Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе

_____ А.Л. Толстик

20.06.2016

Регистрационный № УД- 2071 / уч.

ВВЕДЕНИЕ В КОМПЬЮТЕРНЫЙ И ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Учебная программа учреждения высшего образования
по учебной дисциплине для специальностей:

1-31 03 05 Актуарная математика;

1-31 03 06 Экономическая кибернетика (по направлениям)
направления специальности
1-31 03 06-01 Экономическая кибернетика (математические методы и компьютерное моделирование в экономике);

1-98 01 01 Компьютерная безопасность (по направлениям)
направления специальности
1-98 01 01-01 Компьютерная безопасность (математические методы и программные системы)

2016 г.

Учебная программа составлена на основе образовательных стандартов высшего образования ОСВО 1-31 03 05-2013 и учебного плана G31-168/уч., 30.05.2013, ОСВО 1-31 03 06-2013 и учебного плана G31-166/уч., 30.05.2013, ОСВО 1-98 01 01-2013 и учебного плана Р98-138/уч., 30.05.2013.

Составители:

Н.Н. Труш, профессор кафедры теории вероятностей и математической статистики Белорусского государственного университета, доктор физико-математических наук,
Т.И. Воротницкая, ассистент кафедры теории вероятностей и математической статистики Белорусского государственного университета,
Т.В. Цеховая, доцент кафедры теории вероятностей и математической статистики Белорусского государственного университета, кандидат физико-математических наук.

Рекомендована к утверждению:

Кафедрой теории вероятностей и математической статистики Белорусского государственного университета
(протокол №13 от 24 мая 2016 г.);
Методической комиссией факультета прикладной математики и информатики Белорусского государственного университета
(протокол №6 от 24 мая 2016 г.)

Пояснительная записка

Цели и задачи дисциплины

Введение в компьютерный и интеллектуальный анализ данных – это дисциплина для изучения методов и алгоритмов, которые позволяют описывать, систематизировать большие объёмы данных, делать ранее неизвестные выводы и прогнозы.

Дисциплина «Введение в компьютерный и интеллектуальный анализ данных» имеет следующие основные цели:

1) изучение основных понятий теории вероятностей и математической статистики, описательной статистики и интеллектуального анализа данных;

2) формирование практических навыков решения прикладных задач с использованием свободно доступного современного программного обеспечения, основываясь на языке программирования R для статистической обработки данных и работы с графикой.

Принципы изложения материала и организации лабораторных занятий

Теоретический материал курса «Введение в компьютерный и интеллектуальный анализ данных» представляет собой методы описательной статистики и интеллектуального анализа данных (Data Mining).

В рамках лекционного курса вводятся основные понятия теории вероятностей, математической статистики и случайных процессов.

В разделе описательной статистики рассматриваются вариационные ряды, графическое изображение статистических данных, алгоритмы группировки массивов данных, приводятся формулы основных числовых характеристик.

На примере регрессионного анализа изучаются вопросы прогнозирования и исследования зависимости между двумя переменными. Для вычисления оценок параметров модели приводится метод наименьших квадратов.

Технологии Data Mining используются при исследовании больших объёмов данных для обнаружения скрытых и ранее неизвестных закономерностей и построения предсказательных моделей. В этом разделе изучаются методы поиска ассоциативных правил, поиск типовых последовательностей, классификация, кластеризация, нейронные сети, задачи исследования текстов и Web сайтов.

Поиск ассоциативных правил - это алгоритмы, которые позволяют находить связи, имеющиеся в реальных данных. Ключевым моментом является эффективность алгоритмов на больших объемах данных.

Кластерный анализ позволяет выполнять сбор данных, содержащих информацию о выборке объектов, и затем упорядочивать объекты в сравнительно однородные группы. В курсе лекций рассматриваются иерархические и неиерархические методы кластеризации, вводится понятие дендрограммы – древовидной диаграммы.

Приводятся постановки задач и других новых методов и алгоритмов интеллектуального анализа данных. Например, нейронные сети, Text Mining, Web Mining, классификация.

Решение задач компьютерного и интеллектуального анализа данных требует применения соответствующего программного обеспечения. Лабораторный практикум проводится в компьютерном классе и предполагает применение изученных методов для анализа данных с использованием современного программного обеспечения, имеющегося в свободном доступе, на основе языка программирования R.

Взаимосвязь с другими дисциплинами

Основой для изучения дисциплины «Введение в компьютерный и интеллектуальный анализ данных» являются курсы «Математический анализ», «Дискретная математика». Методы и алгоритмы, излагаемые при изучении дисциплины, являются полезными для лучшего усвоения студентами таких дисциплин как «Теория вероятностей и математическая статистика», «Компьютерный анализ данных с использованием языка R».

