

Интеллектуальный анализ данных

Лекция 1. Наука о данных. Цели, терминология, средства

Зуев С.В. 09-09-2024

Информация и данные

Концепции и терминология

- Информация - знание, относящееся к объектам, которое в рамках определенного контекста имеет конкретное значение (ГОСТ Р ИСО/МЭК 2382-1)
 - Примечание 1 - Примерами объектов являются факты, события, предметы, процессы и идеи, включая концепции.
 - Примечание 2 - Информация это нечто, что является значимым. Данные могут рассматриваться как информация, если выявлено их значение.
- Данные (data): реинтерпретируемое представление информации в формализованном виде, пригодном для коммуникации, интерпретации или обработки. (Тот же источник)
 - Примечание: Данные могут быть обработаны людьми или автоматическими средствами.
 - Понятие данных использует понятие информации, но данные могут не являться информацией (см. Примечание 2) - противоречие (в государственных и международных стандартах).
 - Данные - определенным образом упорядоченные структуры памяти. Все экземпляры одних и тех же данных имеют одну и ту же упорядоченность. Структуры памяти представляют собой множества пар «ключ-значение», где значением может являться число или другая структура памяти.
 - Информация - это доступные связи между данными. Информация может быть интерпретируемой, когда связи сами представляют собой данные, и не-интерпретируемой, когда связи не записаны как данные, но тем не менее используются для преобразования других данных.

Информация и данные

Примеры

- **«Каждый вечер солнце садится за горизонт».** Это не данные. Это информация, так как здесь содержится только связь между понятиями вечер, солнце и горизонт. Для машины это не интерпретируемая информация, так как описывается процесс, представление которого в памяти не определено. Для человека это интерпретируемая информация, так как она связывает понятия (ключи в памяти человека) и действие (значение в его памяти).
- **Расписание занятий в университете.** Это данные и, в то же время, информация. Расписание можно представить в виде набора пар «ключ-значение». Разные экземпляры данных имеют общие ключи или общие значения на разных структурных уровнях и могут быть сгруппированы по этим классам эквивалентности. То есть, связи между данными представляют собой данные - это интерпретируемая информация.
- **Набор событий информационной безопасности.** Это данные, так как эти события обычно представляют собой пары «ключ-значение». Но это не информация до тех пор, пока не будут установлены доступные связи между экземплярами данных.

Данные и их представления

Данные предназначены для машин!

- **Структуры данных.** Это представление данных в памяти. В итоге сводятся к иерархии пар «ключ-значение», но для упрощения используются более понятные человеку конструкции: таблицы, деревья, множества, ... В разных языках программирования используются похожие и, в то же время, немного различающиеся структуры данных.
- **Размещение в памяти.** Делается средствами языка программирования и операционной системы. Для разработчика на языках высокого уровня это автоматизировано. Для языков низкого уровня (ассемблер) и тех, которые позволяют вмешиваться в адресацию, размещение структур в памяти может быть организовано разработчиком.
- **Трансформации данных.** Главное при преобразовании данных до их анализа - не потерять информацию, то есть, не увеличить необоснованно число связей между экземплярами.
- Представление данных для человека называется **визуализацией**.
- В данных могут присутствовать **временные метки** - это пары «время: значение» и **метки классификации** - это пары «класс: значение». Метки являются управляющими данными для анализа.

Данные и их представления

Примеры

- **Табличные данные.** Самый простой вид данных. Ключом является номер (или название) столбца, значением - число в этом столбце. Экземпляр данных - это строка.
- **Тензорные данные.** Используются для машинного обучения. Представляют собой таблицы таблиц. Уровень вложенности может повышаться.
- **Текстовые данные.** Сводятся к древовидной структуре: {текст: {№ предложения: {№ слова: код слова}}}. Могут иметь метки классификации и дополнительные данные о кодировке слов (так называемые эмбеддинги).
- **Графические данные.** Представляют собой структуру вида {адрес: пиксель} или {адрес: кривая} с вариациями. Пиксель, кривая и другие значения представляют собой структуры данных.
- **Видеоданные.** Это последовательность графических данных, то есть, графические данные с временными метками, значения которых дискретны.

Обработка и анализ данных

Обработка - это не анализ!

