

ИНСТИТУТ ИНТЕЛЛЕКТУАЛЬНЫХ КИБЕРНЕТИЧЕСКИХ СИСТЕМ
КАФЕДРА КРИПТОЛОГИИ И ДИСКРЕТНОЙ МАТЕМАТИКИ

ОДОБРЕНО УМС ИИКС

Протокол № УМС-575/01-1

от 30.08.2021 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
АНАЛИЗ ДАННЫХ И МАШИННОЕ ОБУЧЕНИЕ

Направление подготовки
(специальность)

[1] 10.04.01 Информационная безопасность

| Семестр | Трудоемкость, кред. | Общий объем курса, час. | Лекции, час. | Практич. занятия, час. | Лаборат. работы, час. | В форме практической подготовки/В СРС, час. | КСР, час. | Форма(ы) контроля, экс./зач./КР/КП |
|---------|------------------------|----------------------------|--------------|---------------------------|--------------------------|--|-----------|--|
| 2 | 2 | 72 | 30 | 0 | 15 | 27 | 0 | 3 |
| 3 | 3 | 108 | 32 | 0 | 16 | 24 | 0 | Э |
| Итого | 5 | 180 | 62 | 0 | 31 | 0 | 51 | |

АННОТАЦИЯ

В курсе рассматриваются следующие темы:

- основы интеллектуального анализа данных;
- основы методов машинного обучения;
- методы классификации, кластеризации и восстановления регрессии;
- основы методов глубокого обучения, включая искусственные нейронные сети.

В рамках лабораторного практикума студенты получают навыки программирования.

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ УЧЕБНОЙ ДИСЦИПЛИНЫ

Цель дисциплины – освоение комплекса современных методов интеллектуального анализа данных и процессов, включая методы, основанные на машинном обучении.

2. МЕСТО УЧЕБНОЙ ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП ВО

Полученные в результате освоения учебной дисциплины знания, умения, навыки используются в процессе дипломного проектирования.

3. ФОРМИРУЕМЫЕ КОМПЕТЕНЦИИ И ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Универсальные и(или) общепрофессиональные компетенции:

| Код и наименование компетенции | Код и наименование индикатора достижения компетенции |
|--|---|
| ОПК-5 [1] – Способен проводить научные исследования, включая экспериментальные, обрабатывать результаты исследований, оформлять научно-технические отчеты, обзоры, готовить по результатам выполненных исследований научные доклады и статьи | З-ОПК-5 [1] – Знать: теоретические и эмпирические методы научных исследований, порядок проведения научных исследований У-ОПК-5 [1] – Уметь: применять методы научных исследований в научной деятельности, обобщать полученные экспериментальные данные, анализировать и делать выводы В-ОПК-5 [1] – Владеть: теоретическими и эмпирическими методами научного исследования при выполнении научно-исследовательских работ, методикой оформления отчетов по научно-исследовательским работам, статей и тезисов докладов |

Профессиональные компетенции в соответствии с задачами и объектами (областями знаний) профессиональной деятельности:

| Задача профессиональной деятельности (ЗПД) | Объект или область знания | Код и наименование профессиональной компетенции; Основание (профессиональный стандарт-ПС, анализ | Код и наименование индикатора достижения профессиональной компетенции |
|--|---------------------------|---|---|
|--|---------------------------|---|---|

| | | | |
|---|------------------------|--|---|
| | | опыта) | |
| | проектный | | |
| разработка проектных решений по обеспечению безопасности данных с применением криптографических методов | информационные ресурсы | ПК-4.1 [1] - Способен разрабатывать проектные решения по обеспечению безопасности данных с применением криптографических методов <i>Основание:</i> Профессиональный стандарт: 06.032 | 3-ПК-4.1[1] - Знать: методы обеспечения безопасности данных с применением криптографических методов; У-ПК-4.1[1] - Уметь: разрабатывать проектные решения по обеспечению безопасности данных с применением криптографических методов; В-ПК-4.1[1] - Владеть: навыками разработки проектных решений по обеспечению безопасности данных с применением криптографических методов |

