

Белорусский государственный университет

УТВЕРЖДАЮ

Проректор по учебной работе

А.Л. Толстик

06.01.2016

Регистрационный № УД-1854/уч.

ПРИКЛАДНОЙ ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

**Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:**

1-31 03 03 Прикладная математика (по направлениям)

2016 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 03 03-2013 и учебного плана G31-173/уч., 30.05.2013.

Составитель:

М. С. Абрамович, доцент кафедры математического моделирования и анализа данных Белорусского государственного университета, кандидат физико-математических наук, доцент.

Рекомендована к утверждению:

Кафедрой математического моделирования и анализа данных Белорусского государственного университета (протокол № 8 от 01 декабря 2015 г.);

Учебно-методической комиссией факультета прикладной математики и информатики Белорусского государственного университета (протокол № 3 от 17 декабря 2015 г.).



ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Прикладной интеллектуальный анализ данных – это дисциплина для изучения процесса поддержки принятия решений, основанная на поиске в данных скрытых закономерностей (шаблонов информации). Прикладной интеллектуальный анализ данных является мультидисциплинарной областью, возникшей и развивающейся на базе таких наук как математическая и прикладная статистика, распознавание образов, искусственный интеллект.

Изучение дисциплины «Прикладной интеллектуальный анализ данных» преследует следующие основные цели:

1) освоение теоретических основ прикладного интеллектуального статистического анализа данных;

2) формирование практических навыков для решения прикладных задач с использованием программного обеспечения интеллектуального анализа данных.

Теоретический материал курса «Прикладной интеллектуальный анализ данных» основывается на предположении, что данные являются разнотипными: непрерывными, ранжированными и номинальными.

В рамках *лекционного курса* рассматриваются задачи, алгоритмы и методы интеллектуального анализа данных, включающие разведочный анализ данных, кластеризацию, классификацию (методы машинного обучения), прогнозирование, анализ ассоциаций и последовательностей.

В первом разделе дисциплины рассматриваются методы и алгоритмы предварительного (разведочного анализа данных) В практических задачах данных вероятностная модель реальных данных, как правило, частично либо полностью неизвестна исследователю. В связи с этим возникает необходимость в предварительном анализе данных, который позволяет на основе компактного описания данных выявить их структуру и особенности и сформулировать предположительные утверждения (гипотезы) относительно модели данных. В случае отклонения моделей данных от гипотетических (например, в случае наличия аномальных наблюдений в данных) вместо классических методов интеллектуального анализа данных рассматриваются их робастные аналоги.

Второй раздел посвящен исследованию взаимосвязи признаков, измеренных в различных шкалах, и построению, анализу и применению регрессионных моделей статистических зависимостей (линейных и нелинейных) между признаками.

В третьем разделе рассматриваются иерархические и итеративные методы кластерного анализа. С помощью кластерного анализа средства интеллектуального анализа данных выделяют различные однородные группы данных.

Под машинным обучением (классификацией) понимается отнесение наблюдений к одному из заранее известных классов. Классификация является одним из основных разделов интеллектуального анализа данных и относится к стратегии обучения с учителем, которое также именуют контролиру-

емым или управляемым обучением. В рамках четвертого раздела дисциплины рассматриваются методы дискриминантного анализа, логистической регрессии, опорных векторов, деревьев классификации, случайного леса, а также ансамбли алгоритмов классификации.

В рамках пятого раздела рассматриваются методы и алгоритмы анализа ассоциаций и последовательностей. Одним из основных типов закономерностей, которые позволяют выявлять методы интеллектуального анализа данных, является ассоциация – последовательность. В анализе ассоциаций отыскиваются закономерности между связанными событиями в наборе данных, причем поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно. В случае если события рассматриваются во времени, то имеет место анализ последовательностей. В этом разделе рассматривается алгоритм поиска ассоциативных правил Apriori, алгоритмы анализа последовательностей AprioriALL и GSP, а также их применение для решения прикладных задач.

В рамках шестого раздела дисциплины изучаются методы и алгоритмы интеллектуального анализа данных: кластеризации, классификации, анализа ассоциаций и последовательностей для больших объемов данных.

