Методические материалы по практическим работам

А.Е. Самотуга

Доверенный искусственный интеллект

# 1 ПРАКТИЧЕСКИЕ РАБОТЫ

## Общие положения и методика оценивания результатов

В процессе обучения по дисциплине «Доверенный искусственный интеллект» студент обязан выполнить 8 практических работ. Под выполнением практических работ подразумевается подготовка к работе, проведение экспериментальных исследований, подготовка отчета и его защита на практическом занятии. Выполнение практических работ студентами осуществляется в бригадах до 2 человек. Оформление отчета студентами осуществляется индивидуально в соответствии с принятыми в СПбГЭТУ правилами оформления студенческих работ. Отчет оформляется после выполнения экспериментальных исследований и представляется преподавателю на проверку. После проверки отчет либо возвращается (при наличии замечаний) на доработку, либо подписывается к защите. Практические работы защищаются студентами индивидуально. Каждый студент получает вопрос по теоретической и практической части, после чего ему предоставляется время для подготовки ответа. На защите практической работы студент должен показать: понимание и умение объяснять особенности применяемых методов, возможные области их применения и т.д., прогнозировать реакции исследуемого объекта на различные воздействия, навыки и умения, приобретенные при выполнении практических работы. При обсуждении ответа преподаватель может задать несколько уточняющих вопросов. В случае если студент демонстрирует достаточное знание вопроса, работа считается зачтенной. Текущий контроль включает в себя выполнение, сдачу в срок отчетов и их защиту по всем практическим работам, по результатам которой студент получает допуск к экзамену. Критерием оценки работы на практическом занятии является оценка, выставляемая по 5ти балльной шкале в соответствии со следующими критериями: оценка в 5 баллов выставляется при отличном выполнении задания, то есть при наличии полных (с детальными пояснениями и выкладками), оригинальных и правильных решений задач, дополненных при необходимости документами, полученными в результате реализации (проверки) решения, верных ответов и высококачественного оформления работы. Оценка в 4 балла выставляется при правильном выполнении задания, то есть при наличии полных (с пояснениями и выкладками), оригинальных и правильных решений задач, дополненных при необходимости документами, полученными в результате реализации (проверки) решения, верных ответов. Оценка в 3 балла выставляется при наличии отдельных неточностей в ответах (включая грамматические ошибки) или неточностях в решении задач непринципиального характера (описки и случайные ошибки арифметического характера). Оценка в 2 и ниже баллов выставляется в случаях, когда в ответах и в решениях задач имеются неточности и ошибки, свидетельствующие о недостаточном понимании вопросов и требующие дополнительного обращения к тематическим материалам.

## Содержание отчетов

* Титульный лист
* Цель работы
* Формулировка задания
* Краткое описание выполнения основных этапов работы
* Выводы по работе

## Практические работы

### Постановка проблемы доверия искусственному интеллекту. Обсуждение индивидуальных заданий

**Цель работы:** изучениекомпонентов и свойств доверенного искусственного интеллекта, уровней доверия с опорой на теоретический лекционный материал.

**Задание:**

Провести анализ индивидуального домашнего задания (ИДЗ), декомпозицию задач и построить план выполнения ИДЗ.

**Контрольные вопросы:**

1. Почему важно доверять системам искусственного интеллекта?
2. Приведите примеры ответственных задач обработки данных.
3. Сколько уровней доверия принято выделять на данный момент?
4. Перечислите основные компоненты доверенного искуственного интеллекта.
5. Какие вам известны аппаратные платформы для доверенного ИИ?

### Управление рисками искусственного интеллекта

**Цель работы:** изучение процесса управления рисками на основе информации из теоретического материала для конкретной модели искусственного интеллекта.

**Задание:**

Провести идентификацию, оценку рисков для конкретной модели искусственного интеллекта, выполнить анализ результатов оценки и описать необходимые шаги по их обработке.

**Контрольные вопросы:**

1. Какие вам известны принципы управления рисками согласно ИСО 31000:2018?
2. Перечислите, что должна включать в себя приверженность риск-менеджменту?
3. В зависимости от чего при определении рисков систем искусственного интеллекта следует учитывать различные источники рисков?
4. Что такое автономность искусственного интеллекта?
5. Что определяет степень прозрачности принятия решений?

### Реализация объяснимого искусственного интеллекта и метрики объяснимости.

**Цель работы:** получение навыков проектирования объяснимого искусственного интеллекта с использованием методов и подходов к реализации объяснимого ИИ (деревья решений, GA2M, TCAV, LIME, SHAPи другое) и получения метрик объяснимости.

**Задание:**

Выбрать метод или подход к оценке объяснимости. Провести проектирование модели искусственного интеллекта по индивидуальному заданию. Выполнить оценку объяснимости модели ИИ, рассчитать метрики объяснимости согласно выбранному подходу. Проанализировать результат.

