**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Интеллектуальный анализ данных

Data Mining

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 043472

2020

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Формирование навыков исследования больших данных, направленных на решение типичных задач, возникающих во время статистического исследования с использованием пакетов языка R.

**1.2. Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)**

Дисциплина изучается магистрантами по направлению подготовки 01.04.02 «Прикладная математика и информатика» во втором семестре обучения. Для успешного освоения дисциплины обучающийся должен иметь предварительную подготовку в объеме курсов «Теория вероятностей», «Математическое моделирование и обработка данных», «Вычислительные методы и пакеты в статистическом исследовании», изучаемых по направлению подготовки 010400 «Прикладная математика и информатика» согласно Учебному плану.

**1.3. Перечень результатов обучения (learning outcomes)**

В результате освоения дисциплины обучающиеся должны владеть современными методами анализа данных посредством программирования в специализированных статистических пакетах программ (на примере пакета R). Обучающиеся должны демонстрировать развитие навыков самостоятельного решения прикладных статистических задач при помощи специализированных пакетов.  
  
Дисциплина участвует в формировании компетенций обучающихся по образовательной программе, установленных учебным планом для данной дисциплины.

ПКП-9 Способен использовать современные стохастические методы вычислений, строить статистические модели и анализировать данные.

ПКП-10 Способен развивать и обосновывать методы и алгоритмы статистического моделирования и анализа данных сложной структуры с учетом особенностей решаемой задачи.

**1.4. Перечень и объём активных и интерактивных форм учебных занятий**

Ответы на вопросы учащихся во время занятий.  
  
Обсуждение полученных на занятии результатов.  
  
Совместное решение задач, связанных с изученным материалом.  
  
Обсуждение ошибок учащихся и неэффективной реализации методов стохастических вычислений на языке R

**Раздел 2. Организация, структура и содержание учебных занятий**

**2.1. Организация учебных занятий**

**2.1.1 профиль Статистическое моделирование**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся | | | | | | | | | | | | | | | | | | |
| Код модуля в составе дисциплины,  практики и т.п. | Аудиторная работа обучающихся, часов | | | | | | | | | | Самостоятельная работа, часов | | | | | | Объём активных и интерактивных  форм учебных занятий | Трудоёмкость |
| лекции | семинары | консультации | практические  занятия | лабораторные работы | контрольные работы | коллоквиумы | текущий контроль | промежуточная  аттестация | итоговая аттестация | под руководством преподавателя | в присутствии  преподавателя | сам. раб. с использованием  методических материалов | текущий контроль (сам.раб.) | промежуточная аттестация (сам.раб.) | итоговая аттестация  (сам.раб.) |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 2 | 4 | 26 |  |  |  |  |  |  | 2 |  | 31 |  | 30 |  | 15 |  | 26 | 3 |
|  | 1-25 | 1-25 |  |  |  |  |  |  | 1-25 |  | 1-25 |  | 1-25 |  | 1-25 |  |  |  |
| ИТОГО | 4 | 26 |  |  |  |  |  |  | 2 |  | 31 |  | 30 |  | 15 |  | 26 | 3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации | | | | | | |
| Код модуля в составе дисциплины, практики и т.п. | Формы текущего контроля успеваемости | | Виды промежуточной аттестации | | Виды итоговой аттестации  (только для программ итоговой аттестации и дополнительных образовательных программ) | |
| Формы | Сроки | Виды | Сроки | Виды | Сроки |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | |
| Форма обучения очная | | | | | | |
| Семестр 2 |  |  | зачёт, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