Глубина и достоверность выводов зависит от выбора адекватной математической модели и соответствующих методов интеллектуального анализа данных, а также от соблюдения условий их корректного применения. Решение задач интеллектуального анализа данных требует применения соответствующего программного обеспечения. Эти принципы лежат в основе *лабораторного практикума*, который проводится в компьютерном классе и предполагает применение изучение методов интеллектуального анализа различных моделей данных и программного обеспечения.

Учебная дисциплина «Прикладной интеллектуальный анализ данных» относится к циклу дисциплин вузовского компонента и взаимосвязана с учебной дисциплиной «Теория вероятностей и математическая статистика». Методы, излагаемые в дисциплине «Прикладной интеллектуальный анализ данных», могут быть использованы при изучении ряда других дисциплин по специальности «Прикладная математика».

В результате изучения дисциплины студент должен:

-знать

- основные задачи и принципы интеллектуального анализа;
- разведочный (предварительный) анализ данных;
- анализ зависимости признаков;
- регрессионный анализ данных;
- методы машинного обучения: метод опорных векторов, деревья классификации, ансамбли алгоритмов классификации, метод случайного леса;
- анализ ассоциаций;
- анализ последовательностей;

- методы интеллектуального анализа для больших объемов данных;
- структуру и функциональные возможности программного обеспечения прикладного интеллектуального анализа данных;

- уметь

- анализировать структуру данных с целью выбора адекватной модели данных;
- решать задачи исследования зависимостей, кластеризации, классификации, прогнозирования, построения ассоциативных правил, анализа последовательностей, в том числе и для больших объемов данных;
- применять программное обеспечение для решения задач интеллектуального анализа данных;
- интерпретировать результаты применения методов интеллектуального анализа данных;

- владеть

- основными методами решения задач интеллектуального анализа данных;
- навыками применения статистического программного обеспечения интеллектуального анализа данных.

В соответствии с образовательным стандартом специальности 1-31 03 03 «Прикладная математика» (по направлениям) учебная программа предусматривает для изучения дисциплины всего 126 часа, из них 68 аудиторных часа, в том числе лекций – 34 часа, лабораторных занятий – 34 часов (3 курс, 6 семестр).

Форма текущей аттестации по учебной дисциплине – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Введение. Основные задачи и методы интеллектуального анализа данных. Примеры применения методов интеллектуального анализа данных.

Раздел I. Предварительный анализ данных

Тема 1.1. Общая характеристика задач предварительного (разведочно-го) анализа данных. Проверка выборки на однородность, случайность. Обнаружение аномальных наблюдений.

Тема 1.2. Робастные оценки параметров положения и масштаба. Визуализация данных. Определение закона распределения выборки.

Раздел 2. Анализ взаимосвязи и зависимости признаков

Тема 2.1. Меры связи для различных типов данных. Выборочные коэффициенты корреляции. Проверка значимости выборочных коэффициентов корреляции. Робастные оценки коэффициентов корреляции. Меры связи для ранжированных признаков. Таблицы сопряженности. Меры связи для номинальных признаков.

Тема 2.2. Множественная линейная регрессия. Проверка адекватности модели. Анализ остатков. Проблема мультиколлинеарности. Гребневая регрессия. Пошаговый регрессионный анализ. Нелинейная регрессия.

Раздел 3. Кластеризация данных

Тема 3.1. Иерархические методы кластерного анализа. Меры сходства. Методы объединения. Алгоритмы иерархического кластерного анализа.

Тема 3.2. Итеративные методы кластерного анализа. Алгоритм k-средних и его аналоги (k-медоидов, k-медиан, k-мод). Робастный метод k-средних.

Тема 3.3. Пространственно-временной кластерный анализ. Критерии глобальной кластеризации. Алгоритмы построения пространственно-временных кластеров. Современные тенденции в кластерном анализе.

Раздел 4. Методы машинного обучения

Тема 4.1. Дискриминантный анализ. Робастный дискриминантный анализ. Пошаговый дискриминантный анализ.

Тема 4.2. Логистическая регрессия. Логит и пробит модели. Оценивание параметров логистической регрессии. Отбор информативных признаков.

Тема 4.3. Метод опорных векторов. Линейная и нелинейная разделимость классов.

