МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)

ШКОЛА ЦИФРОВОЙ ЭКОНОМИКИ

СОГЛАСОВАНО
Руководитель ОП

Р.И. Дремлюга

«17» июня .2019 г.

УТВЕРЖДАЮ
Директор Школы цифровой экономики

_____ И.Г. Мирин

«17» _____ 2019 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
«ПРИКЛАДНЫЕ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ И АНАЛИЗА ДАННЫХ»
направления 09.04.01 Информатика и вычислительная техника
Магистерская программа «Искусственный интеллект и большие данные»
Форма подготовки очная

курс 1, 2 семестр 2,3,4
лекции 74 час.
практические занятия 92 час.
лабораторные работы 0 час.
всего часов аудиторной нагрузки 166 час.
самостоятельная работа 158 час.
контрольные работы программой не предусмотрены
курсовая работа/проект – не предусмотрено
зачет с оценкой 2 семестр
экзамен – 3, 4 семестр

Рабочая программа составлена в соответствии составлена в соответствии с требованиями Федерального государственного образовательного стандарта высшего образования по направлению подготовки/специальности 09.04.01 Информатика и вычислительная техника, утвержденного приказом Министерства образования и науки Российской Федерации от 19.09.2017 г. № 918.

Рассмотрена и утверждена на заседании Дирекции Школы цифровой экономики «17» июня 2019 года (протокол № 124-01-07-05).

Составитель: ст. пр. Кленин А.С.

Оборотная сторона титульного листа РПД

I. Рабочая программа пересмотрена на заседании Дирекции Школы цифровой экономики:

Протокол от «_____» _____ 20 г. № _____

Зам. директора по
учебной и воспитательной работе _____ _____
(подпись) (И.О. Фамилия)

II. Рабочая программа пересмотрена на заседании Дирекции Школы цифровой экономики:

Протокол от «_____» _____ 20 г. № _____

Зам. директора по
учебной и воспитательной работе _____ _____
(подпись) (И.О. Фамилия)

АННОТАЦИЯ

Б1.В.01.01 ПРИКЛАДНЫЕ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ И АНАЛИЗА ДАННЫХ

Рабочая программа учебной дисциплины «Прикладные методы машинного обучения и анализа данных» предназначена для студентов, обучающихся по направлению подготовки 09.04.01 Информатика и вычислительная техника (уровень магистратуры), профиль «Искусственный интеллект и большие данные».

Дисциплина «Прикладные методы машинного обучения и анализа данных» входит в часть, формируемую участниками образовательных отношений, блока «Дисциплины (модули)» (Б1.В.01) учебного плана подготовки магистров, модуля прикладных методов машинного обучения и искусственного интеллекта.

Общая трудоемкость освоения дисциплины составляет 11 зачетных единиц, 396 часов. Дисциплина реализуется на 1 и 2 курсе в 2, 3 и 4 семестрах.

Семестр	Аудиторные занятия			Самостоя-тельная работа	Контроль	Всего по дисциплине	
	Лекции	Лабораторные работы	Практические занятия			Часы	Зачетные единицы
2 семестр	18	-	36	54	Зачет с оценкой	108	3
3 семестр	36	-	36	72	Экзамен	180	5
4 семестр	20	-	20	32	Экзамен	108	3
Всего	**74**	**-**	**92**	**158**		**396**	**11**

Цель курса – дать понимание внутреннего устройства, механики работы, области применимости существующих решений, осветить сильные и слабые стороны, научить практическим навыкам анализа больших массивов информации. Курс посвящен методам построения систем обработки больших данных и существующим инструментам в этой области.

Структурно курс состоит из четырех разделов: пакетная обработка данных, потоковая обработка данных, хранение данных и особенности анализа данных в разных прикладных сферах (медицине, финансах, государственном и муниципальном управлении и т.п.).

В лекционной части курса рассматриваются такие технологии как HDFS, Hadoop MapReduce, HBase, Cassandra, Spark, Kafka, Spark Streaming, Storm.

Для успешного изучения дисциплины «Прикладные методы машинного обучения и анализа данных» обучающиеся должны обладать базовыми знаниями в следующих теоретических дисциплинах:

- специальные разделы математики, в том числе линейная алгебра, основы статистики, основы дискретной математики, исследование операций и оптимизация;
- технологии и методы программирования, в том числе объектно-ориентированного и начал функционального программирования;
- основы теории автоматов, основы теории вычислений;
- прикладные алгоритмы, а именно алгоритмы на графах и сетях, алгоритмы компьютерной графики, алгоритмы извлечения, обработки и классификации данных.

В результате данной дисциплины у обучающихся формируются следующие общепрофессиональные и профессиональные компетенции (элементы компетенций).

