Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе

_____ А.Л. Толстик

_____

Регистрационный № УД- 1606 / уч.

МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА МНОГОМЕРНЫХ ДАННЫХ

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

1-31 81 12 Прикладной компьютерный анализ данных

2016 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 81 12-2015 и учебного плана G31-198/уч., 11.05.2015.

СОСТАВИТЕЛЬ:

А.Ю. Харин, доцент кафедры теории вероятностей и математической статистики Белорусского государственного университета, кандидат физико-математических наук, доцент.

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой теории вероятностей и математической статистики Белорусского государственного университета
(протокол № 12 от 16 июня 2015 г.)

Научно-методическим советом Белорусского государственного университета (протокол № 6 от 29 июня 2015г.).

Рецензенты:
Б.А. Залесский, заведующий лабораторией обработки и распознавания изображений Объединенного института проблем информатики Национальной академии наук Беларуси, доктор физико-математических наук;
Е.Н. Орлова, доцент кафедры математического моделирования и анализа данных Белорусского государственного университет, кандидат физико-математических наук, доцент

Пояснительная записка

Цели и задачи дисциплины

Методы статистического анализа многомерных данных – это дисциплина для изучения вероятностных моделей, методов и алгоритмов статистического исследования данных, имеющих многомерную структуру.

Дисциплина «Методы статистического анализа многомерных данных» имеет следующие основные цели:

1) изучение теоретических основ – математических моделей и методов статистического анализа данных многомерной структуры;

2) формирование практических навыков решения прикладных задач анализа многомерных данных с использованием свободно доступного современного программного обеспечения в области статистического анализа.

Принципы изложения материала и организации лабораторных занятий

Теоретический материал курса «Методы статистического анализа многомерных данных» представляет собой классические методы анализа данных, имеющих многомерную структуру.

В рамках *лекционного курса* рассматриваются отличительные особенности многомерных данных, основные математические модели и методы их анализа. Изучаемые методы основываются на использовании моделей и методов теории вероятностей и математической статистики.

Существенное отличие многомерных данных от скалярных состоит в предположении возможной зависимости компонент наблюдений, что, с одной стороны, усложняет анализ – требуются специально разработанные методы; с другой стороны, такая модель часто более адекватна на практике, и без ее использования исследование причинно-следственных связей проблематично.

Основные характеристики многомерных данных должны быть рассмотрены в начале курса. Кроме того, необходимо подробное ознакомление с многомерным нормальным распределением как с наиболее часто используемым при решении практических задач, методами оценивания его параметров и их свойствами.

Метод главных компонент позволяет построить несколько основных переменных (компонент), значения которых в основном объясняют свойства конкретного наблюдения.

Факторный анализ позволяет сформировать факторы, несущие основную информацию о наблюдениях, и используемые для дальнейшего анализа.

Дискриминантный и кластер-анализ – статистические методы классификации наблюдений; задачи классификации возникают в большинстве приложений. В зависимости от доступной информации разработаны различные методы решения этих задач. Одним из современных методов является многомерное шкалирование.

Для решения задачи проверки гипотез о равенстве векторов математических ожиданий (средних значений) используется классический метод MANOVA.

При анализе данных высокой размерности возникает модель регрессионной зависимости с большим числом предикторов, классический анализ которой приводит к неадекватным результатам, поэтому необходима соответствующая корректировка методов.

Решение задач статистического анализа многомерных данных требует применения соответствующего программного обеспечения. Этот принцип лежит в основе *лабораторного практикума,* который проводится в компьютерном классе и предполагает применение изученных методов анализа для различных моделей данных с использованием современного программного обеспечения, имеющегося в свободном доступе.