4. СТРУКТУРА И СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Разделы учебной дисциплины, их объем, сроки изучения и формы контроля:

| № п.п | Наименование раздела учебной дисциплины | Недели | Лекции/ Практи. (семинары) / Лабораторные работы, час. | Обязат. текущий контроль (форма*, неделя) | Максимальный балл за раздел** | Аттестация раздела (форма*, неделя) | Индикаторы освоения компетенции |
|-------|---|--------|--|---|-------------------------------|-------------------------------------|---|
| | <i>2 Семестр</i> | | | | | | |
| 1 | Первый раздел | 1-8 | | | 25 | КИ-8 | 3-ОПК-5, У-ОПК-5, В-ОПК-5, 3-ПК-4.1, У-ПК-4.1, В- |

| | | | | | | | |
|---|---|------|---------|--|----|-------|---|
| | | | | | | | ПК-4.1 |
| 2 | Второй раздел | 9-15 | | | 25 | КИ-15 | 3-ОПК-5, У-ОПК-5, В-ОПК-5, 3-ПК-4.1, У-ПК-4.1, В-ПК-4.1 |
| | <i>Итого за 2 Семестр</i> | | 30/0/15 | | 50 | | |
| | Контрольные мероприятия за 2 Семестр | | | | 50 | 3 | 3-ОПК-5, У-ОПК-5, В-ОПК-5, 3-ПК-4.1, У-ПК-4.1, В-ПК-4.1 |
| | <i>3 Семестр</i> | | | | | | |
| 1 | Первый раздел | 1-8 | | | 25 | КИ-8 | 3-ОПК-5, У-ОПК-5, В-ОПК-5, 3-ПК-4.1, У-ПК-4.1, |

| | | | | | | | |
|---|---|------|---------|--|----|-------|---|
| | | | | | | | В-ПК-4.1 |
| 2 | Второй раздел | 9-15 | | | 25 | КИ-15 | З-ОПК-5, У-ОПК-5, В-ОПК-5, З-ПК-4.1, У-ПК-4.1, В-ПК-4.1 |
| | <i>Итого за 3 Семестр</i> | | 32/0/16 | | 50 | | |
| | Контрольные мероприятия за 3 Семестр | | | | 50 | Э | З-ОПК-5, У-ОПК-5, В-ОПК-5, З-ПК-4.1, У-ПК-4.1, В-ПК-4.1 |

* – сокращенное наименование формы контроля

** – сумма максимальных баллов должна быть равна 100 за семестр, включая зачет и (или) экзамен

Сокращение наименований форм текущего контроля и аттестации разделов:

| Обозначение | Полное наименование |
|--------------------|----------------------------|
| КИ | Контроль по итогам |
| З | Зачет |
| Э | Экзамен |