Код и формулировка компетенции	**Этапы формирования компетенции**
ОПК-7. Способен адаптировать зарубежные комплексы обработки информации и автоматизированного проектирования к нуждам отечественных предприятий	ОПК-7.1 Знать: функциональные требования к прикладному программному обеспечению для решения актуальных задач предприятий отрасли, национальные стандарты обработки информации и автоматизированного проектирования ОПК-7.2 Уметь: приводить зарубежные комплексы обработки информации в соответствие с национальными стандартами, интегрировать с отраслевыми информационными системами ОПК-7.3 Владеть: навыками настройки интерфейса, разработки пользовательских шаблонов, подключения библиотек, добавления новых функций
ПК-1 Способен проектировать информационные процессы и системы с использованием инновационных инструментальных средств, адаптировать современные информационные технологии к прикладным задачам	ПК-1.1 Знает: основные стандарты системной и программной инженерии; основные языки, средства и методы разработки программного обеспечения; устройство и принципы функционирования информационных систем; стандарты информационного взаимодействия систем; программные и аппаратные средства и платформы инфраструктуры информационных технологий ПК-1.2 Умеет: описывать задачи и составлять график выполнения работ IT-проекта, исходя из его целей и методов их достижения; оценивать трудоемкость и бюджет разработки программных средств; идентифицировать организационные и технические риски проектов; осуществлять текущее управление группой программистов, в том числе распределение заданий, приемку программного кода, обсуждение и принятие архитектурных решений ПК-1.3

	Владеет: методами работы с инструментами проектирования информационных систем; навыками управления разработкой программных продуктов; навыками управления проектными рисками в IT-проекте; навыками работы в распределенных командах
ПК-2 Способен разрабатывать методики выполнения аналитических работ	ПК-2.1 Знает: основные методики и практики выполнения аналитических работ; методы, применяемые для функционального и оперативного управления предприятиями; методы выбора проектных решений для корпоративных информационных систем в условиях неопределенности и риска ПК-2.2 Умеет: выявлять проблемы и сложности в существующих практиках выполнения аналитических работ в организации; описывать методики выполнения аналитических работ ПК-2.3 Владеет навыками выполнения аналитических работ, их апробации и доработки на выбранных проектах
ПК-3 Способен осуществлять планирование, организацию и контроль аналитических работ в IT-проекте	ПК-3.1 Знает: технологии и методы, используемые в управлении IT-проектами; инструментальные, программные и аппаратные платформы, образующие инфраструктуру анализа больших данных; ПК-3.2 Умеет: разрабатывать архитектуру, осуществлять выбор программных и аппаратных средств для аналитических работ; управлять процессом аналитических работ, в том числе осуществлять сбор информации, определять причины отклонений от планов, выявлять и разрешать проблемные ситуации в ходе аналитических работ ПК-3.3 Владеет навыками работы с инструментами анализа данных как системного, так и прикладного уровня
ПК-4 Способен проектировать и разрабатывать системные и прикладные решения по анализу больших данных	ПК-4.1 Знает основные математические методы анализа данных, компьютерного моделирования, методы машинного обучения; алгоритмы и методы работы с большими данными; полный цикл решения задачи анализа данных (подготовка данных; разработка признаков, выбор метрики качества, выбор и обучение модели, валидация модели и т.д.) ПК-4.2 Умеет: решать задачи анализа данных для конкретных предметных областей; проектировать и разрабатывать системные и прикладные решения по анализу больших данных ПК-4.3

	Владеет: навыками решения сложных и нестандартных задач анализа данных
ПК-5 Способен ставить цели и принимать управленческие решения, основанные на анализе больших данных	ПК-5.1 Знает: основные методы и модели машинного обучения и методы постановки задач на основе данных; общие принципы и методы принятия управленческих решений; основные понятия технологического предпринимательства ПК-5.2 Умеет: определять необходимые ресурсы и инструменты для решения задач с использованием анализа данных; руководить работой команды, вырабатывая командную стратегию на основе анализа данных ПК-5.3 Владеет: навыками принятия управленческих решений, как классическими, так и основанными на анализе больших данных

I. СТРУКТУРА И СОДЕРЖАНИЕ ТЕОРЕТИЧЕСКОЙ ЧАСТИ КУРСА

Тема 1: Вступление, распределенные файловые системы (1 час)

Понятие «большие данные». Постановка задачи обработки и хранения больших данных. Примеры применения больших данных в IT индустрии. Архитектура распределенных файловых систем. Основные проблемы в работе распределенных систем. Виды отказов узлов, связей между узлами. Устройство GFS, HDFS. Процесс восстановления HDFS.

Тема 2: Модель вычислений MapReduce (2 часа)

Математическая модель парадигмы MapReduce. Модель вычислений Map, Shuffle и Reduce фазы. Формальная модель парадигмы MapReduce. Задача подсчета слов в датасете (WordCount). Hadoop MapReduce. Обеспечение отказоустойчивости в MapReduce. Сравнение MapReduce v1 и YARN. История развития MapReduce. MapReduce Streaming на примере Python Расширения модели. Comparator, partitioner, combiner, зачем нужны и когда используются. Часто применяемые техники в обработке данных о Map-side join, reduce-side join. Salting. Способы тюнинга MapReduce. Способы семплирования данных. Итеративные задачи.

