МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ЧЕРЕПОВЕЦКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Институт информационных технологий

Кафедра математики и информатики

УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ

«МАШИННОЕ ОБУЧЕНИЕ»

Направление подготовки (специальность):
01.03.02 Прикладная математика и информатика

Образовательная программа:
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Очная форма обучения

Составители:

Лягинова О.Ю., зав. кафедрой МиИ
канд.пед.наук, доцент

Лягинов Н.М., старший
преподаватель кафедры МиИ

г. Череповец - 2022

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

Основная литература:

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
2. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O'Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.

Дополнительная литература по дисциплине:

1. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля), включая перечень информационных справочных систем (при необходимости)

1. Электронная библиотека «Университетская библиотека online». URL: http://biblioclub.ru/.
2. Информационная система «Единое окно доступа к образовательным ресурсам». URL: http://window.edu.ru/.
3. Образовательный портал Череповецкого государственного университета. URL: https://edu.chsu.ru/.
4. Образовательная платформа Stepik, онлайн курсы: Программирование на Python: https://stepik.org/course/67/promo; Машинное обучение, URL: https://stepik.org/course/8057/promo.
5. Технологический акселератор ML START, онлайн курс. URL: https://youtube.com/playlist?list=PLrSH_ggigfrlXzHj8aLKj1cjPfwORqIxy

Учебно-методические указания и рекомендации к изучению тем лекционных и практических занятий, самостоятельной работе студентов

Лекции

№ п/п	Тема лекции	Количество часов
1	Введение в машинное обучение.	4
2	Исследование данных, их визуализация и интерпретация.	4
3	Методы классификации.	4

4	Методы числового прогнозирования.	4
5	Обнаружение закономерностей на основе ассоциативных правил.	4
6	Методы кластеризации.	4
7	Методы понижения размерности данных.	4
	Итого	28

Лабораторные работы

№ п/п	Тема лекции	Количество часов
1	Исследование данных, их визуализация и интерпретация.	12
2	Методы классификации.	12
3	Методы числового прогнозирования.	12
4	Обнаружение закономерностей на основе ассоциативных правил.	12
5	Методы кластеризации.	12
6	Методы понижения размерности данных.	12
	Итого	72

Раздел 1. Введение в машинное обучение

Содержание:

Понятия «наука о данных», «машинное обучение» (далее англ. machine learning, ML), «интеллектуальный анализ данных». Составляющие ML: хранение данных; абстрагирование; обобщение; оценка. Этапы решения задач с использованием ML: сбор данных; исследование и подготовка данных; обучение модели; оценка модели; улучшение модели. Типы входных данных. Типы алгоритмов машинного обучения. Подбор алгоритмов по входным данным. Библиотеки Python для машинного обучения. Методология ML Ops.

Самостоятельная работа по разделу:

По итогам самостоятельной работы студент готовит отчет, включающий в себя ответы на вопросы и решение заданий, предполагавшихся к выполнению в ходе самостоятельной работы. Отчет сдается преподавателю в электронной форме.

Задания для самостоятельной работы:

1. Приведите понятия «наука о данных», «машинное обучение», «большие данные», «интеллектуальный анализ данных».
2. Как Вы считаете, чем машинное обучение отличается от интеллектуального анализа данных (если эти понятия отличаются друг от друга)?
3. Приведите примеры использования методов машинного обучения.
4. Подготовьте интеллект-карту, включающую в себя представление составляющих машинного обучения: хранение данных; абстрагирование; обобщение; оценка.
5. Приведите описание этапов решения задач с использованием машинного обучения: сбор данных; исследование и подготовка данных; обучение модели; оценка модели; улучшение модели.

