ФЕДЕРАЛЬНОЕ АГЕНТСТВО ВОЗДУШНОГО ТРАНСПОРТА
(РОСАВИАЦИЯ)

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ ГРАЖДАНСКОЙ АВИАЦИИ» (МГТУ ГА)

Кафедра вычислительных машин, комплексов, систем и сетей

А.А. Егорова

ОСНОВЫ РАБОТЫ С БОЛЬШИМИ ДАННЫМИ (DATA SIENCE)

Учебно-методическое пособие
по выполнению лабораторных работ

для студентов I курса
направления 09.03.01
очной формы обучения

Москва
ИД Академии Жуковского
2023

УДК 004.6
ББК 6Ф7.3
Е30

Рецензент:
Феоктистова О.Г. – д-р техн. наук, доцент

Егорова А.А.
Е30 Основы работы с большими данными (Data sience) [Текст] : учебно-методическое пособие по выполнению лабораторных работ / А.А. Егорова. – М.: ИД Академии Жуковского, 2023. – 48 с.

Учебно-методическое пособие издается в соответствии с рабочей программой учебной дисциплины «Основы работы с большими данными (Data sience)» по учебному плану для студентов I курса очной формы обучения.

Рассмотрено и одобрено на заседаниях кафедры 20.06.2023 г. и методического совета 27.06.2023 г.

УДК 004.6
ББК 6Ф7.3

В авторской редакции

Подписано в печать 16.10.2023 г.
Формат 60x84/16 Печ. л. 3 Усл. печ. л. 2,79
Заказ № 985/0621-УМП09 Тираж 30 экз.

Московский государственный технический университет ГА
125993, Москва, Кронштадтский бульвар, д. 20

Издательский дом Академии имени Н. Е. Жуковского
125167, Москва, 8-го Марта 4-я ул., д. 6А
Тел.: (495) 973-45-68
E-mail: zakaz@itsbook.ru

ПРЕДИСЛОВИЕ

Настоящее пособие содержит задания на выполнение лабораторных работ по дисциплине «Основы работы с большими данными (Data Sience)», выполняемых студентами 1 курса направления подготовки 09.03.01 «Информатика и вычислительная техника» направленности «Интеллектуальные системы обработки и анализа данных» в двух семестрах.

Навыки, приобретенные в процессе выполнения лабораторных работ, необходимы студентам в процессе дальнейшей подготовки по ряду дисциплин, а также для самостоятельной работы и самоподготовки, в том числе для подготовки к экзамену.

В пособии отражены организационно-методические аспекты выполнения лабораторных работ, особенности их проведения, цели, достигаемые в процессе выполнения лабораторных работ, формируемые компетенции.

Пособие охватывает дисциплину не полностью, а только материал, отрабатываемый на лабораторных работах, но лабораторные работы представлены по обоим семестрам.

Пособие содержит цель, задание на выполнение каждой лабораторной работы, требования к их выполнению, краткие теоретические сведения, примеры выполнения и контрольные вопросы, а также варианты (темы) для выполнения лабораторных работ, которые относятся к нескольким лабораторным работам.

Теоретический материал, содержащийся в пособии, не претендует на полноту, а содержит только необходимую для выполнения работы информацию. Требования к выполнению работ являются обязательными в рамках выполнения работ по дисциплине «Основы работы с большими данными (Data Science)», сформулированы в соответствии с методикой преподавания дисциплины.

Настоящее пособие может быть использовано и как справочник при самостоятельной работе при программировании прикладных задач, в том числе по другим дисциплинам.

Пособие имеет прикладной характер, что способствует формированию у студентов компетенций в соответствии с требованиями, содержащимися в рабочей программе по дисциплине «Основы работы с большими данными (Data Sience)», и в целом соответствующие модели компетенций по направлению подготовки «Информатика и вычислительная техника» направленности «Интеллектуальные системы обработки и анализа данных».

Оглавление

1. Введение

Big Data или большие данные — это структурированные или неструктурированные массивы данных большого объема, которые обрабатывают при помощи специальных программных инструментов, чтобы использовать для статистики, анализа, прогнозов, принятия решений и т.п.

Термин «большие данные» предложил редактор журнала Nature Клиффорд Линч в спецвыпуске 2008 года. Он говорил о взрывном росте объемов информации в мире. К большим данным Линч отнес любые массивы неоднородных данных более 100 Гб в сутки, но единого критерия до сих пор не существует. Однако определены основные характеристики больших данных (иногда называют правилом «трёх V»):

- Volume — объем данных: от 100 Гб в сутки (или от 150);
- Velocity — скорость накопления и обработки массивов данных; большие данные обновляются регулярно, поэтому необходимы интеллектуальные технологии для их обработки в режиме онлайн;
- Variety — разнообразие типов данных; данные могут быть структурированными, неструктурированными или структурированными частично (например, в соцсетях поток данных не структурирован: это могут быть текстовые посты, фото или видео).

В наши дни к этим трем добавляют еще три признака (или один/два из трех):

- Veracity — достоверность как самого набора данных, так и результатов его анализа;
- Variability — изменчивость; у потоков данных бывают свои пики и спады под влиянием сезонов или социальных явлений; чем нестабильнее и изменчивее поток данных, тем сложнее его анализировать;
- Value — ценность или значимость; как и любая информация, большие данные могут быть простыми или сложными для восприятия и анализа; пример простых данных — это посты в соцсетях, сложных — банковские транзакции.

Настоящее пособие, предназначенное для выполнения лабораторных работ по дисциплине «Основы работы с большими данными (Data Science)» (1 и 2 семестры) с использованием наиболее распространенных и простых инструментов, а именно табличного процессора и однопользовательской СУБД.

Целью проведения лабораторных работ является как закрепление основных теоретических положений, изложенных в лекциях, так и получение практических навыков по извлечению и подготовке данных с помощью Microsoft Excel и Microsoft Access, предварительной обработке также средствами Microsoft Excel, а также получение студентами общего представления об основных направлениях обработки больших данных.

Пособие необходимо студентам на всех этапах, начиная от подготовки до оформления отчета и защиты лабораторной работы.

2. Организационно-методические рекомендации

В соответствии с учебным планом подготовки студентов по направлению 09.03.01 «Информатика и вычислительная техника» (бакалавриат), направленность «Интеллектуальные системы обработки и анализа данных» и рабочей программой по дисциплине «Основы работы с большими данными (Data Science)» и изложенными в них требованиями к уровню подготовки бакалавров для работы в организациях гражданской авиации, студенты должны обладать навыками:

- поверки, оценки используемых моделей больших данных;
- сбора и подготовки для исследования больших данных.

В соответствии с учебной программой продолжительность лабораторных работ – 20 часов (4 лабораторные работы по 4 часа, 2 - в первом семестре, 2- во втором семестре и 2 работы по 2 часа – также в первом и во втором семестрах).

2.1. Компетенции обучающегося, формируемые в результате освоения дисциплины «Основы работы с большими данными (Data Science)»

В результате изучения дисциплины студенты должны знать:

- Содержание и последовательность выполнения этапов аналитического проекта;
- Предметную область анализа;
- Современный опыт использования анализа больших данных;

и уметь:

- Планировать аналитические работы с использованием технологий больших данных;
- Разрабатывать и оценивать модели больших данных.

Дисциплина Основы работы с большими данными (Data Science) направлена на обеспечение студентов необходимыми знаниями, умениями и навыками при освоении следующих дисциплин:

- Алгоритмы обработки и анализа больших данных;
- Методы машинного обучения и нейронные сети;
- Методы и модели классификации Big Data;
- Системы искусственного интеллекта;
- Модели и технологии распределенных вычислений;
- Автоматизированные системы обработки информации в ГА;
- Виртуальная и дополненная реальность в системах ГА (VR и AR технологии).

2.2. Перечень тем лабораторных работ дисциплины

ЛР - 1. Извлечение данных с помощью электронных таблиц (4 часа).
ЛР - 2. Применение СУБД для извлечения данных (4 часа).
ЛР - 3. Поддержка, достоверность, лифт (2 часа).
ЛР - 4. Прогнозирование МВР (4 часа).
ЛР - 5. Разделяющая полоса (4 часа).
ЛР - 6. Программирование генетического алгоритма (2 часа).

2.3. Этапы выполнения лабораторных работ

Выполнение лабораторной работы можно разбить на следующие этапы:

- домашняя подготовка;
- выполнение работы на ПК в соответствии со своим вариантом задания;
- сдача выполненной работы преподавателю;
- оформление отчета;
- защита лабораторной работы.

В процессе домашней подготовки студент изучает лекционный материал, внимательно изучает задание и свой вариант, готовит при необходимости файлы с исходными данными.

Выполнение работы производится во время занятий в компьютерном классе МГТУ ГА в присутствии преподавателя. По окончании работы студент демонстрирует преподавателю выполненную работу и отвечает на вопросы преподавателя. В случае критических замечаний студент устраняет их и демонстрирует работу повторно.

Зачет по лабораторной работе выставляется преподавателем после предоставления оформленного отчета и опроса студента по теоретическому материалу по теме работы.

2.4. Отчет по лабораторной работе

По окончании лабораторной работы студент оформляет отчет, который включает:

- номер работы;
- название работы;
- цель работы;
- задание на выполнение работы;
- вариант выполнения работы;
- последовательное выполнение работы с пояснениями (комментариями), скриншотами, подтверждающими выполнение каждого этапа и промежуточными результатами;
- результат выполнения работы с комментариями;
- выводы по работе.

Скриншоты нужно делать не полного экрана, а фрагмента, захватив актуальную информацию. Размеры букв в распечатке должны быть такими, чтобы текст было удобно читать, не поднимая лист со стола (например, не менее 10 пт для шрифта Times New Roman). Если в отчет вставлены рисунки, выполненные «рукой» (или фотографии), то они должны быть аккуратными, использование линейки обязательно.

Название дисциплины, фамилия и группа студента оформляются на титульном листе (форма в Приложении1). На титульном листе также указывается название вуза, кафедры и фамилия преподавателя.

2.5. Защита лабораторной работы

После сдачи выполненной работы преподавателю и оформления отчета студент защищает работу (на следующем занятии). В процессе защиты преподаватель проверяет правильность оформления отчета, соответствие работы заданию, корректность исходных данных и соответствие результатов выполненной работе. Студент отвечает на вопросы преподавателя как по выполнению работы, так и по теоретической части работы (цель работы и контрольные вопросы являются ориентиром).

3. Лабораторная работа №1

Извлечение данных с помощью электронных таблиц

Продолжительность выполнения работы – 4 часа.

3.1. Цель работы

Целью лабораторной работы является приобретение навыков по:

- сбору и структурированию данных для анализа предметной области,
- обработке данных для улучшения результата,
- выбору атрибутов (признаков) для изучения закономерностей,
- формулированию гипотез.

3.2. Задание на выполнение работы

В соответствии со своим вариантом для заданной предметной области выполнить следующее:

1. Создать модель предметной области в виде таблицы в Excel (не менее 15 признаков).
2. Наполнить модель данными (не менее 20 строк).
3. Получить статистическую информацию на основе введенных данных (суммы, среднее значение, минимальные, максимальные по всему набору и на заданной выборке).

4. Выдвинуть 5-7 гипотез о зависимостях в данной предметной области
5. Сформировать 3-4 переформатированных набора данных с использованием дополнительных переменных для проверки гипотез.
6. Проверить гипотезы путем построения графиков/диаграмм.

3.3. Требования к выполнению

- Набор данных может быть вымышленным.
- «Искусственность» признаков для предметной области допустима;
- Тема может быть студентом изменена, но только по согласованию с преподавателем или студенты могут темами меняться;
- Для более гармоничных гипотез рекомендуется смоделировать процесс и владельца (адресата) аналитического отчета.

3.4. Краткие теоретические сведения

3.4.1. Роль данных

Подготовка и обработка данных – это первый шаг и один из четырех ключевых шагов в исследовании с применением Data Science. Причем одну из главных релей (если не главную) в таком исследовании играют сами данные. При низком качестве данных результаты даже самого прекрасного анализа могут быть весьма далеки от совершенства и не принести ту пользу, которая от них ожидалась (или даже принести вред, например, при принятии решений, которые будут основаны на ложных гипотезах). Таким образом, важность данных заключается в следующем:

- Данные — это основа для любого анализа, моделирования или прогнозирования. Без данных не существует никакого обоснования для принятия каких-либо решений.
- Данные используются для определения лучших практик в различных отраслях, с оптимизацией производства и повышением эффективности работы.
- Данные помогают определить тенденции, обнаружить отклонения и прогнозировать будущие события.
- Данные используются для разработки и применения алгоритмов машинного обучения и искусственного интеллекта.
- Хранение, обработка и управление данными являются наиважнейшей частью Data Science, и использование обоснованных и корректных инструментов является необходимым для управления портфелями данных и повышения достоверности их анализа.

Таким образом, данные играют ключевую роль в Data Science, потому что они являются основным элементом для повышения качества прогнозирования и принятия обоснованных решений.

3.4.2. Формат данных

Обычно для анализа данных используют табличное представление, где каждая строка представляет собой элемент данных с описанием отдельного наблюдения (измерения), а каждый столбец несет переменную для его описания, которая также может называться характеристика, атрибут, параметр, признак или размерность.

Например, вымышленная таблица с информацией о покупках покупателей может иметь вид:

ID	Покупатель	Дата	Кол-во фруктов	Куплена рыба	Потрачено, $
1	Пингвин	1 янв.	1	да	5,3
2	Медведь	1 янв.	4	да	9,7
3	Кролик	1 янв.	6	Нет	6,5
4	Лошадь	2 янв.	6	нет	5,5
5	Пингвин	2 янв.	2	Да	6,0
6	Жираф	3 янв.	5	Нет	4,8
7	Кролик	3 янв.	8	Нет	7,6
8	Кот	3 янв.	?	да	7,4

В зависимости от цели можно изменить представленный в строках тип наблюдений. Например, выборка в таблице позволяет изучать закономерности, рассматривая покупки.