Тема 3. SQL over BigData. Hive. (2 часа)

Hive: мотивация, языковая модель. Проблема смещения данных в обработке больших данных. Применение SQL в IT индустрии. Сравнение решений Hive и MapReduce на примере задач анализа логов. Практика SQL: агрегация данных, фильтрация данных, сортировка, объединение таблиц. Архитектура Hive: Metastore + Hadoop + HDFS. Язык определения данных в Hive (Hive DDL): типы таблиц, разделители. Язык управления данными в Hive (Hive DML): загрузка данных, перезапись данных, CTAS. Hive: расширенные возможности. Парсер данных SerDe. Hive View: особенности, преимущества и недостатки. Пользовательские функции (UDF), пользовательские агрегирующие функции (UDAF), пользовательские функции для генерации таблиц (UDTF). История развития MapReduce. MapReduce Streaming на примере Python. Hive Streaming. Hive Partitioning, Bucketing and Sampling. Особенности Join в Hive. Исправление проблемы смещения в Hive. Поколоночное хранение в Hive (RCFile, ORC, Parquet).

Тема 4. Beyond MapReduce. Spark (2 часа)

Недостатки MapReduce. Costly disk spill, write barrier, job launch overhead. Перекосы в данных и перекосы в планировании. От MR к DAG-ам вычислений: почему это удобней?

Spark. Понятие RDD и Source RDD. MR over Spark, Pregel over Spark. Кеширование RDD, итеративные вычисления о Преобразования и действия. Spark UI и работа в режиме YARN. Spark SQL. Spark DataFrame: особенности

и сравнение с Pandas DataFrame о Spark SQL + Hive. Агрегирование данных в Spark DataFrame. Обработка графов при помощи Spark. Задача подсчета количества общих друзей. Задача подсчета числа треугольников. Пакет GraphFrames. Понятие motif. Использование motif для решения задачи. Решение задачи PageRank при помощи GraphFrames и Spark API. Оптимизация Spark. Управление памятью. Оптимизация UDF. Оптимизация объединений.

Тема 5. Машинное обучение на больших данных (2 часа)

Алгоритмы для работы с большими данными. Методы онлайн обучения. Градиентный спуск. Решение задач кластеризации на больших данных о Задача подсчета слов в датасете (WordCount). API для обучения алгоритмов на больших данных о Библиотеки Spark Mllib и Spark ML. Обработка текстов при помощи Spark ML о Ансамблевые модели на Spark ML. Map-side join, reduce-side join

Тема 6. Потоковая обработка данных (1 час)

Обработка больших данных в режиме реального времени. Подходы к обработке больших данных в режиме реального времени. Семантика доставки (Devilery semantics). Архитектуры Lambda и Kappa. Входные и выходные данные для обработки в режиме реального времени. Apache Spark Streaming: объяснение концепции на практической задаче. Apache Spark Structured Streaming: объяснение концепции на практической задаче. Модель парадигмы Kafka. Понятие интервала в Kafka. Особенности Kafka. Интерфейс командной строки Kafka. Связь Kafka и семантики доставки о Потоки Kafka (Kafka Streams).

Тема 7. Key-value хранилища в больших данных (1 час)

HBase. NoSQL подходы к реализации распределенных баз данных, key-value хранилища. Основные компоненты BigTable-подобных систем и их назначение, отличие от реляционных БД. Чтение, запись и хранение данных в HBase. Minor- и major- компактификация. Надёжность и отказоустойчивость в HBase. Cassandra. Основные особенности. Чтение и запись данных. Отказоустойчивость. Примеры применения HBase и Cassandra. Отличие архитектуры HBase от архитектуры Cassandra.

II. СТРУКТУРА И СОДЕРЖАНИЕ ПРАКТИЧЕСКОЙ ЧАСТИ КУРСА И САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Лабораторные работы (44 час.)

Лабораторная работа № 1 «Основы работы с аналитической платформой Deductor studio» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 2 «Трансформация данных в Deductor Studio» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 3 «Создание, заполнение и использование хранилища данных Deductor Warehouse на базе Firebird» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 4 «Определение представления источника данных в проекте служб Analysis Services» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 5 «Определение и развертывание куба» (4 час.)

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.

3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 6 «Изменение мер, атрибутов и иерархий» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 7. «Ассоциативные правила» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 8. «Основы работы с пакетом STATISTICA» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 9. «Кластерный анализ» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 10. «Регрессионный анализ» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.
3. Обработка результатов, составление отчета, защита лабораторной работы.

Лабораторная работа № 11. «Искусственные нейронные сети» (4 час.).

1. Проработка теоретических вопросов по теме лабораторной работы.
2. Постановка задач и компьютерное моделирование по вопросам практической части лабораторной работы.

3. Обработка результатов, составление отчета, защита лабораторной работы.

III. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

Учебно-методическое обеспечение самостоятельной работы обучающихся по дисциплине «Методы и системы обработки больших данных» представлено в Приложении 1 и включает в себя:

- план-график выполнения самостоятельной работы по дисциплине, в том числе примерные нормы времени на выполнение по каждому заданию;
- характеристика заданий для самостоятельной работы обучающихся и методические рекомендации по их выполнению;
- требования к представлению и оформлению результатов самостоятельной работы;
- критерии оценки выполнения самостоятельной работы.