6. Дайте описание типов входных данных, используемых при решении задач с помощью методов машинного обучения.
7. Перечислите типы алгоритмов машинного обучения.
8. Как подбирается метод машинного обучения для решения конкретной прикладной задачи? Что влияет на выбор метода?
9. Каково назначение и возможности библиотек библиотеки Python для машинного обучения (дайте заключение на основе анализа документации разработчиков библиотек).
10. Перечислите правовые нормы и стандарты в области искусственного интеллекта, действующие в РФ.
11. Каковы этические нормы и стандарты в области искусственного интеллекта?
12. Перечислите основные международные и национальные стандарты и методологии разработки автоматизированных систем и программного обеспечения, стандарты в области информационной безопасности, подходы к управлению и фундаментальные принципы работы, развития и использования технологий искусственного интеллекта.
13. Как осуществляется поиск зарегистрированных результатов интеллектуальной деятельности и средств индивидуализации?
14. Как провести исследование результатов интеллектуальной деятельности и средств индивидуализации при создании инновационных продуктов в профессиональной деятельности?
15. Назовите принципы защиты прав результатов интеллектуальной деятельности и средств индивидуализации при создании инновационных продуктов в профессиональной деятельности.
16. Как осуществляется защита прав результатов интеллектуальной деятельности и средств индивидуализации при создании инновационных продуктов в профессиональной деятельности?
17. Приведите описание критериев эффективности и качества функционирования системы искусственного интеллекта: точность, релевантность, достоверность, целостность, быстрота решения задач, надежность, защищенность функционирования систем искусственного интеллекта.
18. Приведите описание методов постановки задач, проведения и анализа тестовых и экспериментальных испытаний работоспособности систем искусственного интеллекта, в том числе систем машинного обучения.
19. Перечислите методы и критерии оценки качества моделей машинного обучения.
20. Приведите содержание унифицированных и обновляемых методологии описания, сбора и разметки данных, а также механизмов контроля за соблюдением указанных методологий.
21. Что такое ML Ops? Перечислите основные этапы жизненного цикла систем машинного обучения.

Литература:

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
2. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O’Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.

3. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

Раздел 2. Исследование данных, их визуализация и интерпретация

Содержание:

Преобразование данных, построение выводов по данным и оценка результатов. Структуры данных. Числовые переменные. Измерение средних значений: среднее арифметическое и медиана. Измерение разброса: квартили и пятичисловая сводка. Визуализация числовых переменных: диаграммы размаха; гистограммы (разбиения по интервалам и плотность). Интерпретация числовых данных: равномерное и нормальное распределение. Измерение разброса: дисперсия и стандартное отклонение. Категориальные переменные. Мода. Взаимосвязи между переменными. Визуализация отношений: диаграммы разброса. Исследование взаимосвязей: перекрестные таблицы.

Самостоятельная работа по разделу:

По итогам самостоятельной работы студент готовит отчет, включающий в себя ответы на вопросы и решение заданий, предполагавшихся к выполнению в ходе самостоятельной работы. Отчет сдается преподавателю в электронной форме.

Задания для самостоятельной работы:

1. Для каких целей выполняется интерпретация данных?
2. Что такое структура данных?
3. Какие базовые наборы изменений обычно применяются в числовым данным?
4. Почему в ходе исследования данных запрашивают как средние, так и медианные значения числовых переменных?
5. Что такое «пятичисловая сводка»? Для каких целей она используется?
6. Что отображает диаграмма размаха?
7. Что отображает гистограмма?
8. Как выглядит гистограмма равномерного распределения?
9. Как выглядит кривая нормального распределения?
10. Что измеряется стандартным отклонением?
11. Что гласит правило «68–95–99,7»?
12. Что отображает таблица частотности?
13. Для каких целей строится диаграмма разброса?
14. Что показывают перекрестные таблицы (кросс-таблицы, таблицы сопряженности)?

Образцы заданий для лабораторных работ:

По итогам выполнения лабораторной работы студент демонстрирует результаты работы программы преподавателю, предварительно разработав тестовые случаи, а также сдает в электронном виде отчет, содержащий порядок выполнения работы.

Лабораторная работа «Исследование данных, их визуализация и интерпретация».