Но если вместо этого мы хотим исследовать закономерности покупок в зависимости от дня, то нужно представить в строках общий итог. Для всестороннего анализа имеет смысл также добавить новые переменные, такие как погода. Такие таблицы называются переформатированными (или переформатированный набор данных).

Дата	Выручка, $	Число покупателей	Погода	Выходные
1 янв.	21,5	3	солнце	да
2 янв.	11,5	2	дождь	нет
3 янв.	19,8	3	солнце	нет

3.4.3. Типы переменных

Есть четыре главных типа переменных. К ним применимы разные алгоритмы, поэтому важно понимать разницу.

- ***Бинарная***. Это простейший тип переменных только с двумя вариантами значения. Например, в первой таблице бинарная переменная показывает, брал ли покупатель рыбу.
- ***Категориальная***. Если вариантов больше двух, информация может быть представлена категориальной переменной. Например, категориальной переменной можно представить погоду.
- ***Целочисленная***. Такой тип используется, когда информация может быть представлена целым числом. В первой таблице целое число выражает количество купленных каждым покупателем фруктов.
- ***Непрерывная (количественная).*** Это самая подробная переменная. Она содержит числа со знаками после запятой, например, такие переменные показывают количество потраченных покупателем денег.

Хотя в первоначальном наборе данных может быть много разных переменных, применение в алгоритме слишком большого их числа ведет к замедлению вычислений или к ошибочным предсказаниям из-за информационного шума. Поэтому следует остановиться на коротком списке важнейших переменных.

Выбор переменных часто делается методом проб и ошибок. Их имеет смысл добавлять и убирать, учитывая промежуточные результаты. Для начала можно использовать простые графики для выявления корреляций (зависимостей) между переменными, отбирая кажущиеся самыми перспективными для дальнейшего анализа.

3.4.4. Конструирование признаков

Иногда хорошие переменные нужно сконструировать. Например, если мы хотим предсказать, кто из покупателей не будет брать рыбу, то можем посмотреть на переменную «Покупатель» (их вид), заключив, что кролики, лошади и жирафы рыбу не покупают. А если мы сгруппируем виды покупателей в более широкие категории — травоядных, хищников и всеядных, — то получим более универсальный вывод: травоядные рыбу не берут.

Вместо переформатирования одной переменной можно скомбинировать их методом, называемым уменьшением размерности. Уменьшение размерности может использоваться для извлечения самой полезной информации и ее выражения в небольшом наборе переменных для дальнейшего анализа.

3.4.5. Неполные данные

Мы не всегда располагаем полными данными. Например, в первой таблице количество фруктов в последней покупке неизвестно. Неполные данные мешают анализу и при любой возможности с ними нужно разобраться одним из известных способов:

- ***Приближение***. Если пропущено значение бинарного или категориального типа, его можно заменить самым типичным значением (модой) переменной. А для целочисленных или непрерывных переменных используется медиана (среднее значение). Применение этого метода к первой таблице позволит предположить, что кот приобрел 5 фруктов, поскольку, согласно остальным семи записям, именно таково среднее число покупаемых фруктов.
- ***Вычисление***. Пропущенные значения также могут быть вычислены с применением более продвинутых алгоритмов обучения с учителем. Хотя такие вычисления требуют времени, они обычно приводят к более точным оценкам неполных значений. Причина в том, что вместо приближения к самому распространенному значению они оценивают значение по сходным записям. Например, мы видим, что если покупатели берут рыбу, они склонны приобретать меньше фруктов, а это значит, что кот должен был купить 3–4 фрукта.
- ***Удаление***. В качестве последнего средства строки с неполными значениями могут быть удалены. Тем не менее этого обычно избегают, чтобы не уменьшать объем данных, доступных для анализа. Более того, исключение элементов данных может привести к искаженным результатам в отношении отдельных групп. Например, коты могут менее охотно, чем другие, раскрывать информацию о количестве приобретаемых фруктов. Если мы удалим такие покупки, коты будут недостаточно представлены в итоговой выборке.

После того как набор данных обработан, можно заняться его анализом.

3.4.6. Гипотезы

Гипотеза - это предварительное утверждение или объяснение, которое предлагается в качестве возможного ответа на научный вопрос или наблюдение. Она служит отправной

точкой для дальнейшего исследования и является важным компонентом научного метода. По сути, гипотеза - это предложение, которое выдвигается для изучения или проверки. В научных исследованиях формулировка гипотезы является неотъемлемым шагом в процессе получения знаний. Она позволяет ученым предлагать и оценивать потенциальные объяснения наблюдаемых явлений или исследовать взаимосвязи между переменными. Гипотезу можно рассматривать как обоснованное предположение или предсказание, которое стремится обеспечить первоначальное понимание исследуемого явления.

Хорошо построенная гипотеза должна основываться на существующих знаниях и быть проверяемой с помощью экспериментов или наблюдений. Она должна быть сформулирована четко, кратко и конкретно, чтобы облегчить разработку экспериментов или методов сбора данных, которые могут либо подтвердить, либо опровергнуть гипотезу.

Гипотеза обладает несколькими ключевыми характеристиками, которые отличают ее от простого предположения или догадки. К этим характеристикам относятся: проверяемость, фальсифицируемость, ясность и конкретность, соответствие существующим знаниям. Рассмотрим их подробнее.

- ***Проверяемость***: гипотеза должна быть проверяемой и способной быть подтвержденной или опровергнутой с помощью экспериментальных данных. Она должна быть сформулирована таким образом, чтобы обеспечить возможность проведения экспериментов или сбора данных для оценки ее достоверности.
- ***Фальсифицируемость***: гипотеза должна быть фальсифицируемой, что означает, что возможно представить эксперимент или наблюдение, которые могли бы доказать ее ошибочность. Этот критерий имеет решающее значение для обеспечения того, чтобы научные гипотезы подвергались эмпирической проверке и могли быть уточнены или отвергнуты на основе фактических данных.
- ***Ясность и конкретность***: гипотеза должна быть четко сформулирована и конкретна, определяя исследуемые переменные и взаимосвязи. Это помогает в разработке целенаправленного плана исследования и облегчает интерпретацию результатов.
- ***Соответствие существующим знаниям***: гипотеза должна основываться на глубоком понимании предыдущих исследований и существующих теорий в данной области. Она должна основываться на предшествующих знаниях и быть направлена на расширение или уточнение существующих теорий.

Процесс проверки гипотезы включает в себя эмпирическое исследование гипотезы для определения ее достоверности. Обычно это включает в себя разработку экспериментов, проведение наблюдений или анализ существующих данных для сбора доказательств в поддержку или против гипотезы. Результаты этих исследований способствуют накоплению научных знаний и могут привести к принятию, отклонению или модификации первоначальной гипотезы.

Проверка гипотезы часто включает в себя формулирование нулевых и альтернативных гипотез. Нулевая гипотеза представляет собой точку зрения по умолчанию или преобладающую точку зрения, предполагающую отсутствие существенной взаимосвязи или эффекта, в то время как альтернативная гипотеза предлагает конкретную взаимосвязь или эффект, которые противоречат нулевой гипотезе.

Статистический анализ обычно используется для определения вероятности того, что наблюдаемые данные подтверждают нулевую гипотезу или благоприятствуют альтернативной гипотезе.

Гипотезы используются в различных областях исследования, включая естественные науки, социальные науки и гуманитарные науки. В естественных науках гипотезы часто

формулируются для объяснения явлений в физике, химии, биологии и других дисциплинах. В социальных науках гипотезы могут исследовать взаимосвязи между переменными в психологии, социологии, экономике и политологии. Даже в гуманитарных науках исследователи могут выдвигать гипотезы для изучения исторических событий, культурных феноменов или литературных интерпретаций.

Итеративный характер научного исследования означает, что гипотезы подлежат пересмотру и уточнению по мере появления новых доказательств. Хорошо подтвержденная гипотеза может в конечном итоге стать частью более широкой теории, которая обеспечивает более всеобъемлющую основу для понимания конкретного явления. Гипотезы играют решающую роль в научном методе, направляя исследователей в их стремлении понять окружающий нас мир. Они служат строительными блоками научного исследования, обеспечивая основу для разработки экспериментов, сбора данных и анализа. Благодаря тщательной проверке гипотезы способствуют расширению знаний и разработке теорий, объясняющих сложные явления.

3.5. Контрольные вопросы (темы)

1. Виды знаний и способы их представления
2. Модели представления данных и знаний (типы, назначение, способы применения)
3. Модели извлечения данных
4. Методы извлечения данных
5. Базы знаний (и их отличие от баз данных)
6. Инженер знаний (функции, способы работы)
7. Типовой процесс приобретения знаний
8. Проблемы при извлечении знаний
9. Идеальный эксперт (кто такой эксперт, каковы его функции, проблемы использования экспертных знаний)
10. Свойства идеального эксперта
11. Методы объединения знаний
12. Теоретические аспекты извлечения данных
13. Что такое гипотеза
14. Сформулируйте свойства (характеристики) гипотезы

4. Лабораторная работа №2

Применение СУБД для извлечения данных.

Продолжительность выполнения работы – 4 часа.

4.1. Цель работы

Целью лабораторной работы является приобретение навыков по:
- экспорту/импорту данных из/в СУБД ACCESS,
- обработке данных в СУБД ACCESS,
- формированию признаков для анализа данных и формулирования гипотез,
- графического представления данных в СУБД ACCESS.

4.2. Задание на выполнение работы

В соответствии с выполненной лабораторной работой №1 для предметной области (своего варианта) позволяет выполнить следующее:

1. Импортировать данные в Access (одну исходную таблицу).

2. Получить по исходной таблице итоговые значения (сумму, количество, среднее, минимальное и максимальное значение, дисперсию стандартное отклонение).
3. Выполнить конструирование запросов (7-10) для формирования таблиц, которые можно использовать для проверки гипотез. Среди них должны быть:
   - Выборка по константе,
   - Выборка по переменной (с параметром),
   - Группировка по переменной,
   - Использование агрегирующих функций (сумма, среднее значение и т.п.).
4. По одной из переформатированных таблиц (п.3) построить графики (круговую диаграмму, гистограмму).
5. Экспортировать результат обработки (переформатированную таблицу) в Excel и дополнить ее переменными (использовать дополнительные переменные).

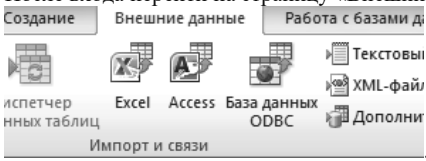
4.3. Требования к выполнению работы

- Импортируемая таблица должна быть одна (проектирование базы данных не выполняется).
- Использование переменных и агрегирующих функций с группировкой обязательно, при этом они подбираются самостоятельно, исходя из логики предметной области и состава исходной таблицы.
- Графиков (круговых диаграмм, гистограмм) должно быть не менее двух (можно использовать разные таблицы).

4.4. Краткие теоретические сведения

4.4.1. Импорт данных

После входа перейти на страницу «Внешние данные» и выбрать Excel



.

Далее следовать указаниям системы, выбрав файл (подготовленный на ЛР №1) из определенной папки.

Файл

	A	B	C	D	E	F
1	ID	Покупатель	Даты	Фрукты	Рыба	Потрачено
2	1	Пингвин	01.янв	1	да	5,3
3	2	Медведь	01.янв	4	да	9,7
4	3	Кролик	01.янв	6	нет	6,5
5	4	Лошадь	02.янв	6	нет	5,5
6	5	Пингвин	02.янв	2	да	6,1
7	6	Жираф	03.янв	5	нет	4,8
8	7	Кролик	03.янв	8	нет	7,6
9	8	Кот	03.янв		да	7,4
10						

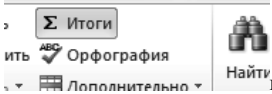
Превратится в таблицу:

Лист1

Код	ID	Покупатель	Даты	Фрукты	Рыба	Потрачено	Ще
1	1	Пингвин	01.01.2022	1	да	5,3	
2	2	Медведь	01.01.2022	4	да	9,7	
3	3	Кролик	01.01.2022	6	нет	6,5	
4	4	Лошадь	02.01.2022	6	нет	5,5	
5	5	Пингвин	02.01.2022	2	да	6,1	
6	6	Жираф	03.01.2022	5	нет	4,8	
7	7	Кролик	03.01.2022	8	нет	7,6	
8	8	Кот	03.01.2022		да	7,4	
(№)							

Столбец ID здесь явно лишний. Его «отсечь» можно на этапе загрузки, если использовать ID из Excel-файла в качестве кода.

