Экзамен по дисциплине «Машинное обучение», состоит из теоретической и практической частей. Балльно-рейтинговая система устанавливает следующую схему выставления баллов промежуточной аттестации:

- 40 баллов текущего контроля определяются по итогам изучения всей дисциплины;
- 20 баллов из 60 отводится для выполнения теоретического задания;
- 40 баллов выставляются за выполнение практического задания.

Экзамен проходит в устной (гибридной) форме в компьютерном классе, с выполнением практического задания на лабораторном компьютере с демонстрацией и устным отчетом преподавателю.

Экзаменационный билет состоит из одного теоретического и одного практического вопроса. Время подготовки студентом ответа на экзаменационный билет - не менее 30 минут. Время ответа одного студента преподавателю - не более 20 минут.

При начале экзамена в аудиторию входит первые 6 студентов, которые последовательно, по одному тянут билет, называют преподавателю свою фамилию, номер выпавшего билета и садятся готовиться на указанное преподавателем место в аудитории.
По истечении 30 минут преподаватель начинает опро первого студента. При желании студента, он может ответить преподавателю досрочно и вне очереди.
После окончания опроса каждого студента преподаватель сразу же оглашает ему его оценку. После того, как один студент ответил билет и получил оценку, он выходит из аудитории и заходит следующий по списку.

При подготовке ответа студент может пользоваться лабораторным компьютером, встроенной документацией к библиотекам. Не допускается использование шпаргалок, собственных компьютеров, ноутбуков, смартфонов, помощи других студентов. При подготовке ответа на вопросы билета студент может делать записи как в письменном, так и в электронном виде, которые он может использовать или демонстрировать преподавателю в процессе ответа.
Преподаватель имеет право задавать дополнительные, уточняющие вопросы по теме экзаменационного вопроса для выяснения глубины и широты знаний студента. Ответ на дополнительный вопрос не предполагает времени на подготовку.

Теоретический вопрос в билете предполагает развернутый и подробный устный ответ преподавателю по заданной теме. Ответ должен раскрыть понимание студентом вопроса билета на теоретическом и практическом

уровнях. При необходимости студент должен продемонстрировать преподавателю формулы, схемы или другой вспомогательный материал

Список теоретических вопросов для подготовки к экзамену:

1. Понятие машинного обучения. Отличие машинного обучения от других областей программирования.
2. Классификация задач машинного обучения. Примеры задач из различных классов.
3. Основные понятия машинного обучения: набора данных, объекты, признаки, атрибуты, модели, параметры.
4. Структура и представление данных для машинного обучения.
5. Инструментальные средства машинного обучения.
6. Задача регрессии: постановка, математическая формализация.
7. Метод градиентного спуска для парной линейной регрессии.
8. Понятие функции ошибки: требования, использование, примеры.
9. Множественная и нелинейная регрессии.
10. Нормализация признаков в задачах регрессии.
11. Задача классификации: постановка, математическая формализация.
12. Метод градиентного спуска для задач классификации.
13. Логистическая регрессия в задачах классификации.
14. Множественная и многоклассовая классификация. Алгоритм “один против всех”.
15. Метод опорных векторов в задачах классификации.
16. Понятие ядра и виды ядер в методе опорных векторов.
17. Метод решающих деревьев в задачах классификации.
18. Метод k ближайших соседей в задачах классификации.
19. Однослойный перцептрон в задачах классификации.
20. Метрики эффективности и функции ошибки: назначение, примеры, различия.
21. Понятие набора данных (датасета) в машинном обучении. Требования, представление. Признаки и объекты.
22. Шкалы измерения признаков. Виды шкал, их характеристика.
23. Понятие чистых данных. Определение, очистка данных.
24. Основные этапы проекта по машинному обучению.
25. Предварительный анализ данных: задачи, методы, цели.
26. Проблема отсутствующих данных: причины, исследование, пути решения.
27. Проблема несбалансированных классов: исследование, пути решения.
28. Понятие параметров и гиперпараметров модели. Обучение параметров и гиперпараметров. Поиск по сетке.
29. Понятие недо- и переобучения. Определение, пути решения.
30. Диагностика модели машинного обучения. Методы, цели.
31. Проблема выбора модели машинного обучения. Сравнение моделей.

