Программа учебной дисциплины «Методы и средства обработки больших данных»

Утверждена

Академическим советом ООП

Протокол № 3 от «29» мая 2018г.

Авторы	Кондрашова Е.В., кандидат ф.-м. наук, elizavetakondr@gmail.com Андрианова О.Г., кандидат ф.-м. наук, oandrianova@hse.ru
Число кредитов	2
Контактная работа (час.)	32
Самостоятельная работа (час.)	44
Курс	2
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Методы и средства обработки больших данных» являются: фундаментальная подготовка в области методов обработки больших данных, овладение средствами обработки больших данных для дальнейшего использования в приложении SAS Enterprise Miner.

Задачи дисциплины состоят в изучении и практическом освоении современных компьютерных технологий для проведения прикладных математических исследований.

Настоящая дисциплина относится к базовой части профессионального цикла и является обязательной дисциплиной.

Изучение данной дисциплины базируется на следующих дисциплинах:

- Основы баз данных, основы прикладной математики, основы теории вероятности.

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- теория вероятностей;
- математическая статистика;
- основы прикладной математики и экономико-математического моделирования;
- основы эконометрики;
- основы программирования.

Основные положения дисциплины могут быть использованы в дальнейшем при изучении следующих дисциплин: научно-исследовательский семинар и выполнение выпускной квалификационной работы.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Обзор Big-Data. Методы и средства. Используемые программы. Особенности

Big-Data. Инструменты. Технологии. Методы анализа.

Тема 2. SAS Interprise Miner. Введение. Возможности. Инструменты.

Меню. Принцип анализа данных SEMMA. Основные инструменты и узлы. Возможности построения моделей.

Тема 3. Создание проекта. Определение источника данных. Исследование источника данных.

Создание проекта, библиотеки и диаграмм SAS. Настройки источника данных. Типы переменных. Изменение размера выборки. Создание диаграмм. Исследование взаимосвязей между переменными.

Тема 4. Прогнозное моделирование

Область прикладных задач с использованием прогнозного моделирования. Проклятие размерности. Избавление от бесполезных и избыточных входных переменных. Создание обучающих и проверочных данных.

Тема 5. Прогнозная модель, использующая дерево решений.

Создание дерева решений: структура. Алгоритм построения. Поиск разбиений. Прогнозная модель использующая дерево решений: построение, создание правила разбиения. Оптимизация сложности деревьев решений. Оценка качества дерева решений.

Тема 6. Прогнозное моделирование: работа с регрессионными моделями

Регрессия. Логистическая регрессия. Полиномиальные регрессии. Оценка параметров. Обработка пропущенных значений. Выбор входных переменных. Оптимизация сложности. Интерпретация регрессии. Регрессии с преобразованными входными переменными. Категориальные переменные в регрессионной модели. Область прикладных задач с использованием прогнозного моделирования. Проклятие размерности. Создание обучающих и проверочных данных.

Тема 7. Прогнозное моделирование: нейронные сети.

Особенности нейронных сетей. Обучение нейронной сети. Инструмент AutoNeural.

Тема 8. Кластерный анализ. Анализ потребительской корзины

Кластерный анализ: методы обучения. Исследование сегментов. Анализ потребительской корзины: инструмент Association.

Тема 9. Оценка моделей. Сравнение моделей.

Статистики подгонки моделей. ROC-индекс, кривые. SBC. Сравнение моделей с помощью сводных статистик. Графики рейтингов. Матрица прибыли. Ансамбль моделей.

III. ОЦЕНИВАНИЕ

Преподаватель оценивает работу студентов на семинарских и практических занятиях: оценивается активность студента на практических занятиях, участие в дискуссиях, правильность решения задач, умение построение модели с

использованием различных данных, понимание сильных сторон и ограничений используемых инструментов. Оценки за работу на практических занятиях преподаватель выставляет в рабочую ведомость. Оценка по 10-ти балльной шкале за работу на практических занятиях определяется перед итоговым контролем — Оауд.

Преподаватель оценивает самостоятельную работу студентов (задания, которые выдаются на семинарских занятиях). Оценки за самостоятельную работу студента преподаватель выставляет в рабочую ведомость. Оценка по 10-ти балльной шкале за работу на практических занятиях определяется перед итоговым контролем — *О*сам.

Накопленная оценка по дисциплине рассчитывается по формуле:

$$Онакопл = 0.5*Оауд+ 0.5*Осам$$

Способ округления накопленной оценки: в пользу студента.

В диплом выставляется результирующая оценка по учебной дисциплине, которая вычисляется по следующей формуле:

$$Орезультат = 0.5*Онакопл+ 0.5*Оэкзамен$$

Способ округления результирующей оценки: в пользу студента.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Экзамен: Студент должен продемонстрировать понимание концепции теоретических основ дисциплины, основных инструментов и приложений данной дисциплины.

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

Примеры вопросов для оценки качества освоения дисциплины

Концепция MapReduce.

Основные методы анализа больших данных.

Продемонстрировать использование окна Explore для изучения распределения переменной. Выявить, какие отклонения выделяются в распределении переменной? Как можно исправить эту ситуацию?

Провести первоначальное исследование данных с использованием узлов Data Partition и Decision Tree. Сделать выводы о модели.

Провести прогнозное моделирование с использованием регрессии Regression. Объяснить, какие переменные являются важными в модели. Какое значение имеет статистика среднеквадратичной ошибки, посчитанная на проверочной выборке?

Нужно ли делать преобразования входных переменных перед их использованием в модели нейронной сети?

Объяснить результаты, полученные на основании Model Comparison. Сделать выводы.

V. РЕСУРСЫ

1. Основная литература

1. Майер-Шенбергер В., Большие данные : революция, которая изменит то, как мы живем, работаем и мыслим, Манн, Иванов и Фербер, 2014
2. Айвазян С. А., Эконометрика : учеб. пособие для вузов, Маркет ДС, 2010
3. Рутковская Д., Нейронные сети,генетические алгоритмы и нечеткие системы, Горячая линия-Телеком, 2008
4. Битюцков В. И., Вероятность и математическая статистика : Энциклопедия, Большая Рос. энцикл., 1999
5. Хайкин С., Нейронные сети : полный курс, Вильямс, 2006
6. Рутковская Д., Нейронные сети,генетические алгоритмы и нечеткие системы, Горячая линия-Телеком, 2008

2. Дополнительная литература

1. Мацкевич И. П., Высшая математика : Теория вероятностей и математическая статистика: Учебник для вузов, Вышэйшая школа, 1993
2. Яхъяева Г. Э., Нечеткие множества и нейронные сети : учеб. пособие, Интернет-Университет Информационных Технологий, 2008
3. Розанов Ю. А., Теория вероятностей, случайные процессы и математическая статистика : учебник для вузов, Наука. Гл. ред. физ.-мат. лит., 1989
4. Осовский С., Нейронные сети для обработки информации, Финансы и статистика, 2004

1. Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	*Из внутренней сети университета*
2.	Google Chrome	*Договор (Свободное лицензионное соглашение)*

2. Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены персональными компьютерами с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.