Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Смоленский государственный университет»

Кафедра математического анализа

«Утверждаю»
Проректор по учебно-
методической работе
_____Ю.А. Устименко
«6» сентября 2022 г.

Рабочая программа дисциплины
Б1.В.ДВ.04.02 Интеллектуальный анализ данных

Направление подготовки: 02.04.01 Математика и компьютерные науки
Направленность (профиль): Методы моделирования в анализе и стохастике
Форма обучения – очная
Курс – 2
Семестр – 4
Всего зачетных единиц – 3, часов – 108
Форма отчетности: зачет – 4 семестр

Программу разработал
старший преподаватель Курицын С.Ю.

Одобрена на заседании кафедры
«30» августа 2022 г., протокол № 11

Смоленск
2022

1. Место дисциплины в структуре ОП

Дисциплина «Интеллектуальный анализ данных» относится к части, формируемой участниками образовательных отношений, и является дисциплиной по выбору. Она изучается в 4 семестре и нацелена на освоение прикладных методов обработки больших массивов данных при научно-исследовательской деятельности магистранта.

При изучении данной дисциплины необходимы компетенции студентов, сформированные при освоении образовательных программ бакалавриата из укрупненных групп направлений подготовки 01.00.00 Математика и механика, 02.00.00 Компьютерные и информационные науки, 09.00.00 Информатика и вычислительная техника, а также при изучении дисциплин «Дискретные и вероятностные модели», «Математические модели в естественных науках», «Прикладные стохастические модели», «Прикладной статистический анализ» и другие.

Изучение курса основано на традиционных методах высшей школы, тесной взаимосвязи со смежными курсами, а также на использовании современных издательских систем.

2. Планируемые результаты обучения по дисциплине

Компетенция	Индикаторы достижения
ПК-1. Способен осуществлять поиск, анализ и систематизацию научной информации в области анализа и стохастики для реализации научно-исследовательских проектов	**Знает**: теоретические основы и технологии организации научно-исследовательской деятельности, современный аппарат, методологическую базу и сферу приложения анализа и стохастики, пути использования имеющихся знаний при проведении научно-исследовательской работы. **Умеет**: осуществлять поиск, анализ, систематизацию научной информации в области анализа и стохастики и их приложений для реализации научно-исследовательских проектов. **Владеет**: навыками организации и проведения научно-исследовательской деятельности в ходе выполнения профессиональных функций.
ПК-2. Способен применять методы стохастического и аналитического математического моделирования для решения прикладных задач	**Знает:** методы стохастического и аналитического математического моделирования. **Умеет:** выбирать методики разработки требований к модели, строить причинно-следственные связи, формулировать требования к модели и цели ее создания, исходя из анализа проблем, потребностей и возможностей, анализировать соответствие требованиям существующих моделей, алгоритмизировать деятельность. **Владеет:** навыками анализа проблемной ситуации, разработки требований к модели, постановки цели, разработки концепции модели, стохастического и аналитического математического моделирования для решения прикладных задач.

3. Содержание дисциплины

Введение в искусственный интеллект. История искусственного интеллекта. Области применения искусственного интеллекта. Кейсы применения.

Анализ данных. Библиотеки numpy и pandas языка Python для обработки больших массивов данных. Очистка данных. Статистическое описание данных. Агрегирование данных. Визуализация данных.

Регрессия и классификация. Постановка задач. Линейная регрессия. Градиентный спуск, стохастический градиентный спуск. Логистическая регрессия. Линейная классификация. Проблема переобучения и методы борьбы с ней. Метрики качества алгоритмов.

Кластеризация. Постановка задачи. Алгоритм k-Means. Алгоритм DBSCAN. Коэффициент силуэта.

Нейронные сети. Постановка задачи. Функции активации и функции ошибки. Обучение нейронных сетей. Оптимизация параметров нейронной сети. Применение нейронных сетей.

Проблемы применения методов машинного обучения. Достоинства и недостатки моделей обработки данных методами машинного обучения. Модельный риск.