Изучите документацию разработчиков библиотек Pandas, Matplotlib и выполните представленные ниже задания:

1. загрузите данные из файла usedcars.csv в dataframe usedcars;
2. отобразите структуру usedcars;
3. запросите статистику по всем числовым переменным usedcars;
4. посчитайте средние значения для всех числовых переменных usedcars;
5. посчитайте медианы для всех числовых переменных usedcars;
6. изучите пятичисловую сводку для переменных price и mileage;
7. постройте диаграммы размаха для переменных price и mileage;
8. постройте гистограмму для данных о цене и пробеге подержанных автомобилей;
9. вычислить дисперсию и стандартное отклонение по векторам price и mileage;
10. постройте таблицу частотности для данных о подержанном автомобиле;
11. вычислите моду переменных year, model и color;
12. ответьте на вопрос о соотношении цены и пробега, построив диаграмму разброса;
13. ответьте на вопрос о том, существует ли связь между моделью и цветом, построив кросс-таблицу.

Литература:

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
2. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O’Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.
3. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

Раздел 3. Методы классификации

Содержание:

Ленивое обучение, классификация с использованием метода ближайших соседей: что такое классификация методом ближайших соседей; алгоритм k-NN; измерение степени сходства с помощью расстояния; выбор подходящего k; подготовка данных для использования в алгоритме k-NN; почему алгоритм k-NN называют ленивым. Вероятностное обучение, классификация с использованием наивного байесовского классификатора: наивный байесовский классификатор; основные понятия байесовских методов; наивный байесовский алгоритм; классификация по наивному байесовскому алгоритму; Критерий Лапласа; использование числовых признаков в наивном байесовском алгоритме. Классификация с использованием деревьев решений и правил: деревья решений; выбор лучшего разделения; сокращение дерева решений.

Самостоятельная работа по разделу:

По итогам самостоятельной работы студент готовит отчет, включающий в себя ответы на вопросы и решение заданий, предполагавшихся к выполнению в ходе самостоятельной работы. Отчет сдается преподавателю в электронной форме.

Задания для самостоятельной работы:

1. В чем заключается суть метода k-NN?
2. Приведите примеры задач, решаемых с использованием метода k-NN.
3. Каковы преимущества метода k-NN?
4. Каковы недостатки метода k-NN?
5. Как измеряется степень сходства между экземплярами набора данных?
6. Каким образом выбирается подходящее k?
7. Что такое «минимаксная» нормализация?
8. Каким образом выполняется стандартизация по z-оценке?
9. Что такое «фиктивное» кодирование?
10. Почему алгоритм k-NN называют ленивым?
11. Изучите документацию разработчиков библиотеки Scikit-learn (https://scikit-learn.org/stable/) в части реализации метода k-NN.
12. Изучите пример использования метода k-NN для классификации данных (https://pythonru.com/uroki/sklearn-kmeans-i-knn).
13. Что такое «вероятностное обучение»?
14. В чем заключается суть работы наивного байесовского классификатора?
15. Приведите примеры задач, решаемых с использованием наивного байесовского классификатора.
16. Каковы преимущества наивного байесовского классификатора?
17. Каковы недостатки наивного байесовского классификатора?
18. Почему алгоритм называют наивным?
19. Изучите документацию разработчиков библиотеки Scikit-learn (https://scikit-learn.org/stable/) в части реализации наивного байесовского классификатора.
20. Изучите пример использования наивного байесовского алгоритма для классификации данных (https://russianblogs.com/article/2703524871/).
21. Для каких целей используются методы деревьев?
22. Почему группа методов получила такое название?
23. Приведите примеры задач, решаемых с использованием деревьев.
24. Что такое «рекурсивное сегментирование»?
25. Каким образом работает алгоритм дерева решений C5.0?
26. Каким образом выбирается лучшее разделение?
27. С какой целью выполняется «сокращение» дерева решений?
28. Изучите документацию разработчиков библиотеки Scikit-learn (https://scikit-learn.org/stable/) в части реализации деревьев решений.
29. Изучите пример использования дерева решений для классификации данных (https://www.machinelearningmastery.ru/scikit-learn-decision-trees-explained-803f3812290d/).

Образцы заданий для лабораторных работ:

По итогам выполнения лабораторной работы студент демонстрирует результаты работы программы преподавателю, предварительно разработав тестовые случаи, а также сдает в электронном виде отчет, содержащий порядок выполнения работы.