**Контрольные вопросы:**

1. Дайте определение объяснимости согласно ГОСТ Р 59276-2020.
2. Какие составляющие включает в себя объяснимость искусственного интеллекта?
3. Какие выделяют методы отбора признаков?
4. Какие методы и подходы к реализации объяснимого искусственного интеллекта и оценке метрик объяснимости на данный момент известны?
5. Назовите несколько программных продуктов для создания объяснимого или повышения объяснимости искусственного интеллекта.

### Оценка робастности модели искусственного интеллекта

**Цель работы:** получение навыков оценки робастности модели искусственного интеллекта с применением методов оценки робастности с опорой на теоретический лекционный материал.

**Задание:**

Выполнить оценки метрик робастности модели искусственного интеллекта, построенной по индивидуальному заданию.

**Контрольные вопросы:**

1. Дайте определение понятию робастность.
2. В чем заключается главный принцип робастности в случае системы искусственного интеллекта?
3. Какие группы методов выделяют среди методов оценки робастности нейронных сетей?
4. Опишите типичный алгоритм для оценки робастности нейронных сетей.
5. Каким образом решают проблему недостаточной обучающей способности нейронных сетей?

### Реализация функциональной безопасности искусственного интеллекта

**Цель работы:** получение навыков управления безопасностью системы искусственного интеллекта.

**Задание:**

Выбрать для исследования систему искусственного интеллекта. Изучить состав технологических элементов системы. Описать методы, используемые для реализации функциональной безопасности в выбранной системе искусственного интеллекта. Привести описание с помощью чего реализован каждый технологический элемент. Выполнить анализ выбранного технологического элемента с точки зрения функциональной безопасности.

**Контрольные вопросы:**

1. Что понимают под функциональной безопасностью искусственного интеллекта?
2. На что направлена функциональная безопасность? Посредством чего достигается эта цель?
3. Опишите трехкомпонентную структуру системы искусственного интеллекта.
4. Какие технологические элементы на основе функциональных уровней экосистемы ИИ содержит процесс создания и выполнения модели искусственного интеллекта?
5. Приведите примеры языка или инструмента реализации каждого технологического элемента.

### Противодействие компьютерным атакам на искусственный интеллект.

**Цель работы:** получение навыков защиты от компьютерных атак на системы искусственного интеллекта.

**Задание:**

Выбрать определенный класс компьютерных атак на искусственный интеллект. Провести эксперимент по моделированию атаки на модель искусственного интеллекта. Выбрать метод противодействия атаке, опираясь на теоретический материал, реализовать выбранный метод для модели. Провести эксперимент по моделированию атаки на защищенную модель искусственного интеллекта.

**Контрольные вопросы:**

1. Какие принято выделять классы характерных атак на искусственный интеллект?
2. Сколько существует разновидностей атак? Назовите их.
3. Какие типы атак выделяют в зависимости от направления вектора атаки?
4. В какой ситуации атаки, направленные на извлечение входных данных, являются актуальными?
5. Дайте определение понятию состязательной атаки на искусственный интеллект? Что понимают под термином зондирование моделей?

### Защита данных, используемых в системе искусственного интеллекта на этапах хранения, обучения и использования.

**Цель работы:** получить навыки по защите данных в системе искусственного интеллекта

**Задание:** Выбрать конкретный вид атаки на данные на одном из этапов их обработки. Студент должен проанализировать материалы лекций, основную и дополнительную литературу, интернет источники, составить перечень способов защиты от атаки на данные, указав ссылки на источники с обоснованием эффективности способов защиты.

**Контрольные вопросы:**

1. Назовите два основных способа предотвращения атаки манипуляции с входными данными.
2. В каком случае атаки типа "ключ под ковриком" и атаки "на решающий бит" станут неосуществимы уже на уровне концепции построения ИИ?
3. Перечислите преимущества федеративного обучения?
4. Какие выделяют причины необходимости применения гомоморфного шифрования при разработке систем искусственного интеллекта?
5. Что является целью определения показателей конфиденциальности?
6. Перечислите несколько программных продуктов, где реализованы методы по защите конфиденциальности и целостности данных ИИ.

### Обнаружение аномалий

**Цель работы:** получение навыков по обнаружению аномалий искусственного интеллекта.

**Задание:**

Согласовать с преподавателем выбор типа аномалии. Студент должен проанализировать материалы лекций, основную и дополнительную литературу, интернет источники, привести подход к обнаружению выбранной аномалии и реализовать его для модели ИИ самостоятельно либо с помощью готового программного продукта. Выполнить эксперимент по моделированию обнаружения аномалий.

**Контрольные вопросы:**

1. Приведите определение дрейфа данных. С чем часто связан дрейф данных? Приведите пример такой ситуации.
2. Дайте определение понятию дрейф концепции. Каким примером можно описать дрейф концепции?
3. Когда утечка данных может стать актуальной угрозой для искусственного интеллекта? Приведите способы защиты от утечки данных.
4. Что должно рассматриваться для разработки приложений мониторинга ИИ?
5. Приведите примеры программных продуктов для обнаружения аномалий в данных.