Основной курс Основная траектория Очная форма обучения

Период обучения (модуль): Семестр 2

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Наименование темы (раздела, части) | Вид учебных занятий | Количество часов |
| 1 | Введение | лекции | 2 |
| семинары | 0 |
| под руководством преподавателя | 2 |
| по методическим материалам | 4 |
| 2 | Линейная регрессия в R | лекции | 2 |
| семинары | 5 |
| под руководством преподавателя | 6 |
| по методическим материалам | 8 |
| 3 | Задачи классификации. Линейные методы в R | лекции | 0 |
| семинары | 5 |
| под руководством преподавателя | 6 |
| по методическим материалам | 8 |
| 4 | Resampling | лекции | 0 |
| семинары | 3 |
| под руководством преподавателя | 3 |
| по методическим материалам | 4 |
| 5 | Снижение размерности данных и кластеризация в R | лекции | 0 |
| семинары | 4 |
| под руководством преподавателя | 4 |
| по методическим материалам | 6 |
| 6 | Линейные модели. Регуляризация в R | лекции | 0 |
| семинары | 3 |
| под руководством преподавателя | 4 |
| по методическим материалам | 6 |
| 7 | Нелинейные модели в задачах регрессии и классификации в R | лекции | 0 |
| семинары | 6 |
| под руководством преподавателя | 6 |
| по методическим материалам | 9 |

I. Введение

Введение. Задачи интеллектуального анализа данных. Примеры.

II. Линейная регрессия в R

Оценивание. Проверка гипотез относительно коэффициентов. Доверительные интервалы для коэффициентов. Категориальные переменные в регрессии. Отбор информативных признаков. Выбросы, гетероскедастичность, high-leverage points.

III. Задачи классификации. Линейные методы в R.

Логистическая регрессия. Линейный и квадратичный дискриминантный анализ. Понятие ROC-кривой и показателя AUC (area under the curve). Метод Naïve Bayes.

IV. Resampling.

Кросс-валидация. Бутстреп. Бутстреп-доверительные интервалы.

V. Снижение размерности данных и кластеризация в R

Метод главных компонент. Оптимальные проекции на пространства меньшей размерности. Методы k-средних и k-медиан. Иерархическая кластеризация. EM-алгоритм и кластеризация на основе модели (оценивание компонент смеси распределений).

VI. Линейные модели. Регуляризация в R

Регуляризация. Ridge regression. Lasso. Elastic net regression. Робастная регрессия.

VII. Нелинейные модели в задачах регрессии и классификации в R

Полиномиальная регрессия. Локальные сплайны. Локальная регрессия. Обобщенные аддитивные модели (Generalized Additive Models, GAM), Projection Pursuit Regression. Деревья регрессии. Алгоритм Random Forest. Подходы bagging и boosting. Алгоритм AdaBoost. Метод опорных векторов (Support Vector Machines, SVM). Постановка задачи. Параметризация. Kernel trick, ядра и спрямляющие пространства. Подбор оптимальных параметров.

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Успешное освоение дисциплины возможно благодаря систематическому и добросовестному посещению лекций и семинаров, участию в обсуждении вопросов, подготовленных к занятию, самостоятельной работе, включающей в себя чтение специальной литературы по разделам темы, а также решению задач.

**3.1.2 Методическое обеспечение самостоятельной работы**

Методические материалы включают в себя следующие типы материалов — учебники, учебные пособия, методические указания для обучающихся, Интернет-ресурсы, электронные учебные пособия, с опорой на которые проводится аудиторная работа.

Роль преподавателя в организации самостоятельной работы состоит в координации действий обучающихся в освоении дисциплины, в методическом и организационном обеспечении учебного процесса. Взаимодействие между преподавателем и учащимся осуществляется в форме консультаций. Преподаватели также оказывают помощь обучающимся по планированию и организации самостоятельной работы.

Методическое обеспечение самостоятельной работы обеспечивается наличием в библиотеке необходимых учебников и монографий, наличием электронных копий другой методической литературы, а также электронных копий комментариев к наиболее сложным частям материала, написанным преподавателем. Материалы к семинару могут выкладываться в Интернет.

**3.1.3 Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания**

Аппарат контроля за усвоением материала включает в себя выполнение домашних заданий и зачет по итогам курса. Зачет проводится в виде опроса и зачет выставляется согласно выработанной шкале баллов. Результаты выполнения домашних заданий (самостоятельной работы) учитываются при формировании итогового балла зачета. Правила текущего контроля успеваемости и промежуточной аттестации (в том числе формат зачета) доводятся до учащихся на первом занятии.

