Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе

_____ А.Л. Толстик

_____ 2016

Регистрационный № УД-1381 / уч.

МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА СЛОЖНЫХ 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 № 3 от 20 октября 2015 г.)

Научно-методическим советом Белорусского государственного университета
(протокол № 2 от 11 ноября 2015 г.).

Пояснительная записка

Цели и задачи дисциплины

Методы статистического анализа сложных данных – это дисциплина для изучения современных вероятностных моделей, методов и алгоритмов статистического исследования данных, относящихся в категорию сложных.

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

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

2) формирование практических навыков решения прикладных задач анализа сложных 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 сложных данных» представляет собой современные методы анализа данных, имеющих сложную структуру: данные большой размерности, потоки данных, длительные (лонгитюдные) данные, данные с искажениями.

В рамках *лекционного курса* рассматриваются различные классы сложных данных, основные математические модели и методы их анализа. Изучаемые методы основываются на использовании современных моделей и методов теории вероятностей и прикладной статистики.

Основные классы сложных данных должны быть рассмотрены в начале курса. Кроме того, необходимо подробное ознакомление с классификацией методов прикладной статистики для различных классов сложных данных.

Байесовские методы анализа данных позволяют учесть априорную информацию об исследуемом явлении или процессе и обеспечивают достаточную точность статистических выводов, основанных на статистических данных малого объема, когда точность классических методов является неудовлетворительной.

Последовательный статистический анализ – подход, позволяющий в среднем экономить число необходимых наблюдений при обеспечении заданных малых значений вероятностей ошибочных решений. Эта особенность обусловлена тем фактом, что момент остановки процесса наблюдений является случайным и зависит от самих наблюдений.

Потребность в анализе неполных данных возникает во многих приложениях. В зависимости от модели пропусков используются различные методы анализа таких данных.

Анализ длительных (лонгитюдных), в том числе так называемых панельных данных позволяет выявить закономерности развития и связи между наблюдаемыми группами объектов. Здесь используются различные современные вероятностные модели, поскольку данные характеризуются часто большой размерностью и малым числом временных отсчетов.

Полупараметрические методы регрессионного анализа, использующие сплайны, позволяют уйти от жесткого параметрического вида модельных предположений, описывающих данные.

Непараметрические модели приходится использовать в ситуациях, когда нет информации о виде функции регрессии.

В данных возможны искажения различных типов: «выбросы», искажения структуры зависимости, функциональные искажения распределений вероятностей, задаваемые окрестностями в различных пространствах. Для анализа данных с такими особенностями используются робастные (устойчивые) статистические методы.

Решение задач статистического анализа сложных данных требует интенсивного применения соответствующего программного обеспечения, поэтому *лабораторный практикум* проводится в компьютерном классе и предполагает применение реализованных методов анализа для различных моделей данных с использованием современного программного обеспечения, имеющегося в свободном доступе.

Взаимосвязь с другими дисциплинами

Дисциплина «Методы статистического анализа сложных данных» относится к компоненту учреждения высшего образования цикла дисциплин специализаций. Основой для изучения дисциплины «Методы статистического анализа сложных данных» является курс первой ступени «Теория вероятностей и математическая статистика» (или «Высшая математика» с включением соответствующих разделов), а также изучаемый ранее курс второй ступени «Методы статистического анализа многомерных данных». Кроме того, лабораторный практикум предполагает предварительное изучение курса «Компьютерный анализ данных с использованием языка R». Дисциплина «Методы статистического анализа сложных данных» будет полезна для изучения дисциплины «Статистическое моделирование и анализ данных в экономике и финансах», а также способствует успешному прохождению практики и написанию магистерских работ.

В результате изучения дисциплины студент должен:

-знать

- современные вероятностные модели сложных данных, возникающие при решении прикладных задач;
- эффективные статистические методы анализа моделей сложных данных;
- правила и свойства применения указанных методов;

- уметь

- подбирать подходящую модель для решения прикладных задач статистического анализа данных;
- оценивать потенциальную эффективность применения конкретного статистического метода для решения задачи анализа данных;

-владеть

- знаниями основных моделей сложных статистических 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 170 часов, из них 66 аудиторных часов, в том числе лекций – 34 часа, лабораторных занятий – 32 часа (магистратура, 2 семестр).

Форма текущей аттестации по учебной дисциплине – экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Некоторые современные модели сложных данных

Тема 1.1. **Особенности анализа сложных данных.** Введение. Примеры современных прикладных задач, показывающие принципиальные сложности статистического анализа данных классическими методами.

Тема 1.2. Модели сложных данных. Классификация современных моделей сложных данных.

Раздел 2. Байесовский подход в статистическом анализе

Тема 2.1. Байесовское оценивание параметров. Схема байесовского подхода в статистике. Байесовское оценивание параметров.

Тема 2.2. Байесовское прогнозирование. Байесовское прогнозирование временных рядов с трендом и авторегрессионных временных рядов.