Далее, используя итоговые функции на главной странице

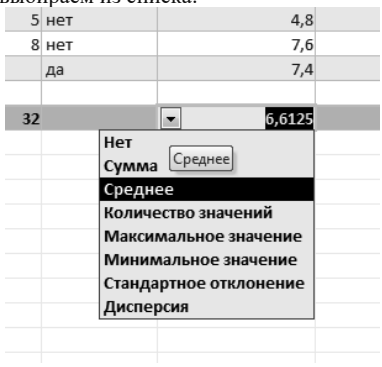


получаем:

Лист1

Код	ID	Покупатель	Даты	Фрукты	Рыба	Потрачено	Щелк
1	1	Пингвин	01.01.2022	1	да	5,3	
2	2	Медведь	01.01.2022	4	да	9,7	
3	3	Кролик	01.01.2022	6	нет	6,5	
4	4	Лошадь	02.01.2022	6	нет	5,5	
5	5	Пингвин	02.01.2022	2	да	6,1	
6	6	Жираф	03.01.2022	5	нет	4,8	
7	7	Кролик	03.01.2022	8	нет	7,6	
8	8	Кот	03.01.2022		да	7,4	
(№)							
Итог				32		2,52125	

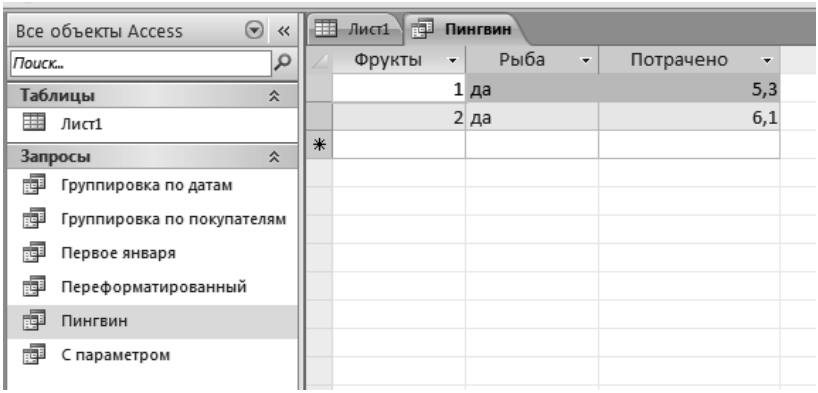
Разные итоговые функции выбираем из списка:



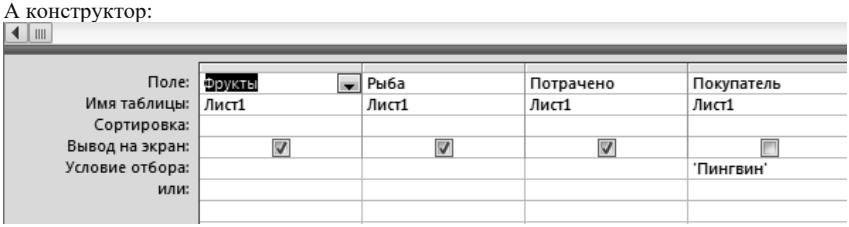
4.4.2. Запросы

1. Например, все покупки Пингвина

Результат запроса выглядит следующим образом:

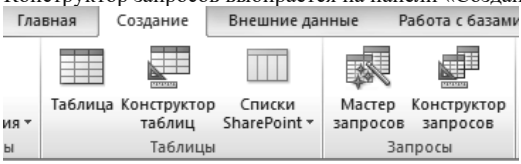


А конструктор:

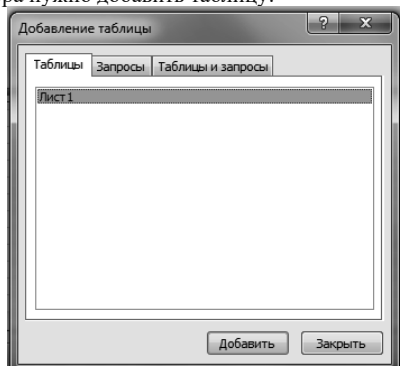


«Покупатель» не выводится (нет «галочки»).

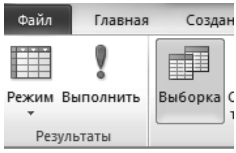
Конструктор запросов выбирается на панели «Создание»:



После запуска конструктора нужно добавить таблицу:



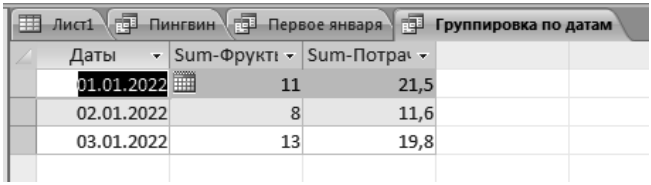
А запрос после его формирования нужно запустить кнопкой «Выполнить»:



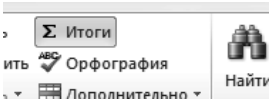
Запрос «Первое января» аналогичен, только условие отбора – даты.

2. Запрос «Группировка по датам»

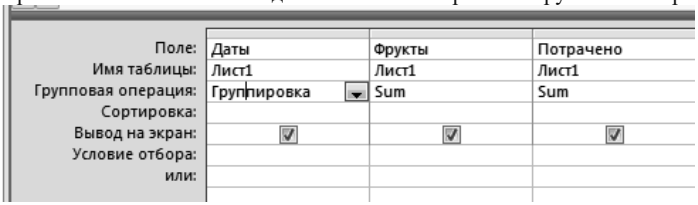
Результат:



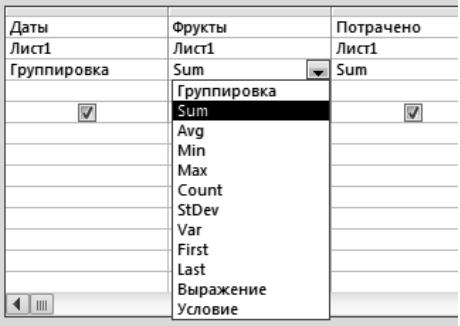
В конструкторе использована та же итоговая функция



При ее нажатии появляется дополнительная строка – «Групповая операция».



Далее в нужном столбце можем выбрать операцию:



Функции можно выбирать для каждого столбца (т.е. разные).

Например, аналогичный запрос с группировкой по покупателям использует кроме суммы еще и COUNT (подсчет количества строк, т.е. в нашем случае количество посещений магазина каждым покупателем).

Лист1 | Первое января | **Группировка по покупателям**

Покупатель	Sum-Фрукты	Sum-Потрач	Count-ID
Жираф	5	4,8	1
Кот		7,4	1
Кролик	14	14,1	2
Лошадь	6	5,5	1
Медведь	4	9,7	1
Пингвин	3	11,4	2

3. Запрос с параметром позволяет получать данные через переменную.

В нашем случае - получать данные на любого покупателя.
Результат:

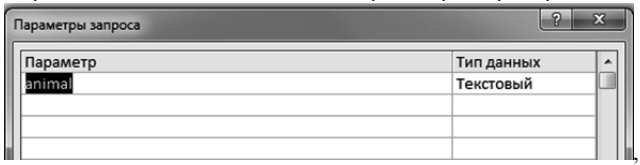
Лист1 | Первое января | **С параметром**

Покупатель	Фрукты	Рыба	Потрачено
Кролик	6	нет	6,5
Кролик	8	нет	7,6
*			

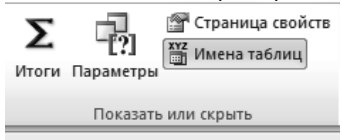
Конструктор:

Поле:	Покупатель	Фрукты	Рыба	Потрачено
Имя таблицы:	Лист1	Лист1	Лист1	Лист1
Сортировка:				
Вывод на экран:	☑	☑	☑	☑
Условие отбора:	[animal]			
или:				

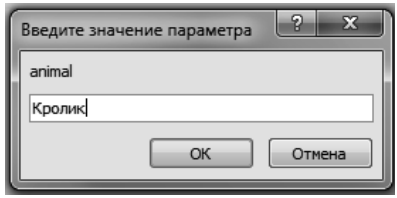
Обратите внимание на «Условие отбора». Параметр запроса вводится через:



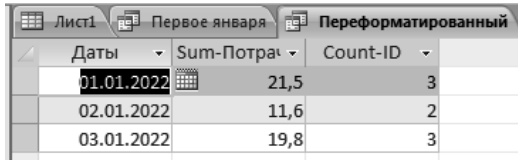
предварительно вызванный с помощью кнопки «Параметры»



При запуске такого запроса нужно вводить значение параметра (в данном случае был введен «Кролик»).

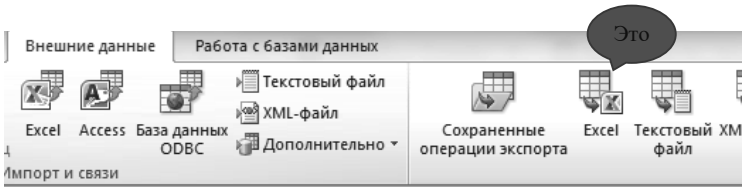


4. Запрос «Переформатированный» - это 3 столбца, которые получились из основной таблицы.



Даты	Sum-Потрач	Count-ID
01.01.2022	21,5	3
02.01.2022	11,6	2
03.01.2022	19,8	3

Сам запрос – еще одна группировка по датам, использованы другие столбцы, выполняется аналогично. Но именно эту таблицу экспортируем в Excel, выбрав опять на вкладке «Внешние данные» для экспорта Excel:

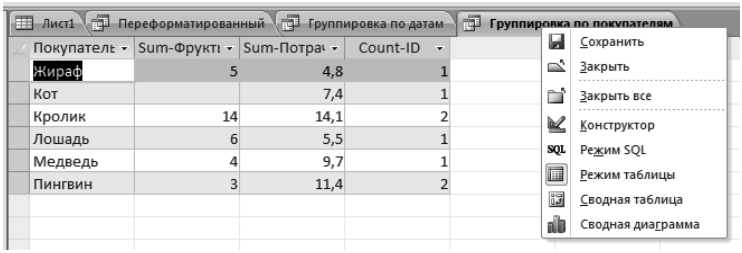


Получаем 3 столбца с данными, а столбцы D и E дописываем:

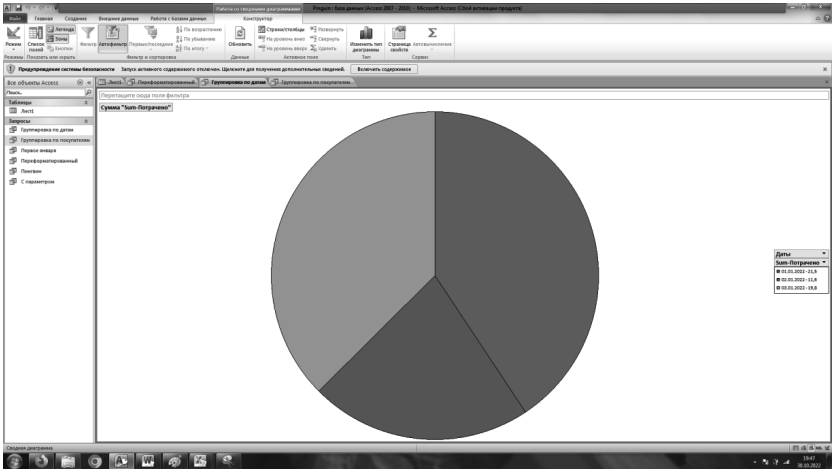
	A	B	C	D	E
1	Даты	Sum-Потрачен	Count-ID	Погода	Выходные
2	01.01.2022	21,5	3	солнечно	да
3	02.01.2022	11,6	2	дождливо	нет
4	03.01.2022	19,8	3	солнечно	нет
5					

4.4.3. Диаграмма

Сводная диаграмма строится для конкретного запроса. Например, для представленного далее.

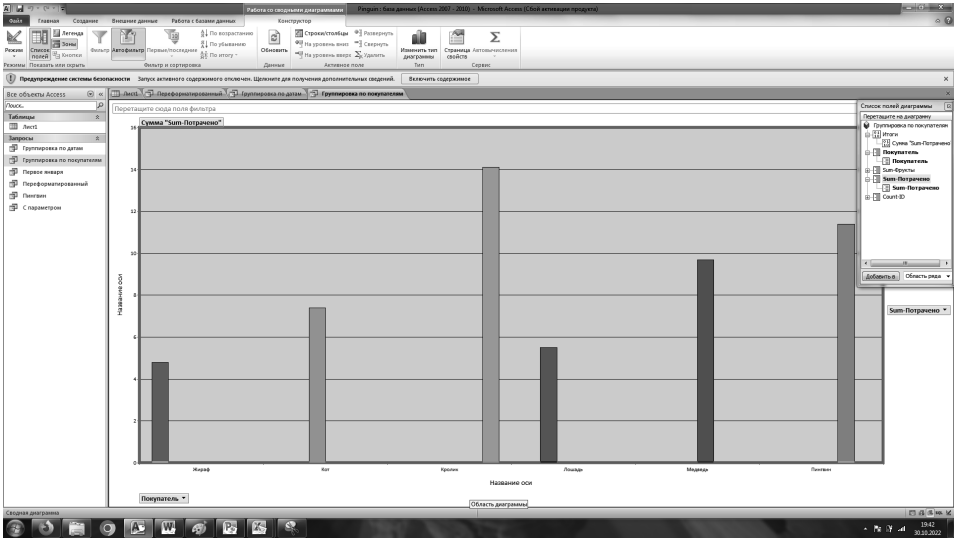


Покупатель	Sum-Фрукт	Sum-Потрач	Count-ID
Жираф	5	4,8	1
Кот		7,4	1
Кролик	14	14,1	2
Лошадь	6	5,5	1
Медведь	4	9,7	1
Пингвин	3	11,4	2

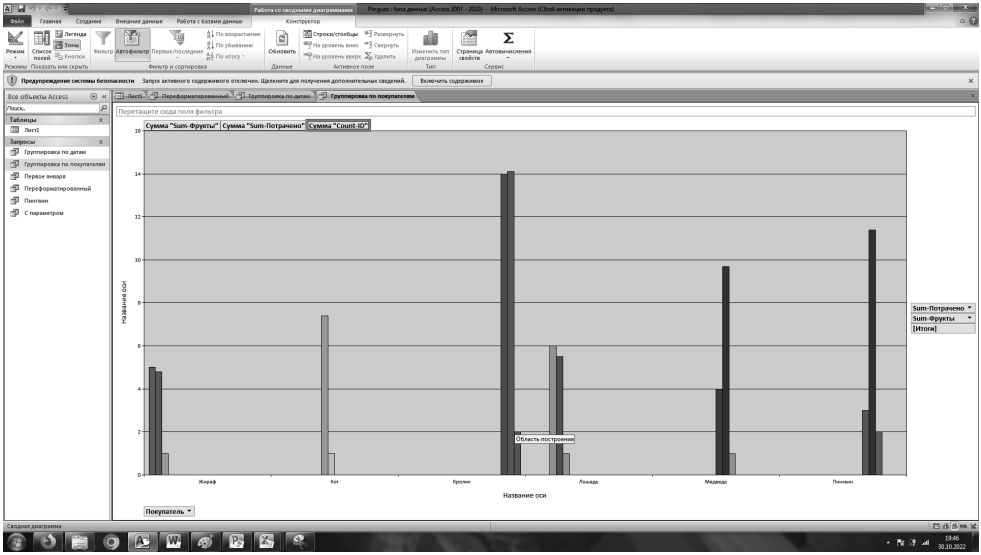


Для построения диаграммы необходимо «перетащить» на указанные позиции поля таблицы.