IV. КОНТРОЛЬ ДОСТИЖЕНИЯ ЦЕЛЕЙ КУРСА

№ п/п	Контролируемые разделы / темы дисциплины	Коды и этапы формирования компетенций		Оценочные средства	
				текущий контроль	промежуточная аттестация
1.	Тема 1: Вступление, распределенные файловые системы	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
2.	Тема 2: Модель вычислений MapReduce	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
3.	Тема 3. SQL over BigData. Hive.	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
4.	Тема 4. Beyond MapReduce. Spark		знает	эссе (ПР-3)	Экзамен

		ОПК-6, ПК-8, УПК-1	умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
5.	Тема 5. Машинное обучение на больших данных	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
6.	Тема 6. Потоковая обработка данных	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
7.	Тема 7. Key-value хранилища в больших данных	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
		ОПК-6, ПК-8, УПК-1	умеет	лабораторная работа (ПР-6)	Экзамен
		ОПК-6, ПК-8, УПК-1	владеет	лабораторная работа (ПР-6)	Экзамен

Обозначения: ПР-3 – эссе (письменная работа); ПР-6 – Отчет по лабораторной работе (письменная работа)

Типовые контрольные задания, методические материалы, определяющие процедуры оценивания знаний, умений и навыков и (или) опыта деятельности, а также критерии и показатели, необходимые для оценки знаний, умений, навыков и характеризующие этапы формирования компетенций в процессе освоения образовательной программы, представлены в Приложении 2.

V. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ

Компьютерный класс:
Проектор DLP, 3000 ANSI Lm, WXGA 1280x800, 2000:1 EW330U Mitsubishi,;
Системный блок с монитором. Процессор: Intel I5-8600k 3.6Ghz, оперативная память: 32gb, жесткий диск: 1ТБ, графический ускоритель: Nvidia GTX 1080
Беспроводные ЛВС для обучающихся обеспечены системой на базе точек доступа 802.11a/b/g/n 2x2 MIMO(2SS).

VI. СПИСОК УЧЕБНОЙ ЛИТЕРАТУРЫ И ИНФОРМАЦИОННО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература

(электронные и печатные издания)

1. Просто о больших данных : пер. с англ. / Джудит Гурвиц, Алан Ньюджент, Ферн Халпер [и др.]. - Москва : Сбербанк, : [Эксмо], 2015. - 395 с. - http://lib.dvfu.ru:8080/lib/item?id=chamo:826169&theme=FEFU

2. Гончарук, С. В. Администрирование ОС Linux [Электронный ресурс] / С. В. Гончарук. — Электрон. текстовые данные. — М. : Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. — 164 c. — 2227-8397. — Режим доступа: http://www.iprbookshop.ru/52142.html

3. Бражук, А. И. Сетевые средства Linux [Электронный ресурс] / А. И. Бражук. — 2-е изд. — Электрон. текстовые данные. — М.: Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. — 147 c. — 2227-8397. — Режим доступа: http://www.iprbookshop.ru/73722.html

4. Алексеева, Т.В. Информационные аналитические системы [Электронный ресурс] : учебник / Т.В. Алексеева [и др.]. – М. : Московский финансово-промышленный ун-т «Синергия», 2013. – 384 c. – Режим доступа : http://www.iprbookshop.ru/17015.html

Дополнительная литература

(печатные и электронные издания)

1.Пальмов С.В. Интеллектуальный анализ данных [Электронный ресурс] : учебное пособие / С.В. Пальмов. – Самара: Поволжский государственный университет телекоммуникаций и информатики, 2017. – 127 c. – 2227-8397. – Режим доступа: http://www.iprbookshop.ru/75376.html

2.Петрунин, Ю. Ю. Информационные технологии анализа данных. Data Analysis : учебное пособие для вузов по управленческим и экономическим специальностям и направлениям / Ю. Ю. Петрунин ; Московский государственный университет, Факультет государственного управления. – 3-е изд. – М. : Университет, 2014 – 291 с. – Каталог НБ ДВФУ: http://lib.dvfu.ru:8080/lib/item?id=chamo:734307&theme=FEFU
http://lib.dvfu.ru:8080/lib/item?id=chamo:417764&theme=FEFU

3. Туманов, В.Е. Проектирование хранилищ данных для систем бизнес-аналитики [Электронный ресурс] : учеб. пособие / Туманов В.Е. – М. : БИНОМ. Лаборатория знаний, Интернет-Университет Информационных Технологий (ИНТУИТ), 2010. – 615 с. – Режим доступа : http://www.iprbookshop.ru/16096.html

4. Интеллектуальные системы принятия решений и управления : учебное пособие для вузов / Ю. И. Еременко. – Старый Оскол : ТНТ, 2015. – 401 с. – Каталог НБ ДВФУ:.http://lib.dvfu.ru:8080/lib/item?id=chamo:813810&theme=FEFU

5. Интеллектуальный анализ данных и систем управления бизнес-правилами в телекоммуникациях: Монография / Р.Р. Вейнберг. – М.: НИЦ ИНФРА-М, 2016. – 173 с.: – Режим доступа: http://znanium.com/catalog/product/520998

6. Нестеров, С.А. Интеллектуальный анализ данных средствами MS SQL Server [Электронный ресурс] / Нестеров С.А. – М.: Интернет-Университет Информационных Технологий (ИНТУИТ), 2012. – 189 с. – Режим доступа : http://www.iprbookshop.ru/16702.html

7. Чубукова И.А. Data Mining [Электронный ресурс]/ Чубукова И.А. – М.: Интернет-Университет Информационных Технологий (ИНТУИТ), 2016. – 470 с. – Режим доступа: http://www.iprbookshop.ru/56315.html.

