

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ  
филиал федерального государственного бюджетного образовательного учреждения  
высшего образования «Мурманский арктический государственный университет»  
в г. Апатиты

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)**

**Б1.В.16 Анализ данных и машинное обучение**

(название дисциплины (модуля) в соответствии с учебным планом)

**образовательной программы  
по направлению подготовки бакалавриата**

**09.03.02 Информационные системы и технологии  
направленность (профиль): Программно-аппаратные комплексы**

(код и наименование направления подготовки  
с указанием направленности (профиля) (наименования магистерской программы))

**высшее образование – бакалавриат**

уровень профессионального образования: высшее образование – бакалавриат / высшее образование –  
специалитет, магистратура / высшее образование – подготовка кадров высшей квалификации

**бакалавр**

квалификация

**заочная**

форма обучения

**2019**

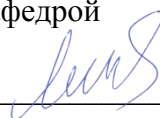
год набора

**Составитель:**

Тоичкин Н.А., к.т.н.,  
доцент кафедры  
информатики и вычислительной  
техники

Утверждено на заседании кафедры  
информатики и вычислительной техники  
(протокол № 13 от «06» июня 2019 г.)

Зав. кафедрой



Яковлев С.Ю.

## **Структура рабочей программы дисциплины**

### **1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)**

Целью изучения дисциплины «Анализ данных и машинное обучение» является: освоение обучающимися современных технологий для обработки и анализа информации и эффективных методов ее обработки с применением современных ЭВМ а также формирование целостной системы знаний в области создания, накопления, обработки и использования информационных ресурсов.

В результате изучения дисциплины обучающийся должен:

*знать:*

- формализацию задачи машинного обучения;
- понятие больших данных и их свойства;
- постановку задачи классификации и регрессии;
- понятие обобщенного метрического классификатора;
- алгоритмы метрической классификации;
- основные принципы построения логических алгоритмов классификации;
- алгоритм построения дерева классификации ID 3;
- линейные методы классификации;

*уметь:*

- осуществлять математическую и информационную постановку задач по обработке информации;
- использовать алгоритмы обработки информации для различных приложений;
- выполнять постановку задачи машинного обучения;
- выбирать методы и средства для решения задач машинного обучения;

*владеть:*

- инструментальными средствами обработки информации;
- информационными технологиями поиска данных и способами их использования;
- приемами программирования на языках анализа данных;
- методами интеллектуального анализа информации.

### **2. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ), СООТНЕСЕННЫХ С ПЛАНИРУЕМЫМИ РЕЗУЛЬТАТАМИ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ**

В результате освоении содержания дисциплины «Анализ данных и машинное обучение» обучающийся должен обладать следующими компетенциями:

- способен применять системный подход и математические методы в формализации решения прикладных задач, моделировать прикладные (бизнес) процессы и предметную область автоматизации организации (ПК-2);
- способен эксплуатировать и сопровождать информационные системы и сервисы, осуществлять ведение информационных хранилищ для решения прикладных задач профессиональной деятельности (ПК-3).

### **3. УКАЗАНИЕ МЕСТА ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ**

Данная дисциплина относится к части формируемой участниками образовательных отношений блока Б1.В, в структуре образовательной программы бакалавриата по направлению подготовки 09.03.02 Информационные системы и технологии, направленность (профиль) Программно-аппаратные комплексы.

Для освоения данной дисциплины обучающиеся используют знания, умения, навыки, способы деятельности и установки, которые они получили в процессе изучения дисциплин:

«Информатика», «Программирование на языках высокого уровня», «Технология программирования», «Управление данными», «Вычислительные системы, сети, телекоммуникации», «Информационная безопасность», «Операционные системы».

В свою очередь, «Анализ данных и машинное обучение» представляет собой методологическую базу для написания выпускной квалификационной работы.

## 5. ОБЪЕМ ДИСЦИПЛИНЫ (МОДУЛЯ) В ЗАЧЕТНЫХ ЕДИНИЦАХ С УКАЗАНИЕМ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ИЛИ АСТРОНОМИЧЕСКИХ ЧАСОВ, ВЫДЕЛЕННЫХ НА КОНТАКТНУЮ РАБОТУ ОБУЧАЮЩИХСЯ С ПРЕПОДАВАТЕЛЕМ (ПО ВИДАМ УЧЕБНЫХ ЗАНЯТИЙ) И НА САМОСТОЯТЕЛЬНУЮ РАБОТУ ОБУЧАЮЩИХСЯ

Общая трудоемкость дисциплины составляет 2 зачетные единицы или 72 часа.  
(из расчета 1 ЗЕТ= 36 часов).