4. Тематический план

№ п/п	Разделы и темы	Всего часов	Формы занятий		
			лекции	лабораторные занятия	самостоятельная работа
1 семестр					
1.	Введение в искусственный интеллект	10	2	–	8
2.	Анализ данных	21	2	4	15
3.	Регрессия и классификация	23	4	6	15
4.	Кластеризация	21	2	2	15
5.	Нейронные сети	23	4	4	15
6.	Проблемы применения методов машинного обучения	10	2	–	8
Итого		108	16	16	76

5. Виды образовательной деятельности

Занятия лекционного типа

1. История искусственного интеллекта. Области применения искусственного интеллекта. Кейсы применения.
2. Библиотеки numpy и pandas языка Python для обработки больших массивов данных. Очистка данных. Статистическое описание данных. Агрегирование данных. Визуализация данных.
3. Постановка задач. Линейная регрессия. Градиентный спуск, стохастический градиентный спуск. Логистическая регрессия. Линейная классификация.
4. Проблема переобучения и методы борьбы с ней. Метрики качества алгоритмов.
5. Постановка задачи кластеризации. Алгоритм k-Means. Алгоритм DBSCAN. Коэффициент силуэта.
6. Нейронные сети. Однослойная и многослойная нейронная сеть. Функции активации и функции ошибки.
7. Обучение нейронных сетей. Оптимизация параметров нейронной сети. Применение нейронных сетей.
8. Достоинства и недостатки моделей обработки данных методами машинного обучения. Модельный риск.

Занятия семинарского типа

Лабораторная работа №1-2. *Анализ данных при помощи языка Python*

Теоретические вопросы

1. Как импортировать библиотеки в Python?
2. Какие библиотеки являются основными для обработки табличных данных?
3. Перечислите основные методы очистки данных.

Задания для аудиторной работы

1. Создайте Series из последовательности 15 значений, равномерно разбивающих отрезок [0, 20] (воспользуйтесь функцией linspace). Определите отношение элементов полученной серии к их предыдущим элементам. В результате необходимо получить среднее полученного вектора, оставив в нём только те значения, которые не более чем 1.5.
2. Для всех последующих заданий будем использовать обезличенные транзакционные банковские данные. Для этого считайте в переменные tr_mcc_codes, tr_types, transactions и gender_train из одноимённых таблиц из папки data.

   2.1. В tr_types выберите произвольные 100 строк с помощью метода sample (указав при этом random_seed равный 242). В полученной на предыдущем этапе подвыборке найдите долю наблюдений (стобец tr_description), в которой содержится подстрока 'плата' (в любом регистре).

   2.2. В датафрейме transactions задайте столбец customer_id в качестве индекса. Выделите клиента с максимальной суммой транзакции (то есть с максимальным приходом на карту). Найдите у него наиболее часто встречающийся модуль суммы приходов/расходов.
3. Соедините transactions с всеми остальными таблицами (tr_mcc_codes, tr_types, gender_train). Причём с customers_gender_train необходимо произвести связь с помощью left join, а с оставшимися датафреймами - через inner. После получения результата таблицы gender_train, tr_types, tr_mcc_codes можно удалить. В результате соединения датафреймов должно получиться 999584 строки.

   3.1. Определите модуль разницы между средними тратами женщин и мужчин (трата - отрицательное значение amount).

   3.2. Создайте новый столбец - mcc_code+tr_type, сконкатенировав значения из соответствующих столбцов. Оставьте только наблюдения с отрицательным значением amount. Посчитайте дисперсию по категориям получившегося столбца mcc_code+tr_type, в которых количество наблюдений >= 10. Определите отношение максимальной дисперсии к минимальной.

   3.3. Выделите из поля tr_datetime относительный день tr_day (первое число до точного времени). Отфильтруйте строки таким образом, чтобы оставить только те транзакции, у которых в соответствующий относительный день tr_day количество уникальных MCC кодов при транзакциях было больше 75 (можно воспользоваться функцией nunique()). Сгруппируйте полученный отфильтрованный датафрейм по MCC коду и полу.

Задания для самостоятельной работы

1. По клиенту получены зашумленные данные (объект s типа Series) по его транзакциям. Для заданного объекта s проделайте следующее: создайте новый Series, значения которого совпадают со значениями s, а индексы - целочисленные значения от 2 до 12, не включая 12; выберите из s элементы с индексами 3 и 5, после чего просуммируйте их, сохранив результат; Выберите из s только целочисленные элементы и вычислите их дисперсию.

   s = pd.Series(data=['1', 2, 3.1, 'hi!', 5, -512, 12.42, 'sber', 10.10, 98], index=range(6, 26, 2))

2. Для всех последующих заданий будем использовать обезличенные транзакционные банковские данные. Для этого считайте в переменные tr_mcc_codes, tr_types, transactions и gender_train из одноимённых таблиц из папки data.