- **Обработка данных** - это сбор и манипулирование данными для получения значимой информации (C. French (1996). Data Processing and Information Technology. Thomson. p.2).
 - То есть, обработка должна выявить допустимые связи в данных. Но! См. определение анализа данных.
- **Анализ данных** - это область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из данных (Энциклопедический словарь Брокгауза и Ефона); процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений (Социология: энциклопедия).
 - В анализе речь идет об извлечении знаний из данных (это грубо говоря то же самое, что информация о принятии решений).
- **Знания** характеризуются значительно меньшим количеством, чем сырья информации. Поэтому:
 - **Обработка** - преобразования без изменения количества информации;
 - **Анализ** - преобразования с уменьшением количества информации в датасете (в результате получается новая информация у аналитика).
- **Обработка:** ввод, парсинг, эмбеддинг, кодировка, преобразование (размерности, вида, представления), вывод данных, хранение, передача и т.д.
- **Анализ:** кластеризация, классификация, прогнозирование, распознавание, оценка рисков, аномалии, и подобные задачи.
- **Интеллектуальный анализ данных.** Называется так, если для решения задач анализа применяются методы искусственного интеллекта: эвристические, нечеткая логика, машинное обучение.

Задачи анализа данных

Подробнее

- **Кластеризация** - это распределение экземпляров данных на группы (кластеры), внутри которых у данных есть какое-то общее свойство. Если в пространстве признаков есть расстояние, то экземпляры данных - это точки, а кластеры - это облака точек, внутри которых все точки принадлежат одному кластеру.
- **Классификация** - это распределение экземпляров данных на группы (классы) по приведенному образцу, причем число классов всегда известно. Данные делятся на обучающие и тестовые, причем в обучающих присутствует метка классификации.
- **Прогнозирование** имеет смысл, когда в данных присутствует временная метка или пространственная метка; задача заключается в построении новых экземпляров данных, соответствующих заданным временным или пространственным меткам.
- **Распознавание** похоже на классификацию, но вместо жесткого отнесения к классу выдается вероятность принадлежности экземпляра к тому или иному классу.
- **Оценка рисков** производится тогда, когда имеется хотя бы один профиль риска, то есть, набор примеров или описание того, как выглядит рисковый экземпляр данных; оценка риска - это распознавание принадлежности экземпляра профилю риска;
- **Аномалиями** называются такие экземпляры данных, которые будучи не бракованными, не являются близкими к подавляющему большинству других данных; в случае данных с временными метками аномалиями считаются также процессы, протекающие с данными так, что они резко отличаются от процессов в других временных промежутках, в том числе, процессы с нетипичным временем начала и окончания; для данных с пространственными метками аномалиями являются экземпляры данных, значительно отличающиеся от других, и можно говорить о пространственной аномалии в секторе ... (указывается набор пространственных меток аномальных экземпляров).
- **Комбинированные задачи.** Например, распознавание аномалий, прогнозирование аномалий, прогнозирование кластерной структуры, классификация рисков, и т.п.

Задачи анализа данных могут решаться с данными любых типов

Машинное обучение

как методология интеллектуального анализа данных

- **Машинное обучение** - это процесс заполнения заданной структуры данных значениями, которые решают задачу интеллектуального анализа данных по заданному алгоритму (упрощенное определение).
 - Для машинного обучения нужны конкретный алгоритм и конкретная структура данных - вместе они называются моделью машинного обучения;
 - Процесс заполнения структуры данных правильными значениями называется процессом обучения; алгоритмы процессов обучения могут быть разными для одной и той же модели.
- **Машинное обучение с учителем** подразумевает наличие в обучающих данных меток классификации - учителем является именно набор данных с такими метками.
 - Самый простой метод обучения с учителем - метод ближайших соседей: структура данных - число (соседей), алгоритм - пробному экземпляру присваивается класс, наиболее распространенный среди указанного числа ближайших соседей. Это - модель. Алгоритм обучения сводится к определению того, с каким числом классификация получается наилучшим образом (перебор).
- **Машинное обучение без учителя** находит модель в самих данных, если задана структура данных для заполнения.
 - Пример - кластеризация в табличных данных. Структура - множество пар «кластер: центр» (в Python - словарь). Алгоритм: экземпляр относится к тому кластеру, центр которого ближе. Алгоритм обучения: после отнесения пробного экземпляра к кластеру, центры кластеров рассчитываются заново как средние координаты точек кластера.
- **Машинное обучение с подкреплением** использует для решения задачи некоторую функцию оценки, которая по одному или двум аргументам (экземпляру данных и, дополнительно, метке классификации) выдает числовую оценку соответствия экземпляра целям обучения.
- **Онлайн-обучение** характеризуется тем, что структура данных модели может изменяться и для решения задачи анализа используется ее текущая версия.

Метод обучения выбирается для конкретной задачи интеллектуального анализа данных, исходя из имеющегося опыта решения подобных задач, а также с учетом возможностей и ограничений, имеющихся в данных.