| Курс          | Семестр | Трудоемкость в ЗЕТ | Общая трудоемкость (час.) | Контактная работа |          |           | Всего контактных часов | Из них в интер-активной форме | Кол-во часов на СРС | Курсовые работы | Кол-во часов на контроль | Форма контроля |
|---------------|---------|--------------------|---------------------------|-------------------|----------|-----------|------------------------|-------------------------------|---------------------|-----------------|--------------------------|----------------|
|               |         |                    |                           | ЛК                | ПР       | ЛБ        |                        |                               |                     |                 |                          |                |
| 4             | 8       | 1                  | 36                        | 4                 | -        | 10        | 14                     | 4                             | 22                  | -               | -                        | -              |
| 5             | 9       | 1                  | 36                        | -                 | -        | -         | -                      | -                             | 32                  | 4               | 4                        | зачет          |
| <b>Итого:</b> |         | <b>2</b>           | <b>72</b>                 | <b>4</b>          | <b>-</b> | <b>10</b> | <b>14</b>              | <b>4</b>                      | <b>54</b>           | <b>-</b>        | <b>-</b>                 | <b>зачет</b>   |

В интерактивной форме часы используются в виде: групповой дискуссии, заслушивании и обсуждении подготовленных студентами докладов по тематике дисциплины.

## 5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ, СТРУКТУРИРОВАННОЕ ПО ТЕМАМ (РАЗДЕЛАМ) С УКАЗАНИЕМ ОТВЕДЕННОГО НА НИХ КОЛИЧЕСТВА АКАДЕМИЧЕСКИХ ИЛИ АСТРОНОМИЧЕСКИХ ЧАСОВ И ВИДОВ УЧЕБНЫХ ЗАНЯТИЙ

| № п/п | Наименование раздела, темы          | Контактная работа |    |    | Всего контактных часов | Из них в интерактивной форме | Кол-во часов на СРС | Кол-во часов на контроль |
|-------|-------------------------------------|-------------------|----|----|------------------------|------------------------------|---------------------|--------------------------|
|       |                                     | ЛК                | ПР | ЛБ |                        |                              |                     |                          |
| 1     | Большие данные и машинное обучение. | 1                 | -  | 2  | 6                      | 1                            | 12                  | -                        |
| 2     | Метрические методы классификации    | 1                 | -  | 2  | 6                      | 1                            | 14                  | -                        |
| 3     | Логические методы классификации     | 1                 | -  | 2  | 6                      | 1                            | 14                  | -                        |
| 4     | Линейные методы классификации       | 1                 | -  | 4  | 10                     | 1                            | 14                  | -                        |

|  |               |          |          |           |           |          |           |          |
|--|---------------|----------|----------|-----------|-----------|----------|-----------|----------|
|  | <b>Зачет</b>  | -        | -        | -         | -         | -        | -         | 4        |
|  | <b>Итого:</b> | <b>4</b> | <b>-</b> | <b>10</b> | <b>14</b> | <b>4</b> | <b>54</b> | <b>4</b> |

## Содержание дисциплины:

### Тема 1. Большие данные и машинное обучение.

Большие данные. Свойства больших данных. Машинное обучение, формализация задачи машинного обучения. Признаковое описание объекта. Ответы и типы задач машинного обучения. Модель алгоритмов. Метод обучения. Этап обучения и этап применения. Функционалы качества. Сведение задачи обучения к задаче оптимизации. Переобучение и обобщение. Пример переобучения (Рунге). Эмпирические оценки обобщающей способности. Примеры задач машинного обучения: задачи классификации и регрессии; задачи ранжирования. Эксперименты в машинном обучении: эксперименты на реальных и синтетических данных.

### Тема 2. Метрические методы классификации

Формализация задачи. Обобщенный метрический классификатор. Метод ближайшего соседа. Метод k взвешенных ближайших соседей. Метод парзеновского окна. Метод потенциальных функций. Отбор эталонных объектов. Понятие отступа объекта. Типы объектов в зависимости от отступа. Отбор эталонов, алгоритм STOLP. Задача выбора метрики. Жадное добавление признаков.

### Тема 3. Логические методы классификации

Логическая закономерность. Основы вопросы построения логических алгоритмов классификации. Виды закономерностей. Критерии информативности: простые критерии, статистический критерий, энтропийный критерий. Где находятся закономерности в  $(p, n)$ -плоскости. Схема локального поиска информативных закономерностей. Определение бинарного решающего дерева. Жадный алгоритм построения дерева ID 3. Варианты критериев ветвления в ID 3. Обработка пропусков, алгоритм обработки пропусков на этапе обучения и этапе классификации. Алгоритм ID3: достоинства и недостатки. Стратегии редукции решающих деревьев. Небрежные решающие деревья. Бинаризация вещественного признака.