32. Измерение эффективности работы моделей машинного обучения. Метрики эффективности.
33. Метрики эффективности моделей классификации. Виды, характеристика, выбор.
34. Метрики эффективности моделей регрессии. Виды, характеристика, выбор.
35. Перекрестная проверка (кросс-валидация). Назначение, схема работы.
36. Конвейеры в библиотеке sklearn. Назначение, использование.
37. Использование методов визуализации данных для предварительного анализа.
38. Исследование коррелированности признаков: методы, цели, выводы.
39. Решкалирование данных. Виды, назначение, применение. Нормализация и стандартизация данных.
40. Преобразование категориальных признаков в числовые.
41. Методы визуализации данных для машинного обучения.
42. Задача выбора модели. Оценка эффективности, валидационный набор.
43. Кривые обучения для диагностики моделей машинного обучения.
44. Регуляризация моделей машинного обучения. Назначение, виды, формализация.
45. Проблема сбора и интеграции данных для машинного обучения.
46. Понятие чистых данных и требования к данным.
47. Основные задачи описательного анализа данных.
48. Полиномиальные модели машинного обучения.
49. Основные виды преобразования данных для подготовки к машинному обучению.
50. Задача выбора признаков в машинном обучении.
51. Задачи обучения без учителя: общая характеристика, особенности, примеры.
52. Задача кластеризации. Формализация, применение, примеры, общая характеристика. Метрики качества кластеризации.
53. Алгоритм кластеризации K-средних.
54. Иерархическая (агломеративная) кластеризация.
55. Плотностные алгоритмы кластеризации. DBSCAN.
56. Задача понижения размерности в машинном обучении. Метод главных компонент.
57. Алгоритм t-SNE. Общая характеристика, применение, особенности.
58. Задача обнаружения аномалий в машинном обучении. Использование многомерного гауссова распределения.
59. Алгоритм DBSCAN для задач обнаружения аномалий.
60. Ансамбли моделей машинного обучения.
61. Случайный лес как ансамблевая модель.
62. Стекинг как вид ансамблирования моделей. Общая характеристика, особенности.

63. Беггинг как вид ансамблирования моделей. Общая характеристика, особенности.
64. Бустинг как вид ансамблирования моделей. Общая характеристика, особенности.
65. Алгоритм градиентного бустинга. Общая характеристика. LightGBM, XGBoost, CatBoost.

Практическая часть представляет собой задачу по выполнению конкретных действий по анализу данных, либо построению, обучению или диагностике моделей машинного обучения.

Для выполнения задачи, рекомендуется использовать язык программирования Python и специализированные библиотеки для анализа данных и машинного обучения.

Студент должен прислать рабочий notebook с решением кейса задачи, комментариями кода и аналитическими выводами до конца экзамена на электронную почту преподавателя.

Критерии оценки практической экзаменационной работы:

1. Выполнение задания. Студент должен написать работающий код, выполняющий действия, указанные в экзаменационной задаче. Код не должен выполняться с ошибками.
2. Наличие выводов. Ответ должен содержать текстовые либо устные замечания, поясняющие каждый шаг работы студента: что делается, зачем и какую информацию это нам дает. Оценивается полнота и адекватность выводов.
3. Дополнительные выводы и действия. Оцениваются все дополнительные комментарии и операции, более полно раскрывающие процесс, описанный в экзаменационной задаче.

Примерные задачи для подготовки к экзамену:

1. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Несколькими способами, в том числе графическим, убедиться в отсутствии пропущенных значений.
2. Загрузить встроенный в библиотеку sklearn датасет “Диабет”. Визуализировать распределение четырех любых признаков, входящих в датасет. Сделать содержательные выводы по полученным данным.
3. Загрузить встроенный в библиотеку sklearn датасет “Рак груди”. Построить модель бинарной классификации любым методом. Вывести несколько первых теоретических и эмпирических значений целевой переменной. Сделать выводы по полученным результатам.

4. Загрузить встроенный в библиотеку sklearn датасет “Вина”. Построить линейную модель обучения с учителем, вывести и проинтерпретировать коэффициенты линейной модели. Коэффициенты должны выводиться вместе с названием соответствующего признака.
5. Загрузить встроенный в библиотеку sklearn датасет “Калифорния”. Построить модель регрессии любым методом. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать выводы.
6. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Построить модель множественной классификации любым методом. Оценить ее эффективность при помощи кросс-валидации. Сделать выводы.
7. Загрузить встроенный в библиотеку sklearn датасет “Диабет”. Построить модель регрессии по методу опорных векторов с линейным ядром. Оценить ее эффективность по метрикам r2, mae, rmse, mape. Сделать выводы о применимости модели.
8. Загрузить встроенный в библиотеку sklearn датасет “Рак груди”. Построить модель бинарной линейной классификации. Задать значения аргументов конструктора объекта модели, отличающиеся от значений по умолчанию. Пояснить смысл каждого аргумента.
9. Загрузить встроенный в библиотеку sklearn датасет “Вина”. Построить модель множественной классификации по методу опорных векторов с полиномиальным ядром. Оценить ее эффективность по метрикам accuracy, precision, recall, f1. Сделать выводы о применимости модели.
10. Загрузить встроенный в библиотеку sklearn датасет “Калифорния”. Построить модель регрессии с регуляризацией. Задать значения аргументов конструктора объекта модели, отличающиеся от значений по умолчанию. Пояснить смысл каждого аргумента.