

МИНОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ  
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Институт прикладной математики и компьютерных наук

УТВЕРЖДАЮ

Директор института прикладной  
математики и компьютерных наук

А.В. Замятин  
« 11 » ноября 2021 г.



## Введение в интеллектуальный анализ данных

### рабочая программа дисциплины

Закреплена за кафедрой	<i>теоретических основ информатики</i>
Учебный план	<i>01.03.02 Прикладная математика и информатика, профиль «Прикладная математика и информатика»</i>
Форма обучения	<i>очная</i>
Общая трудоёмкость	<i>3 з.е.</i>
Часов по учебному плану	<i>108</i>
в том числе:	
аудиторная контактная работа	<i>50,65</i>
самостоятельная работа	<i>57,35</i>
Вид(ы) контроля в семестрах	
<i>экзамен/зачет/зачет с оценкой</i>	<i>Семестр 5 – зачет с оценкой</i>

Программу составил:

д-р техн. наук, профессор,  
заведующий кафедрой теоретических основ информатики

  
А.В. Замятин

Рецензент:

д-р техн. наук, профессор,  
профессор кафедры теоретических основ информатики

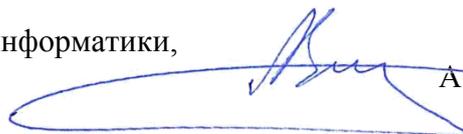
  
Ю.Л. Костюк

Рабочая программа дисциплины «Введение в интеллектуальный анализ данных» разработана в соответствии с самостоятельно устанавливаемым образовательным стандартом высшего образования – бакалавриат – федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский Томский государственный университет» по направлению подготовки 09.03.03 – Прикладная информатика (Утвержден Ученым советом НИ ТГУ, протокол от 27.10.2021 г. № 08).

Рабочая программа одобрена на заседании теоретических основ информатики

Протокол от 04 июня 2021 г. № 05

Заведующий кафедрой теоретических основ информатики,  
д-р техн. наук, профессор

  
А.В. Замятин

Рабочая программа одобрена на заседании учебно-методической комиссии института прикладной математики и компьютерных наук (УМК ИПМКН)

Протокол от 17 июня 2021 г. № 05

Председатель УМК ИПМКН,  
д-р техн. наук, профессор

  
С.П. Сущенко

### Цель освоения дисциплины

**Цель** – Получение знаний в области моделей и методов интеллектуального анализа данных в задачах поиска информации, обработки и анализа данных, а также приобретение навыков исследователя данных (data scientist) и разработчика математических моделей, методов и алгоритмов анализа данных.

### 1. Место дисциплины в структуре ОПОП

Дисциплина «Введение в интеллектуальный анализ данных» относится к вариативной части Блока 1 «Дисциплины», входит в модуль «Введение в искусственный интеллект».

Пререквизиты дисциплины: нет.

Постреквизиты дисциплины: «Нейронные сети», «Технологии высокопроизводительной обработки больших данных».

### 2. Компетенции и результаты обучения, формируемые в результате освоения дисциплины

Таблица 1.

Компетенция	Индикатор компетенции	Код и наименование результатов обучения (планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций)
ПК-3. Способен осуществлять научно-исследовательские и опытно-конструкторские разработки как при исследовании самостоятельных тем, так и разработки по тематике организации.	ИПК-3.1. Осуществляет проведение работ по обработке и анализу научно-технической информации и результатов исследований.	ОР-3.1.1. Знать основные методы научно-практического поиска в задачах интеллектуального анализа данных и других областях с использованием информационных технологий. ОР-3.1.2: Знать существующие методы и подходы к интеллектуальному анализу данных различной природы. ОР-3.1.3. Уметь применять существующие методы интеллектуального анализа данных, обоснованно адаптируя и модифицируя их с учетом особенностей задачи предметной области.

### 3. Структура и содержание дисциплины

#### 3.1. Структура и трудоемкость видов учебной работы по дисциплине

Общая трудоемкость дисциплины составляет 3 зачетные единицы, 108 часов.

Таблица 2.

Вид учебной работы	Трудоемкость в академических часах	
	5 семестр	всего
<b>Общая трудоемкость</b>	108	108
<b>Контактная работа:</b>	50,65	50,65
Лекции (Л):	32	32
Практики (ПЗ)	16	16
Лабораторные работы (ЛР)		
Семинары (СЗ)		
Групповые консультации	2	2
Индивидуальные консультации	0,4	0,4
Промежуточная аттестация	0,25	0,25
<b>Самостоятельная работа обучающегося:</b>	57,35	57,35
- изучение учебного материала	20	20

- подготовка к лабораторным работам	37,35	37,35
<b>Вид промежуточной аттестации (зачет, зачет с оценкой, экзамен)</b>	<b>Зачет с оценкой</b>	<b>Зачет с оценкой</b>

### 3.2. Содержание и трудоемкость разделов дисциплины

Таблица 3.

