

1к, 6к, 8к

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ


УТВЕРЖДАЮ
Проректор по учебной работе
А.Л. Толстик
_____ 2016 г.
Регистрационный № УД- 1776 /уч.

СТАТИСТИЧЕСКИЙ АНАЛИЗ: ЯЗЫК R

Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:
1-31 03 01 Математика (по направлениям)
(1-31 03 01-05 Математика (информационные технологии))

2015 г.

Учебная программа составлена на основе образовательного стандарта ОСВО 1-31 03 08-2012 (30.08.2012), учебного плана специальности Математика (по направлениям) направление специальности Математика (информационные технологии) № G31-105/уч. 30.05.2012.

СОСТАВИТЕЛИ:

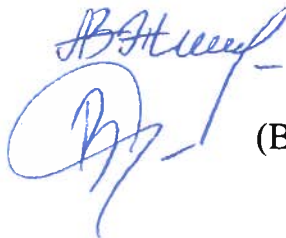
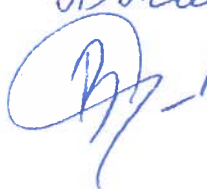
В.С. Романчик, заведующий кафедрой веб-технологий и компьютерного моделирования механико-математического факультета Белорусского государственного университета, кандидат физико-математических наук, доцент;

А.В. Жерело, ведущий научный сотрудник Института математики НАН Беларуси, кандидат физико-математических наук.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой веб-технологий и компьютерного моделирования
(протокол № 10 от 14.05.2015г.);

Учебно-методической комиссией механико-математического факультета
Белорусского государственного университета
(протокол № 6 от 26.05.2015г.).

 (А.В. Жерело)
 (В.С. Романчик)

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Специальный курс «Статистический Анализ: Язык R» разработан для студентов IV курса очной формы обучения специальности 1-31 03 01 Математика (по направлениям) (направление 1-31 03 01-05 Математика (информационные технологии)) механико-математического факультета Белорусского государственного университета.

Рост объемов передаваемой и хранимой информации и влияние ее на протекающие в мире процессы привел к необходимости постоянной обработки данных и их активного использования в процессе принятия решений. Это существенно повысило интерес к статистическим методам анализа данных с использованием специальных программных решений, известным на данный момент как Machine Learning. Данный специальный курс посвящен основам Machine Learning и основан на сочетании базовых понятий статистического анализа и возможностей языка R. Язык R выбран по причине его широкой распространенности в сфере Machine Learning, открытости его исходных кодов и отсутствия ограничений при его некоммерческом распространении.

Курс имеет цели:

- дать студентам основные знания, навыки , принципы и подходы, применяемые при статистическом анализе данных.
- научить студентов работе в среде R и R-Studio и созданию приложений с их использованием.
- подготовить студента к самостоятельной деятельности в области Machine Learning
-
- привить умения в установке и настройке программного обеспечения, умения использовать на начальном профессиональном уровне популярные средства автоматизации тестирования, формировать и выполнять автоматизированные тесты.
- курс предназначен для студентов технических вузов, а также для дополнительного образования специалистов.

Задачей курса является подготовка студента к самостоятельному использованию языка R для решения задач Machine Learning.

По окончании курса студенты будут знать:

- базовые определения в области теории вероятностей и статистического анализа;
- особенности подготовки данных и основные виды применяемых статистических методов;
- базовые конструкции языка R;

уметь:

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Тема 1. Работа с данными в R

Знакомство с R. Основные синтаксические конструкции. Базовые типы и их преобразование. Контейнерные типы. Ввод-вывод данных. Общая структура программы.

Тема 2. Первичная обработка данных

Понятие случайной величины и ее распределения. Основные распределения. Оценка основных параметров случайной величины средствами R. Стандартизация (нормировка) данных. Простейшие методы очистки и пополнения данных. Визуализация результатов: диаграммы рассеяния и размаха, гистограммы, плотности и т.д.

Тема 3. Регрессионные модели

Загрузка и установка дополнительных модулей в R. Знакомство с парной линейной регрессией. Оценка коэффициентов и общей точности модели средствами R.

Тема 4. Множественная линейная регрессия

Понятие множественной линейной регрессии. Выбор переменных в множественной линейной регрессии. Качественные предикторы в множественной линейной регрессии. Интерпретация коэффициентов в парной и множественной линейной регрессии.

Тема 5. Классификационные модели

Логистическая регрессия. Точность и полнота модели. Confusion Matrix. F-мера. ROC-кривая.

Тема 6. Дискриминантный анализ

Линейный дискриминантный анализ для случая двух классов. Линейный дискриминантный анализ для случая многих классов. Квадратичный дискриминант и особенности его применения.

Тема 7. Обучающее и тестовое множество. Перекрестная проверка

Качество данных. Кросс-валидация. Что такое «недообучение» и «переобучение». Регуляризация.

Тема 8. Сокращение размерности

Метод главных компонент как один из подходов к сокращению размерности пространства параметров.

Тема 9. Элементы кластерного анализа

Кластеризация данных. Основные виды используемых расстояний. Использование R для кластерного анализа.

