Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет»

Кафедра компьютерной безопасности и математического обеспечения информационных систем

РАБОЧАЯ ПРОГРАММА

ДИСЦИПЛИНЫ

«Б1.Д.В.Э.4.2 Многомерные статистические методы»

Уровень высшего образования

СПЕЦИАЛИТЕТ

Специальность
10.05.01 Компьютерная безопасность
(код и наименование специальности)

специализация №3 «Разработка защищенного программного обеспечения»
(наименование направленности (профиля)/специализации образовательной программы)

Квалификация
Специалист по защите информации

Форма обучения
Очная

Год набора 2021

Рабочая программа дисциплины «*Б1.Д.В.Э.4.2 Многомерные статистические методы*» рассмотрена и утверждена на заседании кафедры

Кафедра компьютерной безопасности и математического обеспечения информационных систем
наименование кафедры

протокол № 10 от "21" июня 2021 г.

Заведующий кафедрой
Кафедра компьютерной безопасности и математического обеспечения информационных систем — И.В. Влацкая
наименование кафедры *подпись* *расшифровка подписи*

Исполнители:
доцент — И.В. Влацкая
должность *подпись* *расшифровка подписи*

должность *подпись* *расшифровка подписи*

СОГЛАСОВАНО:
Председатель методической комиссии по специальности
10.05.01 Компьютерная безопасность — И.В. Влацкая
код наименование *личная подпись* *расшифровка подписи*

Заведующий отделом комплектования научной библиотеки
Н.Н. Бигалиева
личная подпись *расшифровка подписи*

Уполномоченный по качеству факультета
И.В. Крючкова
личная подпись *расшифровка подписи*

№ регистрации 117857

© Влацкая И.В., 2021
© ОГУ, 2021

1 Цели и задачи освоения дисциплины

Цель (цели) освоения дисциплины:

Формирование у будущих специалистов теоретических знаний и практических навыков по многомерным статистическим методам и методам машинного обучения.

Задачи:

- приобретение умений выбора признаков анализируемых объектов, выбора моделей машинного обучения, их гиперпараметрической оптимизации и контроля качества обучения;
- приобретение базовых навыков решения задач классификации, регрессии и кластеризации с использованием языка программирования Python и различных библиотек.

2 Место дисциплины в структуре образовательной программы

Дисциплина относится к дисциплинам (модулям) по выбору вариативной части блока Д «Дисциплины (модули)»

Пререквизиты дисциплины: *Б1.Д.Б.21 Системы управления базами данных, Б1.Д.Б.44.5 Технология создания прикладного программного обеспечения*

Постреквизиты дисциплины: *Отсутствуют*

3 Требования к результатам обучения по дисциплине

Процесс изучения дисциплины направлен на формирование следующих результатов обучения

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1-В-2 Осуществляет критический анализ и синтез информации, полученной из разных источников УК-1-В-4 Применяет методы сбора, хранения, обработки, передачи, анализа и синтеза информации с использованием компьютерных технологий для решения поставленных задач	**Знать:** - теоретические основы многомерных статистических методов; - теоретические основы моделей машинного обучения для решения различных задач анализа данных и построения предсказательных моделей. **Уметь:** - выбирать признаки объектов для решения различных задач анализа данных и построения предсказательных моделей; - подбирать семейства моделей машинного обучения для конкретных задач, производить их обучение,

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
		гиперпараметрическую оптимизацию, контроль качества обучения. **Владеть:** - навыками написания приложений, использующих различные модели машинного обучения для решения прикладных задач анализа данных.

4 Структура и содержание дисциплины

4.1 Структура дисциплины

Общая трудоемкость дисциплины составляет 4 зачетные единицы (144 академических часа).