Цели интеллектуального анализа данных

Не исчерпывающий перечень

- **Поддержка принятия решений** - самая частая причина применения анализа данных.
 - Решение принимает человек, но ему нужно «мнение» машины;
 - Для сбора данных нужна своя архитектура.
- **Автоматическое управление** - актуально ввиду развития робототехники и углубления автоматизации процессов.
 - Решение принимает машина, могут быть «запрещенные решения»;
 - Возникает вопрос об ответственности за принятые решения.
- **Создание новых объектов** - решается задача прогнозирования на пространственных, временных или комбинированных метках. Эта задача еще называется задачей генерации.
 - Наиболее социально опасная цель, так как генерация не человеком может преследовать цели, отличные от человеческих.
 - Используется dark side хакерами для создания сценариев атак, властными структурами для замены руководителя в общении, обычными людьми для облегчения решения своих творческих задач.
- **Модерация** - решает задачу поддержания соответствия цели коммуникации.
- **Развлечения** - решается задача выбора развлекательного контента за пользователя.

Методы интеллектуального анализа данных составляют современную основу т.н. искусственного интеллекта.

Средства интеллектуального анализа для разработчиков и для обычных пользователей

- **Языки и среды программирования** - средство анализа для разработчиков. Имеется множество библиотек для ИАД для языков
 - Python;
 - Go, Rust;
 - C++, Java, Scala, ...
- **LowCode платформы** - позволяют пользователю осуществлять анализ, не вникая в технические детали, но пользователь должен понимать как работают методы.
 - Loginom;
 - PolyAnalyst.
- **BI системы** - решают бизнес-задачи, которые могут быть решены анализом данных.
 - Как правило, ориентированы на конкретный бизнес.
 - Выдают результаты, пригодные для презентаций, но не помогающие непосредственно в принятии решений.

В каждом случае требуется понимание используемых методов, алгоритмов и структур данных. Без этого результаты анализа получатся, но не те или их использование будет некорректным.

Средства интеллектуального анализа и правила практической работы в настоящем курсе

- **Язык программирования Python** - наиболее распространенное средство анализа для разработчиков. Все существующие библиотеки имеют версии для этого языка. Среды:
 - JupyterLab;
 - JupyterNotebook;
 - Google Colab;
 - Без среды (код в текстовом файле с расширением .py)
- **Выполнение и сдача практических работ:**
 - В средах Jupyter... Сдается блокнот .ipynb с кодом и подробными комментариями в коде. Также сдаются используемые в коде файлы.
 - В среде Google Colab. Сдается ссылка на блокнот .ipynb с кодом и подробными комментариями в коде. Отдельно сдаются используемые в коде файлы.
 - Без среды. Сдается .py файл с используемыми файлами в одном архиве. Исполняемый файл должен запускаться из командной строки в каталоге, где расположен скрипт и служебные файлы. Исполнение программы должно показывать все этапы работы и выдавать необходимые комментарии.
- **Оценка работ** - производится по 100-балльной системе:
 - Сдача работы без нарушения правил оформления - до 30 баллов.
 - Ответ на один вопрос при устной сдаче - до 20 баллов (три вопроса).
 - За творческий подход и качество решения могут быть начислены еще до 10 баллов.

После окончания срока сдачи работы в Moodle, она принимается еще 2 недели. После этого ее нельзя сдать никогда.

Внимание! При выполнении работ студенты пишут свои функции и классы для решения задач! Использование библиотек со средствами, полностью решающими задачи интеллектуального анализа данных (scikit-learn, dsmItf и других) в этом курсе запрещено - приводит к получению 0 за сдачу работы и сильно осложняет ответы на вопросы.

Состав практических работ

с календарным графиком их сдачи

- 1. Упражнения на программирование.** Отчет - 19 сентября. Сдача - 3 октября.
- 2. Проверка статистических гипотез.** Отчет - 26 сентября. Сдача - 10 октября.
- 3. Парсинг веб-ресурсов.** Отчет - 3 октября. Сдача - 17 октября.
- 4. Препроцессинг данных.** Отчет - 17 октября. Сдача - 31 октября.
- 5. Градиентный спуск.** Отчет - 24 октября. Сдача - 7 ноября.
- 6. Линейная регрессия.** Отчет - 31 октября. Сдача - 14 ноября.
- 7. Логистическая регрессия и метод ближайших соседей.** Отчет - 7 ноября. Сдача - 21 ноября.
- 8. Деревья принятия решений.** Отчет - 14 ноября. Сдача - 28 ноября.
- 9. Кластеризация методом k средних.** Отчет - 21 ноября. Сдача - 5 декабря.
- 10. Восходящая кластеризация.** Отчет - Отчет - 28 ноября. Сдача - 12 декабря.
- 11. Нейронные сети.** Отчет - 12 декабря. Сдача - 26 декабря.