Тема 4.4. Деревья классификации. Выбор типа ветвления. Определения момента прекращения ветвлений. Кросс-проверка дерева классификации.

Тема 4.5. Ансамбли алгоритмов классификации. Алгоритмы бэггинг и бустинг. Метод случайного леса. Основные направления развития методов машинного обучения.

Раздел 5 . Анализ ассоциаций и последовательностей

Тема 5.1. Анализ ассоциаций. Основные понятия и определения анализа ассоциаций. Построение ассоциативных правил для поиска закономерностей. Алгоритм Apriori. Применение алгоритмов анализа ассоциаций.

Тема 5.2. Анализ последовательностей. Основные понятия и определения анализа последовательностей. Алгоритм AprioriALL. Алгоритм GSP. Применение алгоритмов анализа последовательностей.

Раздел 6. Методы интеллектуального анализа больших объемов данных

Тема 6.1. Проблемы интеллектуального анализа больших объемов данных. Методы и алгоритмы кластеризации больших объемов данных.

Тема 6.2. Методы и алгоритмы классификации больших объемов данных. Алгоритмы анализа ассоциаций и последовательностей больших объемов данных.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

№п/п	Название раздела, темы	Количество часов				Количество часов УСР	Форма контроля знаний
		Аудиторные					
		Лекции	Практ. и сем. занятия	Лаб. занятия	Иное		
	Введение	2		2			
	Основные задачи и методы интеллектуального анализа данных.	2		2			
1	Предварительный анализ данных	4		4			
1.1	Общая характеристика задач предварительного (разведочного) анализа данных.	2		2			
1.2	Робастные оценки параметров положения и масштаба.	2		2			Тест
2	Анализ взаимосвязи и зависимости признаков	4		4			
2.1	Меры связи для различных типов данных.	2		2			
2.2	Множественная линейная регрессия.	2		2			Контрольная работа №1
3	Кластеризация данных	6		6			
3.1	Иерархические методы кластерного анализа.	2		2			
3.2	Итеративные методы кластерного анализа.	2		2			Контрольная работа №2
3.3	Пространственно-временной кластерный анализ.	2		2			
4	Методы машинного обучения	10		10			
4.1	Дискриминантный анализ.	2		2			
4.2	Логистическая регрессия.	2		2			
4.3	Метод опорных векторов.	2		2			
4.4	Деревья классификации.	2		2			Контрольная работа

							№3
4.5	Ансамбли алгоритмов классификации.	2		2			
5	Анализ ассоциаций и последовательностей	4		4			
5.1	Анализ ассоциаций.	2		2			
5.2	Анализ ассоциаций последовательностей.	2		2			Тест
6	Методы интеллектуального анализа больших объемов данных	4		4			
6.1	Проблемы интеллектуального анализа больших объемов данных.	2		2			
6.2	Методы и алгоритмы классификации больших объемов данных.	2		2			Тест
ИТОГО		34		34			

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Рекомендуемая литература

Основная

1. Харин Ю.С., Малюгин В.И., Абрамович М.С. Математические и компьютерные основы статистического моделирования анализа данных. Минск: БГУ, 2008.
2. Харин Ю.С., Жук Е.Е. Математическая и прикладная статистика. Минск, БГУ, 2005.
3. Афифи С.А., Эйзен. Статистический анализ. Подход с использованием ЭВМ. Мир, 1982.
4. Айвазян С.А. и др. Прикладная статистика. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983
5. Айвазян С.А. и др. Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика, 1985.
6. Айвазян С.А. и др. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989.
7. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики, т. 1. Теория вероятностей и прикладная статистика, М.: ЮНИТИ, 2001.
8. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии: Учебное пособие.– М. : Изд-во МГТУ им. Н.Э. Баумана, 2009 .– 304 с.
9. Степанов Р.Г. Технология DATA MINING: Интеллектуальный анализ данных. – Казань: Изд-во Казанского государственного университета. 2008. – 58 с.
10. Harrington P. Machine Learning in Action. N-Y, *Manning*, 2012, 382 p.
11. Hastie T., Tibshirani R., Friedman J. H. The Elements of Statistical Learning, 3rded. N-Y, *Springer*, 2013, 764 p.
12. Воронцов К.В. Машинное обучение: курс лекций / К.В.Воронцов (<http://www.machinelearning.ru/wiki/index.php>).