8. Чубукова, И.А. Data Mining : учеб. пособие для вузов / И.А. Чубукова / М.Р. Мидлтон ; пер. с англ. [С.Г. Кобелькова]. – М. : Интернет-Университет Информационных Технологий БИНОМ. Лаборатория знаний, 2008. – 282 с. : – Каталог НБ ДВФУ: http://lib.dvfu.ru:8080/lib/item?id=chamo:274659&theme=FEFU

Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. Linux. Карманный справочник. Скотт Граннеман

2. Unix и Linux. Руководство системного администратора. Эви Немет, Гарт Снайдер,Трент Р. Хейн, Бен Уэйли

3. The Official Ubuntu Book Matthew Helmke, Elizabeth K. Joseph, José Antonio Rey, Philip

Ballew, Benjamin Mako Hill

4. Hadoop: The Definitive Guide 3e. Tom White

5. Professional Hadoop Solutions. Boris Lublinsky

Приложение 1

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)

ШКОЛА ЦИФРОВОЙ ЭКОНОМИКИ

УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

по дисциплине

«МЕТОДЫ И СИСТЕМЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ»

направления 09.04.03 Прикладная информатика

Магистерская программа «Искусственный интеллект и большие данные»

Форма подготовки очная

Владивосток
2018

План-график выполнения самостоятельной работы по дисциплине

№ п/п	Дата/сроки выполнения	Вид самостоятельной работы	Примерные нормы времени на выполнение	Форма контроля
1	1-18 неделя обучения	Подготовка лабораторных работ.	35	Отчет по лабораторной работе
2	Сессия	Подготовка к экзамену	10	Экзамен

Методические рекомендации к работе с литературными источниками

В процессе подготовки к практическим занятиям, студентам необходимо обратить особое внимание на самостоятельное изучение рекомендованной учебно-методической (а также научной и популярной) литературы. Самостоятельная работа с учебниками, учебными пособиями, научной, справочной и популярной литературой, материалами периодических изданий и Интернета, статистическими данными является наиболее эффективным методом получения знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала, формирует у студентов свое отношение к конкретной проблеме. Более глубокому раскрытию вопросов способствует знакомство с дополнительной литературой, рекомендованной преподавателем по каждой теме практического занятия, что позволяет студентам проявить свою индивидуальность в рамках выступления на данных занятиях, выявить широкий спектр мнений по изучаемой проблеме.

Критерии оценки выполнения самостоятельной работы

Контроль самостоятельной работы студентов предусматривает:

- соотнесение содержания контроля с целями обучения;
- объективность контроля;
- валидность контроля (соответствие предъявляемых заданий тому, что предполагается проверить);
- дифференциацию контрольно-измерительных материалов.

Формы контроля самостоятельной работы

1. Просмотр и проверка выполнения самостоятельной работы преподавателем.
2. Самопроверка, взаимопроверка выполненного задания в группе.
3. Обсуждение результатов выполненной работы на занятии.
4. Текущее тестирование.

Критерии оценки результатов самостоятельной работы

Критериями оценок результатов внеаудиторной самостоятельной работы студента являются:

- уровень освоения студентов учебного материала;
- умения студента использовать теоретические знания при выполнении практических задач;
- умения студента активно использовать электронные образовательные ресурсы, находить требующуюся информацию, изучать ее и применять на практике;
- обоснованность и четкость изложения ответа;
- оформление материала в соответствии с требованиями;
- умение ориентироваться в потоке информации, выделять главное;
- умение четко сформулировать проблему, предложив ее решение, критически оценить решение и его последствия;
- умение показать, проанализировать альтернативные возможности, варианты действий;
- умение сформировать свою позицию, оценку и аргументировать ее

Критерии оценки выполнения контрольных заданий для самостоятельной работы

Процент правильных ответов	**Оценка**
От 95% до 100%	отлично
От 76% до 95%	хорошо
От 61% до 75%	удовлетворительно
Менее 61 %	неудовлетворительно

Самостоятельная работа при подготовке к экзамену включает изучение теоретического материала с использованием лекционных материалов, рекомендуемых источников, материалов по практическим занятиям и лабораторным работам.

Контрольные вопросы для самостоятельной оценки качества освоения учебной дисциплины:

1. Определите сущность понятия «большие данные».
2. Опишите методики анализа больших данных.
3. Процесс аналитики анализа больших данных.
4. Дайте характеристику Big Data на мировом рынке.
5. Охарактеризуйте Big Data в России.
6. Определите понятие Data Mining.
7. Вопросы безопасности больших данных.
8. В чем состоит когнитивный анализ данных.
9. Какие модели данных вы знаете?
10. Концепция MapReduce.
11. Основные методы анализа больших данных.
12. Продемонстрировать использование окна Explore для изучения распределения переменной. Выявить, какие отклонения выделяются в распределении переменной? Как можно исправить эту ситуацию?
13. Провести первоначальное исследование данных с использованием узлов Data Partition и
14. Decision Tree. Сделать выводы о модели.
15. Провести прогнозное моделирование с использованием регрессии Regression. Объяснить, какие переменные являются важными в модели. Какое значение имеет статистика среднеквадратичной ошибки, посчитанная на проверочной выборке?
16. Нужно ли делать преобразования входных переменных перед их использованием в модели нейронной сети?
17. Объяснить результаты, полученные на основании Model Comparison. Сделать выводы.
18. Регрессия. Логистическая регрессия. Полиномиальные регрессии.
19. Задания по обработке данных и созданию моделей выполняются с использованием данных из набора SAMPSIO