   2.1. Для поля tr_type датафрейма transactions посчитайте частоту встречаемости всех типов транзакций tr_type в transactions.

   2.2. Найдите максимальную разницу между медианами суммы транзакций, посчитанными при заданных ниже условиях по полю amount из таблицы transactions: Медиана суммы транзакций; Медиана суммы транзакций по тем строкам, которые ни в одном из своих столбцов не содержат пустые значения; Медиана суммы транзакций по строкам, отсортированным по полю amount в порядке возрастания, и из которых удалены дублирующиеся по столбцам [mcc_code, tr_type] строки, причём при удалении соответстующих дублей остаются только последние из дублирующихся строк (keep='last').

3. Соедините transactions с всеми остальными таблицами (tr_mcc_codes, tr_types, gender_train). Причём с customers_gender_train необходимо произвести связь с помощью left join, а с оставшимися датафреймами - через inner. После получения результата таблицы gender_train, tr_types, tr_mcc_codes можно удалить. В результате соединения датафреймов должно получиться 999584 строки.

   3.1. По всем типам транзакций рассчитайте максимальную сумму прихода на карту (из строго положительных сумм по столбцу amount) отдельно для мужчин и женщин (назовите ее "max_income"). Оставьте по 5 транзакций для мужчин и для женщин, наименьших среди всех транзакций по полученным значениям "max_income".

   3.2. Измените тип поля tr_day на int. Выберите из transactions все MCC коды, которые встретились в выборке более чем 60000 раз. Сгруппируйте отфильтрованный датафрейм по дню и MCC-коду, получая средние значения суммы amount. Далее отрисуйте зависимость средних сумм (может пригодится метод unstack()) по каждому из MCC-кодов по дням.

Лабораторная работа №3-5. *Регрессия и классификация*

Теоретические вопросы

1. Какой математический аппарат лежит в основе решения задач регрессии?
2. Какие приложения имеют задачи регрессии?
3. Какой математический аппарат лежит в основе решения задач классификации?
4. Какие приложения имеют задачи классификации?
5. Почему задачи регрессии и классификации называют задачами обучения с учителем?

Задания для аудиторной работы

1. Загрузите данные из файла advertising.csv в объект pandas DataFrame. Создайте массивы NumPy X из столбцов TV, Radio и Newspaper и y - из столбца Sales. Используйте атрибут values объекта pandas DataFrame. Отмасштабируйте столбцы матрицы X, вычтя из каждого значения среднее по соответствующему столбцу и поделив результат на стандартное отклонение. Для определенности, используйте методы mean и std векторов NumPy (реализация std в Pandas может отличаться). Обратите внимание, что в numpy вызов функции .mean() без параметров возвращает среднее по всем элементам массива, а не по столбцам, как в pandas. Чтобы произвести вычисление по столбцам, необходимо указать параметр axis. Добавьте к матрице X столбец из единиц, используя методы hstack, ones и reshape библиотеки NumPy. Вектор из единиц нужен для того, чтобы не обрабатывать отдельно коэффициент $w0$ линейной регрессии.

2. Реализуйте функцию mserror - среднеквадратичную ошибку прогноза. Она принимает два аргумента - объекты Series y (значения целевого признака) и y_pred (предсказанные

значения). Не используйте в этой функции циклы - тогда она будет вычислительно неэффективной. Какова среднеквадратичная ошибка прогноза значений Sales, если всегда предсказывать медианное значение Sales по исходной выборке?

3. Напишите функцию stochastic_gradient_step, реализующую шаг стохастического градиентного спуска для линейной регрессии. Функция должна принимать матрицу X, вектора y и w, число train_ind - индекс объекта обучающей выборки (строки матрицы X), по которому считается изменение весов, а также число η (eta) - шаг градиентного спуска (по умолчанию eta=0.01). Результатом будет вектор обновленных весов. Наша реализация функции будет явно написана для данных с 3 признаками, но несложно модифицировать для любого числа признаков, можете это сделать.
4. Напишите функцию stochastic_gradient_descent, реализующую стохастический градиентный спуск для линейной регрессии. Функция принимает на вход следующие аргументы: X - матрица, соответствующая обучающей выборке, y - вектор значений целевого признака, w_init - вектор начальных весов модели, eta - шаг градиентного спуска (по умолчанию 0.01), max_iter - максимальное число итераций градиентного спуска (по умолчанию 10000), max_weight_dist - максимальное евклидово расстояние между векторами весов на соседних итерациях градиентного спуска, при котором алгоритм прекращает работу (по умолчанию 1e-8), seed - число, используемое для воспроизводимости сгенерированных псевдослучайных чисел (по умолчанию 42), verbose - флаг печати информации (например, для отладки, по умолчанию False). На каждой итерации в вектор (список) должно записываться текущее значение среднеквадратичной ошибки. Функция должна возвращать вектор весов w, а также вектор (список) ошибок. Запустите 100000 итераций стохастического градиентного спуска. Укажите вектор начальных весов w_init, состоящий из нулей. Оставьте параметры eta и seed равными их значениям по умолчанию. Теперь посмотрим на зависимость ошибки от номера итерации для всех итераций стохастического градиентного спуска. Алгоритм должен сходиться.
5. Для работы с алгоритмом линейной классификации и изучения метрик качества алгоритма, выполните задание в Google Colab: https://colab.research.google.com/drive/1btKsHQpAE4OQi9EiYqGwX5cWzv2GFq6k.