Фильтр можно не задавать. Вид диаграммы можно менять.



Можно отображать несколько полей.



4.5. Контрольные вопросы

1. Поясните правила импорта/экспорта таблиц в СУБД ACCESS.
2. Поясните правила и назначение SQL-запросов.
3. Поясните правила формирования запросов в конструкторе.
4. Как работает SQL-запрос? Поясните правила его записи.

5. Что такое агрегирующие функции?
6. Каково назначение агрегирующих функций?
7. Для чего используется группировка?
8. Для чего предназначен оператор SELECT?
9. Как выполняется оператор SELECT?
10. Что такое атрибут?
11. Как задать тип атрибута?
12. Для чего используются переформатированные таблицы и как они связаны с подтверждаемыми или опровергаемыми гипотезами.

5. Лабораторная работа №3

Ассоциативные правила

Продолжительность выполнения работы – 2 часа.

5.1. Цель работы

Целью данной лабораторной работы является освоение:

- применения методики ассоциативных правил для поиска покупательских шаблонов,
- мер для определения ассоциаций (поддержка, достоверность, лифт),
- изучение принципа Apriori.

5.2. Задание на выполнение работы

В соответствии с предметной областью лабораторной работой №1 выполнить следующее:

1. Создать «продуктовую таблицу» для 20-25 строк и 8-12 «товаров» (одну исходную таблицу).
2. Определить поддержку для каждого товара.
3. Определить поддержку, достоверность и лифт для пар товаров, оставив в знаменателе только 4-5 товаров с наибольшей поддержкой.
4. Определить поддержку, достоверность и лифт 5-7 для наборов из трех-четырех товаров.
5. Сделать выводы по результатам расчетов о закономерностях в покупках.

5.3. Требования к выполнению

- Для выполнения работы можно использовать электронные таблицы, ACCESS или выполнить программирование этой задачи на языке высокого уровня (например, C++).
- Данные в таблице вымышлены, но они должны соответствовать предметной области и соображениям разумного.
- Таблица может не повторять данные из таблиц 1 или 2 лабораторных работ, так как вариантом является только предметная область.
- Строки в столбце 1 могут повторяться, могут не повторяться, принципиального значения это не имеет, тем не менее хотя бы минимальное соответствие предметной области должно соблюдаться.
- «Продуктовая таблица» содержит информацию о факте покупки покупателем, но не о количестве купленного товара или его цене. Таблица может иметь вид:

	1	2	3	4	5	6	7	8	9	10
1	Столбец1	Столбец	Столбец	Столбец	Столбец5	Столбец	Столбец	Столбец	Столбец	Столбе
2		Рыба	Мясо	Фрукты	Овощи	Хлеб	Молоко	Вода	Сметана	Мёд
3	Пингвин	1	1					1		
4	Медведь	1						1		1
5	Кролик			1	1	1				
6	Лошадь			1	1	1		1		
7	Пингвин	1	1			1		1		
8	Жираф			1	1	1		1		
9	Кролик			1	1	1				
10	Кот	1					1		1	
11	Пингвин	1	1					1		
12	Медведь	1					1			1
13	Кролик			1	1	1				
14	Лошадь			1	1	1		1		
15	Пингвин	1						1		
16	Жираф			1			1			
17	Кролик			1	1	1				
18	Кот	1					1		1	
19	Пингвин	1				1			1	
20	Медведь	1	1							1
21	Кролик			1	1	1				
22	Лошадь			1	1	1		1		
23	Пингвин	1						1		
24	Жираф				1					
25	Кролик			1	1	1				
26	Кот						1		1	

5.4. Краткие теоретические сведения

5.4.1. Поиск покупательских шаблонов

Ассоциативные правила представляют собой механизм нахождения логических закономерностей между связанными элементами (событиями или объектами). Поиск покупательских шаблонов часто используют как другое название этого метода, принимая определение «закономерности» как «Кто купил x, также купил y».

Таким образом:

- ассоциативные правила позволяют находить закономерности между связанными событиями;
- алгоритмы поиска ассоциативных правил - один из популярных методов обнаружения знаний;
- Associations rules learning (ARL) – обучение на ассоциативных правилах – ряд популярных алгоритмов для извлечения знаний. В основе лежит анализ транзакций, внутри каждой из которых лежит свой уникальный itemset из набора items. При помощи ARL алгоритмов находятся те самые «правила» совпадения items внутри одной транзакции, которые потом сортируются по их силе.

Выделяют три вида правил:

- ***полезные правила***, содержащие действительную информацию, которая ранее была неизвестна, но имеет логическое объяснение;
- ***тривиальные правила***, содержащие действительную и легко объяснимую информацию, отражающую известные законы в исследуемой области, и поэтому не приносящие какой-либо пользы;
- ***непонятные правила***, содержащие информацию, которая не может быть объяснена (такие правила или получают на основе аномальных исходных данных, или они содержат глубоко скрытые закономерности, и поэтому для интерпретации непонятных правил нужен дополнительный анализ).

Поиск ассоциативных правил обычно выполняют в два этапа:

1. в пуле имеющихся признаков A находят наиболее часто встречающиеся комбинации элементов T;

2. из этих найденных наиболее часто встречающихся наборов формируют ассоциативные правила.

5.4.2. Основные меры для определения ассоциаций

Существуют три основные меры для определения ассоциаций.

Рассмотрим таблицу «покупок»:

Покупатель 1	Яблоко, сок, рис, курица
Покупатель 2	Яблоко, сок, рис
Покупатель 3	Яблоко, сок
Покупатель 4	Яблоко, груша
Покупатель 5	Молоко, сок, рис, курица
Покупатель 6	Молоко, сок, рис
Покупатель 7	Молоко, сок
Покупатель 8	Молоко, груша

Мера 1: Поддержка. Поддержка показывает то, как часто данный товарный набор появляется. Это измеряется долей покупок, в которых он присутствует. Например, если {яблоко} появляется в четырех из восьми покупок, значит, его поддержка 50 %. Товарные наборы могут содержать и несколько элементов. Например, поддержка набора {яблоко, сок, рис} — два из восьми, то есть 25 %. Для определения часто встречающихся товарных наборов может быть установлен порог поддержки. Товарные наборы, встречаемость которых выше заданного числа, будут считаться частотными.

Мера «Поддержка» => Поддержка {Яблоко}= 4/8.

Мера 2: Достоверность. Достоверность показывает, как часто товар Y появляется вместе с товаром X, что выражается как {X->Y}. Это измеряется долей их одновременных появлений. Согласно таблице, достоверность {яблоко->сок} соответствует трем из четырех, то есть 75 %.

Мера «Достоверность» =>

Достоверность {Яблоко -> сок} = Поддержка {Яблоко, сок} / Поддержка {Яблоко}.

Одним из недостатков этой меры является то, что она может исказить степень важности предложенной ассоциации. Рассмотренный пример Достоверности принимает во внимание только то, как часто покупают яблоки, но не то, как часто покупают сок. Если сок тоже довольно популярно, что и видно из таблицы, то неудивительно, что покупки, включающие яблоки, нередко содержат и сок, таким образом увеличивая меру достоверности. Тем не менее мы можем принять во внимание частоту обоих товаров, используя третью меру.

Мера 3: Лифт. Лифт отражает то, как часто товары X и Y появляются вместе, одновременно учитывая, с какой частотой появляется каждый из них.

Таким образом,

Лифт {яблоко->сок} = Достоверность {яблоко->сок} / частота {сока}.

или

Лифт {яблоко->сок} = Поддержка {яблоко, сок} / (Поддержка {яблоко} * Поддержка {сок}).

Согласно таблице, лифт для {яблоко->сок} равен единице, что означает отсутствие связи между товарными позициями***. Значения лифта больше единицы означают, что товар Y вероятно купят вместе с товаром X, а значение меньше единицы — что их совместная покупка маловероятна.***

5.4.3. Принцип Apriori

Одним из способов снизить количество конфигураций рассматриваемых товарных наборов является использование принципа Apriori. Принцип Apriori утверждает, что если какой-то товарный набор редкий, то и большие наборы, которые его включают, тоже должны быть редки. Это значит, что если редким является, скажем, {сок}, то редким должно быть и сочетание {сок, конфеты}. Таким образом, составляя список частотных товарных наборов, мы уже не будем рассматривать ни пару {сок, конфеты}, ни какую-либо другую с содержанием сока.

С применением принципа Apriori можно получить список частотных товарных наборов, используя следующие этапы.

1. начать с товарных наборов, содержащих всего один элемент, таких как {яблоки} или {груши}.
2. вычислить поддержку для каждого товарного набора. Оставить наборы, удовлетворяющие порогу, и отбросить остальные.
3. увеличить размер анализируемого товарного набора на единицу и сгенерировать все возможные конфигурации, используя товарные наборы из предыдущего шага.
4. повторять шаги 2 и 3, вычисляя поддержку для возрастающих товарных наборов до тех пор, пока они не закончатся.

Таким образом, если у элемента {яблоки} низкая поддержка, то он будет удален из списка анализируемых товарных наборов вместе со всем, что его содержит, тем самым это сократит число наборов для анализа более чем вдвое.

Кроме определения товарных наборов с высокой поддержкой, принцип Apriori также может помочь найти товарные ассоциации с высокой достоверностью или лифтом. Поиск этих ассоциаций требует меньше вычислений, поскольку если товарные наборы с высокой поддержкой известны, то достоверность и лифт вычисляются уже с использованием значения поддержки.

5.4.4. Ограничения алгоритма Apriori

Требует долгих вычислений. Хотя принцип Apriori и снижает число потенциальных товарных наборов для рассмотрения, оно все еще может быть достаточно значительным, если список товаров большой или указан низкий порог поддержки. В качестве альтернативного решения можно сократить число сравнений, используя расширенные структуры данных, чтобы отобрать потенциальные товарные наборы с большей эффективностью.

Ложные ассоциации. В больших наборах данных ассоциации могут быть чистой случайностью. Чтобы убедиться, что обнаруженные ассоциации масштабируемы, их нужно оценить.

Несмотря на эти ограничения, ассоциативные правила остаются интуитивно-понятным методом обнаружения закономерностей в наборах данных с управляемым размером.

5.4.5. Пример

Рассмотрим пример выполнения лабораторной работы. Определим исходную таблицу покупок:

Столбец1	Столбец	Столбец	Столбец	Столбец5	Столбец	Столбец	Столбец	Столбец	Столбе
Пингвин	1	1					1		
Медведь	1						1		1
Лошадь			1	1	1		1		
Пингвин	1	1			1		1		
Жираф			1	1	1		1		
Пингвин	1	1					1		
Лошадь			1	1	1		1		
Пингвин	1						1		
Лошадь			1	1	1		1		
Пингвин	1						1		
0	6	3	4	4	5	0	10	0	1
	Рыба	Мясо	Фрукты	Овощи	Хлеб	Молоко	Вода	Сметана	Мёд

При этом поддержка составит:

Рыба	Мясо	Фрукты	Овощи	Хлеб	Молоко	Вода	Сметана	Мёд
Поддержка								
0,25	0,125	0,166666667	0,166666667	0,208333333	0	0,416666667	0	0,041667

Далее определяем пары, где поддержка более 30%.

Достоверность пар, где поддержка более 30%							
Рыба	Мясо	3	12	Фрукты	Мясо	0	
Рыба	Фрукты	0		Фрукты	Рыба	0	
Рыба	Овощи	0		Фрукты	Овощи	10	60
Рыба	Хлеб	2	8	Фрукты	Хлеб	10	
Рыба	Молоко	3	12	Фрукты	Молоко	1	
Рыба	Вода	6	24	Фрукты	Вода	4	
Рыба	Сметана	3	12	Фрукты	Сметана	0	
Рыба	Мёд	3	12	Фрукты	Мёд	0	

Овощи	Мясо	0		Хлеб	Мясо	2		Вода	Мясо	6
Овощи	Рыба	0		Хлеб	Рыба	1		Вода	Рыба	3
Овощи	Фрукты	10		Хлеб	Фрукты	10		Вода	Фрукты	4
Овощи	Хлеб	10		Хлеб	Овощи	10		Вода	Овощи	4
Овощи	Молоко	0		Хлеб	Молоко	0		Вода	Молоко	0
Овощи	Вода	4		Хлеб	Вода	5		Вода	Хлеб	5
Овощи	Сметана	0		Хлеб	Сметана	1		Вода	Сметана	0
Овощи	Мёд	0		Хлеб	Мёд	0		Вода	Мёд	1

Далее определяем лифт и делаем выводы в соответствии с правилом, определенным на стр. 25 настоящего пособия.

5.4.6. Другие алгоритмы

Существует ряд часто используемых классических алгоритмов, позволяющих находить правила в itemsets согласно перечисленным выше понятиям — Наивный или Брутфорс-алгоритм, Apriori-алгоритм, ECLAT-алгоритм, FP-growth алгоритм и другие.