Как правило, опрос включает в себя пять вопросов по программе курса. Каждый вопрос оценивается от 0 до 2 баллов, в зависимости от полноты и правильности ответа. Таким образом, за исчерпывающие и правильные ответы на все вопросы опроса можно получить 10 баллов. Невыполненные или неправильно выполненные домашние задания снижают итоговый балл за опрос пропорционально своей доли (например, при общем количестве домашних заданий в 7 и невыполнении 2 домашних заданий итоговое число баллов уменьшается на 28%).

**Критерии оценивания:**

* Оценка «зачтено» («A» по системе ECTS) ставится, если обучающийся показал глубокое знание предмета, дал исчерпывающие ответы на поставленные вопросы, способен без подготовки или после небольших затрат времени ответить на дополнительные вопросы, общее количество набранных баллов: 9.0-10.0.
* Оценка «зачтено» («B» по системе ECTS) ставится, если обучающийся уверенно владеет материалом, но при ответе неуверенно себя чувствует при ответах на дополнительные вопросы, общее количество набранных баллов: 8.0-8.9.
* Оценка «зачтено» («C» по системе ECTS) ставится, если обучающийся уверенно владеет материалом, но при ответе упускает отдельные существенные моменты; неуверенно себя чувствует при ответах на дополнительные вопросы, общее количество набранных баллов: 7.6-7.9.
* Оценка «зачтено» («D» по системе ECTS) ставится, если обучающийся ориентируется в поставленных вопросах, может сформулировать основные моменты, но теряется при ответах на дополнительные вопросы, общее количество набранных баллов: 7.1-7.5.
* Оценка «зачтено» («E» по системе ECTS) ставится, если обучающийся ориентируется в поставленных вопросах, может сформулировать основные моменты, но не способен дать полный ответ; теряется при ответах на дополнительные вопросы, общее количество набранных баллов: 6.6-7.0.
* Если обучающийся не способен дать ответ на поставленные вопросы, ему выставляется оценка «не зачтено» («F» по системе ECTS), общее количество набранных баллов: менее 6.6.

**3.1.4 Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)**

Темы домашних заданий и вопросы для опроса.

Оценивание. Проверка гипотез относительно коэффициентов. Доверительные интервалы для коэффициентов. Категориальные переменные в регрессии. Отбор информативных признаков. Выбросы, гетероскедастичность, high-leverage points.

Логистическая регрессия. Линейный и квадратичный дискриминантный анализ. Понятие ROC-кривой и показателя AUC (area under the curve). Метод Naïve Bayes.

Кросс-валидация. Бутстреп. Бутстреп-доверительные интервалы.

Метод главных компонент. Оптимальные проекции на пространства меньшей размерности. Методы k-средних и k-медиан. Иерархическая кластеризация. EM-алгоритм и кластеризация на основе модели (оценивание компонент смеси распределений).

Регуляризация. Ridge regression. Lasso. Elastic net regression. Робастная регрессия.

Полиномиальная регрессия. Локальные сплайны. Локальная регрессия. Обобщенные аддитивные модели (Generalized Additive Models, GAM), Projection Pursuit Regression. Деревья регрессии. Алгоритм Random Forest. Подходы bagging и boosting. Алгоритм AdaBoost. Метод опорных векторов (Support Vector Machines, SVM). Постановка задачи. Параметризация. Kernel trick, ядра и спрямляющие пространства. Подбор оптимальных параметров.

Пример домашнего задания  
Провести анализ многомерных данных с помощью линейных моделей, используя методы регуляризации и методы кросс-валидации.

**3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса**

Для оценки содержания и качества учебного процесса может применяться анкетирование или опрос в соответствии с методикой и графиком, утверждаемым в установленном порядке.

**3.2. Кадровое обеспечение**

**3.2.1 Образование и (или) квалификация штатных преподавателей и иных лиц, допущенных к проведению учебных занятий**

К преподаванию дисциплины могут быть допущены преподаватели, имеющие диплом о высшем образовании по соответствующему направлению.

**3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом**

Не предусмотрено.