Лабораторная работа «Классификация методом k-NN»

Обычный скрининг рака позволяет диагностировать и вылечить это заболевание до того, как появятся заметные симптомы. Процесс раннего выявления включает в себя

исследование ткани на наличие аномальных уплотнений или новообразований. Если такое уплотнение обнаружится, то выполняется аспирационная биопсия с использованием полой тонкой иглы, которой из этого новообразования извлекают небольшое количество клеток. Затем врач рассматривает клетки под микроскопом и определяет, злокачественное это новообразование или доброкачественное. Интеллектуальная система, позволяющая автоматизировать идентификацию раковых клеток, принесла бы значительную пользу системе здравоохранения. Автоматизированные процессы, очевидно, повысят эффективность процесса выявления рака, что сократит время диагностики и позволит уделять больше внимания лечению заболевания. Интеллектуальная программа скрининга могла бы также обеспечить большую точность диагностики, исключив из процесса субъективный человеческий фактор. Напишите программу для выявления рака, применив алгоритм k-NN к исследованиям клеток, полученных при биопсии.

Лабораторная работа «Классификация с использованием наивного байесовского алгоритма»

По мере роста популярности мобильных телефонов во всем мире появились новые возможности для распространения рекламы по почте, используемые недобросовестными маркетологами. Такие рекламодатели используют короткие текстовые сообщения (СМС), чтобы привлечь потенциальных потребителей нежелательной рекламой, известной как СМС-спам. Этот тип спама является особенно опасным, поскольку, в отличие от почтового спама, СМС может причинить больше ущерба из-за широкого использования мобильных телефонов. Разработка интеллектуальной программы классификации, которая бы фильтровала СМС-спам, стала бы полезным инструментом для операторов сотовой связи. Поскольку наивный байесовский алгоритм успешно применялся для фильтрации спама в электронной почте, вполне вероятно, что он также может быть применен к СМС-спаму. Однако в отличие от спама в электронной почте СМС-спам создает дополнительные проблемы для автоматических фильтров. Размер СМС часто ограничен 160 символами, что сокращает объем текста, по которому можно определить, является ли сообщение нежелательным. Такое ограничение привело к тому, что сформировался своеобразный сокращенный СМС-язык, что еще больше стирает грань между обычными сообщениями и спамом. Напишите программу для фильтрации СМС-спама, используя наивный байесовский алгоритм.

Лабораторная работа «Классификация с использованием деревьев решений»

Мировой финансовый кризис 2007–2008 годов показал, как важна прозрачность и строгость в принятии банковских решений. Когда кредиты стали менее доступными, банки ужесточили систему кредитования и обратились к машинному обучению для более точного определения рискованных кредитов. Благодаря высокой точности и возможности формулировать статистическую модель на понятном человеку языке деревья решений широко применяются в банковской сфере. Поскольку правительства многих стран тщательно следят за справедливостью кредитования, руководители банков должны быть в состоянии объяснить, почему одному заявителю было отказано в получении займа, в то время как другому одобрили выдачу кредита. Эта информация полезна и для клиентов, желающих узнать, почему их кредитный рейтинг оказался неудовлетворительным. Автоматические модели оценки кредитоспособности используются для рассылок по кредитным картам и мгновенных онлайн-процессов одобрения кредитов. Разработайте простую модель принятия решения о предоставлении кредита с использованием

алгоритма построения деревьев решений. Настройте параметры модели, чтобы свести к минимуму ошибки, которые могут привести к финансовым потерям.

Литература:

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
2. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O'Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.
3. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

Раздел 4. Методы числового прогнозирования

Содержание:

Прогнозирование числовых данных, регрессионные методы: понятие регрессии; простая линейная регрессия; оценка методом наименьших квадратов; корреляции; множественная линейная регрессия.

Самостоятельная работа по разделу:

По итогам самостоятельной работы студент готовит отчет, включающий в себя ответы на вопросы и решение заданий, предполагавшихся к выполнению в ходе самостоятельной работы. Отчет сдается преподавателю в электронной форме.