Код занятия	Наименование разделов и тем и их содержание	Вид учебной работы, занятий, контроля	Семестр	Часы в электронной форме	Всего (час.)	Литература	Код (ы) результата(ов) обучения
	<b>Раздел 1. Основные проблемы построения систем</b>		1		16,4	1	ОР-3.1.1, ОР-3.1.2 ОР-3.1.3.
1.1.	Актуальность, базовая терминология и тенденции развития. Основные задачи, этапы и классификация методов анализа данных.	Лекции	1		2		
1.2.	Предварительная обработка данных. Классификация.	Лекции	1		2		
1.3.	Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация.	лабораторные	1		2		
1.4.	Высокопроизводительная обработка данных. Программные среды для интеллектуального анализа данных.	лабораторные	1		2		
1.5	Форма СРС	СРС	1		8,4		
	Текущий контроль успеваемости		1				
	<b>Раздел 2. Предварительная обработка данных. Классификация.</b>		1		18	1, 2,3	ОР-3.1.1, ОР-3.1.2 ОР-3.1.3.
2.1	Основные методы и предварительная обработка данных.	Лекции	1		2		
2.2	Оптимизация признакового пространства без трансформации пространства признаков.	Лекции	1		2		
2.3	Контролируемая непараметрическая нейросетевая классификация.	лабораторные	1		2		
2.4	Классификация по методу машины опорных векторов. Деревья решений.	лабораторные	1		2		
	Форма СРС		1		10		
	Текущий контроль успеваемости		1				
	<b>Раздел 3. Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация</b>		1		18	1, 2,3	ОР-3.1.1, ОР-3.1.2 ОР-3.1.3.
3.1	Понятие регрессии. Основные этапы регрессионного анализа.	Лекции	1		2		
3.2	Описание алгоритма ассоциации.	Лекции	1		2		
3.3	Алгоритмы семейства «Априори». Алгоритм GSP.	лабораторные	1		2		
3.4	Обнаружение аномалий и методы визуализации.	лаборатор	1		2		

		ные					
	Форма СРС		1		10		
	Текущий контроль успеваемости		1				
	<b>Раздел 4. Высокопроизводительная обработка данных. Программные среды для интеллектуального анализа данных</b>		1		18	1, 2,3	ОП-3.1.1, ОП-3.1.2 ОП-3.1.3.
4.1	Принципы организации высокопроизводительных вычислений. SMP-системы.	Лекции	11		2		
4.2	Модели параллельных вычислений MPMD, SPMD.	Лекции	1		2		
4.3	Платформа программирования и выполнения распределённых вычислений Hadoop MapReduce, Mahout, Cassandra, Spark. Нереляционные базы данных HBase и язык NoSQL. Среда и язык программирования Python, R.	лабораторные	11		4		
	Форма СРС	СРС	11		10		
	Консультации в период теоретического обучения и промежуточной аттестации				3,6		
	Промежуточная аттестация	ЗаО	1		2,3		

#### 4. Образовательные технологии, учебно-методическое и информационное обеспечение для освоения дисциплины

Каждый студент реализует индивидуальный или групповой проект как последовательность лабораторных работ. Темы проектов имеют следующий шаблон:

1. Реализовать алгоритм анализа данных.
2. Предложить и реализовать технологии повышения производительности вычислений, выполняемых алгоритмом.

**Лабораторная работа №1.** Индивидуальное задание по теме «Анализ предметной области, формулировка целей и задач исследования. Извлечение и первичное сохранение данных».

**Цель работы** – научить студентов решать задачи анализа предметной области, ее адаптации для методов анализа данных с учетом принципиальных особенностей предметной области.

**Лабораторная работа №2.** Индивидуальное задание по теме «Предварительная обработка данных: очистка, интеграция, преобразование».

**Цель работы** – научить студентов решать задачи предварительной обработки данных, предполагающей трудоемкую процедуру очистки (исключение противоречий, случайных выбросов и помех, пропусков), интеграции (объединение данных из нескольких возможных источников в одном хранилище), преобразования (может включать агрегирование и сжатие данных, дискретизацию атрибутов и сокращение размерности и т.п.).

**Лабораторная работа №3.** Индивидуальное задание по теме «Содержательный анализ данных методами Data Mining».

**Цель работы** – научить студентов обоснованно применять базовые методы интеллектуального анализа данных, учитывая особенности как теоретического построения применяемых методов, так и выбранной предметной области.

**Лабораторная работа №4.** Индивидуальное задание по теме «Визуализация и интерпретация полученных результатов».

**Цель работы** – научить студентов выполнять визуализацию и интерпретация полученных результатов в виде, пригодном для принятия управленческих решений.

Самостоятельная работа студентов по предмету организуется в следующих формах:

- 1) самостоятельное изучение основного теоретического материала, ознакомление с дополнительной литературой, Интернет-ресурсами;
- 2) выполнение индивидуальных проектов, решение профессиональных задач из реальной предметной области.

В качестве учебно-методического обеспечения самостоятельной работы используется основная и дополнительная литература по предмету, Интернет-ресурсы, материал лекций, указания, выданные преподавателем при проведении лабораторных работ.

