

УТВЕРЖДЕНА

Приказом Ректора АНО ВО
«Центральный университет»
Ивашкевич Е.В.
от «19» января 2024 г. № 0119.37

**Рабочая программа дисциплины (модуля)
«Machine Learning (Машинное обучение)»
дополнительной профессиональной программы – программы
профессиональной переподготовки «Академия data science»**

Траектория: Продуктовая аналитика

**Москва
2024**

Содержание

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

1. Краткая характеристика дисциплины (модуля)

Изучение дисциплины (модуля) Machine Learning (Машинное обучение) позволяет слушателям 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.

Цель изучения дисциплины (модуля): формирование у слушателей теоретических знаний и практических навыков по основам машинного обучения, овладение слушателями 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):

- понимание ключевых понятий и терминов (например, обучающая и тестовая выборки, переобучение, регуляризация);
- знание различных типов машинного обучения (обучение с учителем, обучение без учителя, обучение с подкреплением);
- знание основных алгоритмов машинного обучения (линейные модели, деревья решений, ансамбли моделей, нейронные сети и др.);
- понимание принципов работы алгоритмов и их применимости к различным типам данных;
- изучение методов оценки производительности моделей (точность, полнота, F-мера, ROC-AUC и др.);
- знание принципов кросс-валидации и настроек гиперпараметров;
- ознакомление с популярными библиотеками для машинного обучения (например, Scikit-Learn, TensorFlow, Keras, PyTorch);
- знание основ работы с языком программирования Python и его библиотеками для работы с данными (NumPy, Pandas, Matplotlib);
- умение собирать, очищать и преобразовывать данные для машинного обучения;
- способность визуализировать данные для выявления закономерностей и аномалий; умение выбирать и применять подходящие алгоритмы машинного обучения для решения конкретных задач;
- способность настраивать гиперпараметры моделей и проводить их оценку;
- умение интерпретировать результаты работы моделей и делать выводы на основе полученных данных, способность выявлять и устранять проблемы, такие как переобучение и недообучение;
- уверенное владение языком программирования Python и умение работать с библиотеками для анализа данных и машинного обучения;
- умение критически анализировать результаты и обосновывать выбор алгоритмов и методов, используемых в проекте.

2. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контр оль	Самосто ятельна я работа	
Лекции	Практиче ские занятия					
1	Введение в машинное обучение, постановка задач	8	9		23	Домашнее задание Тест
2	Модели обучения на размеченных данных	8	8		23	Домашнее задание Соревнование
3	Контроль качества и сложность алгоритмов	8	8		24	Домашнее задание Тест
4	Обучение на неразмеченных данных и подготовка данных	9	9		24	Домашнее задание Тест
	<i>Экзамен</i>			4		
	Итого:	33	34	4	94	
	Объем дисциплины (модуля) (в ак. ч.)	165				

3. Содержание дисциплины (модуля)

№ п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Введение в машинное обучение, постановка задач	Постановка основных задач машинного обучения
2	Модели обучения на размеченных данных	Линейная регрессия Метрические алгоритмы. Контроль качества и выбор модели. Функции ошибки и функционалы качества. Линейные модели классификации. Решающие деревья. Ансамбли алгоритмов
3	Контроль качества и сложность алгоритмов	Случайные леса. Градиентный бустинг. Сложность алгоритмов, смещение и разброс
4	Обучение на неразмеченных данных и подготовка данных	Кластеризация. Отбор признаков. Генерация признаков. Искусство визуализации

4. Учебно-методическое обеспечение

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

Каждый слушатель 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.

Слушателям 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.

5. 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 к 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	Катастрофы, стихийные бедствия, аварии, эпидемии. Солнечная и геомагнитная активность. /ежедневный обзор	http://www.disasters.chat.ru
2	Каталог по безопасности жизнедеятельности	http://www.eun.chat.ru
3	Научная электронная библиотека eLibrary.ru библиотека	https://elibrary.ru/defaultx.asp
4	База данных для IT-специалистов	https://habr.com
5	База данных ScienceDirect	https://www.sciencedirect.com
6	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
7	Федеральный портал «Российское образование»	https://www.edu.ru/
8	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
9	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
10	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
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

6. 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 лекции, практические занятия, домашние задания, соревнование, тесты, 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 практическом занятии (аудиторная работа) – активная работа слушателя на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре слушателям 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.

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 слушателями 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 слушателей, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы слушателя взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи слушателя 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)

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

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме *экзамена*.

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

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Слушатель полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину (модуль). 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. Слушатель хорошо знаком с методами исследования, необходимыми
9	Отлично	
8	Отлично	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Слушатель 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. Слушатель хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	
5	Удовлетворительно	Слушатель обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Слушатель способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	
3	Не сдан	Слушатель 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) «Machine Learning (Машинное обучение)» оценивается следующим образом:

Активность	Вес	Количество	Описание
Накопительная оценка			
Домашние задания	60%	13	Письменная работа с набором задач, которые нужно решить за ограниченное время
Соревнование		1	Kaggle-style соревнование с задачей на ML
Тесты по лекциям		3	В течение изучения дисциплины (модуля) слушателям будут предложены несколько тестов по пройденному материалу
Промежуточная аттестация			
Экзамен	40%	1	Письменная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Формула расчёта итоговой оценки по дисциплине (модулю) «Machine Learning (Машинное обучение)»: « $0,6 \times$ накопительная оценка ($0,5 \times$ среднее за домашние задания + $0,25 \times$ соревнование + $0,25 \times$ среднее за тесты по лекциям) + $0,4 \times$ экзамен»

Итоговая оценка выставляется по накопительной при условии, если средний балл слушателя за домашние задания составляет 4 и более баллов.

