Лекция 2. Сбор и подготовка данных

Стремительное проникновение информационных технологий в различные сферы жизнедеятельности человека определило актуальность и тенденцию развития методов искусственного интеллекта для анализа данных и принятия решений. Особую актуальность для решения профессиональных задач приобретают методы машинного обучения, позволяющие по имеющимся данным, отражающим характеристики объектов исследования, обобщать, прогнозировать развитие ситуаций и принимать решения на основе обработки статистических данных. Математические методы и их реализация в виде алгоритмов машинного обучения позволяет лицу, принимающему решения для определения оптимальной стратегии управления, опираться на результаты решения задач с использованием различных алгоритмов машинного обучения, их сравнения посредством оценки точности. Использование математических методов и базирующихся на них алгоритмах машинного обучения для решения прикладных задач принятия решений является актуальным развитием информационных технологий.

Цель машинного обучения – анализ данных.

Данные – зарегистрированная информация; представление фактов, понятий или инструкций в форме, приемлемой для общения, интерпретации, или обработки человеком или с помощью автоматических средств (ISO/IEC/IEEE 24765-2010).

В информатике и информационных технологиях:

Данные — формы представления информации, с которыми имеют дело информационные системы и их пользователи (ISO/IEC 10746-2:1996).

Данные – поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи или обработки (ISO/IEC 2382:2015).

Данные в машинном обучении – это представление информации об исследуемой задаче в виде множеств исследуемых объектов и множеств их

характеристик, на основе которых строятся модели, разрабатываются подходы, методы и алгоритмы анализа для принятия решений.

Для Аналитика (Data Scientist, Data Analyst, Data Mining Engineer) очень важно обладать правильными данными, что гарантирует эффективность обработки и построении прогнозов.

Качество данных – важный аспект машинного обучения. На рисунке 1 представлены основные требования к данным.

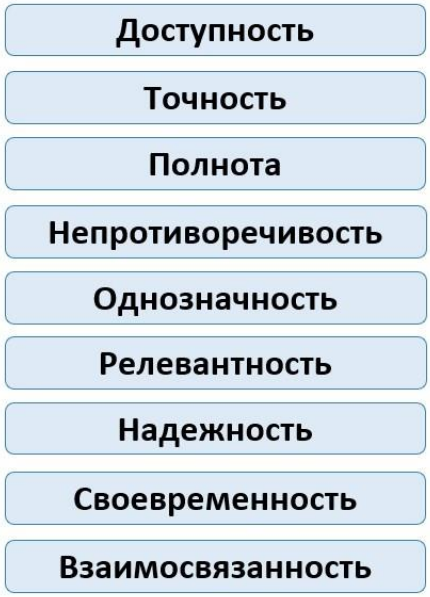

Рисунок 1 – Требования к данным

Доступность

У аналитика должен быть доступ к данным. Это предполагает не только разрешение на их получение, но также наличие соответствующих инструментов, обеспечивающих возможность их использовать, обрабатывать и анализировать.

Точность

Данные должны отражать истинные значения или положение дел. Например, показания неправильно настроенного датчика, ошибка в дате рождения или устаревший адрес — это все примеры неточных данных.

Полнота

Под неполными данными может подразумеваться как отсутствие части информации (например, в сведениях о клиенте не указано его имя), так и полное отсутствие единицы информации (например, в результате ошибки при сохранении в базу данных потерялась вся информация о клиенте).

Непротиворечивость

Данные должны быть согласованными. Например, адрес конкретного клиента в одной базе данных должен совпадать с адресом этого же клиента в другой базе. При наличии разногласий один из источников следует считать основным или вообще не использовать сомнительные данные до устранения причины разногласий.

Однозначность

Каждое поле, содержащее индивидуальные данные, имеет определенное, недвусмысленное значение. Четко названные поля в совокупности со словарем базы данных помогают обеспечить качество данных.

Релевантность

Данные зависят от характера анализа. Релевантность – степень соответствия результатов поиска результатам запроса.

Надежность

Данные должны быть одновременно полными (то есть содержать все сведения, которые вы ожидали получить) и точными (то есть отражать достоверную информацию).

Своевременность

Между сбором данных и их доступностью для использования в аналитической работе всегда проходит время. На практике это означает, что аналитики получают данные как раз вовремя, чтобы завершить анализ к необходимому сроку.

Взаимосвязанность

Должна быть возможность точно связать одни данные с другими. Например, заказ клиента должен быть связан с информацией о нем самом, с

товаром или товарами из заказа, с платежной информацией и информацией об адресе доставки. Этот набор данных обеспечивает полную картину заказа клиента. Взаимосвязь обеспечивается набором идентификационных кодов или ключей, связывающих воедино информацию из разных частей базы данных. Ошибка всего в одном из этих аспектов может привести к тому, что данные окажутся частично или полностью непригодными к использованию или, хуже того, будут казаться достоверными, но приведут к неправильным выводам.