Задания для самостоятельной работы:

1. Для каких целей используются методы регрессии?
2. Приведите понятие регрессии.
3. Приведите примеры задач, решаемых с использованием регрессии.
4. Как определяется простая линейная регрессия?
5. Приведите описание оценки методом наименьших квадратов.
6. Как рассчитывается коэффициент корреляции Пирсона?
7. Приведите описание множественной линейной регрессии. В чем заключаются преимущества и недостатки данного метода?
8. Изучите документацию разработчиков библиотеки Scikit-learn (https://scikit-learn.org/stable/) в части реализации линейной регрессии.
9. Изучите пример использования линейной регрессии для числового прогнозирования (https://pythonru.com/uroki/linear-regression-sklearn).

Образцы заданий для лабораторных работ:

По итогам выполнения лабораторной работы студент демонстрирует результаты работы программы преподавателю, предварительно разработав тестовые случаи, а также сдает в электронном виде отчет, содержащий порядок выполнения работы.

Лабораторная работа «Прогнозирование числовых данных, регрессия»

Для того чтобы медицинская страховая компания могла зарабатывать деньги, необходимо, чтобы сумма ежегодных взносов превышала расходы на медицинское обслуживание бенефициаров. Следовательно, страховщики вкладывают много времени и денег в разработку моделей, которые точно прогнозируют медицинские расходы застрахованного населения. Медицинские расходы трудно оценить, поскольку самые дорогостоящие случаи происходят редко и кажутся случайными. Тем не менее некоторые ситуации являются более распространенными для определенных слоев населения. Например, рак легких чаще встречается у курильщиков, чем у некурящих, а от болезней сердца чаще страдают тучные люди. Целью этого анализа является использование данных о пациентах для прогнозирования средних расходов на медицинское обслуживание для подобных групп населения. Эти оценки могут быть использованы для создания страховых таблиц, согласно которым сумма ежегодных взносов устанавливается выше или ниже в зависимости от ожидаемых затрат на лечение. Используя регрессию, напишите программу, дающую прогноз стоимости медицинской страховки для конкретного клиента.

Литература:

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
4. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O’Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.
5. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

Раздел 5. Обнаружение закономерностей на основе ассоциативных правил

Содержание:

Ассоциативные правила. Типы задач, решаемых с использованием ассоциативных правил. Алгоритм Apriori для поиска ассоциативных правил, преимущества и недостатки алгоритма. Измерение интересности правила: поддержка и доверие. Построение набора правил по принципу Apriori. Выявление часто покупаемых продуктов в соответствии с ассоциативными правилами.

Самостоятельная работа по разделу:

По итогам самостоятельной работы студент готовит отчет, включающий в себя ответы на вопросы и решение заданий, предполагавшихся к выполнению в ходе самостоятельной работы. Отчет сдается преподавателю в электронной форме.

Задания для самостоятельной работы:

1. В чем заключается суть метода ассоциативных правил?
2. Какие задачи решаются с использованием данного метода?
3. К какому типу методов машинного обучения относится метод ассоциативных правил?
4. В чем заключается суть метода Apriori?
5. В каких библиотеках Python реализован метод ассоциативных правил?
6. Проанализируйте документацию разработчиков библиотек. Каким образом производится обучение модели? Какие параметры необходимо указать для запуска обучения? Как проверить эффективность модели?
7. Что необходимо сделать, чтобы повысить эффективность модели?
8. Как сохранить ассоциативные правила в файл или фрейм данных?
9. Изучите пример решения задачи с использованием метода ассоциативных правил (http://datascientist.one/apriori-algorithm/).

Образцы заданий для лабораторных работ:

По итогам выполнения лабораторной работы студент демонстрирует результаты работы программы преподавателю, предварительно разработав тестовые случаи, а также сдает в электронном виде отчет, содержащий порядок выполнения работы.