Если слушатель не выполняет условие для получения оценки по накопительной системе, ему необходимо сдать экзамен. В данном случае оценка за дисциплину (модуль) выставляется по формуле: « $0,6 \times$ накопительная оценка за семестр + $0,4 \times$ оценка за экзамен».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1

Задание 1. [0.5 балла]

Откройте файл с таблицей (не забудьте про её формат) в переменную `data`.

Положите в новую переменную `data_spb` все имеющиеся данные по объектам недвижимости Санкт-Петербурга.

Сколько записей в получившейся таблице `data_spb`?

Подсказка `region 2661` - это Санкт-Петербург. Все коды регионов можно найти в первом семинаре ML

Задание 2. [0.5 балла]

Проведите анализ данных: сделайте следующие шаги и ответьте на вопросы.

- Есть ли в данных по Санкт-Петербургу пропущенные значения?
- Удалите те данные (в `data_spb` и `data`), для которых значение цены (`price`) ≤ 0
- Сколько строк осталось в `data` и `data_spb`?

Задание 3. [2 балла]

Поисследуйте, как распределены числовые признаки. Для этого изобразите на одном полотне `box plot` всех числовых признаков.

Подсказка В данном задании, возможно, это будет сделать чуть удобнее через `matplotlib` или `pandas`

Подсказка Значения будут лучше видны, если вы сделаете логарифмирование по оси `y` для отображения графика

Ответьте на следующие вопросы:

- 1) [0.25 балла] В каких столбцах наблюдается больше всего выбросов?
- 2) [0.25 балла] Начиная с каких границ для каждого из признаков начинаются выбросы? (можно определить программно)
- 3) [0.25 балла] Какой процент выбросов в каждом столбце?
- 4) [0.25 балла] Как вы думаете, почему в этих столбцах наблюдаются выбросы?
- 5) [0.5 балла] Как вы думаете, являются ли они действительно выбросами или это могли быть реальные данные? Исследования приветствуются!
- 6) [0.5 балла] Как вы думаете, в каких задачах было бы необходимо удалять данные объекты? А в каких - нет? Приведите примеры постановок таких задач

Задание 4. [1 балл]

Ответьте на все следующие вопросы.

- Какая средняя цена квартиры по Санкт-Петербургу?
 - А минимальная?
 - Максимальная?
- Все эти цены выше или ниже, чем средняя цена по стране?
- Все эти цены выше или ниже, чем средняя цена по Москве?

- Совпадает ли это с вашими ожиданиями? Чем вы можете объяснить данные значения?

Задание 5. [1 балл]

Удалите из датасетов `data` и `data_spb` те объекты, которые не входят в отрезок [0.05 квантиль; 0.95 квантиль] по признаку `price`.

Сколько строк осталось в каждом из датасетов?

Задание 6. [1 балл]

Постройте гистограмму распределения признака "цена квартиры за квадратный метр" (создайте его) в Санкт-Петербурге.

Есть ли какие-то аномалии на графике? Соответствует ли такое распределение вашим ожиданиям?

Задание 7. [1 балл]

Есть ли зависимость цены за квадратный метр от района в Санкт-Петербурге? Изобразите это (программно) на карте.

Подсказка Построение на карте с помощью `plotly` было на семинаре.

Подсказка Если график не отрисовывается, возьмите случайную выборку меньшего размера

Задание 8. [2 балла]

Ответьте на следующие вопросы:

- Есть ли зависимость цены квартиры от площади квартиры?
- А от жилой площади?
- А от этажа?

Изобразите это на графиках

Подсказка Можно выбрать случайную часть данных (провести сэмплирование), если график строится долго

Задание 9. [1 балл]

- Какие выводы можно сделать, по получившимся результатам выше?
- Какие переменные можно было бы использовать для предсказания стоимости объекта недвижимости?
- Какие гипотезы проверили бы еще с помощью первоначального анализа данных?

Домашнее задание 2

Задание 1 [5 баллов]

Вот основные этапы обучения и оценки качества случайного леса:

1. Формирование обучающей выборки для каждого решающего дерева:

Случайным образом выбирается набор объектов и их характеристик (признаков) для обучения дерева принятия решения. Целевая переменная (то, что необходимо предсказать) также входит в обучающую выборку.

2. Обучение решающих деревьев:

На основе выбранной обучающих объектов и признаков строится решающее дерево, которое пытается наилучшим образом предсказать целевую переменную.

Шаги 1-2 повторяем для обучения `n_estimators` независимых решающих деревьев.