Приложение 2

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)

ШКОЛА ЦИФРОВОЙ ЭКОНОМИКИ

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

по дисциплине

«МЕТОДЫ И СИСТЕМЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ»

направления 09.04.03 Прикладная информатика

Магистерская программа «Искусственный интеллект и большие данные»

Форма подготовки очная

Владивосток
2018

Паспорт фонда оценочных средств

Код и формулировка компетенции	Этапы формирования компетенции	
ОПК-6 – способность к профессиональной эксплуатации современного электронного оборудования в соответствии с целями основной образовательной программы магистратуры	Знает	основные принципы работы с современным электронным оборудованием; методы эксплуатации современного электронного оборудования в задачах интеллектуального анализа и хранилищ данных
	Умеет	использовать современное электронное оборудование в задачах интеллектуального анализа и хранилищ данных
	Владеет	навыками работы с современным электронным оборудованием в целях обеспечения задач интеллектуального анализа и хранилищ данных
ПК-8 – способность анализировать данные и оценивать требуемые знания для решения нестандартных задач с использованием математических методов и методов компьютерного моделирования	Знает	основные математические методы анализа данных и методы компьютерного моделирования
	Умеет	анализировать данные и оценивать требуемые знания для решения нестандартных задач
	Владеет	математическими методами и методами компьютерного моделирования для анализа данных и оценки требуемых знаний для решения нестандартных задач
УПК-1 - способность проектировать и разрабатывать системные и прикладные решения по анализу больших данных	Знает	основные методы и модели машинного обучения и их применение для анализа данных; полный цикл решения задачи анализа данных: подготовка данных; разработка признаков, выбор метрики качества, выбор и обучение модели, валидация модели и т.д.
	Умеет	решать задачи анализа данных для конкретных предметных областей
	Владеет	навыками решения сложных и нестандартных задач анализа данных

Контроль достижений целей курса

№ п/п	Контролируемые разделы / темы дисциплины	Коды и этапы формирования компетенций		Оценочные средства	
				текущий контроль	промежуточная аттестация
1.	Тема 1: Вступление, распределенные файловые системы	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
2.	Тема 2: Модель вычислений MapReduce	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
3.	Тема 3. SQL over BigData. Hive.	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
4.	Тема 4. Beyond MapReduce. Spark	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
5.	Тема 5. Машинное обучение на больших данных	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
6.	Тема 6. Потоковая обработка данных	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен
			умеет	лабораторная работа (ПР-6)	Экзамен
			владеет	лабораторная работа (ПР-6)	Экзамен
7.	Тема 7. Key-value хранилища в больших данных	ОПК-6, ПК-8, УПК-1	знает	эссе (ПР-3)	Экзамен

		ОПК-6, ПК-8, УПК-1	умеет	лабораторная работа (ПР-6)	Экзамен
		ОПК-6, ПК-8, УПК-1	владеет	лабораторная работа (ПР-6)	Экзамен

Методические рекомендации, определяющие процедуры оценивания результатов освоения дисциплины

Текущая аттестация студентов. Текущая аттестация студентов по дисциплине проводится в соответствии с локальными нормативными актами ДВФУ и является обязательной.

Текущая аттестация по дисциплине проводится в форме контрольных мероприятий (защита эссе, защита лабораторных работ, тестирование) по оцениванию фактических результатов обучения студентов осуществляется ведущим преподавателем.

Объектами оценивания выступают:

- учебная дисциплина (активность на занятиях, своевременность выполнения различных видов заданий, посещаемость всех видов занятий по аттестуемой дисциплине);

- степень усвоения теоретических знаний;

- уровень овладения практическими умениями и навыками по всем видам учебной работы;

- результаты самостоятельной работы.

Оценивание результатов освоения дисциплины на этапе текущей аттестации проводится в соответствии с используемыми оценочными средствами и критериями.

Процедура и критерии оценивания эссе

Оценивание защиты эссе проводится при представлении эссе в электронном виде, по двухбалльной шкале: «зачтено», «незачтено».

Оценка «зачтено» выставляется студенту, если он представляет к защите эссе, удовлетворяющее поставленным к эссе требованиям (использование данных отечественной и зарубежной литературы, источников Интернет, информации нормативноправового характера и передовой практики, представление краткого терминологического словаря по теме), по оформлению, если студент демонстрирует владение методами и приемами

теоретических аспектов работы, не допускает фактических ошибок, связанных с пониманием проблемы.

Оценка «незачтено» выставляется студенту, если он не владеет методами и приемами теоретических аспектов работы, допускает существенные ошибки в работе, связанные с пониманием проблемы, представляет эссе с существенными отклонениями от правил оформления письменных работ.