Алгоритм Брутфорс

Брутфорс-алгоритм (BFS) самый простой и, в то же время, самый неэффективный способ. Brute-force (атака полным перебором) – метод решения математических задач, сложность которого зависит от количества всех возможных решений. Для того чтобы найти

все возможные Association rules применяя брутфорс-алгоритм необходимо перечислить все подмножества X из набора I и для каждого подмножества X рассчитать Поддержку {X}.

Алгоритм выглядит следующим образом:

Вход: Датасет D, содержащий список транзакций.

Выход: Наборы itemsets F_1, F_2, …, F_q, где— набор F_i itemsets размера i, которые встречаются как минимум S раз в D.

Подход:

1. R — целочисленный массив, содержащий в себе все комбинации items в D, размера $2^{|D|}$.
2. Для n, принадлежащих [1, |D|] делаем:
   F — все возможные комбинации из D_n
   Увеличить каждое значение в R согласно значениям в каждом F[]
3. Вернуть Все itemsets в R >= s.

Сложность брутфорс-алгоритма очевидна. Для серьезных вычислений он не пригоден.

Алгоритм ECLAT

Идея алгоритма ECLAT (Equivalence CLAss Transformation) заключается в ускорении подсчета Поддержки {X}. Для этого необходимо проиндексировать базу данных D так, чтобы это позволило быстро рассчитывать Поддержку {X}.

Легко заметить, что если t(X) обозначает множество всех транзакций, где встречается подмножество X, то

t(XY) = t(X) and t(Y) и

Поддержка {XY} = |t(XY)|,

то есть Поддержка {XY} равна кардинальности (размеру) множества t(XY).

Данный подход может быть значительно усовершенствован путем уменьшения размера промежуточных множеств идентификаторов транзакций (tidsets). А именно, можно хранить не все множество транзакций на промежуточном уровне, а только множество различий этих транзакций.

В отличие от Apriori-алгоритма, ECLAT производит поиск в глубину. Иногда его называют «вертикальным» (в отличие от «горизонтального» для Apriori).

Ключевым понятием для ECLAT-алгоритма является I-префикс. В начале генерируется пустое множество I, это позволяет на первом проходе выделить все частотные itemsets. Затем алгоритм будет вызывать сам себя и увеличивать I на 1 на каждом шаге до тех пор, пока не будет достигнута заданная пользователем длина I.

Для хранения значений используется префиксное дерево (trie). Вначале строится нулевой корень дерева (то самое пустое множество I), затем по мере прохода по itemsets алгоритм прописывает содержащиеся в каждом itesmsets items, при этом самая левая ветвь является child нулевого корня и далее вниз. При этом ветвей столько, сколько items встречается в itemsets. Такой подход позволяет записывать itemset в памяти только один раз, что делает ECLAT быстрее Apriori.

Алгоритм FP-роста

FP-Growth (Frequent Pattern Growth) алгоритм самый молодой из представленных в этом пособии, впервые он описан в 2000 году.

FP-Growth предлагает радикальную вещь — отказаться от генерации кандидатов (напомним, генерация кандидатов лежит в основе Apriori и ECLAT). Теоретически, такой подход позволит еще больше увеличить скорость алгоритма и использовать еще меньше памяти.

Это достигается за счет хранения в памяти префиксного дерева (trie) не из комбинаций кандидатов, а из самих транзакций.

При этом FP-Growth генерирует таблицу заголовков для каждого item, чья Поддержка выше заданного пользователем. Эта таблица заголовков хранит связанный список всех однотипных узлов префиксного дерева. Таким образом, алгоритм сочетает в себе плюсы BFS за счет таблицы заголовков и FP-роста за счет построения дерева. Псевдокод алгоритма схож с ECALT.

5.5. Контрольные вопросы (темы)

1. Покупательский шаблон
2. Поиск покупательских шаблонов
3. Поддержка
4. Достоверность
5. Лифт
6. Ассоциативное правило
7. Граф ассоциаций
8. Принцип Apriori
9. Алгоритм Apriori
10. Ограничения принципа Apriori
11. Алгоритм Брутфорс
12. Алгоритм ECLAT
13. Алгоритм FP-роста

6. Лабораторная работа №4

Прогнозирование

Продолжительность выполнения работы – 4 часа.

6.1. Цель работы

Целью лабораторной работы является приобретение навыков по:

- выполнению парного регрессионного анализа,
- использованию Excel для проведения регрессионного анализа, в том числе множественного,
- визуализации результатов,
- объяснению результатов.

6.2. Задание на выполнение работы

I. В соответствии со своим вариантом для заданной предметной области:

1. По данным наблюдений двух величин выполнить регрессионный анализ, найти коэффициент корреляции и уравнение линии регрессии. Построить ее график, а также диаграмму рассеяния. Построение выполнить с помощью Excel, но без использования «Анализа данных».
2. По данным наблюдений трех величин (одной зависимой и двух независимых) выполнить регрессионный анализ, найти коэффициент детерминации и уравнение линии регрессии. Построить диаграмму рассеяния для каждого регрессора.
3. Добавить третью независимую величину, выполнить регрессионный анализ, найти коэффициент детерминации и уравнение линии регрессии.
4. Осуществить проверку полученных результатов и их интерпретацию.

II. На основе реальных статистических данных выполнить корреляционный анализ для трех величин (одной зависимой и двух независимых) и интерпретацию результатов.

6.3. Требования к выполнению

- набор данных может быть вымышленным, регрессоры можно менять;
- наблюдений должно быть не менее 20;
- реальные статистические данные (желательно выбирать близко к предметной области, определенной в варианте) брать здесь:
  https://rosstat.gov.ru/storage/mediabank/Ejegodnik_2022.pdf
  https://rosstat.gov.ru/storage/mediabank/Strani_mira_2022.pdf
  https://showdata.gks.ru
  https://www.gks.ru/free_doc/new_site/population/demo/progn1.htmrfr

6.4. Краткие теоретические сведения

6.4.1. Прогнозирование

Прогноз – это научно обоснованное суждение о возможных состояниях объекта в будущем. Прогнозирование — это разработка прогноза (или специальное научное исследование конкретных перспектив дальнейшего развития какого-либо процесса).

Необходимость прогноза обусловлена желанием знать события будущего (знать достоверно в принципе невозможно), исходя из статистических (ошибки текущих оценок), вероятностных (многовариантность следствий), эмпирических (методологические ошибки моделей), философских (ограниченность текущих знаний) принципов.

Прогностика — научная дисциплина, изучающая общие принципы и методы прогнозирования развития объектов любой природы, закономерности процесса разработки прогнозов.

Существует целый ряд методов прогнозирования. В настоящее время разрабатываются методы прогнозирования, использующие положения теории хаоса и фракталов, но они пока мало проработаны как с теоретической точки зрения, так и в плане практической реализации. Отдельные моменты иногда применяются при анализе финансовых рынков: трейдеры, как правило, первыми испытывают все новые методы прогнозирования. В результате могут быть получены методы довольно точного прогнозирования резких и внезапных изменений: экономических кризисов, скачкообразной динамики спроса, банкротств и т.д. Но рассмотрим более традиционные методы прогнозирования.

6.4.2. Классификация методов прогнозирования

Чтобы получить общее представление о методах прогнозирования, необходимо для начала классифицировать эти методы. Для начала их принято разделять на количественные и качественные.

Рассмотрим другие классификационные признаки.

По горизонту прогноза (срокам):
- краткосрочные (как правило, в пределах года или нескольких месяцев);
- среднесрочные (несколько лет);
- долгосрочные (более пяти лет);
- дальнесрочные (более 10 лет).

По типу прогнозирования:
- эвристические (использующие субъективные данные, оценки и мнения);
- поисковые (в свою очередь делятся на экстраполяционные, проецирующие прошлые тенденции в будущее, и альтернативные, учитывающие возможности скачкообразной динамики явлений и различные варианты их развития);
- нормативные (оценка тенденций проводится исходя из заранее установленных целей и задач).

По масштабу:
- частные,
- местные,
- региональные,
- отраслевые,
- страновые,
- мировые (глобальные).

По ответственности (авторству):
- личные,
- на уровне предприятия (организации),
- на уровне государственных органов.

По степени вероятности событий:
- вариантные (подразумевают вероятностный характер будущего и предлагают несколько сценариев развития событий);
- инвариантные (предполагается единственный сценарий).

По способу представления результатов:
- точечные (прогнозируется точное значение показателя);
- интервальные (прогнозируется диапазон наиболее вероятных значений).

По степени однородности:
- простые;
- комплексные (сочетают в себе несколько взаимосвязанных простых методов).

По характеру базовой информации:
- фактографические (основываются на имеющейся информации о динамике развития явления или объекта, бывают статистическими и опережающими);
- экспертные (индивидуальные и коллективные, в зависимости от числа экспертов);
- комбинированные (использующие разнородную информацию).

По инструментальному подходу:
- статистические методы;

- экспертные оценки;
- методы моделирования, в том числе имитационного;
- интуитивные (то есть выполненные без применения технических средств, экспромтом, «в уме» специалистом, имеющим опыт ранее применяемых научных методов в данном типе прогнозов).

6.4.3. Временные ряды

Временной ряд представляет собой набор данных, описывающих объект в последовательные равноотстоящие моменты времени. Если исходные данные относятся к различным моментам времени, традиционный подход состоит в аппроксимации данных и использовании интерполированных отсчетов на равномерной сетке.

Для оценивания качества прогноза предложено использовать коэффициент расхождения (или коэффициент несоответствия), представляющий собой отношение среднеквадратической ошибки прогноза и среднеквадратической оценки рассеяния исходного ряда.

Методы прогнозирования по своему информационному основанию делятся на три класса:

- Фактографические методы, которые базируются на имеющемся информационном материале об объекте прогнозирования и его прошлом развитии.
- Экспертные методы, которые базируются на информации, обеспечиваемой систематизированными процедурами выявления и обобщения мнений специалистов-экспертов.
- К комбинированным относятся методы со смешанной информационной основой, использующие как фактографическую, так и экспертную информацию.

В действительности, любой прогноз использует экспертную информацию, хотя бы в части предположений о неизменности условий протекания изучаемого процесса на каком-то временном отрезке в будущем.

Наиболее распространенными и разработанными при фактографическом прогнозировании являются методы экстраполяции тенденций, в основе которых лежит предположение о том, что рассматриваемый процесс изменения переменной x (t) представляет собой сочетание нескольких составляющих, регулярных и случайных:

$$x(t) = \sum_{i=1}^{r} f_i(t) + \xi(t).$$

Считается, что регулярные составляющие f_i (t) представляют собой достаточно гладкие функции от аргумента t (в большинстве случаев – времени), которые сохраняют свой вид на промежутке упреждения процесса. Они отвечают интуитивному представлению о какой-то очищенной от помех сущности исследуемого процесса. Сумма регулярных составляющих образует тренд исследуемого процесса. Экстраполяционные методы прогнозирования делают основной упор на выявление наилучшего в том или ином смысле описания тренда и получение прогнозных значений путем его экстраполяции. Регулярную часть ряда оценивают в виде разложения по некоторому ортогональному базису. Этот базис обычно стараются задать на основе априорных предположений о природе изучаемого процесса, но наибольшую ценность имело бы использование базиса, непосредственно порождаемого самим исходным временным рядом. Случайная составляющая обычно проявляется в повышенной изменчивости значений временного ряда.

6.4.4. Регрессионный анализ

Регрессионным анализом называется раздел математической статистики, объединяющий практические методы исследования корреляционной зависимости между случайными величинами по результатам наблюдений над ними. Сюда включаются методы выбора модели изучаемой зависимости и оценки ее параметров, методы проверки статистических гипотез о зависимости.

Целями регрессионного анализа являются:

- предсказание значения зависимой переменной с помощью независимых переменных;
- определение вклада отдельных независимых переменных в вариацию зависимой переменной;

Регрессионный анализ нельзя использовать для определения наличия связи между переменными, поскольку наличие такой связи и есть предпосылка для применения этого вида анализа.

Пусть между случайными величинами X и Y существует линейная корреляционная зависимость. Это означает, что математическое ожидание Y линейно зависит от значений случайной величины X. График этой зависимости (линия регрессии Y на X) имеет уравнение $M(Y) = \rho X + b,$

линейная модель пригодна в качестве первого приближения и в случае нелинейной корреляции, если рассматривать небольшие интервалы возможных значений случайных величин.

Допущения:

- переменные модели должны иметь распределение близкое к нормальному;
- зависимые и независимые переменные должны быть измерены в метрической шкале;
- для построения линейных регрессий зависимая и независимая переменные должны иметь линейную связь.

Любая регрессионная модель позволяет обнаружить только количественные зависимости, которые не обязательно отражают причинные зависимости, т.е. влияние одного фактора на другой.

Рассмотрим парную (простую регрессию):

$$Y_i = a + bX_i$$

где

- Y_i – зависимая переменная,
- X_i – независимая переменная,
- a – константа, определяет точку пересечения прямой с осью Y (экономически не интерпретируется);
- b – угловой коэффициент, характеризует наклон прямой (slope) (коэффициент регрессии b показывает, на какую величину в среднем изменится результативный признак Y_i, если переменная X_i увеличится на единицу своего измерения).

Коэффициент эластичности (Э) показывает, на сколько процентов в среднем изменится Y_i при изменении X_i на 1%. Для простой линейной регрессии: Э = b*(X/Y).