Раздел 3. Последовательный статистический анализ

Тема 3.1. Последовательный подход в статистике. Построение последовательного критерия отношения правдоподобия. Анализ его характеристик. Обобщение на случай данных, образующих цепи Маркова.

Тема 3.2. Последовательная проверка сложных гипотез. Последовательные тесты проверки сложных гипотез. Применение в медицине.

Раздел 4. Анализ неполных данных

Тема 4.1. Модели пропусков. Различные модели пропусков в данных. Цензурирование.

Тема 4.2. EM-алгоритм. EM-алгоритм и его свойства.

Раздел 5. Анализ лонгитюдных данных

Тема 5.1. Кросс-секционные и лонгитюдные данные. Примеры кросс-секционных и длительных статистических исследований. Панели. Когортный анализ.

Тема 5.2. Модели с фиксированными эффектами. Оценивание параметров, анализ точности.

Тема 5.3. Модели со случайными эффектами. Оценивание параметров, анализ точности.

Раздел 6. Полупараметрические и непараметрические модели

Тема 6.1. Полупараметрические модели. Полупараметрические модели (сплайны).

Тема 6.2. Непараметрические модели. Модель ядерной регрессии.

Раздел 7. Робастные статистические выводы

Тема 7.1. Понятие робастности. Классификация типов искажений. Характеристики робастности.

Тема 7.2. Робастность оценивания параметров. Методы робастного статистического оценивания параметров модели.

Тема 7.3. Робастность проверки гипотез. Робастная статистическая проверка гипотез.

Тема 7.4. Робастность прогнозирования. Робастность в статистическом прогнозировании. 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	**Некоторые современные модели сложных данных**	**4**		**2**			
1.1	Особенности анализа сложных данных	2					
1.2	Модели сложных данных	2		2			Отчет по заданию с устной защитой
2	**Байесовский подход в статистическом анализе**	**4**		**4**			
2.1	Байесовское оценивание параметров	2		2			
2.2	Байесовское прогнозирование	2		2			Отчет по заданию с устной защитой
3	**Последовательный статистический анализ**	**4**		**4**			
3.1	Последовательный подход в статистике	2		2			
3.2	Последовательная проверка сложных гипотез	2		2			Контрольная работа 1
4.	**Анализ неполных данных**	**4**		**4**			
4.1	Модели пропусков	2		2			
4.2	EM-алгоритм	2		2			Коллоквиум
5	**Анализ лонгитюдных данных**	**6**		**6**			
5.1	Кросс-секционные и лонгитюдные данные	2		2			
5.2	Модели с фиксированными эффектами	2		2			
5.3	Модели со случайными эффектами	2		2			Отчет по заданию с устной защитой
6	**Полупараметрические и непараметрические модели**	**4**		**4**			
6.1	Полупараметрические модели	2		2			
6.2	Непараметрические моде-	2		2			Кон-

	ли						трольная работа 2
7	**Робастные статистические выводы**	**8**		**8**			
7.1	Понятие робастности	2		2			
7.2	Робастность оценивания параметров	2		2			
7.3	Робастность проверки гипотез	2		2			Отчет по заданию с устной защитой
7.4	Робастность прогнозирования	2		2			
ИТОГО		**34**		**32**			

ИНФОМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Рекомендуемая литература

Основная

1. Харин А.Ю. Робастность байесовских и последовательных статистических решающих правил / А.Ю. Харин. – Минск : БГУ. – 2014. – 207 с.
2. Харин Ю.С. Эконометрическое моделирование / Ю.С. Харин, В.И. Малюгин, А.Ю. Харин. – Минск : БГУ. – 2004. – 313 с.
3. Hastie T. The elements of statistical learning / T. Hastie, R. Tibshirani, J. Friedman. – New York : Springer. – 2008. – 758 p.

Дополнительная

4. Харин Ю.С. Теория вероятностей, математическая и прикладная статистика / Ю.С. Харин, Н.М. Зуев, Е.Е. Жук. – Минск : БГУ. – 2011. – 463 с.
5. 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.
6. Johnson R.A. Applied multivariate statistical analysis / R.A. Johnson, D.W. Wichern. – New York : Pearson. – 2007. – 800 p.
7. Everitt B. An introduction to applied multivariate analysis with R / B. Everitt, T. Hothorn. – New York: Springer. – 2011. – 274 p.

Рекомендации по контролю качества усвоения знаний и проведению аттестации

На лекционных занятиях по учебной дисциплине «Методы статистического анализа сложных 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 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.

Перечни используемых средств диагностики результатов учебной деятельности

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

– устная форма: экзамен;

– письменная форма: контрольные работы, коллоквиум;

– устно-письменная форма: отчеты по заданиям 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 заданиям и оценки за коллоквиум.

Итоговая аттестация предусматривает проведение 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)
Статистическое моделирование и анализ данных в экономике и финансах	Кафедра теории вероятностей и математической статистики	нет	Оставить содержание учебной дисциплины без изменения, протокол № 3 от 20 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)