Вид работы	Трудоемкость, академических часов	
	10 семестр	всего
Общая трудоёмкость	**144**	**144**
Контактная работа:	**61,25**	**61,25**
Лекции (Л)	30	30
Лабораторные работы (ЛР)	30	30
Консультации	1	1
Промежуточная аттестация (зачет, экзамен)	0,25	0,25
Самостоятельная работа: *- выполнение индивидуального творческого задания (ИТЗ);* *- самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий;* *- подготовка к лабораторным занятиям;* *- подготовка к рубежному контролю и т.п.)*	**82,75**	**82,75**
Вид итогового контроля (зачет, экзамен, дифференцированный зачет)	**экзамен**	

Разделы дисциплины, изучаемые в 10 семестре

№ раздела	Наименование разделов	Количество часов				
		всего	аудиторная работа			внеауд. работа
			Л	ПЗ	ЛР	
1	Введение в машинное обучение и анализ многомерных данных	40	6		10	24
2	Методы обучения с учителем	54	12		10	42
3	Методы обучения без учителя	50	12		10	28
	Итого:	144	30		30	84
	Всего:	144	30		30	84

4.2 Содержание разделов дисциплины

1 Введение в машинное обучение и анализ многомерных данных. *Понятие машинного обучения. Соотнесение областей искусственного интеллекта, машинного обучения, нейронных сетей и глубокого обучения. Задача обучения по прецедентам. Признаковое описание объектов. Задание ответов. Типы задач машинного обучения.*

Предсказательная модель. Примеры задач классификации и регрессии. Этапы обучения и применения моделей машинного обучения. Функционалы качества. Сведение задачи обучения к задаче оптимизации.

Объекты, признаки и особенности задач медицинской диагностики, кредитного скоринга, предсказания оттока клиентов, категоризации текстовых документов, прогнозирования стоимости недвижимости, прогнозирования объемов продаж, предсказания прибыли ресторана, ранжирования поисковой выдачи, ранжирования в рекомендательных системах, предсказания перехода по контекстной рекламе.

Исследовательский анализ данных с помощью библиотеки Pandas. Тип DataFrame. Загрузка данных, получение информации о первых или последних пяти записях, колонках, типах. Описательная статистика. Операции доступа к отдельным колонкам, строчкам, элементам DataFrame, их группам. Получение уникальных значений в конкретных колонках, подсчет частот уникальных значений в конкретных колонках. Сортировка и фильтрация данных, применение функций к отдельным ячейкам, колонкам и строчкам. Маппирование, изменение типа колонок данных. Добавление, удаление, замена колонок данных. Группировка данных. Получение Summary Tables.

Библиотека Numpy. Многомерные гомогенные массивы. Создание массивов. Атрибуты для получения их размерности, формы, количества элементов, типа данных. Печать массивов. Доступ к элементам и срезам, итерирование. Изменение формы. Базовые операции над массивами. Понятие и правила броадкастинга. Индексирование с помощью массива индексов, логических массивов, выражений.

Библиотеки Matplotlib и Seaborn. Визуализация количественных одномерных признаков с помощью гистограмм и графиков ядерной оценки плотности, графиков «ящиков с усами», скрипичного графика. Визуализация одномерных категориальных и бинарных признаков с помощью столбчатых диаграмм. Визуализация пары количественных признаков (корреляционных матриц, диаграммы разброса). Попарная визуализация всех признаков данных. Использование категориальных признаков для раскраски при визуализации пары количественных признаков. Использование графика «ящика с усами» и скрипичного графика для визуализации количественных признаков в зависимости от различных значений категориальных признаков. Визуализация «Categorical vs. Categorical».

Понятие t-SNE (t-distributed Stochastic Neighbor Embedding). Принцип работы и визуализация многомерных данных на двумерной плоскости.

Библиотека Plotly. Отличие от других библиотек. Ее использование для визуализации линейных графиков, столбчатых диаграмм и графиков «ящиков с усами».

Понятие переобучения. Примеры переобучения в задачах классификации и регрессии. Эмпирические оценки обобщающей способности моделей. Валидационные кривые и кривые обучения. Понятия обучения и переобучения. Необходимость добавления новых данных. Кросс-валидация и проверка на контрольных данных. Оптимизация гиперпараметров моделей.