Взаимосвязь с другими дисциплинами

Основой для изучения дисциплины «Методы статистического анализа многомерных данных» является курс первой ступени «Теория вероятностей и математическая статистика» (или «Высшая математика» с включением соответствующих разделов). Кроме того, лабораторный практикум предполагает дополнение курсом «Компьютерный анализ данных с использованием языка R», преподаваемым параллельно. Дисциплина «Методы статистического анализа многомерных данных» способствует успешному освоению курса «Методы статистического анализа сложных данных», а результаты ее изучения используется при прохождении практики и написании магистерских работ.

В результате изучения дисциплины студент должен:

-знать

- основные вероятностные модели, применяемые для статистического анализа многомерных данных;
- статистические методы анализа многомерных данных;
- правила применения методов статистического анализа многомерных данных, их свойства;

- уметь

- подбирать подходящую модель для решения конкретной задачи статистического анализа многомерных данных;
- исследовать потенциальную эффективность применения конкретного статистического метода для решения задачи анализа многомерных данных;

-владеть

- знаниями основных моделей статистического анализа многомерных данных;
- навыками выбора и обоснования модели при решении конкретной задачи;

- умениями компьютерной реализации основных методов решения задач.

В соответствии с образовательным стандартом и учебным планом специальности 1-31 81 12 «Прикладной компьютерный анализ данных» учебная программа предусматривает для изучения дисциплины всего 172 часа, из них 68 аудиторных часов, в том числе лекций – 34 часа, лабораторных занятий – 34 часа (магистратура, 1 семестр).

Форма текущей аттестации по учебной дисциплине – зачет и экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Основные понятия и определения многомерного статистического анализа

Тема 1.1. **Особенности многомерного анализа.** Введение. Примеры задач, показывающие принципиальные отличия от задач анализа скалярных данных.

Тема 1.2. Основные понятия. Основные понятия и определения, содержательный смысл характеристик.

Раздел 2. Многомерное нормальное распределение и его свойства

Тема 2.1. Определения и простейшие свойства. Определение многомерного нормального распределения, характеристики. Условные распределения. Линейные преобразования гауссовских случайных векторов.

Тема 2.2. Функция регрессии. Функция регрессии, ее оптимальные свойства. Частный и множественный коэффициенты корреляции.

Тема 2.3. Статистическое оценивание параметров. Статистическое оценивание параметров многомерного нормального распределения. Распределение Уишарта.

Раздел 3. Метод главных компонент

Тема 3.1. Построение главных компонент. Метод главных компонент. Построение, свойства, геометрическая интерпретация.

Тема 3.2. Выборочные главные компоненты. Главные компоненты для стандартизованных переменных. Выборочные главные компоненты. Построение доверительных интервалов.

Тема 3.3. Аппроксимация. Графическое представление результатов. Аппроксимация с использованием главных компонент. Связь с ортогональной регрессией.

Раздел 4. Факторный анализ

Тема 4.1. Ортогональная модель. Факторный анализ. Ортогональная модель. Анализ ковариационной структуры факторной модели. Неоднозначность факторных нагрузок.

Тема 4.2. Методы построения факторов. Использование метода главных компонент в факторном анализе. Использование метода максимального правдоподобия в факторном анализе.

Тема 4.3. Вращение факторов. Факторный анализ – вращение факторов.

Раздел 5. Дискриминантный и кластер-анализ

Тема 5.1. Решающие правила и их построение. Дискриминантный анализ. Ожидаемая стоимость ошибки. Построение решающего правила. Минимизация полной вероятности ошибки. Классификация наблюдений из нормального распределения (случай общей ковариационной матрицы). Дискриминантная функция Фишера.

Тема 5.2. Методы кластер-анализа. Кластерный анализ. Матрица различий. Кластеризация методом *K* средних. Разделение вокруг медоидов (PAM-алгоритм).

Тема 5.3. Графическая интерпретация. Графическое представление результатов кластер-анализа.

Раздел 6. Проверка гипотез о равенстве векторов математических ожиданий

Тема 6.1. Алгоритм MANOVA. Сравнение векторов математических ожиданий – многомерный анализ дисперсий (MANOVA).