3. Агрегирование решений деревьев и формирование прогноза:

Для каждого наблюдения каждое дерево в лесу выдает свой прогноз. Эти прогнозы агрегируются (например, путем голосования) для получения окончательного предсказания случайного леса.

4. Оценка качества модели:

Производится оценка точности прогнозов случайного леса на тестовых данных, не использованных при обучении. При необходимости, модель может быть дообучена или настроена для повышения качества.

- Выберите, какую реализацию хотите сделать - для **регрессии** или для **классификации**. В случае выбора **классификации** делайте реализацию с учетом возможности многоклассовой классификации

- **Можно использовать** уже реализованные классы `sklearn.tree.DecisionTreeClassifier` и `tree.DecisionTreeRegressor`.

- **Нельзя использовать** уже реализованные классы `sklearn.ensemble.BaggingClassifier` и `ensemble.BaggingRegressor`.

- **Необходим** код реализации. Без него задание засчитываться не будет

Комментарии

- Выберите дефолтные значения для приведенных в классе параметров и вставьте их
 - **Можно добавлять** новые параметры в метод `self` для более широкой реализации класса

- Методы `fit`, `predict` должны **обязательно присутствовать**. От них ожидается стандартный привычный функционал и вывод по аналогии с `sklearn`

- В случае выбора задачи классификации рекомендуется также сделать и метод `predict_proba`, но это необязательно

Задание 2 [0.75 балла]

Выберите и загрузите датасет в соответствии с вашей выбранной реализацией (регрессия или классификация). Датасеты можно искать, например, вот тут:

- [На kaggle](#)

- [На google](#)

Загружать датасет можно как и напрямую в среду `google colab`, так и через утилиту `gdown`, загрузив их предварительно на свой гугл-диск (по аналогии с семинарами)

- При необходимости, выделите тестовую выборку при помощи `train_test_split`

Подсказка: рекомендуем выбирать не "игрушечные"/"обучающие" датасеты. Брать слишком много данных тоже не нужно, чтобы у вас ноутбук не считался очень долго.

Задание 3 [0.75 балла]

Выберите метрику качества и обоснуйте ее выбор.

Задание 4 [1 балла]

Предобработайте датасет так, как вы хотите - можете поресерчить, почистить пропуски, шумы, аномалии и пр.

Задание 5 [2 балла]

Подберите оптимальные гиперпараметры и обучите 2 модели:

- Ваша реализация `random forest` [1 балл]

- Реализация `random forest` из `sklearn` [1 балл]

Важно: необходимо перебрать перебрать хотя бы 3 различных гиперпараметра. Сетки для каждого из них выбирайте разумные

`addКод`

`addТекст`

Задание 6 [0.5 балла]

Сравните 2 модели:

- У какой получилось выше качество на тесте?

- Какие оптимальные параметры получились у двух моделей?

- Как вы думаете, чем обусловлена разница в качестве?

Примерные задания для теста

Тест 1: Решающие деревья. Сложность алгоритмов. Ансамбли алгоритмов.

Градиентный бустинг. Аплифт-моделирование

1. Какой из следующих алгоритмов машинного обучения является примером ансамбля?
- a) Решающее дерево
 - b) Градиентный бустинг
 - c) Линейная регрессия
 - d) 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 аплифта-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 аплифта-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 аплифта-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.

Примерные вопросы для подготовки к экзамену

1. Введение в машинное обучение, постановка задач

1. Что такое машинное обучение?
2. Назовите три основных типа задач машинного обучения.
3. Какое различие между задачами классификации и регрессии?
4. Что такое переобучение (overfitting)?
5. Какова цель обучения с учителем?
6. В чем заключается задача кластеризации?
7. Что такое обучение без учителя?
8. Какова роль валидации в машинном обучении?
9. Что такое тестовая выборка и зачем она нужна?
10. Каковы основные этапы процесса машинного обучения?

2. Модели обучения на размеченных данных

11. Какова формула линейной регрессии?
12. Что такое функция ошибки в контексте линейной регрессии?
13. Какой алгоритм используется для минимизации функции ошибки в линейной регрессии?
14. Что такое метрика в машинном обучении?
15. Каковы основные метрики для оценки качества классификации?
16. Как работает алгоритм решающего дерева?
17. В чем преимущество ансамблей алгоритмов?
18. Какой метод объединяет результаты нескольких моделей для улучшения качества предсказаний?
19. Что такое градиентный бустинг?

20. Какова роль контроля качества модели в процессе обучения?

3. Контроль качества и сложность алгоритмов

21. Что такое случайный лес и как он работает?

22. Какова основная идея метода "бустинга"?

23. Каковы основные причины смещения (bias) и разброса (variance) в алгоритмах?

24. Как можно оценить сложность алгоритма?

25. Что такое кросс-валидация и как она помогает в контроле качества?

26. Каковы преимущества использования ансамблей по сравнению с одиночными моделями?

27. Как можно уменьшить переобучение в модели?

28. Какова роль гиперпараметров в обучении моделей?

29. Что такое ROC-кривая и как она используется для оценки качества классификации?

30. Какой метод можно использовать для выбора лучших признаков в модели?