МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ПЕРМСКИЙ ГОСУДАРСТВЕННЫЙ НАЦИОНАЛЬНЫЙ

ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Фонды оценочных средств по дисциплине

«СОВРЕМЕННЫЕ ТЕХНОЛОГИИ МАШИННОГО ОБУЧЕНИЯ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА»

1. Формируемые дисциплиной компетенции

ОПК.2.4 Способен разрабатывать и применять автоматизированные технологии обработки больших информационных потоков (массивов) финансовой и/или экономической информации в режиме реального времени

Индикаторы:

ОПК.2.4.1 Формирует разделы технических заданий на создание автоматизированных технологий обработки больших массивов экономических и финансовых данных

ОПК.2.4.2 Применяет автоматизированные технологии обработки больших информационных потоков (массивов) финансовой и/или экономической информации в режиме реального времени

2. Планируемые результаты обучения

Коды индикаторов	Планируемый результат
ОПК.2.4.1	Знает разделы технических заданий на создание автоматизированных технологий обработки больших массивов экономических и финансовых данных, понимает требования к данным, умеет делать выбор алгоритмов и технологий, анализ и обоснование архитектуры системы
ОПК.2.4.2	Умение применять автоматизированные технологии обработки больших информационных потоков (массивов) финансовой и/или экономической информации в режиме реального времени, используя методы машинного обучения и искусственного интеллекта для анализа, интерпретации и визуализации данных.

3. Спецификация теста

Тест по дисциплине «Современные технологии машинного обучения и искусственного интеллекта» представляет собой перечень примерных вопросов, предлагаемых студентам с учетом тем и заданий контрольных мероприятий, предусмотренных по дисциплине.

Вариант 1

Вопрос 1. В чем отличие задачи классификации от задачи регрессии?
A. Целевая переменная в задаче классификации имеет дискретный характер, а в задаче регрессии – непрерывный
B. Суть задачи классификации состоит в том, что нужно сгруппировать объекты выборки, выделив схожие объекты, а задачи регрессии – отнести каждый объект к той или иной заранее известной группе
C. Задача классификации относится к обучению с учителем, а задача регрессии – к обучению без учителя
D. Задача регрессии относится к задаче предсказания, а задача классификации – нет

Вопрос 2. Есть выборка из n объектов, для каждого из которых известны значения нескольких его признаков (X1, X2, X3). Необходимо разделить эту выборку на 3 группы, используя признаки X1 и X2. Как называется эта задача?
A. задача кластеризации
B. задача регрессии
C. задача классификации
D. задача определения тесноты связи между признаками

Вопрос 3. Выберите ошибочное утверждение:
A. Функция потерь максимизируется при обучении модели.
B. Функция потерь минимизируется при обучении модели.
C. Чем меньше ошибка прогноза модели на тестовой выборке, тем лучше эта модель.
D. В качестве функции потерь и метрики качества модели может использоваться один и тот же показатель.

Вопрос 4. В чем отличие задачи регрессии от задачи классификации?
A. задача регрессии, в отличие от задачи классификации, предполагает числовой непрерывный целевой признак
B. регрессия относится к обучению с учителем, а классификация – к обучению без учителя
C. задача классификации, в отличие от задачи регрессии, предполагает числовой непрерывный целевой признак
D. регрессия предполагает несколько целевых переменных, а классификация – только две (поэтому она еще называется бинарной)

Вопрос 5. Что может стать проблемой при построении линейной регрессионной модели?
A. высокая корреляция между факторными признаками
B. высокая корреляция между факторными и целевым признаками
C. очень большой объем обучающей выборки
D. отсутствие пропусков и выбросов в данных

Вопрос 6. Переобучение – это явление, при котором:
A. ошибка на тестовой выборке значительно больше ошибки на обучающей выборке
B. ошибка на обучающей выборке значительно больше ошибки на тестовой выборке
C. ошибки на тестовой и на обучающей выборке одинаково малы
D. ошибки на тестовой и на обучающей выборке одинаково велики

Вопрос 7. Для чего в процессе построения моделей машинного обучения следует делить имеющиеся размеченные данные на обучающую и тестовую выборки?
A. чтобы проверить предсказательную способность модели на данных, которых она не видела при обучении
B. чтобы минимизировать функцию потерь
C. чтобы отобрать наиболее значимые факторные признаки
D. чтобы вычислить как можно больше показателей качества модели (метрик)

Вопрос 8. Имеется набор данных, характеризующихся миллионом признаков. Учитывая, что объектов в наборе тоже весьма немало, при обработке (и даже при хранении) этих данных возникает проблема нехватки машинных ресурсов. Какая задача машинного обучения предназначена для решения подобных проблем?
A. задача понижения размерности
B. задача ранжирования
C. задача кластеризации
D. задача классификации