Задача регрессионного анализа сводится к поиску коэффициентов a и b. Рассмотрим на примере. Получена выборка 10 значений величин X и Y.

n	10									
X	3	8	4	4	7	8	2	5	6	3
Y	4	5	2	5	6	8	3	4	5	5

Проведен предварительные вычисления и получим:

n	10										Суммы
X	3	8	4	4	7	8	2	5	6	3	50
Y	4	5	2	5	6	8	3	4	5	5	47
X^2	9	64	16	16	49	64	4	25	36	9	292
X*Y	12	40	8	20	42	64	6	20	30	15	257

Для вычисления коэффициентов a и d используем соответственно формулы:

$$\frac{n\sum_{k=1}^{n} X_k Y_k - \sum_{k=1}^{n} X_k \sum_{k=1}^{n} Y_k}{n\sum_{k=1}^{n} X_k^2 - \left(\sum_{k=1}^{n} X_k\right)^2} \qquad \frac{\sum_{k=1}^{n} X_k^2 \sum_{k=1}^{n} Y_k - \sum_{k=1}^{n} X_k \sum_{k=1}^{n} X_k Y_k}{n\sum_{k=1}^{n} X_k^2 - \left(\sum_{k=1}^{n} X_k\right)^2}$$

Тогда

a avr	0,52381
b avr	2,080952

, а оценка линии регрессии примет вид Y = 0,52X + 2,08.

С учетом средних значений

X avr	5
Y avr	4,7

, коэффициент корреляции примет вид: $\tilde{r}_{xy} = 0{,}52 \cdot \frac{2{,}16}{1{,}64} = 0{,}68$. Теперь можно изобразить линию регрессии.

Теоретической линией регрессии называется линия, вокруг которой группируются точки корреляционного поля и которая указывает основную тенденцию связи.

Теоретическая линия регрессии должна отображать изменение средних величин результативного признака Y по мере изменения величин факторного признака X при условии полного взаимопогашения всех прочих, случайных по отношению к фактору X, причин.

6.4.5. Множественная регрессия

Множественная регрессия - это статистический метод, который может быть использован для анализа взаимосвязи между одной зависимой переменной и несколькими независимыми переменными. Широко используемый метод, который может включать как непрерывные, так и категориальные предикторы. Можно использовать также для описания последствий изменения зависимых переменныхв зависимости от изменений, которые могут иметь независимые переменные, и прогнозирования будущих значений зависимых переменных, когда значения независимых переменных определены заранее.

Множественная регрессия – это уравнение связи с несколькими регрессорами:

$$y = f(x_1, x_2, \ldots, x_k).$$

Существует ряд моделей множественной регрессии, но в данном пособии рассмотрим только модель множественной линейной регрессии, которая имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \varepsilon$$

Коэффициент β_j, j=1,2,...,m, называется j-м теоретическим коэффициентом регрессии, характеризует чувствительность величины Y к изменению X_j (отражает влияние на условное математическое ожидание $M(Y|x_1,x_2,\ldots,x_m)$ зависимой переменной Y, объясняющей переменной X_j при условии, что все другие объясняющие переменные остаются постоянными.

Коэффициент β_0 определяет значение Y, в случае, когда все объясняющие переменные X_j равны нулю.

Случайная ошибка ε удовлетворяет тем же предпосылкам, что и в модели с парной регрессией. Предполагается, что объясняющие переменные некоррелированы друг с другом (в модели отсутствует мультиколлинеарность).

Для оценки параметров регрессии используется метод наименьших квадратов (МНК), в соответствии с которым минимизируется сумма квадратов остатков:

$$F(b_0, b_1, b_2, ..., b_m) = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - ... - b_m x_{im})^2 \to \min .$$

Рассмотрим пример регрессии с двумя объясняющими переменными: предположим, нужно определить, влияет ли количество часов, потраченных на учебу, и количество сданных подготовительных экзаменов на балл, который студент получает на определенном вступительном экзамене в колледж.

Чтобы исследовать эту взаимосвязь, можно выполнить множественную линейную регрессию, используя часы обучения и подготовительные экзамены, взятые в качестве объясняющих переменных (столбцы A и B), и экзаменационный балл в качестве переменной ответа (столбец C). Используем Пакет «Анализ данных» Microsoft Excel, который рассмотрен далее в п.6.4.7.

	A	B	C	D	E	F	G	H	I
1	ЧАСЫ	ПРЕДВАР	БАЛЛЫ						
2	1	1	76	ВЫВОД ИТОГОВ					
3	2	3	78						
4	2	3	85	*Регрессионная статистика*					
5	4	5	88	Множественный R	0,874559958				
6	2	2	72	R-квадрат	0,76485512				
7	1	2	69	Нормированный R-квадрат	0,737191016				
8	5	1	94	Стандартная ошибка	5,045170133				
9	4	1	94	Наблюдения	20				
10	2	9	88						
11	4	3	92	Дисперсионный анализ					
12	4	4	90		*df*	*SS*	*MS*	*F*	*Значимость F*
13	3	3	75	Регрессия	2	1407,486392	703,7431958	27,64792717	4,53265E-06
14	6	2	96	Остаток	17	432,7136084	25,45374167		
15	5	4	90	Итого	19	1840,2			
16	3	4	82						
17	4	4	85		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*
18	6	5	99	Y-пересечение	65,27472887	2,790093603	23,39517527	2,27522E-14	59,38814592
19	2	1	83	Переменная X 1	4,968445147	0,743971097	6,678277111	3,88615E-06	3,398803337
20	1	0	62	Переменная X 2	0,956782386	0,580295543	1,648784654	0,117547062	-0,267534191
21	2	1	76						

Интерпретация результатов:

– R-квадрат: 0,765. Это известно как коэффициент детерминации или доля дисперсии переменной отклика, которая может быть объяснена объясняющими переменными. В этом примере 76,5% вариаций в экзаменационных баллах можно объяснить количеством часов обучения и количеством сданных подготовительных экзаменов.

- Стандартная ошибка: 5,045. Это среднее расстояние, на которое наблюдаемые значения отходят от линии регрессии. В этом примере наблюдаемые значения отклоняются от линии регрессии в среднем на 5,045 единицы.
- F: 27,65. Это общая F-статистика для регрессионной модели, рассчитанная как MS регрессии / остаточная MS (703,7431958 / 25,45374167) .
- Значимость F: 0,000004. Это p-значение, связанное с общей статистикой F. Он говорит нам, является ли регрессионная модель в целом статистически значимой. Другими словами, он говорит нам, имеют ли объединенные две объясняющие переменные статистически значимую связь с переменной отклика. В этом случае p-значение меньше 0,05, что указывает на то, что независимые переменные количество часов обучения и сданных подготовительных экзаменов вместе имеют статистически значимую связь с экзаменационным баллом.
- P-значения. Отдельные p-значения говорят нам, является ли каждая независимая переменная статистически значимой. Мы можем видеть, что изученные часы статистически значимы ($p = 0,00$), в то время как пройденные подготовительные экзамены ($p = 0,12$) не являются статистически значимыми при $\alpha = 0,05$. Поскольку сданные подготовительные экзамены не являются статистически значимыми, мы можем принять решение удалить их из модели.
- Коэффициенты: коэффициенты для каждой независимой переменной говорят нам о среднем ожидаемом изменении переменной отклика при условии, что другая независимая переменная остается постоянной. Например, ожидается, что за каждый дополнительный час, потраченный на учебу, средний экзаменационный балл увеличится на 4,97 при условии, что количество сданных подготовительных экзаменов останется неизменным.

  Еще один способ подумать об этом: если учащийся А и учащийся Б сдают одинаковое количество подготовительных экзаменов, но учащийся А учится на один час больше, то ожидается, что учащийся А получит результат на 4,97 балла выше, чем учащийся Б.
- Мы интерпретируем коэффициент для перехвата как означающий, что ожидаемая оценка экзамена для студента, который учится ноль часов и сдает нулевые подготовительные экзамены, составляет 65,27.

Коэффициенты из выходных данных модели можно использовать, чтобы создать следующее ***расчетное уравнение регрессии***:

экзаменационный балл = 65,23 + 4,97*(часы) + 0,96*(подготовительные экзамены).

6.4.6. Прогнозирование МВР

Процесс прогнозирования многомерного временного ряда (МВР) основан на алгоритме множественной регрессии. Пусть имеется случайный вектор

$$Z = [x_1, \ldots, x_k, y_1, \ldots, y_r]^T,$$

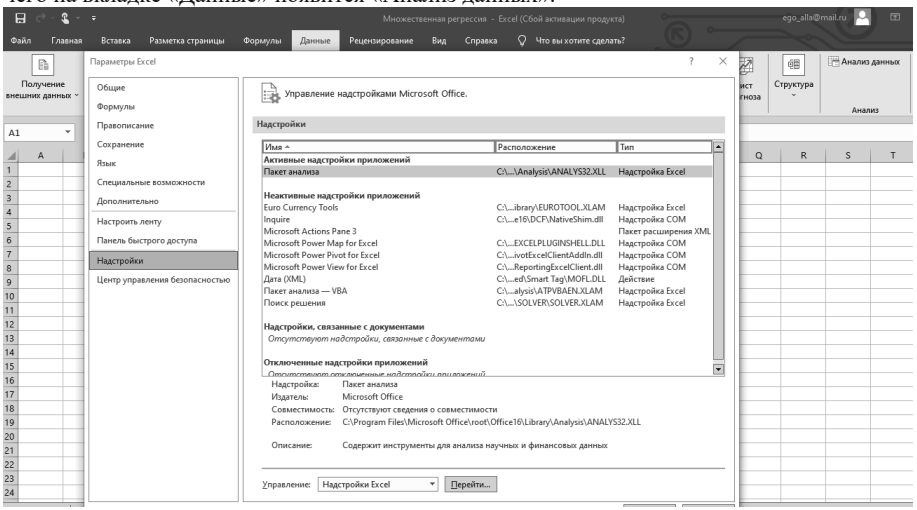
который подчиняется $(k + r)$-мерному нормальному закону $N_{k+r}(a, \Sigma)$ с известными вектором средних a и ковариационной матрицей Σ. Рассмотрим случай, когда первые k компонент $x_1, \ldots, x_k$ вектора Z наблюдаются в эксперименте, а оставшиеся компоненты $y_1, \ldots, y_r$ являются ненаблюдаемыми. Требуется получить оценки ненаблюдаемых компонент.

Согласно поставленным условиям, средние и ковариационная матрица вектора Z имеют блочную структуру: $a = [a_x, a_y]$, $\Sigma = [\Sigma_{xx}\ \Sigma_{xy}\ \Sigma_{xy}^t\ \Sigma_{yy}]$. Требуется определть матрицу

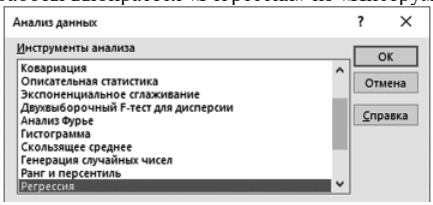
С размерности rxk, доставляющую минимум функции потерь, которая представляет собой функцию дисперсии погрешностей прогноза. В итоге требуется получить на основе процедуры множественной регрессии прогноз компонент ряда на r шагов вперед.

6.4.6. Пакет «Анализ данных» Microsoft Excel

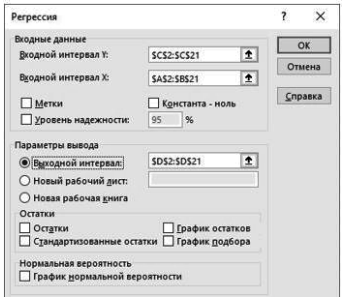
Пакет «Анализ данных» Microsoft Excel подключается на вкладку «Данные» через «Надстройки» «Параметров» Excel. Для этого нужно выбрать «Пакет анализа», после чего на вкладке «Данные» появится «Анализ данных».



Для выполнения работы выбирается «Регрессия» из «Инструментов анализа»:



И настраивается под предварительно размещенные данные (входные интервалы и параметры вывода).



6.5. Контрольные вопросы (темы)

1. Понятие регрессии
2. Линия тренда
3. Предиктор
4. Градиентный спуск
5. Коэффициенты регрессии
6. Уравнение регрессии
7. Коэффициенты корреляции
8. Веса
9. Искажение веса при корреляции предикторов
10. Линейная регрессия
11. Множественная регрессия
12. Возможности Excel для проведения регрессионного анализа
13. Трактование результатов регрессионного анализа

7. Лабораторная работа №5

Разделяющая полоса

Продолжительность выполнения работы – 4 часа.

7.1. Цель работы

Целью лабораторной работы является изучение ключевых концепций методов опорных векторов, а также приобретение навыков:

– по использованию Excel для проведения расчетов методом опорных векторов для линейно-разделяемых данных в двумерном пространстве;
– визуализации результатов;
– объяснению результатов.

7.2. Задание на выполнение работы

В соответствии со своим вариантом для заданной предметной области:

1. Сформулировать классификационный признак (вопрос, на который нужно получить ответ) и подобрать 4-7 критериев.
2. Сформировать набор данных (не менее 30 объектов, из них не менее половины – для обучающего набора).
3. Провести попарное исследование методом опорных векторов на обучающем наборе. Построить 3-4 графика по парам критериев, найти опорные вектора (точки) (возможны варианты), расстояние между ними, вычислить линейную функцию (также возможны варианты) - коэффициенты, построить гиперплоскость (ее график).
4. Проанализировать результаты и сделать выводы.
5. Осуществить проверку полученных результатов на оставшихся данных.

7.3. Требования к выполнению

– Набор данных может быть вымышленным.
– Для подготовки данных можно использовать любое программное обеспечение, в том числе и использовать языки программирования высокого уровня.
– Для выполнения работы предлагается использовать табличный процессор (например, Microsoft Excel), можно использовать языки программирования для расчетов и визуализации, но использование библиотечных функций недопустимо.

– Критерии, а также их комбинации подбираются, исходя из предметной области, соображений разумного и особенностей процесса, на который ориентируетесь. Каждый критерий может использоваться как многократно, так и однократно.

7.4. Краткие теоретические сведения

7.4.1. Метод опорных векторов

Метод опорных векторов (support vector machine - SVM) выявляет оптимальную границу для классификации, которая может быть использована для разделения объектов на две группы (то есть здоровых и нездоровых, успешных и неуспешных, работающих исправно и находящихся в предаварийном состоянии и т.п.).