Лабораторная работа «Ассоциативные правила»

Анализ потребительской корзины применяется рекомендательными системами, используемыми во многих обычных и интернет-магазинах. Выявленные ассоциативные правила указывают на сочетания товаров, которые часто покупаются вместе. Знание этих паттернов позволяет создать новые способы оптимизации товаров в сети продуктовых магазинов, рекламных акций или раскладки товаров в магазине. Например, если покупатели часто приобретают на завтрак кофе или апельсиновый сок вместе с выпечкой, то, возможно, удастся повысить прибыль, если разместить выпечку поближе к кофе и сокам. Однако эти методы можно применять ко многим другим типам задач, от рекомендаций фильмов до обнаружения опасных зависимостей между лекарствами. При этом алгоритм Apriori способен эффективно обрабатывать потенциально большие наборы ассоциативных правил. Выполните анализ потребительской корзины на основе данных о транзакциях продуктового магазина.

Литература:

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
2. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O’Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.
3. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

Раздел 6. Методы кластеризации

Содержание:

Что такое кластеризация. Кластеризация как задача машинного обучения. Алгоритм кластеризации методом k-средних: преимущества и недостатки метода; использование расстояния для разбиения на кластеры и внесения изменений; выбор количества кластеров. Сегментация рынка для подростков с использованием кластеризации методом k-средних.

Самостоятельная работа по разделу:

По итогам самостоятельной работы студент готовит отчет, включающий в себя ответы на вопросы и решение заданий, предполагавшихся к выполнению в ходе самостоятельной работы. Отчет сдается преподавателю в электронной форме.

Задания для самостоятельной работы:

1. Что такое «кластеризация»? Чем кластеризация отличается от классификации?
2. Какие задачи решаются с использованием методов кластеризации?
3. Перечислите известные Вам методы кластеризации.
4. В чем заключатся суть метода k-средних?
5. Перечислите достоинства и недостатки метода k-средних
6. В каких библиотеках Python реализован метод k-средних?
7. Проанализируйте документацию разработчиков библиотек. Каким образом производится обучение модели? Какие параметры необходимо указать для запуска обучения? Как проверить эффективность модели?
8. Что необходимо сделать, чтобы повысить эффективность модели?
9. Изучите пример решения задачи с использованием метода k-средних (https://coderlessons.com/tutorials/python-technologies/uznaite-mashinnoe-obuchenie-s-python/ml-algoritm-klasterizatsii-k-srednikh).

Образцы заданий для лабораторных работ:

По итогам выполнения лабораторной работы студент демонстрирует результаты работы программы преподавателю, предварительно разработав тестовые случаи, а также сдает в электронном виде отчет, содержащий порядок выполнения работы.

Лабораторная работа «Кластеризация методом k-средних»

Общение с друзьями в социальных сетях, таких как Facebook, ВКонтакте, Instagram и др. стало для подростков всего мира обычным делом. Имея достаточное количество наличных денег, подростки являются желанной социально-демографической группой для компаний, которые продают закуски, напитки, электронику и средства гигиены. Миллионы подростков, посещающих такие сайты, привлекли внимание маркетологов, стремящихся найти свою нишу на все более высококонкурентном рынке. Один из способов найти такую нишу — выявление среди подростков групп, имеющих схожие вкусы, чтобы клиенты, не заинтересованные в этих товарах, не получали рекламу, ориентированную на подростков. Например, скорее всего, будет трудно продать спортивную одежду тем подросткам, которые не интересуются спортом. Исходя из информации на страницах подростков в социальных сетях, можно выделить группы с общими интересами, такими

как спорт или музыка. Кластеризация может автоматизировать процесс обнаружения естественных сегментов в этой социально-возрастной группе. Однако только нам решать, насколько эти кластеры интересны и как их можно использовать для рекламы. Используя алгоритм кластеризации k-средних, напишите программу, выполняющую сегментацию рынка для подростков.0000000000

Литература:

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
2. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O'Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.
3. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

Раздел 7. Методы понижения размерности данных

Содержание:

Для чего понижают размерность данных. Метод главных компонент, новая система координат, достоинства и ограничения метода. Использование метода главных компонент для понижения размерности данных успеваемости школьников.

Самостоятельная работа по разделу:

По итогам самостоятельной работы студент готовит отчет, включающий в себя ответы на вопросы и решение заданий, предполагавшихся к выполнению в ходе самостоятельной работы. Отчет сдается преподавателю в электронной форме.