Задания для самостоятельной работы

1. Реализуйте функцию normal_equation, которая по заданным матрицам (массивам NumPy) X и y вычисляет вектор весов w согласно нормальному уравнению линейной регрессии. Какие продажи предсказываются линейной моделью с весами, найденными с помощью нормального уравнения, в случае средних инвестиций в рекламу по ТВ, радио и в газетах? (то есть при нулевых значениях масштабированных признаков TV, Radio и Newspaper). Напишите функцию linear_prediction, которая принимает на вход матрицу X и вектор весов линейной модели w, а возвращает вектор прогнозов в виде линейной комбинации столбцов матрицы X с весами w. Какова среднеквадратичная ошибка прогноза значений Sales в виде линейной модели с весами, найденными с помощью нормального уравнения?
2. Какова среднеквадратичная ошибка прогноза значений Sales в виде линейной модели с весами, найденными с помощью градиентного спуска?

Лабораторная работа №6. *Кластеризация*

Теоретические вопросы

1. Какой математический аппарат лежит в основе решения задач кластеризации?
2. Почему задачу кластеризации называют задачей обучения без учителя?
3. Какие приложения имеют задачи кластеризации?

Задания для аудиторной работы

1. Представим, что международное круизное агентство Carnival Cruise Line решило себя разрекламировать с помощью баннеров и обратилось для этого к вам. Чтобы протестировать, велика ли от таких баннеров польза, их будет размещено всего 20 штук по всему миру. Вам надо выбрать 20 таких локаций для размещения, чтобы польза была большой и агентство продолжило с вами сотрудничать. Агентство крупное, и у него есть несколько офисов по всему миру. Вблизи этих офисов оно и хочет разместить баннеры — легче договариваться и проверять результат. Также эти места должны быть популярны среди туристов. Для поиска оптимальных мест воспользуемся базой данных крупнейшей социальной сети, основанной на локациях — Foursquare. Часть открытых данных есть, например, на сайте archive.org: https://archive.org/details/201309_foursquare_dataset_umn. Теперь необходимо кластеризовать данные координаты, чтобы выявить центры скоплений туристов. Поскольку баннеры имеют сравнительно небольшую площадь действия, нам нужен алгоритм, позволяющий ограничить размер кластера и не зависящий от количества кластеров. Применяя алгоритмы k-Means и DBSCAN, выберите 20 локаций для размещения баннеров. С помощью коэффициента силуэта оцените оптимальность количества локаций.

Задания для самостоятельной работы

1. На задаче для аудиторной работы примените алгоритм кластеризации MeanShift.

Лабораторная работа №7-8. *Нейронные сети*

Теоретические вопросы

1. Какой математический аппарат лежит в основе нейронных сетей?
2. Что такое персептрон?
3. Что такое функция активации?

Задания для аудиторной работы

1. Для работы с нейронными сетями и их применении к задаче многоклассовой классификации, выполните задание в Google Colab: https://colab.research.google.com/drive/18ZKhfeHuarXHkkqR1uqMIDZz7WDvSygi.
2. Для работы с нейронными сетями и их применении для генерации текста, выполните задание в Google Colab: https://colab.research.google.com/drive/1stX85TonfftftY10WRrNIvc1d5YcB9wJ.

Задания для самостоятельной работы

3. Для работы с нейронными сетями для анализа изображений и текстов, выполните задание в Google Colab: https://colab.research.google.com/drive/1IB5j5dIblvs1QAD2LHZJafFBP8JcSr4Q.