Эксперименты на реальных данных – на конкретной прикладной задаче или на наборах прикладных задач. Эксперименты на модельных данных – синтетических и полусинтетических. Межотраслевой стандарт решения задач интеллектуального анализа данных CRISP-DM. Основные этапы решения задач машинного обучения.

2 Методы обучения с учителем. *Понятие дерева решений. Их использование для решений задач классификации и регрессии. Понятие прироста информации. Метрики качества разбиения в задачах классификации и регрессии. Процедуры построения деревьев классификации и регрессии.*

Понятие деревьев решений. Использование стандартных классов DecisionTreeClassifier и DecisionTreeRegressor из библиотеки scikit-learn. Визуализация построенных деревьев решений. Проблема переобучения деревьев решений, подбор гиперпараметров. Достоинства и недостатки деревьев решений..

Метод k-NN (k-Nearest Neighbors). Работа метода на этапах обучения и использования. Использование k-NN в реальных задачах. Использование стандартного класса KNeighborsClassifier из библиотеки scikit-learn. Подбор гиперпараметров. Достоинства и недостатки k-NN.

Задача линейной регрессии. Метод наименьших квадратов. Метод максимального правдоподобия. Разложение ошибки на смещение и разброс. Регуляризация линейной регрессии.

Линейный классификатор. Логистическая регрессия как линейный классификатор. Принцип максимального правдоподобия. L2-регуляризация логистической функции потерь.

Использование классов LogisticRegresson и LinearRegression из библиотеки scikit-learn. Регуляризация и оптимизация гиперпараметров.

Метод опорных векторов (Support Vector Machine). Понятие отступа, оптимальной разделяющей гиперплоскости. Постановка задачи SVM. Ядра и спрямляющие пространства (ядерные функции). Регуляризация. Преимущества и недостатки SVM.

Метрики оценки качества моделей классификации и регрессии.

Метод градиентного спуска в решении оптимизационных задач. Стохастический, пакетный методы градиентного спуска при решении задач классификации и регрессии. Задачи машинного обучения, в которых применяется стохастический градиентный спуск. Основные возможности инструмента Vowpal Wabbit, примеры использования.

Ансамбли алгоритмов машинного обучения, случайных лес. Понятие и примеры ансамбля. Понятия boostrap, bagging, Out-Of-Bag Error. Алгоритмы обучения и работы обученной модели случайного леса. Сравнение случайного леса с деревом решений и bagging. Реализация случайного леса в библиотеке scikit-learn – классы RanfomForestClassifier и RamdomForestRegressor, их основные параметры. Достоинства и недостатки алгоритма случайного леса.

Понятие бустинга, общая схемы работы бустинга. Алгоритм AdaBoost. Градиентный бустинг.

Библиотека CatBoost, принципы работы, сравнение с другими библиотеками бустинга деревьев. Кодирование категориальных признаков. Использование библиотеки в задачах классификации и регрессии, выполнение кросс-валидации, детектирование переобучения, ранняя остановка, настройка гиперпараметров. Оценка важности признаков.

Библиотека LightGBM, принципы работы, сравнение с другими библиотеками бустинга деревьев. Понятие дерева решений с повышением градиента. Стратегии для вычисления деревьев. Нахождение оптимальных точек разделения. Использование библиотеки в задачах классификации и регрессии, настройка гиперпараметров. Оценка важности признаков.

Использование библиотек ELI5, SHAP, LIME для интерпретации результатов обучения различных алгоритмов машинного обучения (XGBoost, Random Forest, k-NN). Основные методы, лежащие в основе данных работы данных библиотек.

Понятие временного ряда. Примеры рядов. Понятие скользящего среднего. Экспоненциальное сглаживание. Двойное экспоненциальное сглаживание. Метод Holt-Winters. Кросс-валидация временных рядов. Метрики качества моделей – MAE, MAPE, WAPE.