КАЛЕНДАРНЫЙ ПЛАН

| Недел и | Темы занятий / Содержание | Лек., час. | Пр./сем. , час. | Лаб., час. |
|-------------|--|------------------------|--------------------|---------------|
| | <i>2 Семестр</i> | 30 | 0 | 15 |
| 1-8 | Первый раздел | 20 | | 8 |
| 1 - 2 | Предмет анализа данных. Основные методы анализа данных. Отличительные особенности интеллектуального анализа данных. Основная терминология. Типы данных. Типы признаков. Приложения интеллектуального анализа данных. Жизненный цикл процесса анализа данных. | Всего аудиторных часов | | |
| | | 5 | | 2 |
| | | Онлайн | | |
| | | | | |
| 3 - 4 | Предобработка и постобработка данных. Задачи, возникающие в связи с анализом данных. Взаимосвязь интеллектуального анализа данных с другими видами обработки данных. Способы хранения данных. Инструментарий анализа данных. Основные задачи предобработки данных. Очистка данных. Интеграция данных. Редукция данных. Дискретное вейвлет-преобразование. Вейвлеты Хаара. Визуализация данных. | Всего аудиторных часов | | |
| | | 5 | | 2 |
| | | Онлайн | | |
| | | | | |
| 5 - 6 | Поиск шаблонов и ассоциативных правил. Постановка задачи поиска часто встречающихся шаблонов и ассоциативных правил. Алгоритм Apriori. Генерация ассоциативных правил из часто встречающихся подмножеств. Алгоритм FP-Growth. | Всего аудиторных часов | | |
| | | 5 | | 2 |
| | | Онлайн | | |
| | | | | |
| 7 - 8 | Введение в машинное обучение. Формальная постановка задачи машинного обучения. Задание объектов в машинном обучении. Задание ответов в машинном обучении. Этапы обучения и применения модели. Функционалы качества. Сведение задачи обучения к задаче оптимизации. Эмпирические оценки обобщающей способности. Способы проверки обобщающей способности на практике. Переобучение. | Всего аудиторных часов | | |
| | | 5 | | 2 |
| | | Онлайн | | |
| | | | | |
| 9-15 | Второй раздел | 10 | | 7 |
| 9 - 10 | Классификация. Формальная постановка задачи классификации. Логические методы классификации. Бинарные решающие деревья. Алгоритм построения решающего дерева по обучающей выборке. Варианты критериев ветвления. Индекс Джини. Достоинства и недостатки решающих деревьев. Проблемы, связанные с «жадным» алгоритмом ID3. | Всего аудиторных часов | | |
| | | 3 | | 2 |
| | | Онлайн | | |
| | | | | |
| 11 - 12 | Оценка качества классификаторов. Метрики для оценки качества классификаторов. ROC-кривая. Композиции алгоритмов классификации. Простое голосование классификаторов. | Всего аудиторных часов | | |
| | | 3 | | 2 |
| | | Онлайн | | |
| | | | | |
| 13 - 14 | Кластеризация. Формальная постановка задачи кластеризации. Расстояния между объектами в признаковом пространстве. Типы кластерных структур. Методы кластеризации. Метод k средних. Агломеративная иерархическая кластеризация. Формула Ланса – Уильямса. Визуализация кластерной | Всего аудиторных часов | | |
| | | 2 | | 2 |
| | | Онлайн | | |
| | | | | |

| | | | | |
|-------|---|------------------------|---|----|
| | структуры. Основные свойства иерархической кластеризации. Плотностные методы кластеризации. Алгоритм DBSCAN. | | | |
| 15 | Обнаружение аномалий. Аномалии: определение и пример. Методы обнаружения аномалий. Методы, основанные на измерении близости объектов. Метод вложенных циклов. Сеточный метод. Кластерные методы. Примеры применения кластерных методов обнаружения аномалий в задачах информационной безопасности. | Всего аудиторных часов | | |
| | | 2 | | 1 |
| | | Онлайн | | |
| | 3 Семестр | 32 | 0 | 16 |
| 1-8 | Первый раздел | 16 | | 8 |
| 1 - 2 | Метрические методы классификации и восстановления регрессии. Формальные постановки задач классификации и восстановления регрессии. Обобщенный метрический классификатор. Метод k ближайших соседей. Метод окна Парзена. Метод потенциальных функций. Метрические методы в задачах восстановления регрессии. Непараметрическая регрессия. Формула ядерного сглаживания Надарая – Ватсона. Часто используемые ядра. Выбор ядра и ширины окна. Проблема выбросов. Локально взвешенное сглаживание. Алгоритм LOWESS. Вероятностные (байесовские) методы классификации. Теорема Байеса. Наивный байесовский классификатор. Примеры применения. Метод коррекции Лапласа для случая отсутствия объектов с требуемыми значениями атрибутов. | Всего аудиторных часов | | |
| | | 4 | | 2 |
| | | Онлайн | | |
| 3 - 4 | Линейные классификаторы. Определение и формальное описание линейного классификатора. Обучение линейного классификатора. Непрерывные аппроксимации пороговой функции потерь. Обучение регрессии. Градиент функции. Градиентный метод численной минимизации. Метод стохастического градиента. Обоснование оценки функционала в методе стохастического градиента. Метод стохастического усредненного градиента. Достоинства и недостатки метода стохастического градиента. Эвристики: варианты инициализации весов, варианты порядка предъявления объектов. Проблема мультиколлинеарности. Проявления переобучения. Регуляризация. | Всего аудиторных часов | | |
| | | 4 | | 2 |
| | | Онлайн | | |
| 5 - 6 | Логистическая регрессия. Обоснование логарифмической функции потерь. Оптимизация параметров логистической регрессии. Скоринговые карты. Регуляризованная логистическая регрессия. Композиции алгоритмов машинного обучения. Простое голосование классификаторов. Бэггинг и метод случайных подпространств. Метод случайного леса. Градиентный бустинг. Параметрическая аппроксимация градиентного шага. Стохастический градиентный бустинг. Алгоритм AdaBoost. Частные случаи при различных | Всего аудиторных часов | | |
| | | 4 | | 2 |
| | | Онлайн | | |