6. Критерии оценивания результатов освоения дисциплины (модуля)

6.1. Оценочные средства и критерии оценивания для текущей аттестации

1. Нормы оценивания каждой лабораторной работы:

№п/п	Структурная часть работы	Количество баллов (*)
1	Ответ на теоретические вопросы по теме лабораторной работы	1 балл
2	Демонстрация выполнения конкретного задания, предложенного для самостоятельного решения к лабораторной работе	2 балла

(*) с возможностью градации до 0,25 балла.

2. Шкала оценивания. Оценка «зачтено» за лабораторную работу выставляется, если набрано не менее 3 баллов, в противном случае за работу выставляется «не зачтено».

6.2. Оценочные средства и критерии оценивания для промежуточной аттестации

Промежуточная аттестация включает зачет.

Образец задания к зачету

Дан датасет, необходимо реализовать на нём один из методов машинного обучения и оценить качество работы алгоритма.

Зачет выставляется по результатам работы студента в течение семестра согласно Положению о текущем контроле успеваемости и промежуточной аттестации студентов в федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Смоленский государственный университет» (утверждено приказом и.о. ректора № 01-113 от 26.09.2019 г., внесены дополнения приказом ректора № 01-48 от 30.04.2020 г.).
Для получения зачета студент должен:

- уметь отвечать на теоретические вопросы, рассмотренные на лекциях;
- уметь решать задачи, предложенные на лабораторных занятиях.

7. Перечень основной и дополнительной учебной литературы

7.1. Основная литература

1. Платонов, А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. — Москва : Издательство Юрайт, 2022. — 85 с. — (Высшее образование). — ISBN 978-5-534-15561-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://www.urait.ru/bcode/508804.
2. Миркин, Б. Г. Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — Москва : Издательство Юрайт, 2022. — 174 с. — (Высшее образование). — ISBN 978-5-9916-5009-0. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://www.urait.ru/bcode/469306.
3. Анализ данных : учебник для вузов / В. С. Мхитарян [и др.] ; под редакцией В. С. Мхитаряна. — Москва : Издательство Юрайт, 2022. — 490 с. — (Высшее образование). — ISBN 978-5-534-00616-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://www.urait.ru/bcode/489100.
4. Воронов, М. В. Системы искусственного интеллекта : учебник и практикум для вузов / М. В. Воронов, В. И. Пименов, И. А. Небаев. — Москва : Издательство Юрайт, 2022. — 256 с. — (Высшее образование). — ISBN 978-5-534-14916-6. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://www.urait.ru/bcode/485440.

7.2 Дополнительная литература.

1. Дж. Вандер Плас. Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с.: ил. — (Серия «Бестселлеры O’Reilly»).
2. Харрисон, Мэтт. Машинное обучение: карманный справочник. Краткое руководство по методам структурированного машинного обучения на Python.: Пер. с англ. - СПб. : ООО "Диалектика", 2020 - 320 с. : ил. - Парал. тит. англ.

7.3. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. Система дистанционного обучения СмолГУ https://cdo.smolgu.ru
2. Национальная платформа открытого образования https://openedu.ru
3. Яндекс-практикум https://practicum.yandex.ru
4. СберУниверситет https://partner.sberuniversity.online/
5. Открытый курс машинного обучения https://ods.ai/
6. Онлайн-лаборатория Google Colaboratory https://colab.research.google.com

8. Материально-техническое обеспечение

Учебная аудитория для проведения занятий лекционного типа, оснащенная стандартной учебной мебелью, интерактивной доской, мультимедиапроектором, ноутбуком и колонками.

Учебная аудитория для проведения занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная стандартной учебной мебелью, компьютерами с выходом в Интернет.

Помещение для самостоятельной работы – компьютерный класс с доступом к сети «Интернет» и ЭИОС СмолГУ.

9. Программное обеспечение

Для осуществления образовательного процесса по дисциплине используется Информационно-вычислительный центр физико-математического факультета.

При осуществлении образовательного процесса по дисциплине используются:

1.Система дистанционного обучения СмолГУ. URL: http://www.cdo.smolgu.ru. (СДО Русский Moodle 3KL Norm с техническим обслуживанием, Акт на передачу прав №УТДЮ0001785 от 06.12.2016)

2. Microsoft Open License (Windows XP, 7, Office 2003-2016) - Лицензия 66975477 от 03.06.2016 – в составе: ОС Windows, MS Excel 2003/2007.

Anaconda (дистрибутив Python), бесплатная лицензия.