Вопрос 9. Как (на базе чего) рассчитываются практически все метрики качества регрессионной модели?
A. сравниваются предсказанные данные и фактические данные, на основе разницы вычисляется метрика
B. сравниваются предсказанные данные на обучающей и тестовой выборках, на основе разницы вычисляется метрика
C. метрика сильно зависит от объема выборки: чем меньше объем выборки, тем ниже качество модели
D. метрика сильно зависит от качества разбиения выборки на обучающую и тестовую: чем четче они выделены, тем выше качество модели

Вопрос 10. Выборка значений некоторых показателей была разбита на 3 равных сегмента (I, II, III). Было обучено 3 модели с одинаковыми параметрами на объединенных сегментах I+II, I+III, II+III, их качество было оценено на оставшихся сегментах. В итоге за оценку качества модели с заданными параметрами было принято усредненное значение полученных метрик. Как называется этот подход к оценке качества модели?
A. кросс-валидация
B. скользящее окно
C. стекинг
D. бутстрэп

Вопрос 11. «Проклятие размерности» – это эффект, при котором:
A. с увеличением числа независимых признаков происходит падение качества модели и рост трудоемкости вычислений
B. с увеличением числа независимых признаков происходит рост качества модели и рост трудоемкости вычислений
C. с уменьшением числа независимых признаков происходит рост качества модели и падение трудоемкости вычислений
D. с уменьшением числа независимых признаков происходит падение качества модели и падение трудоемкости вычислений

Вопрос 12. Лес решений – это пример реализации следующей техники:

A. бэггинг
B. бустинг
C. стекинг
D. SVD-разложение

Вопрос 13. Гиперпараметры модели - это:
A. параметры модели, которые не подбираются в процессе минимизации функции потерь, их необходимо задавать вручную
B. веса (коэффициенты) модели
C. коэффициенты модели, которые минимизируются при ее обучении
D. некоторые показатели, описывающие модель

Вопрос 14. Точка a называется глобальным минимумом функции f(x), если:
A. f(a) меньше f(b) для любой точки b из области определения функции
B. существует окрестность точки a такая, что f(a) меньше f(b) для любой точки b из этой окрестности
C. f'(x)=0
D. производная функции f(x) меняет знак при переходе точки a из множества отрицательных чисел в множество положительных и наоборот

Вопрос 15. В чем идея метода градиентного спуска?
A. чтобы найти точку минимума функции, нужно двигаться в направлении, противоположном направлению ее наибольшего возрастания, задаваемом ее антиградиентом
B. в том, что каждая последующая модель концентрируется на ошибках предыдущей и пытается их компенсировать
C. в минимизации суммы квадратов отклонений некоторых функций от фактических данных
D. в максимизации функции правдоподобия

Вопрос 16. Выберите наиболее продвинутую из перечисленных модификацию градиентного спуска:
A.Adam
B. Momentum
C. Adagrad
D. RMSProp

Вопрос 17. Алгоритм обратного распространения заключается в:
A. вычислении производной сложной функции (суперпозиции функций) посредством движения от всей функции целиком к ее отдельным аргументам
B. распространении сигналов ошибки от входов нейронной сети к ее выходам
C. минимизации функции потерь на обучающей выборке
D. том, что каждая последующая модель концентрируется на ошибках предыдущей и пытается их компенсировать

Вопрос 18. Формулу искусственного нейрона можно записать так:
A.y=ф(w1*x1+w2*x2+…+wm*xm+w0)
B. y=w1*x1+w2*x2+…+wm*xm+w0
C. y=ф(wm*x^m+…+w1*x+w0)

D. y=ф(w1*x1+w2*x2+…+wm*xm)+w0

Вопрос 19. Пусть имеется полносвязная нейронная сеть из 3-х слоев. Скрытый слой состоит из 5-ти нейронов. Сеть используется для решения задачи регрессии, количество факторных признаков - 3. Сколько весов (параметров) будет у такой сети?
A. 26
B. 15
C. 32
D. 8

Вопрос 20. Пусть имеется полносвязная нейронная сеть из 4-х слоев. Первый скрытый слой состоит из 5-ти нейронов, второй - из 10-ти. Сеть используется для решения задачи бинарной классификации, количество факторных признаков - 3. Сколько весов (параметров) будет у такой сети?
A. 102
B. 150
C. 91
D. 101