| | | | | |
|---------|---|------------------------|--|---|
| | функциях потерь. Градиентный бустинг над деревьями. | | | |
| 7 - 8 | Частичное обучение. Задача частичного обучения. Отличия от задач классификации и кластеризации. Метод самообучения. Метод совместного обучения. Метод обучения композиции алгоритмов. Методы частичного обучения на основе кластеризации: графовый подход, метод Ланса – Уильямса, метод k средних. | Всего аудиторных часов | | |
| | | 4 | | 2 |
| | | Онлайн | | |
| 9-15 | Второй раздел | 16 | | 8 |
| 9 - 10 | Метод опорных векторов. Постановка задачи для метода опорных векторов. Первый подход – аппроксимация и регуляризация эмпирического риска. Второй подход – оптимальная разделяющая гиперплоскость. Переход к линейно неразделимой выборке. Условия Каруша – Куна – Таккера. Двойственная задача и её нелинейное обобщение для метода опорных векторов. Ядра для нелинейного обобщения метода опорных векторов. Примеры ядер. Классификация с разными ядрами. | Всего аудиторных часов | | |
| | | 4 | | 2 |
| | | Онлайн | | |
| 11 - 12 | Введение в искусственные нейронные сети. Линейная модель нейрона Мак Каллока – Питтса. Часто используемые функции активации нейронов. Нейронная реализация логических функций. Приближение произвольной функции нейронной сетью. | Всего аудиторных часов | | |
| | | 4 | | 2 |
| | | Онлайн | | |
| 13 - 14 | Обучение искусственных нейронных сетей. Многослойная нейронная сеть. Алгоритм стохастического градиента в применении к нейронным сетям. Задача дифференцирования суперпозиции функций. Быстрое вычисление градиента. Алгоритм обратного распространения ошибки. Эвристики для обучения нейронных сетей. Ускорение сходимости. Динамическое наращивание сети. Прореживание сети (метод OBD). Увеличение числа уровней. | Всего аудиторных часов | | |
| | | 4 | | 2 |
| | | Онлайн | | |
| 15 - 16 | Введение в глубокое обучение. Свёрточные нейронные сети. Архитектура свёрточной сети. Отличия от полносвязной нейронной сети. Слой свёртки. Слой ReLU. Слой субдискретизации. Преимущества и недостатки свёрточных нейронных сетей. Применение свёрточных нейронных сетей в фотофорефике. | Всего аудиторных часов | | |
| | | 4 | | 2 |
| | | Онлайн | | |

Сокращенные наименования онлайн опций:

| Обозначение | Полное наименование |
|-------------|-------------------------|
| ЭК | Электронный курс |
| ПМ | Полнотекстовый материал |
| ПЛ | Полнотекстовые лекции |
| ВМ | Видео-материалы |
| АМ | Аудио-материалы |
| Прз | Презентации |
| Т | Тесты |

| | |
|-----|----------------------------------|
| ЭСМ | Электронные справочные материалы |
| ИС | Интерактивный сайт |

5. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Основными образовательными технологиями в освоении дисциплин профессионального цикла являются традиционные технологии лекций и лабораторных работ. Интерактивные методики обеспечиваются решением индивидуальных задач студентами и коллективным обсуждением результатов и методов решения.