Задания для самостоятельной работы:

1. В чем заключается принцип работы алгоритма понижения размерности данных t-SNE?
2. Какие задачи решаются с использованием данного алгоритма?
3. В каких библиотеках Python реализован данный алгоритм?
4. Изучите документацию разработчиков по оценщику TSHE, реализующему алгоритм понижения размерности данных t-SNE (https://scikit-learn.org/stable/modules/manifold.html#t-sne).
5. Каким образом можно выполнить визуализацию результата работы оценщика TSHE? Проанализируйте информацию разработчиков средств визуализации.

Образцы заданий для лабораторных работ:

По итогам выполнения лабораторной работы студент демонстрирует результаты работы программы преподавателю, предварительно разработав тестовые случаи, а также сдает в электронном виде отчет, содержащий порядок выполнения работы.

Лабораторная работа «Понижение размерности данных. Метод главных компонент»

В наборе данных содержится информация о 200 школьниках в США: их поле, этнической принадлежности, социально-экономическом статусе, типе школы, программе обучения и оценкам по пяти предметам (чтение, письмо, математика, естественные науки и социальные науки).

```
##    id female race ses schtyp prog read write math science socst
## 1  70      0    4   1      1    1   57    52   41      47     57
## 2 121      1    4   2      1    3   68    59   53      63     61
## 3  86      0    4   3      1    1   44    33   54      58     31
## 4 141      0    4   3      1    3   63    44   47      53     56
## 5 172      0    4   2      1    2   47    52   57      53     61
```

Постройте парные диаграммы рассеяния для предметов, как скоррелированы оценки между собой? Примените метод главных компонент, передав в него оценки по пяти предметам. Что описывает первая главная компонента? Какой вклад вносят предметы в первую главную компоненту? Что представляет собой вторая главная компонента? Проанализируйте связь успеваемости с категориальными переменными.

Литература:

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
2. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O’Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-5-4461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.
3. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

Средства контроля качества обучения

Вопросы к экзамену:

1. Понятия «наука о данных», «машинное обучение» (далее *англ*. machine learning, ML), «большие данные», «интеллектуальный анализ данных».
2. Составляющие ML: хранение данных; абстрагирование; обобщение; оценка.
3. Этапы решения задач с использованием ML: сбор данных; исследование и подготовка данных; обучение модели; оценка модели; улучшение модели.
4. Типы входных данных.

5. Типы алгоритмов машинного обучения.
6. Подбор алгоритмов по входным данным.
7. Библиотека Scikit-Learn.
8. Методология ML Ops. Этапы жизненного цикла систем машинного обучения.
9. Преобразование данных, построение выводов по данным и оценка результатов.
10. Структуры данных. Числовые переменные.
11. Измерение средних значений: среднее арифметическое и медиана.
12. Измерение разброса: квартили и пятичисловая сводка.
13. Визуализация числовых переменных: диаграммы размаха; гистограммы (разбиения по интервалам и плотность).
14. Интерпретация числовых данных: равномерное и нормальное распределение.
15. Измерение разброса: дисперсия и стандартное отклонение.
16. Категориальные переменные. Мода.
17. Взаимосвязи между переменными.
18. Визуализация отношений: диаграммы разброса.
19. Исследование взаимосвязей: перекрестные таблицы.
20. Ленивое обучение, классификация с использованием метода ближайших соседей.
21. Вероятностное обучение, классификация с использованием наивного байесовского классификатора.
22. Классификация с использованием деревьев решений и правил.
23. Прогнозирование числовых данных, регрессионные методы.
24. Ассоциативные правила. Типы задач, решаемых с использованием ассоциативных правил.
25. Алгоритм Apriori для поиска ассоциативных правил, преимущества и недостатки алгоритма.
26. Измерение интересности правила: поддержка и доверие.
27. Построение набора правил по принципу Apriori.
28. Кластеризация как задача машинного обучения.
29. Алгоритм кластеризации методом k-средних.
30. Понижение размерности данных. Метод главных компонент, новая система координат, достоинства и ограничения метода.