Процедура и критерии оценивания отчетов по лабораторным работам

Оценивание защиты лабораторной работы проводится при представлении отчета в электронном виде, по двухбалльной шкале: «зачтено», «незачтено».

Оценка «зачтено» выставляется студенту, если он представляет к защите отчет по лабораторной работе, удовлетворяющий требованиям по поставленным заданиям, по оформлению, демонстрирует владение методами и приемами теоретических и/или практических аспектов работы.

Оценка «незачтено» выставляется студенту, если он не владеет методами и приемами теоретических и/или практических аспектов работы, допускает существенные ошибки в работе, представляет отчет с существенными отклонениями от правил оформления письменных работ.

Процедура и критерии оценивания тестирования

Тест включает 50 заданий, максимальная оценка по тесту - 100.

В рамках текущего контроля уровня усвоения знаний по дисциплине допускается результат тестирования, не ниже 61 балла.

Промежуточная аттестация студентов. Промежуточная аттестация студентов по дисциплине «Интеллектуальный анализ на основе хранилищ данных» проводится в соответствии с локальными нормативными актами ДВФУ и является обязательной.

Промежуточная аттестация по дисциплине проводится в виде экзамена, форма экзамена - «устный опрос в форме ответов на вопросы», «практические задания по типам».

Порядок проведения экзамена, форма экзаменационного билета определены локальным нормативными актом ДВФУ «Положение о текущем контроле успеваемости, текущей и промежуточной аттестации студентов, обучающихся по программам высшего образования (бакалавриата, специалитета и магистратуры) в ДВФУ».

В экзаменационный билет входят два вопроса (1-й – по темам 1-3, 2-й – по темам 4-6) и одно практическое задание.

Критерии выставления оценки студенту на экзамене по дисциплине

Баллы (рейтинговой оценки)	Оценка экзамена (стандартная)	Требования к сформированным компетенциям
86 -100	«отлично»	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, свободно справляется с задачами, вопросами и другими видами применения знаний, причем не затрудняется с ответом при видоизменении заданий, использует в ответе материал монографической литературы, правильно обосновывает принятое решение, владеет разносторонними навыками и приемами выполнения практических задач.
76 - 85	«хорошо»	Оценка «хорошо» выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении практических вопросов и задач, владеет необходимыми навыками и приемами их выполнения.
61 -75	«удовлетворительно»	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, испытывает затруднения при выполнении практических работ.
0 -60	«неудовлетворительно»	Оценка «неудовлетворительно» выставляется студенту, который не знает значительной части программного «не материала, допускает существенные ошибки, неуверенно, с большими затруднениями выполняет практические работы. Как правило, оценка «неудовлетворительно» «неудовлетворительно» ставится студентам, которые не могут продолжить обучение без дополнительных занятий по соответствующей дисциплине.

Оценочные средства для промежуточной аттестации

Вопросы к экзамену

1. Модели и их свойства. Аналитический и информационный походы к моделированию.
2. Формы представления, типы и виды анализируемых данных.
3. Обучение моделей «с учителем» и «без учителя». Обучающее и тестовое множество. Ошибки обучения. Эффект переобучения.
4. Общая схема анализа данных. Требования к алгоритмам анализа данных.
5. Основные принципы сбора (формализации) данных. Требования к объемам анализируемых данных.
6. Характеристика этапов технологии KDD.
7. Data Mining. Характеристика классов задач, решаемых методами Data Mining.
8. Программный инструментарий для выполнения анализа данных.
9. Цели, задачи и основное содержание консолидации данных. Обобщенная схема процесса консолидации.
10. Характеристика OLTP-систем.
11. Предпосылки появления систем поддержки принятия решений DSS. Понятие ESS, EIS и GDSS.
12. Основные положения концепции хранилищ данных (DW).
13. Реляционные хранилища данных (ROLAP).
14. Технология OLAP. Сущность многомерного представления данных.
15. Структура многомерного куба. Работа с измерениями.
16. Многомерные хранилища данных (MOLAP).
17. Гибридные хранилища данных (HOLAP).
18. Виртуальные хранилища данных.
19. Цели, задачи и основное содержание процесса ETL.
20. Основные виды проблем в данных, из-за которых они нуждаются в очистке.
21. Организация процесса загрузки данных в хранилище. Постзагрузочные операции.
22. Причины отказа от использования хранилищ данных. Особенности загрузки данных из локальных источников.
23. Обогащение данных.

24. Цели, задачи и основное содержание трансформации данных. Трансформация данных на разных этапах аналитического процесса. Типичные средства трансформации.

25. Особенности трансформации временных рядов. Скользящее окно. Преобразование даты и времени.

26. Группировка и разгруппировка данных.

27. Способы слияния данных.

28. Квантование данных.

29. Нормализация и кодирование данных.

30. Цели, задачи и основное содержание визуализации данных. Группы методов визуализации.

31. Визуализаторы общего назначения. OLAP-анализ.

32. Манипуляции с измерениями OLAP-куба.

33. Визуализаторы, применяемые для оценки качества моделей.

34. Визуализаторы, применяемые для интерпретации результатов анализа.