Метод опорных векторов - набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа, который позволяет выявить оптимальную границу при распределении, которая может быть использована для разделения объектов на 2 группы. Набор контролируемых методов обучения, используемых для классификации, регрессии и обнаружения выбросов.

Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как метод классификатора с максимальным зазором.

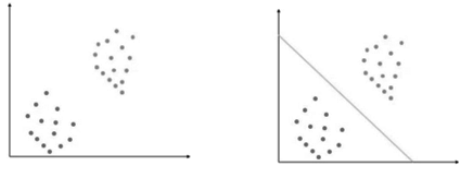
Идея метода:

– Основная идея — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве.
– Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы.
– Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей.
– Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

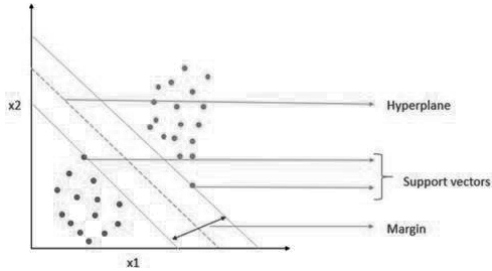
7.4.2. Построение оптимальной границы

Линейно-разделяемые данные - два класса, между которыми можно найти границу, которая разделяет эти классы разделяет. В одномерном пространстве разделение выглядит следующим образом (например, на красный и зеленый):

Точно так же в двухмерном пространстве можно нарисовать линию, которая действует как граница между двумя классами:



Гиперплоскости - это границы решений, которые классифицируют данные. Алгоритм SVM находит лучшую гиперплоскость, которая находится посередине двух классов с максимальным отсупом с обеих сторон. Точки данных, ближайшие к гиперплоскости в обоих классах, известны как опорные векторы. Если точка данных, которая является опорным вектором, удаляется, положение гиперплоскости изменяется.



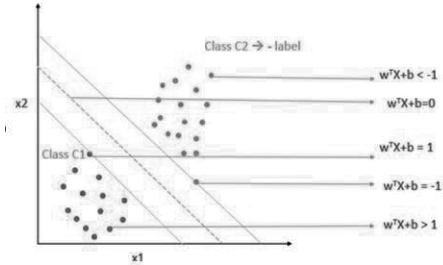
Как найти лучшую гиперплоскость? Предположим, нужно разделить два класса C1 и C2 с помощью SVC в двумерном пространстве. Затем предсказать класс неизвестного вектора признаков X как класс C1 или класс C2.

Можно использовать линейное уравнение: $g(X)=w^TX+b=0$, где

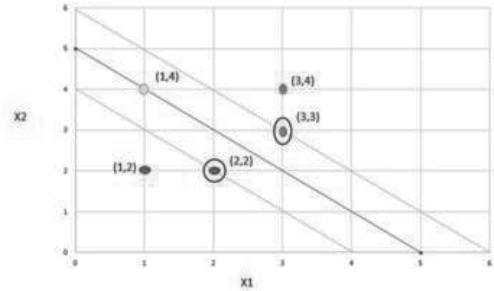
- w - вектор веса, перпендикулярный гиперплоскости (он представляет ориентацию гиперплоскости в d-мерном пространстве, где d - размерность вектора признаков).
- b – определяет положение гиперплоскости в d-мерном пространстве.

Это линейное уравнение в двух измерениях представляет собой прямую линию, плоскость - в трехмерном пространстве и гиперплоскость в более чем трех измерениях.

Для каждого вектора признаков нужно вычислить линейную функцию таким образом, что если вектор признаков лежит на положительной стороне гиперплоскости, то G(xi) >0, а если вектор признаков лежит на отрицательной стороне гиперплоскости, то G(xi) <0.



Пример. Возьмем несколько точек данных, нарисуем гиперплоскость и вычислим значение wx + b для всех точек данных. Вычислим линейную функцию.



1. Точка (1,4) лежит на гиперплоскости

g (X) = wX + b

g (X) = (- 1) (1,4) +5 => (-1) (1) + (- 1) (4) +5

g (X) = -1–4 + 5 = 0 r (X) = 0

2. Вектор поддержки (2,2) принадлежит метке «+»
   g (X) = (- 1) (2,2) +5 => (- 1) (2) + (- 1) (2) +5
   g (X) = - 2–2 + 5 = 1 r (X) = 1
3. Вектор поддержки (3,3) принадлежит метке «-»
   g (x) = (- 1) (3,3) +5 = (- 1) (3) + (- 1) (3) +5
   g (x) = - 3–3 + 5 = -1 r (x) = - 1
4. Точка данных (3,4) принадлежит отрицательной стороне гиперплоскости.
   g (x) = (- 1) (3,4) = (- 1) (3) + (- 1) (4) +5
   g (x) = - 3–4 + 5 = -2, т.е. g (x) <- 1
5. Точка данных (1,2) принадлежит положительной стороне гиперплоскости.
   g (x) = (- 1) (1,2) +5 = (- 1) (1) + (- 1) (2) +5
   g (x) = - 1–2 + 5 = 3, т.е. g (x)> 1

Как найти лучшую гиперплоскость:

– На этапе обучения метод запускается с какой-то случайной гиперплоскости и проверяет, нет ли ошибки.
– Если точка данных, принадлежащая классу C1, прогнозируется как означающая класс C2, тогда она изменит значение m и повернет гиперплоскость таким образом, чтобы точка данных ошибки вернулась в правильную сторону.
– На этапе обучения модель найдет правильные m и b, которые дают нулевую ошибку обучения.

7.4.3. Ограничения

Хотя метод опорных векторов является адаптивным и быстрым инструментом, он может не подходить в следующих случаях:

– **Малые наборы данных**. Поскольку для определения границ метод опирается на опорные векторы, то небольшой набор данных сокращает их число и отрицательно влияет на точность расчета.
– **Множество групп**. Метод опорных векторов способен классифицировать данные только на две группы за раз. Если групп три и более, то необходимо применять итеративно для выявления каждой отдельной группы метод, который называется многоклассовая классификация (multi-class SVM).
– **Большое перекрытие данных**. Метод опорных векторов классифицирует элементы данных исходя из того, с какой стороны границы разграничения они оказались. Когда элементы данных сильно перекрываются обеими группами, то те из них, которые находятся ближе к границе, могут быть классифицированы ошибочно. Более того, метод не дает информации о вероятности ошибочной классификации для отдельного элемента данных.

7.5. Контрольные вопросы

1. Машины опорных векторов
2. Метод классификации с максимальным зазором
3. Разделяющая прямая
4. Опорные точки
5. Алгоритм разделения полосой на плоскости
6. Оптимальная разделяющая гиперплоскость
7. Линейно-отделимый класс
8. Угловой коэффициент оптимальной разделяющей прямой
9. Случай отсутствия линейной отделимости

8. Лабораторная работа №6

Программирование генетического алгоритма

Продолжительность выполнения работы – 2 часа.

8.1. Цель работы

Целью лабораторной работы является:

- изучение ключевых концепций методов эволюционного моделирования;
- приобретению навыков по использованию Excel для проведения расчетов на основе генетического алгоритма;
- изучение этапов генетического алгоритма;
- приобретению навыков по формированию функции приспособленности.

8.2. Задание на выполнение работы

1. В соответствии с заданной предметной областью (лабораторная работа №1) сформулировать функцию приспособленности на основе задачи коммивояжера (в соответствии с вариантом).
2. Сформулировать правила скрещивания и уточнить правила мутации (в соответствии с вариантом).
3. В соответствии с вариантом сформировать начальную популяцию из четырех особей.
4. Вычислить функции приспособленности.
5. Сформировать четырех потомков.
6. Сформировать новую популяцию из четырех особей.
7. Повторить п.4-5 несколько раз (не менее 4).
8. Сформировать оптимальное решение.
9. Проанализировать результаты и сделать выводы.

Варианты мутации:

Нечетные варианты – инверсия
Четные варианты - транслокация

Примечание

Считаем, что расстояние от начальной точки до всех точек одинаково.

8.3. Требования к выполнению

- В отчете должны быть сформулированные правила скрещивания и мутации, а формирование поколений должно иллюстрировать выполнение этих правил.
- Прекращение формирования новых поколений должно быть обосновано, проиллюстрировано на примере и отражено в выводе.
- Для скрещивания используется не менее двух особей.
- Функция приспособленности (формулировка) должна отражать предметную область, но при этом ни к каким другим лабораторным работам быть привязана не должна. При этом в основе функции приспособленности может лежать принцип максимизации, минимизации или оптимизации с учетом особенностей предметной области.

8.4. Краткие теоретические сведения

8.4.1. Определение генетического алгоритма

Идея генетических алгоритмов (ГА) «подсмотрена» у систем живой природы, у систем, эволюция которых развертывается в сложных системах достаточно быстро.

Генетический алгоритм – это алгоритм, основанный на имитации генетических процедур развития популяции в соответствии с принципами эволюционной динамики. Часто используется для решения задач оптимизации (в т. ч. многокритериальной), поиска, управления.

Данные алгоритмы адаптивны, развивают решения, развиваются сами. Особенность этих алгоритмов – их успешное использование при решении сложных проблем (проблем, для которых невозможно построить алгоритм с полиномиальной алгоритмической сложностью).

Концепция ГА:

- Воспроизводится новая популяция допустимых решений, выбирая лучших представителей предыдущего поколения, скрещивая их и получая множество новых особей.
- Это новое поколение содержит более высокое соотношение характеристик, которыми обладают хорошие члены предыдущего поколения.
- Таким образом, из поколения в поколение хорошие характеристики распространяются по всей популяции. Скрещивание наиболее приспособленных особей приводит к тому, что исследуются наиболее перспективные участки пространства поиска.
- В конечном итоге, популяция будет эволюционировать к оптимальному решению задачи.

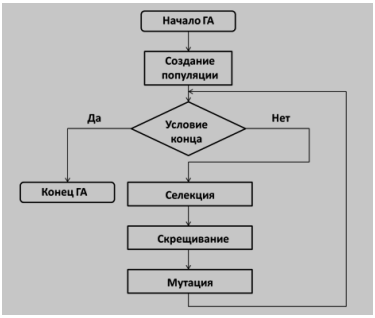
ГА часто используются в задачах оптимизации. Цель в оптимизации состоит в том, чтобы найти лучшее возможное решение задачи по одному или нескольким критериям. Чтобы реализовать генетический алгоритм, нужно сначала выбрать подходящую структуру для представления этих решений. В постановке задачи поиска экземпляр этой структуры данных представляет точку в пространстве поиска всех возможных решений.

Структура данных генетического алгоритма состоит из одной или большего количества хромосом (обычно из одной). Как правило, хромосома – это битовая строка, так что термин строка часто заменяет понятие «хромосома».

Каждая хромосома (строка) представляет собой конкатенацию (объединение) ряда подкомпонентов, называемых генами. Гены располагаются в различных позициях или локусах хромосомы и принимают значения, называемые аллелями. В представлениях с бинарными строками ген – бит, локус – его позиция в строке и аллель – его значение (0 / 1).

8.4.2. Схема генетического алгоритма

Существует несколько схем ГА, но в основном он делится на четыре следующих этапа: Создание популяции (инициализация), размножение, мутации и отбор.



8.4.3. Генетические операторы

Создание популяции (инициализация).

Перед первым шагом нужно случайным образом создать некую начальную популяцию; даже если она будет совершенно неконкурентоспособной, генетический алгоритм всё равно достаточно быстро приведёт её в жизнеспособную популяцию. Таким образом, на первом шаге можно особенно не стараться сделать слишком уж приспособленных особей, достаточно, чтобы они соответствовали формату особей популяции, и на них можно было подсчитать функцию. Итогом первого шага является популяция X, состоящая из N особей.

Размножение (скрещивание).

Для размножения (чтобы произвести потомка) в генетических алгоритмах обычно нужны несколько родителей (чаще – два). Размножение в различных алгоритмах определяется по-разному – оно зависит от представления данных. Главное требование к размножению – чтобы потомок или потомки имели возможность унаследовать черты обоих родителей, «смешав» их каким-либо разумным способом.

В классическом генетическом алгоритме операция скрещивания представляет собой, так называемое точечное скрещивание. При точечном скрещивании выбираются пары хромосом из родительской популяции. Далее для каждой пары отобранных таким образом родителей разыгрывается позиция гена (локус) в хромосоме, определяющая так называемую точку скрещивания – lk. Если хромосома каждого из родителей состоит из L генов, то очевидно, что точка скрещивания lk представляет собой натуральное число, меньшее L. Поэтому фиксация точки скрещивания сводится к случайному выбору числа из интервала [1, L-1]. В результате скрещивания пары родительских хромосом получается следующая пара потомков: первый потомок, хромосома которого на позициях от 1 до lk состоит из генов первого родителя, а на позициях от до L – из генов второго родителя; второй потомок, хромосома которого на позициях от 1 до lk состоит из генов второго родителя, а на позициях от lk + 1 до L – из генов первого родителя.

К другим видам скрещивания относятся:

- Двухточечное скрещивание – отличается от точечного скрещивания тем, что родительские хромосомы обмениваются участком генетического кода, который находится между двумя случайно выбранными точками скрещивания.
- Многоточечное скрещивание представляет собой обобщение предыдущих операций и характеризуется соответственно большим количеством точек скрещивания.
- Равномерное скрещивание, иначе называемое монолитным или одностадийным, выполняется в соответствии со случайно выбранным эталоном, который указывает, какие гены должны наследоваться от первого родителя (остальные гены берутся от второго родителя).

Для некоторых разновидностей ГА оператор размножения также называется кроссовер.

Мутации.

К мутациям относится все то же самое, что и к размножению: есть некоторая доля мутантов m, являющаяся параметром генетического алгоритма, и на шаге мутации нужно выбрать N особей, а затем изменить их в соответствии с заранее определенными операциями мутации.