Вариант 2

Вопрос 1. Всегда ли предпочтительно применение нейронных сетей в качестве алгоритмов для решения задачи предсказания?
A. не всегда в силу ряда причин (сложность качественной настройки сети, повышенные требования к вычислительным ресурсам, и т.д.)
B. да, потому что нейронные сети всегда выдают более точные прогнозы относительно алгоритмов классического машинного обучения
C. не всегда, потому что нейронные сети склонны к переобучению в отличие от алгоритмов классического машинного обучения
D. да, потому что нейронные сети позволяют добиться точности прогноза, сопоставимой с точностью классических алгоритмов, при меньших затратах ресурсов

Вопрос 2. В чем отличие стохастического градиентного спуска от его классической реализации?
A. В случае классического ГС на каждой итерации минимизируется вся функция потерь целиком (все ее слагаемые), а при СГС на каждой итерации минимизируется одно случайное слагаемое.
B. В случае классического ГС на каждой итерации минимизируется одно случайное слагаемое функции потерь, а при СГС на каждой итерации минимизируется вся функция целиком.
C. В случае классического ГС минимизируется вся функция потерь целиком, а при СГС - случайное количество ее слагаемых.
D. В случае классического ГС ищется минимум функции потерь, а в случае СГС - максимум.

Вопрос 3. Эпоха обучения - это:
A. этап обучения, на котором были минимизированы все слагаемые функции потерь, то

есть в обновлении весов сети поучаствовали все объекты обучающей выборки
B. количество слагаемых функции потерь, минимизируемых за одну итерацию
C. длительность обучения сети в секундах
D. количество шагов алгоритма обратного распространения ошибки, за которое вычисляется сложная функция

Вопрос 4. В качестве функции потерь при решении задачи регрессии используют:
A.MSE
B. accuracy
C. бинарную кросс-энтропию
D. категориальную кросс-энтропию

Вопрос 5. В качестве функции потерь при решении задачи многоклассовой классификации используют:
A. категориальную кросс-энтропию
B. accuracy
C. бинарную кросс-энтропию
D. MSE

Вопрос 6. Дропаут- это:
A. преднамеренная деактивация части нейронов на шаге обучения
B. нормализация данных во внутренних слоях сети
C. техника подбора начальных весов сети
D. техника, направленная на борьбу с переобучением, состоящая в добавлении в функцию потерь дополнительных слагаемых

Вопрос 7. Для чего нужна функция EarlyStopping() в библиотеке Keras?
A. для предотвращения переобучения нейронной сети
B. для остановки процедуры градиентного спуска (если не применить эту функцию, алгоритм зациклится)
C. для перехода в другое признаковое пространство
D. для регуляризации путем применения техники «dropout»

Вопрос 8. Расположите архитектуры нейронных сетей в хронологическом порядке (том, в котором они появлялись).
A. полносвязная (многослойный перцептрон), RNN, LSTM, трансформер
B. RNN, полносвязная (многослойный перцептрон), LSTM, трансформер
C. трансформер, полносвязная (многослойный перцептрон), LSTM, RNN
D. полносвязная (многослойный перцептрон), LSTM, RNN, трансформер

Вопрос 9. Каково основное назначение полносвязных нейронных сетей (многослойных перцептронов)?
A. они являются универсальными аппроксиматорами функций
B. обработка временных последовательностей
C. распознавание объектов на изображениях
D. уменьшение размерности данных

Вопрос 10. Что является особенностью рекуррентных нейронных сетей (RNN)?
A. наличие памяти о предыдущих состояниях

B. использование сверток для обработки изображений
C. наличие слоев без обратной связи
D. наличие полносвязных слоев

Вопрос 11. Какая из перечисленных разновидность слоев (layers) типична для сверточной нейронной сети?
A.Max Pooling Layer
B. Fully Connected Layer
C. Recurrent Layer
D. LSTM Layer

Вопрос 12. Назовите преимущество сверточных нейронных сетей при обработке изображений.
A. инвариантность к сдвигу
B. непрерывное обучение
C. способность моделировать временные зависимости
D. уменьшение переобучения

Вопрос 13. Какая функция активации обычно (часто) используется в полносвязных нейронных сетях?
A.ReLU
B. softmax
C. tanh (гиперболический тангенс)
D. сигмоида

Вопрос 14. Какая разновидность рекуррентных сетей способна решать задачи по обработке данных с длинными временными зависимостями?
A. GRU
B. CNN (сверточная нейронная сеть)
C. FNN (нейронная сеть прямого распространения)
D. перцептрон

Вопрос 15. Какой слой (layer) используется для уменьшения размерности в сверточных сетях?
A.Pooling Layer
B. ReLU Layer
C. Dropout Layer
D. Dense Layer

Вопрос 16. Для чего используется операция softmax в нейронных сетях?
A. для такого преобразования выходных сигналов, чтобы их можно было интерпретировать как оценки вероятностей принадлежности объекта тому или иному классу
B. для обработки временных данных
C. для нормализации входных данных
D. для генерации случайных чисел

Вопрос 17. Какой из перечисленных терминов связан со спецификой RNN (рекуррентных нейронных сетей)?