Учебно-методическая карта учебной дисциплины

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов				Количество часов УРС	Литература	Формы контроля знаний
		лекции	семинарские занятия	лабораторные занятия	Иное			
1.	Работа с данными в R	2					[1]	Индивидуальные задания в форме программных проектов
2.	Первичная обработка данных	2					[1,2,5]	
3.	Регрессионные модели	2	2				[1-5]	
4.	Множественная линейная регрессия	2		2			[1-5]	Опрос
5.	Классификационные модели	2	2				[1,3-5]	
6	Дискриминантный анализ	2		2			[1,3-5]	Опрос
7	Обучающее и тестовое множество. Перекрестная проверка	2	2			2	[1-5]	Индивидуальные задания в форме программных проектов
8	Сокращение размерности	2		2			[1-5]	
9	Элементы кластерного анализа	2		2			[1,5]	Опрос
	Итого	18	6	8		2		

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Литература

Основная

1. G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R. Springer Science+Business Media, New York, 2013
2. Айвазян С.А. и др. Прикладная статистика: Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983.
3. Айвазян С. А. и др. Прикладная статистика: Исследование зависимостей. М.: Финансы и статистика, 1985.
4. Кашьяп Р.Л., Рао А.Р. Построение динамических стохастических моделей по экспериментальным данным. М.: Наука, 1983.
5. Max Kuhn, Kjell Johnson. Applied Predictive Modeling. Springer, New York, 2013

Дополнительная

1. <https://www.r-project.org>
2. <https://www.rstudio.org>
3. <http://www.machinelearning.ru>

СРЕДСТВА ДИАГНОСТИКИ РЕЗУЛЬТАТОВ УЧЕБНОЙ ДЕЯТЕЛЬНОСТИ

Рекомендуются следующие формы диагностики компетенций:

1. Опрос.
2. Индивидуальные задания в форме программных проектов.
3. Зачет.

ПЕРЕЧЕНЬ ЗАДАНИЙ И КОНТРОЛЬНЫХ МЕРОПРИЯТИЙ УПРАВЛЯЕМОЙ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ

Контрольные мероприятия УСР по дисциплине «Статистический Анализ: Язык R» проводятся преподавателем, как правило, во время аудиторных занятий и осуществляется в виде.

Полученные студентом количественные результаты УСР учитываются как составная часть итоговой оценки по дисциплине в рамках рейтинговой системы.

Полученные студентом ~~количественные~~ ~~результаты~~ УСР учитываются как составная часть ~~итоговой оценки по дисциплине~~ в рамках рейтинговой системы.

Результативность самостоятельной работы студентов во многом определяется наличием активных методов ее контроля. Существуют следующие виды контроля:

- входной контроль знаний и умений студентов при 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.

Примерный перечень заданий для программных проектов

Примерный перечень заданий для программных проектов

1. Создание проекта в среде R Studio.
2. Загрузка и преобразование данных с последующей визуализацией.
3. Создание проекта, реализующего этапы загрузки, первичной обработки данных, применения к данным элементов дискриминантного и кластерного анализа.

**ПРОТОКОЛ
СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ ПО ИЗУЧАЕМОЙ
УЧЕБНОЙ ДИСЦИПЛИНЕ С ДРУГИМИ ДИСЦИПЛИНАМИ
СПЕЦИАЛЬНОСТИ**

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)
Web-программирование	Веб-технологий и компьютерного моделирования	нет	Вносить изменения не требуется. Протокол №10, 14.05.2015
Теория вероятностей	Математической кибернетики	нет	Вносить изменения не требуется. Протокол №10, 14.05.2015
Математическая статистика	Математической кибернетики	нет	Вносить изменения не требуется. Протокол №10, 14.05.2015

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ
на 2016/2017 учебный год

№ пп	Дополнения и изменения	Основание
	В пояснительную записку внести требования к компетенциям специалиста: академические – АК-1, АК-2, АК-3, АК-4, АК-5, АК-6, АК-7, АК-8, АК-9; социально-личностные – СЛК-1, СЛК-2, СЛК-3, СЛК-4, СЛК-5, СЛК-6; профессиональные – ПК-1, ПК-2, ПК-3, ПК-4, ПК-5, ПК-6, ПК-7, ПК-8, ПК-9, ПК-10, ПК-11, ПК-12, ПК-13, ПК-14, ПК-15, ПК-16, ПК-17, ПК-18, ПК-19, ПК-20, ПК-21, ПК-22, ПК-23, ПК-24, ПК-25, ПК-26, ПК-27.	

Учебная программа пересмотрена и одобрена на заседании кафедры веб-технологий и компьютерного моделирования (протокол № 9 от 27.05.2016г.)

Заведующий кафедрой

канд. физ.-мат. наук, доцент
(ученая степень, ученое звание)


(подпись)

В.С. Романчик
(И.О.Фамилия)

УТВЕРЖДАЮ

Зам. Декан факультета

канд. физ.-мат. наук, доцент
(ученая степень, ученое звание)


(подпись)

Д.Г. Медведев
(И.О.Фамилия)