

МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
Ярославский государственный университет им. П.Г.Демидова

УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ
ПО ДИСЦИПЛИНЕ

ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ И DATA MINING

Направление подготовки (специальность):

02.04.02 Фундаментальная информатика и информационные технологии

Образовательная программа

Искусственный интеллект и компьютерные науки

очная форма обучения

Составитель:

БОГОМОЛОВ ЮРИЙ ВИКТОРОВИЧ,
К.Ф-М.Н, ДОЦЕНТ Ф-ТА ИВТ ЯРГУ ИМ. П.Г. ДЕМИДОВА

г. Ярославль

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

Основная литература:

1. Лесковец Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман, Дж. Ульман – Пер. с англ. Слинкин А.А. – М.: ДМК Пресс, 2016.

Дополнительная литература:

2. Боровиков, В. П., Популярное введение в современный анализ данных в системе STATISTICA : методология и технология современного анализа данных : учеб. пособие для вузов / В. П. Боровиков, М., Горячая линия - Телеком, 2015, 288с
3. Федин Ф.О. Анализ данных. Часть 1. Подготовка данных к анализу [Электронный ресурс]: учебное пособие / Ф.О. Федин, Ф.Ф. Федин. — Электрон. текстовые данные. — М.: Московский городской педагогический университет, 2012. — 204 с. — 2227-8397. Режим доступа: <http://www.iprbookshop.ru/26444.htm> (по паролю).
4. Федин Ф.О. Анализ данных. Часть 2. Инструменты Data Mining [Электронный ресурс]: учебное пособие / Ф.О. Федин, Ф.Ф. Федин. — Электрон. текстовые данные. — М. : Московский городской педагогический университет, 2012. — 308 с. — 2227-8397. — Режим доступа: <http://www.iprbookshop.ru/26445.html> (по паролю).
5. Воронова Л.И. Big Data. Методы и средства анализа [Электронный ресурс]: учебное пособие / Л.И. Воронова, В.И. Воронов – М.: Московский технический университет связи и информатики, 2016. 33с. - Режим доступа: <http://www.iprbookshop.ru/61463.html> (по паролю).

Учебно-методическое обеспечение самостоятельной работы обучающихся по дисциплине (модулю)

Для самостоятельной работы особенно рекомендуется использовать учебную литературу.

Также для подбора учебной литературы рекомендуется использовать широкий спектр Интернет-ресурсов:

1. Электронно-библиотечная система «Университетская библиотека online» (www.biblioclub.ru) - электронная библиотека, обеспечивающая доступ к наиболее востребованным материалам-первоисточникам, учебной, научной и художественной литературе ведущих издательств (регистрация в электронной библиотеке – только в сети университета. После регистрации работа с системой возможна с любой точки доступа в Internet.).
2. Информационная система "Единое окно доступа к образовательным ресурсам" (<http://window.edu.ru/library>).

Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля), включая перечень информационных справочных систем (при необходимости)

1. Электронные каталоги Научной библиотеки ЯрГУ им. П.Г. Демидова (http://www.lib.uni-yar.ac.ru/opac/bk_one_find.php)

2. Электронная картотека «Книгообеспеченность» Научной библиотеки ЯрГУ им. П.Г. Демидова (http://www.lib.uniyl.ac.ru/opac/bk_bookreq_find.php)
3. Электронно-библиотечная система «Университетская библиотека online» (www.biblioclub.ru)

Перечень информационных технологий, используемых при изучении дисциплины, включая программное обеспечение

В процессе осуществления образовательного процесса используются:

- для формирования текстов материалов для промежуточной и текущей аттестации, для разработки документов, презентаций, для работы с электронными таблицами – программы OfficeStd 2013 RUS OLP NL Acdmc 021-10232, LibreOffice (свободное), издательская система LaTeX;
- аналитическая платформа Deductor Studio Academic v. 6.2. и выше (свободно распространяемая фирмой-разработчиком BaseGroup Labs. Доступ <https://basegroup.ru/deductor/download>);
- для поиска учебной литературы библиотеки ЯрГУ – Автоматизированная библиотечная информационная система "БУКИ-NEXT" (АБИС "Буки-Next")

Учебно-методические указания и рекомендации к изучению тем лекционных и практических занятий, самостоятельной работе студентов

Содержание дисциплины

Наименование раздела дисциплины	Название темы с кратким содержанием
Задачи интеллектуального анализа данных (DataMining).	РАЗДЕЛ 1. Задачи интеллектуального анализа данных (Data Mining) Формы представления данных Типы данных (векторные, категориальные, порядковые, не-структурированные). Представления наборов данных. Особенности данных, накопленных в компаниях. Формализация данных. Корреляционный анализ числовых и ранжированных данных. Задачи интеллектуального анализа данных в маркетинговых и социологических исследованиях, прогнозирования, технической и медицинской диагностики.
Задачи классификации данных	РАЗДЕЛ 2. Задачи классификации данных Формальная постановка задачи классификации Алгоритмы классификации векторных данных (kNN – метод «к ближайших соседей», линейные классификаторы). Линейная регрессия. Классификация категориальных данных (деревья решений). Вероятностная классификация (байесовский классификатор). Нейросетевые алгоритмы классификации.