В результате изучения дисциплины студент должен:

-знать

- основы теории вероятностей и математической статистики, описательной статистики;
- методы и алгоритмы для анализа, обработки и систематизирования данных;

- уметь

- подбирать необходимую модель или алгоритм для решения конкретной задачи анализа данных;
- исследовать эффективность применения конкретного статистического метода для решения задачи;

-владеть

- знаниями основных подходов Data Mining;
- навыками выбора и обоснования модели при решении конкретной задачи;
- умениями компьютерной реализации основных методов решения задач.

Освоение курса должно обеспечить формирование академических компетенций, включающих знания и умения по изученным дисциплинам, профессиональных компетенций, включающих способность формулировать постановки и решать задачи в различных сферах профессиональной деятельности.

В соответствии с образовательным стандартом и учебным планом специальностей 1-31 03 05 «Актуарная математика», 1-31 03 06 «Экономическая кибернетика», 1-98 01 01 «Компьютерная безопасность» учебная программа предусматривает для изучения дисциплины всего 54 часа, из них 34 аудиторных часа, в том числе лекций – 18 часов, лабораторных занятий – 14 часов, 2 часа управляемой самостоятельной работы.

Дисциплина изучается на втором курсе в третьем семестре. Рекомендуемая форма текущей аттестации – зачет. Форма получения высшего образования – очная.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел I. Основные понятия теории вероятностей, математической статистики и случайных процессов

Пространство элементарных событий. Случайные события и их вероятность. Классическое и геометрическое определение вероятности. Условная вероятность и независимость событий. Формулы полной вероятности и Байеса. Случайные величины и функции распределения, их классификация. Плотность распределения. Основные характеристики случайных величин. Случайные процессы. Генерирование случайных последовательностей.

Раздел II. Описательная статистика

Тема 2.1. Вариационные ряды и графическое изображение статистических данных. Термин «статистика». Генеральная совокупность, выборка и их характеристики. Описательная и аналитическая статистика. Шкалы отношений. Дискретные и интервальные вариационные ряды. Табличное распределение частот. Алгоритм группировки массива данных. Полигон, кумулятивная кривая, кривая концентрации, гистограмма. Кривая Лоренца. Диаграммы. Пакет R.

Тема 2.2. Числовые характеристики одномерных признаков. Характеристики положения. Средние величины и степенные средние. Балансовое равенство (тождество). Среднее арифметическое, геометрическое, гармоническое и их свойства. Структурные средние. Мода, медиана, квантили, квартили. Ящик с усами. Робастные статистики для оценки среднего арифметического, усеченного среднего.

Тема 2.3. Характеристики рассеяния и формы распределения. Вариационный размах. Среднее линейное отклонение. Среднеквадратическое отклонение. Дисперсия. Основные свойства дисперсии. Коэффициент вариации. Коэффициент вариации по размаху. Коэффициент вариации по среднему линейному отклонению. Квартильное отклонение. Характеристики формы распределения вариационного ряда. Коэффициент асимметрии. Коэффициент эксцесса. Моменты распределения вариационного ряда. Сравнение эмпирического и теоретического распределений вариационных рядов. Оценивание параметров распределений по выборке. Доверительные интервалы. Линейная регрессия.

Тема 2.4. Виды связи между явлениями и компьютерный анализ данных. Виды связи между явлениями, двумерные частотные распределения и маргинальные распределения. Среднее. Дисперсия. Ковариация. Виды за-

висимостей. Парный линейный коэффициент корреляции Пирсона и его свойства. Другие коэффициенты корреляции.

Раздел III. Интеллектуальный анализ данных

Тема 3.1. Введение в Data Mining. Что такое Data Mining? Методы и стадии интеллектуального анализа данных. Задачи и сферы применения Data Mining. Процесс открытия знаний. Отличие Data Mining от других методов анализа данных.

Тема 3.2. Методы поиска ассоциативных правил. Задача анализа корзины покупок. Таблица информации. Ассоциативное правило. Поддержка и достоверность ассоциативного правила. Типы ассоциативных правил. Бинарные ассоциативные правила. Количественные ассоциативные правила. Размерность исследуемых данных. Одномерные и многомерные ассоциативные правила. Степень абстракции исследуемых данных. Одноуровневые и многоуровневые ассоциативные правила. Примеры. Поиск бинарных ассоциативных правил – общая постановка задачи. Поддержка и достоверность как меры ассоциативных правил. Сильные ассоциативные правила. Другие меры оценки ассоциативных правил. Наборы частые. Алгоритм наивный нахождения наборов частых. Основные алгоритмы поиска ассоциативных правил. Алгоритм Apriori. Генерация наборов кандидатов в наборы частые различной размерности. Пример для иллюстрации работы алгоритма Apriori. Свойство монотонности меры поддержки. Генерирование частных наборов. Генерирование правил. Разновидности алгоритма Apriori. Идея алгоритма FP Growth.