Раздел 7. Регрессия с большим числом предикторов

Тема 7.1. Методы анализа. Регрессия с большим числом предикторов – методы лассо и регуляризации.

Тема 7.2. Многомерное шкалирование. Многомерное шкалирование. Заключение.

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 и определения многомерного статистического анализа**	**4**		**4**			
1.1	Особенности многомерного анализа	2		2			
1.2	Основные понятия	2		2			Отчет по заданию с устной защитой
2	**Многомерное нормальное распределение и его свойства**	**6**		**6**			
2.1	Определения и простейшие свойства	2		2			
2.2	Функция регрессии	2		2			
2.3	Статистическое оценивание параметров	2		2			Контрольная работа 1
3	**Метод главных компонент**	**6**		**6**			
3.1	Построение главных компонент	2		2			
3.2	Выборочные главные компоненты	2		2			
3.3	Аппроксимация	2		2			Коллоквиум
4.	**Факторный анализ**	**6**		**6**			
4.1	Ортогональная модель	2		2			
4.2	Методы построения факторов	2		2			
4.3	Вращение факторов	2		2			Отчет по заданию с устной защитой
5	**Дискриминантный и кластер-анализ**	**6**		**6**			
5.1	Решающие правила и их построение	2		2			
5.2	Методы кластер-анализа	2		2			
5.3	Графическая интерпретация	2		2			Кон-

							трольная работа 2
6	**Проверка гипотез о равенстве векторов математических ожиданий**	**2**		**2**			
6.1	Алгоритм MANOVA	2		2			Отчет по заданию с устной защитой
7	**Регрессия с большим числом предикторов**	**4**		**4**			
7.1	Методы анализа	2		2			Отчет по заданию с устной защитой
7.2	Многомерное шкалирование	2		2			
ИТОГО		**34**		**34**			

ИНФОМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Рекомендуемая литература

Основная

1. Харин Ю.С. Теория вероятностей, математическая и прикладная статистика / Ю.С. Харин, Н.М. Зуев, Е.Е. Жук. – Минск : БГУ. – 2011. – 463 с.
2. Харин Ю.С. Математические и компьютерные основы статистического моделирования и анализа данных / Ю.С. Харин, В.И. Малюгин, М.С. Абрамович. – Минск : БГУ. – 2008. – 455 с.
3. Лобач В.И. Имитационное и статистическое моделирование / В.И. Лобач, В.П. Кирлица, В.И. Малюгин, С.Н. Сталевская. – Минск : БГУ. – 2004. – 189 с.

Дополнительная

1. Харин Ю.С. Эконометрическое моделирование / Ю.С. Харин, В.И. Малюгин, А.Ю. Харин. – Минск : БГУ. – 2004. – 313 с.
2. Johnson R.A. Applied multivariate statistical analysis / R.A. Johnson, D.W. Wichern. – New York : Pearson. – 2007. – 800 p.
3. Everitt B. An introduction to applied multivariate analysis with R / B. Everitt, T. Hothorn. – New York: Springer. – 2011. – 274 p.
4. Hastie T. The elements of statistical learning / T. Hastie, R. Tibshirani, J. Friedman. – New York : Springer. – 2008. – 758 p.

Рекомендации по контролю качества усвоения знаний и проведению аттестации

На лекционных занятиях по учебной дисциплине «Методы статистического анализа многомерных данных» рекомендуется использовать элементы проблемного обучения: проблемное изложение некоторых аспектов, использование частично-поискового метода.

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, экзамен;

– 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 и экзамена. 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 теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения, протокол № __ от __октября 2015 г.
Компьютерный анализ данных с использованием языка R	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения, протокол № __ от __октября 2015 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на _____/_____ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры теории вероятностей и математической статистики (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой

_____ _____ _____

(ученая степень, звание) (подпись) (И.О. Фамилия)

УТВЕРЖДАЮ

Декан факультета

_____ _____ _____

(ученая степень, звание) (подпись) (И.О.Фамилия)