Понятие временного ряда. STL-декомпозиция и компоненты временного ряда – тренд и сизональности, остаточные значения. Понятие стационарного ряда. Критерий Дики-Фуллера. Автокорреляция и частичная автокорреляция. Подходы к избавлению от нестационарности. Модель SARIMAX, ее компоненты, параметры, способы выбора параметров. Анализ остаточных рядов. Библиотека StatsModels.

3 Методы обучения без учителя. *Понятие обучения без учителя. Метод главных компонент. Методы кластеризации: k-means, affinity propagation, спектральная кластеризация, агломеративная кластеризация. Метрики качества кластеризации. Основные возможности библиотеки scikit-learn для реализации данных методов.*

Автоматический отбор признаков – понятие и потребности его использования. Методы отбора признаков - методы фильтрации, встроенные методы, wrapper methods, их достоинства и недостатки. Библиотека FeatureSelector, ее основные возможности и примеры использования.

4.3 Лабораторные работы

№ ЛР	№ раздела	Наименование лабораторных работ	Кол-во часов

№ ЛР	№ раздела	Наименование лабораторных работ	Кол-во часов
1	1	Исследование и визуализация данных	2
2	2	Деревья решений и метод ближайших соседей	2
3	2	Линейная классификация и регрессия	2
4	2	Методы градиентного бустинга деревьев	4
5	2	Работа с временными рядами	2
6	3	Метод главных компоненты и методы кластеризации данных	2
		Итого:	30

5 Учебно-методическое обеспечение дисциплины

5.1 Основная литература

1. Айвазян, С. А. Прикладная статистика. Основа эконометрики : в 2 т.: учеб. для вузов / С. А. Айвазян, В. С. Мхитарян . - М. : ЮНИТИ-ДАНА, 2001.. - ISBN 5-238-00304-8. Т. 1 : Теория вероятностей и прикладная статистика. - , 2001. - 656 с.

5.2 Дополнительная литература

1. Теория статистики [Текст] : учебник / под ред. Г. Л. Громыко.- 2-е изд., перераб. и доп. - М. : ИНФРА-М, 2006. - 476 с.

5.3 Периодические издания

1. Журнал «Информационные технологии».
2. Журнал «Открытые системы. СУБД».
3. Журнал «Программная инженерия».

5.4 Интернет-ресурсы

1. Open Machine Learning Course mlcourse.ai. [Электронный ресурс]. – Режим доступа: https://mlcourse.ai/
2. Kaggle [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/
3. Введение в машинное обучение [Электронный ресурс]. – Режим доступа: https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie

5.5 Программное обеспечение, профессиональные базы данных и информационные справочные системы

1. Операционная система Microsoft Windows текущей версии. Доступна в рамках подписки Microsoft DreamSpark Premium. Разработчик: компания Microsoft. Режим доступа: https://e5.onthehub.com/WebStore/ProductsByMajorVersionList.aspx?cmi_mnuMain=bdba23cf-e05e-e011-971f-0030487d8897&ws=58727022-4bac-e211-88b7-f04da23e67f4&vsro=8

2. Офисный пакет Microsoft Office (Word, Excel, Power Point) текущей версии. Доступен в рамках лицензионного соглашения OVS-ES. Разработчик: компания Microsoft. Режим доступа: https://products.office.com/en/home

6 Материально-техническое обеспечение дисциплины

Учебные аудитории для проведения занятий лекционного типа, семинарского типа, курсового проектирования, для проведения групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации.

Аудитории оснащены комплектами ученической мебели, техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Для проведения лабораторных занятий используется компьютерный класс, оснащенный персональными компьютерами с установленным программным обеспечением.

Помещение для самостоятельной работы обучающихся оснащены компьютерной техникой, подключенной к сети "Интернет", и обеспечением доступа в электронную информационно-образовательную среду ОГУ.

К рабочей программе прилагаются:

- Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине;
- Методические указания для обучающихся по освоению дисциплины.