Маркетинговая аналитика

Раздел 5. Big Data

5.2. Процесс анализа данных. Data Science. Data Mining. Machine Learning

Мы уже узнали, что объемы данных растут с невероятной скорость, и для работы с ними придумали такой термин, как Big Data. Однако, это далеко не единственный способ работы с данными, и следует знать, какие ещё существует. В это связи, давайте познакомимся с основными терминами и их отличиями. Дадим определения трем понятиям. Data Science (рус. Наука о данных, сокращенно DS) – раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объемов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных. Data Mining (рус. добыча данных, сокращенно DM) – совокупность методов обнаружения в данных раннее неизвестных, нетривиальных, практически полезных знаний. Machine Learning (рус. машинное обучение, сокращенно ML) – класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения множества сходных задач. Но рассматривать сухие термины не очень интересно, поэтому давайте попарно сравним каждое понятие.

|  |  |  |
| --- | --- | --- |
|  | Data Science | Data Mining |
| Что это? | Сфера | Техника |
| Фокус | Наука | Бизнес |
| Цель | Анализ общества, построение предиктивных моделей, открытие неизвестных фактов | Нахождение раннее неизвестных закономерностей |
| С чем работает | Со всеми видами данных | Обычно – структурированные данные |

DM является часть DS, в то время, как DS помимо DM включает в себя ещё кучу разных разделов.

|  |  |  |
| --- | --- | --- |
|  | Data Science | Machine Learning |
| Сфера применения | Нахождение скрытых зависимостей в данных без отрыва от контекста реального мира, включая понимание требование, добычу данных и т. д. | Точная классификация или предсказание для новых данных, посредством обучения на исторических данных, используя математические модели |
| Входные данные | Большинство данных – сгенерированы людьми и являются читабельными, например, таблицы или картинки | Должны быть преобразованы в специальный вид для использования алгоритмов. Например, приведение к стандартному виду. |

ML также является частью DS.

|  |  |  |
| --- | --- | --- |
|  | Data Mining | Machine Learning |
| Цель | Нахождение скрытых зависимостей в данных | Обучение компьютера нахождению этих зависимостей с помощью исторических данных |
| Настройка | Ручная, требует непосредственно участия человека | Не требует человеческих усилий, как только настроена. Автоматическая. |
| Сфера применения | Ограничена | Практически неограничена |

DM является частью ML.

Важные вещи, которые стоит осознать следующие: Во-первых, вложенность понятий. По сути DS включает в себя всё, DM является узко специализированным понятием, а ML более широким, поэтому часто включает в себя DM. Во-вторых, DS занимается изучением вообще всего, что связано с данными и требует довольно ручного исследования зависимостей в данных, ML менее обширный и занимается обработкой данных и построением моделей, а DM это вообще только про данные, их сбор и обработку, с построением моделей связи не имеет.

На самом деле, границы понятий довольно размыты, и сложно понять разницу между ними, особенно, когда, по сути, они решают одинаковые задачи. Но всё же, основные различия в том, какие технологии и средства они применяют для их решения.

Говоря о задачах, два основных типа – это классификация и регрессия. Популярные примеры из сферы маркетинга: прогнозирование продаж в будущем (задача регрессии), допустим нам известны продажи какого-либо конкретного товара каждый день на протяжение 2019 года и мы хотели бы предсказать, как будет продаваться этот товар (каждый день в штуках) в 2020 году, предсказание вот этого числа (количество штук проданных 5 марта 2020 года и т. д., и есть задача регрессия). Популярным примером задачи классификации может являться следующее предсказание: отпишется человек (класс 1) от нашего сервиса (кино/музыка/что угодно) или нет (класс 2). Понятно, что для решения такой задачи используется данные о поведение человека, и для поиска закономерностей в этих данных используется DM. Если мы хотим ручное предсказание, то можем этим и ограничиться, но если мы хотим построить модель, которая сама найдет закономерности, то нам следует использовать методы ML.

Конечно, множество различных методов и их огромное разнообразие это хорошо. Но золотое правило DS говорит о том, что без «хороших» данных, то есть данных, которые являются значимыми, существуют в достаточном объеме и др. невозможно построить хорошую модели или сделать точное предсказание. Поэтому так важен DM и специалист в первую очередь должен позаботиться о релевантности данных (качественно собрать их, разметить, проверить на адекватность и т. д.). Также от данных зависит то, какие способы и методы их обработки выбрать, будь то DM, ML или Big Data.

## ***Литература***

1. https://www.educba.com/data-science-vs-machine-learning/
2. https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining
3. Thomas H. Davenport, D. J. Patil. [Data Scientist: The Sexiest Job of the 21st Century](http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/pr)