Тема 3.3. Поиск типовых последовательностей. Задача обнаружения шаблонов последовательностей. Выражения последовательности, длина и размер последовательности, максимальная последовательность. Нахождение частых эпизодов в последовательности. Анализ временных рядов. Основные определения и понятия. Основной алгоритм GSP состоящий из 5 шагов: сортировка, нахождение событий частых, нахождение всех частых последовательности и максимальных последовательностей. Исследование типовых последовательностей. Алгоритм Prefix Span. Области применения.

Тема 3.4. Другие методы интеллектуального анализа данных. Нейронные сети. Модели нейронных сетей. Искусственный нейрон. Задачи Data Mining решаемые с помощью нейронных сетей. Двухслойный персептрон. Пример решения задачи. Методы классификации и прогнозирование. Анализ текстовой информации. Визуальный анализ данных. Кластеризация. Иерархические и итеративные алгоритмы

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		
1	**Основные понятия теории вероятностей, математической статистики и случайных процессов**	**2**					
2	**Описательная статистика**	**8**		**8**			
2.1	Вариационные ряды и графическое изображение статистик	2		2			отчет по лабораторной работе
2.2	Числовые характеристики одномерных признаков	2		2			отчет по лабораторной работе
2.3	Характеристики рассеяния и формы распределения	2		2			отчет по лабораторной работе
2.4	Виды связи между явлениями и компьютерный анализ данных	2		2			отчет по лабораторной работе
3	**Интеллектуальный анализ данных**	**8**		**6**		**2**	
3.1	Введение в Data Mining	2		2			отчет по лабораторной работе
3.2	Методы поиска ассоциативных правил	2		2			отчет по лабораторной работе
3.3	Поиск типовых последовательностей	2		2			отчет по лабораторной работе
3.4	Другие методы интеллектуального анализа данных	2				2	Коллоквиум, завершающий отчет по лабораторному практикуму
ИТОГО		**18**		**14**		**2**	

ИНФОМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Рекомендуемая литература

Основная

1. Феллер В. Введение в теорию вероятностей и её приложения / В. Феллер – Т. 1, 2, М. – 1967.
2. Хацкевич Г.А. Статистика. Описательный подход / Г.А. Хацкевич. – Минск: НИУП. – 2002.
3. А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP

Дополнительная

4. Елисеева И.И. Общая теория статистики / И.И. Елисеева, М.М. Юзбашев. – М. – 1996.
5. Тюрин Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров
6. Torgo L. Data Mining with R: learning by case studies / L. Torgo - LIACC-FEP, University of Porto. – 2003.

Рекомендации по контролю качества усвоения знаний и проведению аттестации

На лекционных занятиях по учебной дисциплине «Введение в компьютерный и интеллектуальный анализ данных» рекомендуется использовать частично-поисковый метод.

Перечни используемых средств диагностики результатов учебной деятельности

Для аттестации обучающихся на соответствие их персональных достижений поэтапным и конечным требованиям образовательной программы создаются фонды оценочных средств, включающие лабораторные работы и коллоквиум. 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: устный итоговый зачет;

– письменная форма: коллоквиум, отчет по лабораторным работам.

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.

Тематика задания на управляемой самостоятельной работе «Алгоритм Apriori поиска ассоциативных правил». Сформировать файл, описывающий некоторую потребительскую корзину. Вывести информационную (бинарную) матрицу, описывающую транзакции созданного файла данных. Написать программу (скрипт) на языке R, осуществляющую извлечение правил и расчет основных параметров этих правил, исходя из сформированного файла.

Оценка текущей успеваемости рассчитывается как среднее оценок за коллоквиум и оценки за заключительный отчет по лабораторному практикуму.

Текущая аттестация предусматривает проведение зачета.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

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 теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения, протокол №13 от 24 мая 2016 г.
Компьютерный анализ данных с использованием языка R	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения, протокол №13 от 24 мая 2016 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на _2017/2018 учебный год

№№ Пп	Дополнения и изменения	Основание
	Дополнений и изменений нет	

Учебная программа пересмотрена и одобрена на заседании кафедры теории вероятностей и математической статистики (протокол № 17 от 27.06. 2017 г.)

Заведующий кафедрой
доктор физ.-мат. наук,
профессор _____ _Н.Н.Трущ_____
(ученая степень, звание) (подпись) (И.О. Фамилия)

УТВЕРЖДАЮ
Декан факультета
кандидат физ.-мат. наук,
доцент _____ П.А.Мандрик_____
(ученая степень, звание) (подпись) (И.О.Фамилия)