**3.3. Материально-техническое обеспечение**

**3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий**

Учебные аудитории для проведения учебных занятий, оснащенные стандартным оборудованием, используемым для обучения в СПбГУ в соответствии с требованиями материально-технического обеспечения/

**3.3.2 Характеристики аудиторного оборудования, в том числе неспециализированного компьютерного оборудования и программного обеспечения общего пользования**

Стандартное оборудование, используемое для обучения в СПбГУ. MS Windows, MS Office, Mozilla FireFox, Google Chrome, Acrobat Reader DC, WinZip, Антивирус Касперского.

**3.3.3 Характеристики специализированного оборудования**

Не требуется.

**3.3.4 Характеристики специализированного программного обеспечения**

Пакет статистических программ R (версии не ниже 3.1) и оболочка-редактор к нему (например, RStudio), пакет LaTeX (сборка MikTeX версии не ниже 2.9) и редактор-оболочка (например, TeXMaker).

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Белая бумага формата А4 для печати на принтере в размере 500 листов. Для использования доски необходим либо мел, либо специальные фломастеры, в зависимости от типа доски.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. А. Шипунов, Е. Балдин, П. Волкова, А. Коробейников, С. Назарова, С. Петров, В. Суфиянов. Наглядная статистика. Используем R!. ДМК-Пресс, 2012. 298 стр. - ЭБС «Лань» по подписке СПбГУ: <https://proxy.library.spbu.ru/login?url=https://e.lanbook.com/>

2. Р. Кабаков. R в действии. Анализ и визуализация данных на языке R. ДМК-Пресс, 2013. 580 стр. - ЭБС «Лань» по подписке СПбГУ: <https://proxy.library.spbu.ru/login?url=https://e.lanbook.com/>

1. М. Лагутин. Наглядная математическая статистика. Бином, 2009. 472 стр.
2. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Introduction to Statistical Learning (with Applications in R). Springer, 2014.
3. Trevor Hastie, Robert Tibshirani. Elements of Statistical Learning. Springer, 2010.

**3.4.2 Список дополнительной литературы**

1. Ермаков С.М. Метод Монте-Карло и смежные вопросы, М., «Наука» 1975.
2. Peter Dalgaard. Introductory Statistics with R. Springer, 2002.
3. Рао С.Р. Линейные статистические методы и их применение. М., 1968.
4. Голяндина Н.Э. Метод «Гусеница»-SSA: анализ временных рядов: Учеб. пособие. СПб., 2004.
5. Golyandina, N. and Zhigljavsky, A. Singular Spectrum Analysis for Time Series. Springer, 2013.
6. Ермаков С.М., Жиглявский А.А. Математическая теория оптимального эксперимента. М., 1987.
7. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. М., 2001.
8. J. Faraway. Practical Regression and ANOVA using R. 2002.
9. W. Krijnen. Applied Statistics for Bioinformatics Using R. 2009

**3.4.3 Перечень иных информационных источников**

* Сайт Научной библиотеки им. М. Горького СПбГУ:
* <http://www.library.spbu.ru/>
* Электронный каталог Научной библиотеки им. М. Горького СПбГУ:
* <http://www.library.spbu.ru/cgi-bin/irbis64r/cgiirbis_64.exe?C21COM=F&I21DBN=IBIS&P21DBN=IBIS>
* Перечень электронных ресурсов, находящихся в доступе СПбГУ:
* <http://cufts.library.spbu.ru/CRDB/SPBGU/>
* Перечень ЭБС, на платформах которых представлены российские учебники, находящиеся в доступе СПбГУ:
* <http://cufts.library.spbu.ru/CRDB/SPBGU/browse?name=rures&resource_type=8>
* ACM Digital Library: <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/12>
* Institute of Electrical and Electronics Engineers (IEEE): <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/375>
* MathSciNet - электронная коллекция Американского математического сообщества (AMS): <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/415>
* O’Reilly: <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/483>
* **Zentralblatt MATH:** <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/86>

**Раздел 4. Разработчики программы**

Коробейников А.И. e-mail a.korobeynikov@spbu.ru