Сама структура метода генетических алгоритмов имеет своей целью достижение практического позитивного результата на основании имеющихся данных, ресурсов, что предопределяет сравнительно легкую практическую адаптацию алгоритмов и широкий спектр направлений и специализаций их применения (аспект биологического воспроизведения — создание более эффективной системы через разрешение текущих противоречий в рамках наличных возможностей).

Генетический алгоритм оперирует категориями цикличности, позволяя «многократно ошибаться» в рамках группы циклов и все равно прийти к необходимому результату, а также учитывает фактор времени, который может быть опционально использован в соответствующих вычислениях многократно повышая их практичность и приближенность к реальности.

Для бинарных хромосом могут использоваться следующие мутации:

- В простейшем случае в каждой хромосоме, которая подвергается мутации, каждый бит с вероятностью Pm изменяется на противоположный (это так называемая одноточечная мутация):

  Особь до мутации: 1 0 0 1 0 1 1 0 0 1 1 1
  Особь после мутации: 1 0 0 1 0 1 0 0 0 1 1 1
- Более сложной разновидностью мутации являются операторы инверсии и транслокации.
- Инверсия – это перестановка генов в обратном порядке внутри произвольно выбранного участка хромосомы:

  Особь до инверсии: 1 0 0 1 1 1 1 0 0 1 1 1
  Особь после инверсии: 1 0 0 1 0 0 1 1 1 1 1 1
- Транслокация – это перенос какого-либо участка хромосомы в другой сегмент этой же хромосомы:

  Особь до транслокации: 1 0 0 1 1 1 1 0 0 1 1 1
  Особь после транслокации: 1 1 1 0 0 0 1 1 0 1 1 1

Селекция (отбор).

На этапе отбора нужно из всей популяции выбрать определенную её долю, которая останется «в живых» на этом этапе эволюции и пойдет дальше. Вероятность выживания особи должна зависеть от значения функции приспособленности. Сама доля выживших обычно является параметром генетического алгоритма, и ее просто задают заранее. По итогам отбора из N особей популяции X. Остальные остаются за чертой.

Части селекции:

- **Оценка пригодности** (Пригодность, приспособленность, годность, соответствие). Функция пригодности необходима для определения пригодности конкретного индивидуума, это попытка выработать оптимальное решение проблемы, при этом нужно иметь возможность дать численную оценку.
- **Тиндер** (пул спаривания). Здесь можно использовать несколько разных подходов.

Методы селекции:

- **Принцип колеса рулетки** считается для генетических алгоритмов основным методом отбора особей. Родительские особи выбираются пропорционально значениям их функций приспособленности: каждой хромосоме сопоставлен сектор колеса рулетки, величина которого устанавливается пропорциональной значению функции приспособленности данной хромосомы, поэтому, чем больше значение функции приспособленности, тем больше сектор на колесе рулетки и тем больше шанс того, что данная особь будет учувствовать в скрещивании.
- При **турнирной селекции** все особи популяции разбиваются на подгруппы с последующим выбором в каждой из них особи с наилучшей приспособленностью. Различаются два способа такого выбора: детерминированный выбор и случайный выбор. Детерминированный выбор осуществляется с вероятностью, равной 1, а случайный выбор – с вероятностью, меньшей 1. Подгруппы могут иметь произвольный размер, но чаще всего популяция разделяется на подгруппы по 2-3 особи в каждой.
- При **ранговой селекции** особи популяции ранжируются по значениям их функции приспособленности от наиболее приспособленных к наименее приспособленным (или

наоборот), в котором каждой особи приписывается число, определяющее ее место в списке и называемое рангом.

8.4.4. Функция приспособленности

Для правильного функционирования генетического алгоритма необходима так называемая функция приспособленности (которую еще называют функцией пригодности, а также Fitness function). Функция производит числовую оценку, для определения пригодности конкретного индивидуума. В реальном мире, существа просто выживают или нет.

Чтобы оптимизировать структуру, используя ГА, нужно задать некоторую меру качества для каждой структуры в пространстве поиска. Для этой цели используется функция приспособленности или выживаемости. Часто в качестве функции приспособленности выступает сама целевая функция.

Функция приспособленности — вещественная или целочисленная функция одной или нескольких переменных, подлежащая оптимизации в результате работы генетического алгоритма, направляет эволюцию в сторону оптимального решения (минимум/максимум).

8.5. Контрольные вопросы

1. Концепция эволюционного моделирования
2. Свойства биологической системы
3. Понятие генетического алгоритма
4. Применение генетических алгоритмов
5. Модель квазивидов
6. Естественный отбор в природе
7. Приспособленность «особей» в моделях
8. Мутации «особей»
9. Кроссовер
10. Наследования и мутация
11. Схемы мутации
12. Задача коммивояжера
13. Аллель
14. Функция приспособленности

9. Варианты на выполнение лабораторных работ

Варианты для выполнения лабораторных работ представлены в таблице. Первый столбец определяет предметную область для выполнения всех работ, второй столбец определяет зависимые и независимые переменные для выполнения лабораторной работы №4.

Вариант и предметная область	Переменные: З.п. – зависимая; Н.п. – независимая
1. продовольственный магазин (карта покупателя)	З.п. – прибыль Н.п. – расширение ассортимента товаров, рост количества пользователей
2. кредитное учреждение	З.п. – прибыль Н.п. – процент по кредитам, число заемщиков
3. банк (транзакции)	З.п. – деньги вкладчиков Н.п. – ставка по вкладам, число предложений по типам вкладов
4. страхование жизни	З.п. – прибыль от страхования жизни по городам Н.п. -стоимость страховой премии, вариативность страхования (число программ), продолжительность жизни
5. население стран мира	З.п. – численность (по стране)

	Н.п. -бюджет здравоохранение, бюджет по программам демографии, уровень безработицы
6. туры на отдых	З.п. – общая стоимость проданных туров в страну… Н.п. – количество турфирм, стоимость авиабилетов
7. сеть лингвистических центров	З.п. – количество проданных экземпляров курса Н.п. – продолжительность, уровень, стоимость
8. видеосервис (потоковый)	З.п. – количество подписчиков Н.п. – стоимость подписки, количество сервисов в целом
9. каршеринг	З.п. – доход Н.п. – количество автомобилей, стоимость аренды
10. сетевая доставка пиццы	З.п. – доход Н.п. – стоимость пиццы, количество конкурентов рядом
11. прокат велосипедов (самокатов)	З.п. – доход Н.п. – количество велосипедов, стоимость аренды, количество парковок
12. ресторан (кафе)	З.п. – доход Н.п. – средний чек, количество мест
13. телефонная связь (для провайдера)	З.п. – доход Н.п. – количество пользователей, продолжительность связи (разговора)
14. интернет (для провайдера по пользователям)	З.п. – доход Н.п. – количество пользователей, наличие (количество) собственных потоковых сервисов
15. театральная касса	З.п. – количество проданных билетов Н.п. – количество премьер, погода (температура/осадки)
16. туристический клуб (организатор пеших маршрутов)	З.п. – доход Н.п. -количество маршрутов, продолжительность маршрутов
17. агрегатор такси	З.п. – доход Н.п. – количество водителей, погода (температура/осадки)
18. общественный транспорт (пассажиропоток)	З.п. – доход на маршрутах Н.п. – расстояние между остановками (время в пути), количество остановок, количество крупных торговых точек на маршруте
19. занятость парковки	З.п. – количество машин (%) Н.п. – стоимость парковки, наличие охраны, расстояние до станции метро
20. погода в регионе	З.п. – средняя температура Н.п. – удаленность от экватора, удаленность от моря
21. кинотеатр	З.п. – доход Н.п. – количество мест в зале, количество залов, количество премьер
22. автобусный парк (междугородний)	З.п. – доход Н.п. – количество автобусов, количество маршрутов
23. Библиотека (читатели)	З.п. – количество читателей Н.п. – количество мероприятий, стоимость рекламы
24. Библиотека (книжный фонд)	З.п. – количество книг (по библиотекам) Н.п. – год открытия, год последнего ремонта
25. Ремонтная мастерская	З.п. – доход Н.п. – стоимость рекламных мероприятий, стоимость работ
26. Заказ обедов на дом	З.п. – доход Н.п. – ассортимент, стоимость обедов

27. Школьная успеваемость	З.п. – средняя успеваемость по предмету по классам (школам) Н.п. – количество часов дополнительных занятий в неделю/месяц, стаж работы учителя
28. Продажа товаров на рейсах ВС	З.п. – доход от продажи на рейсе Н.п. – продолжительность рейса, время суток
29. Экскурсии по городу (например, Москве)	З.п. – % заполненности экскурсий Н.п. – количество экскурсий, продолжительность, время начала
30. Продажа автомашин	З.п. – доход Н.п. – количество проданных автомашин, количество менеджеров в салоне
31. Продажа авиаперевозок	З.п. – % пассажиров на рейсе (на заданном маршруте) Н.п. – время вылета, стоимость авиабилета, наличие промежуточных посадок
32. Продажа квартир	З.п. – доход от продажи квартир в доме Н.п. – количество проданных квартир, процент однокомнатных квартир
33. население города	ЗП – численность жителей НП – жилищное строительство, количество рабочих мест
34. служба занятости	ЗП – количество трудоустроенных граждан НП – количество заявок от предприятий, количество зарегистрированных безработных
35. компьютерные курсы	ЗП – доход НП – стоимость курсов, продолжительность курсов, количество программ
36. кадровое агентство	ЗП – количество закрытых заявок НП – количество заявок от предприятий, количество заявок от граждан
37. строительство типовых загородных домов	ЗП - доход НП – количество заказов, курс доллара, стоимость проектов
38. результаты ЕГЭ по региону	ЗП – средний балл (по населенным пунктам) НП – количество учеников, количество школ, количество вузов в населенном пункте
39. движение автомашин по магистрали (например, транспондеры)	ЗП – доход НП – день недели, погода
40. книжный магазин (издательство)	ЗП - доход НП – количество наименований книг, количество новых изданий каждый месяц
41. социальная сеть	ЗП – количество подписчиков НП – доход от рекламы, появление новых инструментов

10. Ссылки

1. https://rosstat.gov.ru/storage/mediabank/ - банк данных (для исходной информации)
2. https://www.gks.ru/free_doc/new_site/ - банк данных (для исходной информации)
3. https://rosstat.gov.ru/storage/mediabank/Ejegodnik_2022.pdf
4. https://rosstat.gov.ru/storage/mediabank/Strani_mira_2022.pdf
5. https://showdata.gks.ru

6. https://www.gks.ru/free_doc/new_site/population/demo/progn1.htmrfr
7. https://new-science.ru/gipoteza/?ysclid=lk0q9r6t5779751286
8. https://loginom.ru/blog/associative-rules?ysclid=lb6ih2v5pg400749483
9. https://habr.com/ru/companies/ods/articles/353502/

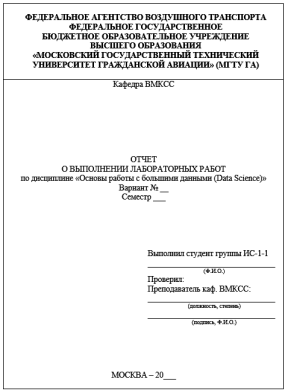
11. Список рекомендуемой литературы

1. Андреас Вайгенд	BIG DATA. Вся технология в одной книге: Эксмо, 2018 г.
2. Ын А., Су К.	Теоретический минимум по Big Data. Всё что нужно знать о больших данных. Издательство: Питер, 2019 г.
3. Макшанов А.В., Журавлев А.Е., Тындыкарь Л.Н.	Большие данные. Big Data. Издательство: Лань, 2021 г.
4. Петрова В. А.	Программирование и решение сложных задач в EXCEL – Екатеринбург, Изд-во Урал. ун-та, 2016. — 88 с.
5. Радченко И.А., Николаев И.Н.	Технологии и инфраструктура BIG DATA. СПб.-Университет ИТМО, 2018. – 52 с.
6. Тутубалин В.Н.	Теория вероятностей / В.Н. Тутубалин. - М.: Academia, 2018. – 210 с.
7. Кузин А. В., Демин В.М.	Разработка баз данных в системе Microsoft Access, - Издательство Форум, 2023г. -224 с.

12. Заключение

Поскольку лабораторные работы, а соответственно и данное учебное пособие имеют практическую направленность и предназначены в первую очередь для отработки навыков, то и защита каждой лабораторной работы проводится с демонстрацией полученных навыков, в том числе по пошаговому выполнению задания. Так как материал, изложенный в этом пособии, является базовым, составляет основу работы с большими данными как при их подготовке, так и при обработке, то его нужно знать безукоризненно, ответы на вопросы давать на уровне автоматизма. Поэтому выполнение каждой работы – это не просто четкое выполнение задания, но и практическое изучение материала, связанного с заданием. Вариативность выполнения каждого задания (несколькими способами) позволит лучше понять изучаемый теоретический материал и приемы работы. В рамках лабораторной работы приветствуется дополнительная активность, использование не указанных в пособии операций, выполнение дополнительных заданий.

Титульный лист для оформления отчетов по лабораторным работам



ФЕДЕРАЛЬНОЕ АГЕНТСТВО ВОЗДУШНОГО ТРАНСПОРТА
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ ГРАЖДАНСКОЙ АВИАЦИИ» (МГТУ ГА)

Кафедра ВМКСС

ОТЧЕТ
О ВЫПОЛНЕНИИ ЛАБОРАТОРНЫХ РАБОТ
по дисциплине «Основы работы с большими данными (Data Science)»
Вариант № __
Семестр ___

Выполнил студент группы ИС-1-1
_____
(Ф.И.О.)
Проверил:
Преподаватель каф. ВМКСС:
_____
(должность, степень)
_____
(подпись, Ф.И.О.)

МОСКВА – 20___