6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Фонд оценочных средств по дисциплине обеспечивает проверку освоения планируемых результатов обучения (компетенций и их индикаторов) посредством мероприятий текущего, рубежного и промежуточного контроля по дисциплине.

Связь между формируемыми компетенциями и формами контроля их освоения представлена в следующей таблице:

| Компетенция | Индикаторы освоения | Аттестационное мероприятие (КП 1) | Аттестационное мероприятие (КП 2) |
|-------------|---------------------|-----------------------------------|-----------------------------------|
| ОПК-5 | З-ОПК-5 | З, КИ-8, КИ-15 | Э, КИ-8, КИ-15 |
| | У-ОПК-5 | З, КИ-8, КИ-15 | Э, КИ-8, КИ-15 |
| | В-ОПК-5 | З, КИ-8, КИ-15 | Э, КИ-8, КИ-15 |
| ПК-4.1 | З-ПК-4.1 | З, КИ-8, КИ-15 | Э, КИ-8, КИ-15 |
| | У-ПК-4.1 | З, КИ-8, КИ-15 | Э, КИ-8, КИ-15 |
| | В-ПК-4.1 | З, КИ-8, КИ-15 | Э, КИ-8, КИ-15 |

Шкалы оценки образовательных достижений

Шкала каждого контрольного мероприятия лежит в пределах от 0 до установленного максимального балла включительно. Итоговая аттестация по дисциплине оценивается по 100-балльной шкале и представляет собой сумму баллов, заработанных студентом при выполнении заданий в рамках текущего и промежуточного контроля.

Итоговая оценка выставляется в соответствии со следующей шкалой:

| Сумма баллов | Оценка по 4-ех балльной шкале | Оценка ECTS | Требования к уровню освоению учебной дисциплины |
|--------------|-------------------------------|-------------|---|
| 90-100 | 5 – «отлично» | A | Оценка «отлично» выставляется студенту, если он глубоко и прочно усвоил программный материал, исчерпывающе, последовательно, четко и логически стройно его излагает, умеет тесно увязывать теорию с практикой, использует в ответе материал монографической литературы. |
| 85-89 | 4 – «хорошо» | B | Оценка «хорошо» выставляется студенту, если он твёрдо знает |
| 75-84 | | C | |

| | | | |
|---------|------------------------------|---|---|
| 70-74 | | D | материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос. |
| 65-69 | 3 – «удовлетворительно» | E | Оценка «удовлетворительно» выставляется студенту, если он имеет знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, нарушения логической последовательности в изложении программного материала. |
| 60-64 | | | |
| Ниже 60 | 2 – «неудовлетворительно» | F | Оценка «неудовлетворительно» выставляется студенту, который не знает значительной части программного материала, допускает существенные ошибки. Как правило, оценка «неудовлетворительно» ставится студентам, которые не могут продолжить обучение без дополнительных занятий по соответствующей дисциплине. |

Оценочные средства приведены в Приложении.

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

ОСНОВНАЯ ЛИТЕРАТУРА:

1. 004 3-17 Использование методов машинного обучения и языка Python для анализа данных. Ч.1 , Москва: НИЯУ МИФИ, 2019
2. ЭИ Д 95 Логический анализ данных : учебное пособие, Санкт-Петербург: Лань, 2020

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА:

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ:

Специальное программное обеспечение не требуется

LMS И ИНТЕРНЕТ-РЕСУРСЫ:

<https://online.mephi.ru/>

<http://library.mephi.ru/>

8. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Специальное материально-техническое обеспечение не требуется

9. УЧЕБНО-МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ДЛЯ СТУДЕНТОВ

Анализ данных и машинное обучение

10. УЧЕБНО-МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ДЛЯ ПРЕПОДАВАТЕЛЕЙ

приложено

Автор(ы):

Запечников Сергей Владимирович, д.т.н., доцент