Наименование раздела	Название темы с кратким содержанием
Поиск ассоциативных правил.	РАЗДЕЛ 3. Поиск ассоциативных правил. Базы транзакций, ассоциативные правила, показатели достоверности и поддержки ассоциативных правил (на примере анализа рыночной корзины). Алгоритм Apriori построения ассоциативных правил. Определение значимости и полезности ассоциативных правил, показатели их характеризующие.
Задача кластеризации данных	РАЗДЕЛ 4. Задача кластеризации данных. Постановка задачи кластеризации. Графовые алгоритмы кластеризации. Алгоритмы k-means и графовые алгоритмы. Иерархическая кластеризация данных, основные подходы. Агломеративные и дивизионные методы. Метрики в пространстве кластеров. Кластеризация категориальных данных, алгоритм CLOPE.
Анализ текстовой информации, классификация и кластеризация текстов.	РАЗДЕЛ 5. Анализ текстовой информации Постановка задач классификация и кластеризация текстов. Алгоритмы кластеризации и классификации текстовой информации. Задача аннотирования текстов.
Концепции больших данных (Big Data)	РАЗДЕЛ 6. Концепции больших данных (Big Data). Понятие и примеры больших данных. Базовые принципы обработки больших данных (горизонтальная масштабируемость и др.). Модель распределённых вычислений MapReduce, используемая для параллельных вычислений над большими наборами данных (Big Data).

Задания для самостоятельной работы

Задания №1 по разделу 2.

Нейросетевые алгоритмы классификации

Задания №2 по разделу 3.

Определение значимости и полезности ассоциативных правил, показатели их характеризующие..

Задания №3 по разделу 4.

Кластеризация категориальных данных, алгоритм CLOPE.

Задания №4 по разделу 5.

Задача аннотирования текстов.

Задания №5 по разделу 6.

Модель распределённых вычислений MapReduce.

Критерии оценивания

Показатели	Критерии	Оценка
Знать: – алгоритмы построения деревьев решений и ассоциативных правил; – базовые алгоритмы кластеризации числовых и категориальных данных; задачи анализа текстов, алгоритмы классификации,	Студентом дан полный, в логической последовательности развернутый ответ на поставленный вопрос, где он продемонстрировал знания предмета в полном объеме учебной программы, достаточно глубоко осмысливает дисциплину. Студент самостоятельно, и	«Зачтено»

<p>кластеризации и аннотирования текстов</p> <p>Уметь:</p> <ul style="list-style-type: none"> – формулировать задачи анализа данных разного типа; – проводить кластеризацию многомерных (векторных) и категориальных данных; – строить деревья решений и ассоциативные правила для обнаружения логических закономерностей в данных <p>Знать области прикладной информатики, связанные с интеллектуальным анализом больших массивов данных.</p>	<p>исчерпывающе отвечает на дополнительные вопросы, приводит собственные примеры по проблематике поставленного вопроса, решил предложенные практические задания без ошибок.</p> <p>Студентом дан ответ, который содержит ряд серьезных неточностей, обнаруживающий незнание процессов изучаемой предметной области, отличающийся неглубоким раскрытием темы, незнанием основных вопросов теории, несформированными навыками анализа явлений, процессов, неумением давать аргументированные ответы, слабым владением монологической речью, отсутствием логичности и последовательности. Выводы поверхностны. Решение практических заданий не выполнено. Т.е студент не способен ответить на вопросы даже при дополнительных наводящих вопросах преподавателя.</p>	<p>Неудовлетворительно (уровень не сформирован)</p>
---	--	---

Список заданий к зачету

1. Коэффициенты корреляции числовых и ранжированных данных
2. Уравнение линейной регрессии.
3. Формальная постановка задачи классификации на основе обучающей выборки. Алгоритм классификации kNN («к ближайших соседей»).
4. Классификация категориальных данных. Деревья решений. Выбор атрибута разбиения в узле. Пример построения дерева решений алгоритмом CART
5. Алгоритм вероятностной классификации (метод Байеса). Решение простой задачи медицинской диагностики методом Байеса.
6. Нейросетевые алгоритмы классификации
7. Ассоциативные правила, их характеристики. Решение задачи поиска ассоциативных правил для небольшой базы транзакций.
8. Формальная постановка задачи кластеризации данных, цель кластеризации. Описание алгоритма k-means.
9. Алгоритм иерархической кластеризации.
10. Алгоритм CLOPE кластеризации категориальных данных, в т.ч. транзакций
11. Классификация текстовых данных. Метод Байеса для классификации текстов.
12. Модель распределённых вычислений MapReduce