Данные содержат ошибки и пропуски. Ошибки могут появиться в данных по многим причинам и на любом этапе сбора информации.

Во многих случаях аналитики лишены возможности контролировать сбор и первичную обработку данных. Обычно они бывают одним из последних звеньев в длинной цепочке по генерации данных, их фиксированию, передаче, обработке и объединению. Тем не менее важно понимать, какие проблемы с качеством данных могут возникнуть и как их потенциально можно разрешить.

Как оценить качество данных и подготовить данные в цифровом виде для применения алгоритмов искуственного интеллекта для их обработки один из важных вопросов машинного обучения. Не только интуитивно формальные требования рассмотренные выше предьявляются к данным, но и требования которые накладывают особенности алгоритмов обработки данных. Аналитик сталкивается с вопросом качества данных, способами их оценки и подготовки для применения технологий искуственного интеллекта при решении любой прикладной задачи.

Рассмотрим постановку задач машинного обучения и определим требования к представлению данных. Пусть задано множество объектов X, множество допустимых ответов Y, и существует целевая функция (target function) $y^*: X \to Y, \text{значения которой } y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $\{x_1, \ldots, x_l\} \subset X$. Пары «объект-ответ» (x_i, y_i) называются прецедентами. Совокупность пар $X^l = (x_i, y_i)_{i=1}^l$ называется обучающей выборкой (training sample).

Задача обучения по прецедентам заключается в том, чтобы по выборке X^l восстановить зависимость y^*, то есть построить решающую функцию (decision function) $f: X \to Y$, которая приближала бы целевую функцию $y^*(x)$, причем не только на объектах обучающей выборки, но и на всем множестве X.

Решающая функция f должна допускать эффективную компьютерную реализацию, следовательно, ее также можно назвать алгоритмом.

Рассмотрим некоторые важные вопросы, возникающие при работе с данными и подготовки данных для применения алгоритмов машинного обучения. Остановимся на основных этапах задач машинного обучения.

Этапы решения задач машинного обучения:

1. Постановка задачи.
2. Сбор и подготовка данных.
3. Предобработка данных и выделение ключевых признаков.
4. Выбор алгоритмов машинного обучения.
5. Обучение модели (моделей).
6. Оценка качества.
7. Эксплуатация модели при достижении требуемого качества, либо возврат к одному из предыдущих шагов (перенастройка модели, добыча новых данных и т. п.).

Сбор данных является одним из начальных и значимых этапов. При сборе данных возможны пропуски значений, выбросы данных за допустимые интервалы. Некоторые алгоритмы машинного обучения не могут обрабатывать пропуски и будет выдаваться ошибка в данных. Алгоритмы чувствительны к выбросам, которые сильно влияют на результат обучения. Наличие в данных категориальных признаков, также, ставит вопрос о выборе алгоритмов или переводе данных в числовые значения.

При подготовке данных можно применять следующие операции:

– структурирование – приведение данных к табличному (матричному) виду;

– заполнение пропусков;

– отбор – исключение записей с отсутствующими или некорректными значениями, если нет возможности заполнения и устранения противоречивости;

– нормализация – приведение числовых значений к определенному диапазону, например к диапазону 0...1;

– кодирование – это представление категориальных данных в числовой форме. Например, при бинарной классификации один из классов можно представить числом «0», а другой класс – числом «1». При множественной классификации система кодирования несколько усложняется: создается несколько числовых полей по количеству классов в выборке данных, каждый класс кодируется проставлением числа «1» в соответствующем поле.

Многие алгоритмы машинного обучения работают только с численными данными – целыми и вещественными числами. Рассмотрим таблицу 1 с данными.

Таблица 1 – Начальные данные

№	ФИО	Возраст	Пол	Семейное положение	Стаж работы	Доход
1	Антонова Антонина	34	ж	замужем	13	55000
2	Борисов Борис	26	м	не женат	3	45000
3	Владимиров Владимир	45	м	женат	28	150000
4	Григорьев Григорий	22	м	не женат	0	15000
5	Дмитриев Дмитрий	34	м	женат	12	40000
6	Ильин Илья	32	м	не женат	10	50000
7	Косанов Константин	54	м	женат	35	60000
8	Маринина Марина	29	ж	не замужем	5	30000
9	Миронов Мирон	20	м	не женат	0	20000
10	Янова Яна	40	ж	замужем	21	100000

Многие алгоритмы машинного обучения работают только с численными данными – целыми и вещественными числами.

Категориальные признаки.

В представленной таблице 1 присутствую категориальные признаки. Как перевести категориальные признаки в числовые? В нашем случае категориальный признак «пол» можно перевести к значениям «ж» - 0, «м» - 1. Такую замену можно применить и к семейному положению «в браке» - 1, «нет» - 0. В таблице 2 представлен результат.