35. Технологии и методы оценки качества данных. Профайлинг.

36. Очистка и предобработка данных.

37. Типичный набор инструментов предобработки данных в аналитическом приложении.

38. Фильтрация данных. Обработка дубликатов и противоречий.

39. Выявление аномальных и восстановление пропущенных значений.

40. Алгоритмы и методы сокращения числа признаков.

Оценочные средства для текущей аттестации

Темы эссе

1. Технологии анализа данных:

1) Аналитический и информационный походы к моделированию.
2) Формы представления, типы и виды анализируемых данных.
3) Источники данных для анализа.

2. Визуализация данных:

1) Визуализаторы общего назначения. *OLAP*-анализ.

2) Визуализаторы, применяемые для оценки качества моделей.

3) Визуализаторы, применяемые для интерпретации результатов анализа.

3. Инструменты Data mining:

1) Поиск ассоциативных правил
2) Кластеризация
3) Классификация и регрессия

Типовые задания к лабораторным работам

Лабораторная работа № 1 «Основы работы с аналитической платформой Deductor studio»

Цель работы: овладеть основами работы с аналитической платформой Deductor studio».

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Программно-аппаратное обеспечение: ПЭВМ IBM PC (операционная система Windows 10), аналитическая платформа Deductor Studio.

Лабораторная работа № 2 «Трансформация данных в Deductor Studio».

Цель работы: овладеть навыками трансформации данных в Deductor Studio.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Программно-аппаратное обеспечение: ПЭВМ IBM PC (операционная система Windows 10), аналитическая платформа Deductor Studio.

Лабораторная работа № 3 «Создание, заполнение и использование хранилища данных Deductor Warehouse на базе Firebird»

Цель работы: овладеть навыками создания, заполнения и использования хранилища данных Deductor Warehouse на базе Firebird.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;

– SQL Server составление отчета, защита работы.

Программно-аппаратное обеспечение: ПЭВМ IBM PC (операционная система Windows 10), аналитическая платформа Deductor Studio Academic.

Лабораторная работа № 4 «Определение представления источника данных в проекте служб Analysis Services».

Цель работы: овладеть навыками представления источника данных в проекте служб Analysis Services.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Программно-аппаратное обеспечение: ПЭВМ IBM PC (операционная система Windows 10 Professional), SQL Server Developer.

Лабораторная работа № 5 «Определение и развертывание куба».

Цель работы: овладеть навыками определения и развертывания куба.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Программно-аппаратное обеспечение: ПЭВМ IBM PC (операционная система Windows 10 Professional), SQL Server Developer.

Лабораторная работа № 6 «Изменение мер, атрибутов и иерархий».

Цель работы: овладеть навыками изменения мер, атрибутов и иерархий при компьютерном моделировании задач.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Программно-аппаратное обеспечение: ПЭВМ IBM PC (операционная система Windows 10 Professional), SQL Server Developer.

Лабораторная работа № 7. «Ассоциативные правила».

Цель работы: овладеть навыками применения ассоциативных правил при компьютерном моделирования задач.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Лабораторная работа № 8. «Основы работы с пакетом STATISTICA».

Цель работы: овладеть основами работы с пакетом STATISTICA.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Лабораторная работа № 9. «Кластерный анализ»

Цель работы: овладеть навыками компьютерного моделирования задач на основе кластерного анализа.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Лабораторная работа № 10. «Регрессионный анализ».

Цель работы: овладеть навыками компьютерного моделирования задач методами регрессионного анализа.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Лабораторная работа № 11. «Искусственные нейронные сети».

Цель работы: овладеть навыками компьютерного моделирования задач искусственной нейронной сети.

Программа работы

- задание исходных данных;
- разработка модели;
- компьютерное моделирование;
- анализ полученных данных;
- составление отчета, защита работы.

Типовые тестовые задания

Укажите номер правильного ответа

МОДЕЛИ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ

1) семантические сети
2) логические подходы
3) когнитивные карты

ПРИМЕРОМ ИНТЕЛЛЕКТУАЛЬНОЙ ЗАДАЧИ ЯВЛЯЕТСЯ

1) расчет % по кредиту
2) выбор партнера по бизнесу
3) расчет годового баланса

ИНФОРМАЦИОННОЕ ХРАНИЛИЩЕ ПРЕДНАЗНАЧЕНО ДЛЯ

1) обработки больших объемов информации
2) обеспечения управляющего персонала аналитическими данными для принятия решений
3) обработки больших объемов информации и обеспечения управляющего персонала аналитическими данными для принятия решений

В ОТЛИЧИЕ ОТ ИНТЕЛЛЕКТУАЛЬНОЙ БАЗЫ ДАННЫХ ИНФОРМАЦИОННОЕ ХРАНИЛИЩЕ ПРЕДСТАВЛЯЕТ СОБОЙ САМООБУЧАЮЩУЮ ИИС, КОТОРАЯ

1) в качестве единиц знаний хранит примеры решений и позволяет по запросу подбирать и адаптировать наиболее похожие решения
2) позволяет извлекать знания из баз данных и создавать специально-организованные базы знаний
3) на основе обучения по примерам реальной практики строит ассоциативную сеть понятий (нейронов) для параллельного поиска на ней решений