Темы для изучения	Формы выполнения заданий
Актуальность, базовая терминология и тенденции развития. Основные задачи, этапы и классификация методов анализа данных	Самостоятельное изучение темы, предложенной преподавателем. Самостоятельное выполнение лабораторной работы №1
Предварительная обработка данных. Классификация.	Самостоятельное изучение темы, предложенной преподавателем. Самостоятельное выполнение лабораторной работы №2
Регрессия. Ассоциация, последовательная ассоциация, аномалии и визуализация	Самостоятельное изучение темы, предложенной преподавателем. Самостоятельное выполнение

	лабораторной работы №3
Высокопроизводительная обработка данных. Программные среды для интеллектуального анализа данных	Самостоятельное изучение темы, предложенной преподавателем. Самостоятельное выполнение лабораторной работы №4

#### Примеры тем для самостоятельного изучения:

- Нейросетевые методы анализа данных, сверточные сети (convolution neural networks). глубинное обучение (deep learning).
- Методы интеллектуального анализа медиа (social media data mining).
- Методы машинного обучения в задачах финансовой аналитики.
- Методы машинного обучения в задачах ранней медицинской диагностики.
- Комбинирование моделей в анализе данных, бустинг.
- Метод анализа независимых компонент (independent component analysis).
- Методы визуализации данных высокой размерности.

Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения, характеризующих этапы формирования компетенций, и методические материалы, определяющие процедуры оценивания результатов обучения, приведены в Приложении 1 к рабочей программе «Фонд оценочных средств».

#### 4.1. Рекомендуемая литература и учебно-методическое обеспечение

№ п/п	Авторы / составители	Заглавие	Издательство	Год издания, количество страниц
Основная литература				
1.	Замятин А.В.	А.В. Введение в интеллектуальный анализ данных	Издательский Дом государственного университета	2016
2.	Mohamed Medhat Gaber, Frederic Stahl, João Bártolo Gomes.	Pocket Data Mining electronic resource : Big Data on Small Devices	Springer International Publishing : Imprint: Springer	2014
3.	Max Bramer	Principles of Data Mining electronic resource	Springer London : Imprint: Springer	2013
Дополнительная литература				
4.	Max Bramer	Principles of Data Mining electronic resource	Springer London: Imprint: Springer	2013, 440 с.
5.	Mohamed Medhat Gaber, Frederic Stahl, João Bártolo Gomes.Gaber, Mohamed Medhat.	Pocket Data Mining electronic resource : Big Data on Small Devices	Imprint: Springer	2014, 108 с.
6.	Миркин Б. Г.	Введение в анализ данных: учебник и практикум для бакалавриата и магистратуры: [для студентов вузов, обучающихся по инженерно-техническим, естественно-научным и экономическим	Москва, Юрайт	2015, 173 с.

		направлениям и специальностям]		
7.	Кулаичев А.П.	Методы и средства комплексного анализа данных: учебное пособие	Москва: Форум	2014, 511 с.

#### **4.2. Базы данных и информационно-справочные системы, в том числе зарубежные**

1. Электронная библиотека (репозиторий) ТГУ [Электронный ресурс] / Электронная библиотека (репозиторий) ТГУ: [сайт]. – [Томск, 2011–2016]. – URL: <http://vital.lib.tsu.ru/vital/access/manager/Index>.
2. Data Mining for Service electronic. Berlin, Heidelberg, Imprint: Springer, Springer eBooks VIII, 291 p. 2014 (edited by Katsutoshi Yada) [Электронный ресурс]. – Режим доступа: <http://dx.doi.org/10.1007/978-3-642-45252-9>
3. Data Mining for Geoinformatics electronic resource: Methods and Applications / /edited by Guido Cervone, Jessica Lin, Nigel Waters. New York, NY: : Springer New York : : Imprint: Springer, , 2014, 166 p. [Электронный ресурс]. – Режим доступа: <http://dx.doi.org/10.1007/978-1-4614-7669-6>

#### **4.3. Перечень лицензионного и программного обеспечения**

Средства и среды программирования C, C++, C#, Python, R-Studio, Rapid Miner, MS Azure.

#### **4.4. Оборудование и технические средства обучения**

Для реализации дисциплины необходимы лекционные аудитории и аудитории для проведения практических занятий. Специальные технические средства (проектор, компьютер и т.д.) требуются для демонстрации материала в рамках изучаемых разделов, проведения защиты проектов в конце семестра. Вся основная и дополнительная литература, необходимая для самостоятельной работы и подготовки к экзамену, имеется в научной библиотеке ТГУ.

#### **5. Методические указания обучающимся по освоению дисциплины**

Для успешного освоения дисциплины студенты должны посещать занятия, прорабатывать указанные материалы для самостоятельной работы студентов, выполнять лабораторные работы.

#### **6. Преподавательский состав, реализующий дисциплину**

Замятин Александр Владимирович, д-р техн. наук, профессор, заведующий кафедрой теоретических основ информатики ТГУ.

#### **7. Язык преподавания – русский язык.**