### Тема 4. Линейные методы классификации.

Задача построения разделяющей поверхности. Задача построения разделяющей поверхности. Минимизация эмпирического риска. Непрерывные аппроксимации пороговой функции потерь. Линейный классификатор. Персептрон. Устройство нервной клетки. Линейная модель нейрона МакКаллока-Питтса. Алгоритм Stochastic Gradient. Дельта-правило ADALINE. Правило Хебба. SG: инициализация весов. SG: проблемы переобучения. Принцип максимума правдоподобия. Оптимальная разделяющая гиперплоскость. Метод SVM. Нелинейное обобщение SVM.

## 6. ПЕРЕЧЕНЬ УЧЕБНО-МЕТОДИЧЕСКОГО ОБЕСПЕЧЕНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

### Основная литература:

1. Архипенков С. Я. , Голубев Д. , Максименко О. Хранилища данных: от концепции до внедрения. М.: Диалог-МИФИ, 2002, 528 с. Режим доступа: [https://biblioclub.ru/index.php?page=book\\_red&id=89285&sr=1](https://biblioclub.ru/index.php?page=book_red&id=89285&sr=1)
2. Чубукова И. А. Data Mining. М.: Интернет-Университет Информационных Технологий, 2008, 383 с. Режим доступа: [https://biblioclub.ru/index.php?page=book\\_red&id=233055&sr=1](https://biblioclub.ru/index.php?page=book_red&id=233055&sr=1)

#### **Дополнительная литература:**

3. Введение в анализ данных с помощью Pandas. Режим доступа: <https://habrahabr.ru/post/196980/>

### **7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)**

В образовательном процессе используются:

- учебные аудитории для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации (мебель аудиторная (столы, стулья, доска аудиторная), комплект мультимедийного оборудования, включающий мультимедиапроектор, экран, переносной ноутбук для демонстрации презентаций; учебно-наглядные пособия; обеспечивающие тематические иллюстрации);
- помещения для самостоятельной работы (оснащены компьютерными столами, стульями, доской аудиторной, компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета);
- помещения для хранения и профилактического обслуживания учебного оборудования (оснащены наборами инструментов, оборудованием, расходными материалами для монтажа, ремонта и обслуживания информационно-телекоммуникационной сети филиала и вычислительной техники);
- лаборатория информационных технологий (оснащена компьютерными столами, стульями, мультимедийным проектором, экраном проекционным, компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета).

#### **7.1 ПЕРЕЧЕНЬ ЛИЦЕНЗИОННОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ**

1. Microsoft Windows XP – операционная система.
2. Open Office Writer – текстовый редактор.
3. Anaconda – пакет разработки на Python.
4. Любой web – браузер.

#### **7.2 ЭЛЕКТРОННО-БИБЛИОТЕЧНЫЕ СИСТЕМЫ**

1. ЭБС «Издательство Лань»[Электронный ресурс]: электронная библиотечная система / ООО «Издательство Лань». – Режим доступа: <https://e.lanbook.com/>;
2. ЭБС «Электронная библиотечная система ЮРАЙТ» [Электронный ресурс]: электронная библиотечная система / ООО «Электронное издательство ЮРАЙТ». – Режим доступа: <https://biblio-online.ru/>;
3. ЭБС «Университетская библиотека онлайн»[Электронный ресурс]: электронно-периодическое издание; программный комплекс для организации онлайн-доступа к лицензионным материалам / ООО «НексМедиа». – Режим доступа: <https://biblioclub.ru/>.

#### **7.3 СОВРЕМЕННЫЕ ПРОФЕССИОНАЛЬНЫЕ БАЗЫ ДАННЫХ**

1. Электронная база данных SCOPUS.
2. Электронная база данных РИНЦ.

#### **7.4 ИНФОРМАЦИОННЫЕ СПРАВОЧНЫЕ СИСТЕМЫ**

1. Справочно-правовая информационная система Консультант Плюс  
<http://www.consultant.ru/>
2. Электронный справочник "Информιο" для высших учебных заведений  
<http://www.informio.ru/>

#### **8. ИНЫЕ СВЕДЕНИЯ И МАТЕРИАЛЫ НА УСМОТРЕНИЕ ВЕДУЩЕЙ КАФЕДРЫ**

Не предусмотрено.

#### **9. ОБЕСПЕЧЕНИЕ ОБРАЗОВАНИЯ ДЛЯ ЛИЦ С ОВЗ**

Для обеспечения образования инвалидов и лиц с ограниченными возможностями здоровья реализация дисциплины может осуществляться в адаптированном виде, с учетом специфики освоения и дидактических требований, исходя из индивидуальных возможностей и по личному заявлению обучающегося.