Дополнительная

1. Харин Ю.С., Степанова М.Д. Практикум на ЭВМ по математической статистике. Мн.: Университетское, 1987.
2. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М.: ИНФРА-М, 1995.

3. Боровиков В.П. STATISTICA. Статистический анализ и обработка данных в среде Windows. М.: Филинь, 1997.
4. Бериков В.Б. Современные тенденции в кластерном анализе / В.Б. Бериков, Г.С. Лбов (<http://www.ict.edu.ru/ft/005638/62315e1-st02.pdf>)

ОРГАНИЗАЦИЯ УПРАВЛЯЕМОЙ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

Управляемая самостоятельная работа (УСР) студентов – это самостоятельная работа, выполняемая по заданию и при методическом руководстве преподавателя, а также контролируемая преподавателем на определенном этапе обучения. Целью УСР является целенаправленное обучение студентов основным навыкам и умению индивидуальной самостоятельной работы.

По второму, третьему и четвертому разделам учебного материала дисциплины проводятся контрольные работы, по остальным разделам контроль осуществляется в виде тестирования.

Результативность самостоятельной работы студентов во многом определяется наличием активных методов ее контроля:

- входной контроль знаний и умений студентов в начале изучения очередной дисциплины;
- текущий контроль, то есть регулярное отслеживание уровня усвоения материала на лекциях, практических и лабораторных занятиях;
- промежуточный контроль по окончании изучения раздела или модуля курса;
- самоконтроль, осуществляемый студентом в процессе изучения дисциплины при подготовке к контрольным мероприятиям;
- итоговый контроль по дисциплине в виде зачета или экзамена;
- контроль остаточных знаний и умений спустя определенное время после завершения изучения дисциплины.

РЕКОМЕНДАЦИИ ПО КОНТРОЛЮ КАЧЕСТВА УСВОЕНИЯ ЗНАНИЙ И ПРОВЕДЕНИЮ АТТЕСТАЦИИ

На лекционных занятиях по учебной дисциплине «Многомерный статистический анализ данных» рекомендуется использовать элементы проблемного обучения: проблемное изложение некоторых аспектов, использование частично-поискового метода.

Перечень рекомендуемых форм диагностики компетенций

Для аттестации обучающихся на соответствие их персональных достижений поэтапным и конечным требованиям образовательной программы создаются фонды оценочных средств, включающие типовые задания, контрольные работы и тесты. Оценочными средствами предусматривается оценка способности обучающихся к творческой деятельности, их готовность вести поиск решения новых задач, связанных с недостаточностью конкретных специальных знаний и отсутствием общепринятых алгоритмов.

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

- устная форма: собеседования, устные промежуточные и итоговый зачеты.
- письменная форма: тесты, контрольные опросы, контрольная работа.
- устно-письменная форма: отчеты по домашним практическим упражнениям с их устной защитой.

Контрольные мероприятия проводятся в соответствии с учебно-методической картой дисциплины. В случае неявки на контрольное мероприятие по уважительной причине студент вправе по согласованию с преподавателем выполнить его в дополнительное время. Для студентов, получивших неудовлетворительные оценки за контрольные мероприятия, либо не явившихся по неуважительной причине, по согласованию с преподавателем и с разрешения заведующего кафедрой мероприятие может быть проведено повторно.

Оценка текущей успеваемости рассчитывается как среднее оценок за каждую из письменных контрольных работ, оценки за отчеты по домашним практическим упражнениям и оценки за итоговый тест.

Итоговая аттестация предусматривает проведение зачета. При этом рекомендуется использовать оценивание успеваемости на основе модульно-рейтинговой системы.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Теория вероятностей и математическая статистика	Кафедра математического моделирования и анализа данных	нет	Оставить содержание учебной дисциплины без изменения, протокол № 8 от 01 декабря 2015 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на ____ / ____ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры дискретной математики и алгоритмики (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой

(ученая степень, звание)

(подпись)

(И.О. Фамилия)

УТВЕРЖДАЮ

Декан факультета

(ученая степень, звание)

(подпись)

(И.О.Фамилия)