Таблица 2 – Категориальные признаки представлены числовыми значениями

№	ФИО	Возраст	Пол, Ж-0 М-1	Семейное положение, в браке -1, нет -0	Стаж работы	Доход
1	Антонова Антонина	34	0	1	13	55000
2	Борисов Борис	26	1	0	3	45000
3	Владимиров Владимир	45	1	1	28	150000
4	Григорьев Григорий	22	1	0	0	15000
5	Дмитриев Дмитрий	34	1	1	12	40000
6	Ильин Илья	32	1	0	10	50000
7	Косанов Константин	54	1	1	35	60000
8	Маринина Марина	29	0	0	5	30000
9	Миронов Мирон	20	1	0	0	20000
10	Янова Яна	40	0	1	21	100000

Рассмотрим подходы к заполнению пропусков в данных. Например, в таблице 3 представлен пропуск значения возраст. Пропуски в данных могут встречаться при ошибках заполнения, например, пропустили дату, так и при переводе данных в цифровой формат (распознование рукописных анкет).

Таблица 3 – Пропуск значения

№	ФИО	Возраст
1	Антонова Антонина	34
2	Борисов Борис	26
3	Владимиров Владимир	45
4	Григорьев Григорий	22
5	Дмитриев Дмитрий	34
6	Ильин Илья	32
7	Косанов Константин	54
8	Маринина Марина	
9	Миронов Мирон	20
10	Янова Яна	40

Для заполнения пропуска можно использовать среднее значение признака или медиану. Медиа́на набора чисел — число, которое находится в середине этого набора, если его упорядочить по возрастанию, то есть такое число, что половина из элементов набора не меньше него, а другая половина не больше. Использование медианы позволяет понизить влияние выброса в данных. Способ расчета медианы сравнительно прост: необходимо упорядочить выборку статистических данных по возрастанию и заполнить пропуск медианным значением, т.е. таким, которое либо находится ровно посередине выборки, либо между ближайшими к этой середине объектами выборки. В примере из таблицы 3, медианное значение составит 34.

При решении задач и построении модели необходимо понимать, что некоторые признаки могут снизить точность оценки из различии в диапазоне изменения, например, возраст макс =100, доход макс=500000. Для повышения точности модели рекомендуется переводить значения в заданный диапазон, например, от [0, 1]. Для нормировки данных воспользуемся формулой:

$$x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)},$$

где x_i – значение признака исследуемого объекта;

$\min(x_i)$ – минимальное значение признака в массиве;

$\max(x_i)$ – максимальное значение признака в массиве.

В таблице 4 приведен пример пересчета данных.

Таблица 4 – Перевод значений в диапазон [0, 1].

№	ФИО	Возраст	
1	Антонова Антонина	34	(34-20)/(54-20)=0,41
2	Борисов Борис	26	(26-20)/(54-20)=0,18
3	Владимиров Владимир	45	0,74
4	Григорьев Григорий	22	0,06
5	Дмитриев Дмитрий	34	0,41
6	Ильин Илья	32	0,35
7	Косанов Константин	54	1,00
8	Маринина Марина	29	0,26
9	Миронов Мирон	20	0,00
10	Янова Яна	40	0,59

Аналогичные действия выполним со столбцами «Стаж работы», «Доход». В таблице 5 представлен результат обработки данных и подготовки к применению алгоритмов машинного обучения. Категориальные признаки переведены в числовые и произведена нормировка данных.

Таблица 5 – Данные после обработки

№	ФИО	Возраст	Пол	Семейное положение, в браке -1, нет -0	Стаж работы	Доход
1	Антонова Антонина	0,41	0	1	0,37	0,30
2	Борисов Борис	0,18	1	0	0,09	0,22
3	Владимиров Владимир	0,74	1	1	0,80	1,00
4	Григорьев Григорий	0,06	1	0	0,00	0,00
5	Дмитриев Дмитрий	0,41	1	1	0,34	0,19
6	Ильин Илья	0,35	1	0	0,29	0,26
7	Косанов Константин	1,00	1	1	1,00	0,33
8	Маринина Марина	0,26	0	0	0,14	0,11
9	Миронов Мирон	0,00	1	0	0,00	0,04

10	Янова Яна	0,59	0	1	0,60	0,63

Этап подготовки данных является важным для последующего применения алгоритмов машинного обучения, позволяет увеличить точность обработки данных.

Контрольные вопросы по теме:

1. Из каких этапов состоит процесс подготовки данных к анализу?
2. Сформулируйте определение понятия данных в машинном обучении?
3. Назовите ключевые требования к данным.
4. Приведите основные этапы решения задачи машинного обучения.
5. Почему необходимо производить обработку данных?
6. Какие операции можно применять при подготовке данных?
7. Охарактеризуйте ключевые операции, которые применяют при подготовке данных.
8. Что показывают категориальные признаки?
9. Можно ли перевести категориальные признаки к числовым значениям?
10. Приведите известные Вам методы нормировки данных. Для чего они нужны?