R^2 = 1$)	ПК-2
3	Назовите тип параметров в простой линейной регрессии (без регуляризации).	Веса и смещение (коэффициенты)	ПК-2
4	Назовите метрику для оценки качества в задаче бинарной классификации.	Точность / Precision / Recall / F1 / ROC-AUC	ПК-1
5	Назовите преимущество дерева решений по умолчанию в библиотеке scikit-learn.	Простота интерпретации	УК-1
6	Назовите метод ансамблирования в случайном лесе.	Бутстрэппинг (Bagging)	ОПК-3

7	Назовите роль learning rate в градиентном бустинге XGBoost.	Скорость обучения (шаг градиентного спуска)	ОПК-3
8	Назовите алгоритм кластеризации для метода k-средних.	K-means	ОПК-3
9	Назовите цель метода рекурсивного исключения для отбора признаков.	Снижение размерности (устранение шумовых признаков)	ОПК-4
10	Назовите метод генерации новых признаков с порогом корреляции.	Полиномиальные признаки / Комбинирование признаков	ПК-2
11	Назовите тип палитры для визуализации данных в matplotlib.	Дискретная / Непрерывная палитра	ОПК-4
12	Назовите библиотеку, рекомендуемую для машинного обучения на Python.	Scikit-learn	ОПК-4
13	Назовите технологию для параллельной обработки в случайном лесе.	Многопоточность (Multi-processing)	ОПК-3
14	Назовите назначение батча в нейронных сетях Keras.	Мини-батч градиентный спуск (оптимизация памяти)	УК-1
15	Назовите технику для предотвращения переобучения при обучении.	Ранняя остановка / Регуляризация / Dropout	ОПК-3
16	Назовите баланс bias-variance tradeoff для оптимальной модели.	Минимальная общая ошибка (баланс смещения и разброса)	ПК-2
17	Назовите тип иерархической кластеризации для визуализации дендрограммы.	Агломеративная кластеризация	ПК-1
18	Какие правовые нормы регулируют использование персональных данных при обучении моделей в РФ?	152-ФЗ «О персональных данных»	ПК-8
19	Какие ресурсы необходимо учесть при выборе между глубокой сетью и линейной моделью?	Вычислительная мощность / Время обучения / Объем данных	ПК-8
20	Как системно подойти к выбору метрики безопасности при развертывании модели?	Анализ угроз / Контроль доступа / Аудит логов	ПК-8