A.Backpropagation Through Time
B. Convolution
C. Batch Normalization
D. Max Pooling

Вопрос 18. Обработка естественного языка (NLP) - это:
A. область искусственного интеллекта, которая фокусируется на возможности машин читать, понимать и извлекать смысл из человеческих языков
B. область искусственного интеллекта, использующая алгоритмы для генерации разнообразного текстового контента
C. процесс конвертации текста в некоторый числовой вид
D. метод самосовершенствования, который позволяет освоить навыки общения, изменить модели поведения и добиться успеха в любой области

Вопрос 19. Tf-idf- это:
A. метрика, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса
B. последовательность букв или слов из N элементов
C. модификация "мешка слов", учитывающая порядок слов
D. модификация "мешка слов", учитывающая связь слов между собой (контекст)

Вопрос 20. ChatGPT- это:
A. чат-бот с генеративным искусственным интеллектом, разработанный компанией OpenAI и способный работать в диалоговом режиме, поддерживающий запросы на естественных языках
B. разновидность искусственного интеллекта, способного генерировать текст, изображения или другие медиаданные в ответ на подсказки
C. семейство моделей нейронных сетей, использующих архитектуру трансформеров и являющихся на сегодня ключевым достижением в области генеративного искусственного интеллекта
D. генеративная языковая модель, создающая тексты, разработанная компанией Яндекс

Ключ

Номер задания	Вариант 1	Вариант 2
1	A	A
2	A	A
3	A	A
4	A	A
5	A	A
6	A	A
7	A	A
8	A	A
9	A	A
10	A	A

11	A	A
12	A	A
13	A	A
14	A	A
15	A	A
16	A	A
17	A	A
18	A	A
19	A	A
20	A	A

Типовые задания по дисциплине

Задание 1
Подобрать одномерный или многомерный временной ряд (длиной не менее 300 значений).
Разбить ряд на обучающую и тестовую выборки.
Выполнить моделирование и прогнозирование ряда. Используя все известные вам техники и модели, попытаться добиться как можно более высокой точности прогноза на тестовой выборке.
Ссылку на блокнот (она должна открываться!) отправить преподавателю на почту.

Задание 2
На табличном датасете из первой контрольной точки решить задачи регрессии и классификации с использованием полносвязных нейронных сетей в качестве моделей.
Вычислить те же метрики качества, что в первой контрольной точке.
Сравнить полученные результаты с полученными в первой контрольной точке результатами.
Путем подбора архитектур нейронных сетей и их гиперпараметров попытаться добиться как можно более высокого значения метрик на тестовой выборке.

Задание 3
1) При помощи предобученной нейронной сети выполнить классификацию 10 изображений (картинки найти и загрузить самостоятельно). По результатам классификации вычислить метрику accuracy.
2) При помощи предобученной нейронной сети YOLO-World выполнить распознавание объектов на изображении (картинку найти и загрузить самостоятельно, на ней должно быть не менее 10 объектов). Для оценки точности распознавания вычислить метрику accuracy.
3) Используя transfer learning, дообучить нейронную сеть ResNet50 для распознавания специфических изображений, которых нет в датасете ImageNet. Датасет с изображениями найти самостоятельно.

Задание 4
Поставить и решить задачи:
* регрессии
* и классификации.
1) Выбрать датасет.

2) Сформулировать задачу, отобрать целевой признак.
3) Решить задачу с помощью слабого алгоритма, т.е. получить baseline-решение. В качестве слабой модели можно использовать регуляризацию. Точность должна быть приемлемой (например, $R2 \geq 0,5$).
4) Добиться максимально возможного улучшения точности прогнозирования, используя все техники из лекции:
* пайплайны
* отбор признаков
* понижение размерности
* балансировка выборки (для задачи классификации)
* подбор гиперпараметров при кросс-валидации (Grid Search и/или Randomized Search)
* и любые известные вам методы и подходы (относящиеся к классическим алгоритмам ML, т.е. без нейросетей).
Качество моделей оценивать при кросс-валидации по K блокам.

Задание 5
Найти размеченный датасет не менее чем из 10 тыс. текстов. Пример датасета – вопросы и ответы"Своей игры": https://github.com/evrog/Russian-QA-Jeopardy/tree/main
Решить на этих данных одну из трех задач: регрессии, классификации или кластеризации. Оценить качество решения. Добиться максимально высокой точности прогноза, используя любыеизученные техники из арсенала классических методов машинного обучения.
Дать интерпретацию полученным результатам, сделать выводы, оформить отчет.