Data Science — что это такое и зачем она нужна?

Что такое data science?

История вопроса

Человечество всегда мечтало приподнять занавес с грядущего и желало знать что будет. И если для сбора данных использовались точные методы (таблицы, архивы, летописи), то для предсказания все шло в ход — шаманы впадали в транс и общались с потусторонним миром, сообщая новости оттуда; пифии, будучи опять же в трансе, делали малосвязные предсказания, которые потом трактовались жрецами в нужном смысле; астрологи пытались применить околонаучный подход и рассчитывали гороскопы для мероприятий и людей. Многое из этого набора до сих пор живо используется, но этим прогнозы не обоснованы и к ним нет доверия у научного сообщества.

Сбор данных можно смело считать началом статистики. Первая статистическая информация — глиняные таблички шумерского царства (III–II тысячелетие до н. э.). В них содержалась экономическая информация — сделки, количество собранного урожая, налоги и пр.

В Римской республике, а затем и в империи, была развитая финансовая и налоговая система, которая требовала ведения точного учета и сбора данных по сделкам, земельным владениям, товарам, услугам и т. д. Официальная отчетность наносилась на доски: мраморные, бронзовые, медные, свинцовые и побеленные деревянные. Текущие записи велись на деревянных та-

бличках, скрепленных вместе с одного края по две, три и больше — кодексы (лат. *code* — дерево). После завоевания Римом Египта появился папирус. Около 180 г. до н. э. был изобретен пергамент (изготовлялся из телячьей кожи, был дорог, но прочен). На развитие учета влияли техника письма и система счета. Для вычислений использовался абак, заимствованный древними греками у египтян.

Бухгалтерский учет велся в Памятных книгах, или Мемориалах, куда записывались ежедневные факты хозяйственной деятельности. Также велась кассовая книга — первый кодекс и книга системной записи — второй кодекс.

Бюджетный учет велся в государственных масштабах. В отдельных провинциях велась книга Бревариум, в которой отражались как сметные ассигнования, так и их исполнение. Такой регистр получил название Книги имперских счетов, которую можно рассматривать как первый баланс государственного бюджета.

Развивался и налоговый учет, который требовал классификации и оценки имущества для начисления налога.

И хотя учет в Древнем Риме носил контрольный характер, уже тогда, по мнению древнеримского ученого Колумеллы, важнейшей функцией учета становилось умение предвидеть результат хозяйствования.

В Средневековье функции сбора данных остались те же — контрольный учет для сбора налогов и ведения хозяйственной деятельности.

С возникновением теории вероятностей в XVII веке были совершены первые попытки обработки накопленных данных и построения первых моделей для прогнозирования. Например, изучалась частота рождения мальчиков и девочек. Своим появлением теория вероятностей обязана азартным играм. Исследуя вероятность выигрыша, Пьер Ферми и Блез Паскаль открыли первые вероятностные закономерности. Независимо от них, но под влиянием их работ, Христиан Гюйгенс в 1657 г. опубликовал работу, в которой дал основные понятия теории

вероятностей (понятие вероятности как величины шанса; математическое ожидание для дискретных случаев, в виде цены шанса) и теоремы сложения и умножения вероятностей.

В 1794 г. (по другим данным — в 1795 г.) немецкий математик формализовал один из методов современной математической статистики. Данный метод стал основой для построения регрессионных моделей, цель которых — предсказание заданной величины. В XIX веке получил развитие анализ больших данных, который дал новый толчок к развитию статистических моделей.

В XX веке пошло быстрое развитие статистики и математической статистики как науки. В начале XX века была развита параметрическая статистика, созданы методы сравнения групп данных, оценки параметров групп и т. д.

Цель сбора данных кардинально изменилась к XX веку и перешла от контрольного учета к созданию математических предсказательных моделей.

Теперь перейдем ближе к современности и к науке о данных.

1974 г. впервые введен термин data science датским ученым в области информатики и компьютерной науки Питером Науром. Он считал, что наука о данных — дисциплина, изучающая жизненный цикл цифровых данных от появления до преобразования для представления в других областях знаний.

В начале 2000-х гг. data science выделяется как отдельная дисциплина.

Определения

Определение науки о данных вполне точно приведено в Wikipedia.

Наука о данных (*data science*; иногда *даталогия — datalogy*) — раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме*.

* С сайта http:www.ru.wikipedia.org.

В принципе такое определение достаточно полно описывает цели и суть науки о данных. Основной целью науки о данных является вывод новых знаний из имеющегося набора данных и получение новых зависимостей, часто неявных. Кроме того, одним из важных разделов науки о данных является визуализация больших данных.

Суть и цели

Остановимся подробнее на сути науки о данных. Исходной точкой в науке о данных являются собственно данные, и чем их больше — тем лучше. Далее нужно на основе этих данных найти взаимосвязи в них или убедиться, что их нет.

Для чего все это нужно? А целей — множество.

Во-первых, на основе полученных закономерностей можно построить прогноз для заданной величины. Например, на основе данных об урожае пшеницы за последние 10 лет в заданном регионе можно построить прогноз урожайности на следующий год.

Во-вторых, можно провести классификацию объектов на основе данных о них. Например, можно на основе клинических данных классифицировать методику лечения как эффективную или неэффективную.

В-третьих, можно визуализировать данные. Визуализация помогает выбрать стратегию анализа данных, а иногда она сама является целью анализа. Например, визуализация данных по движению городского транспорта в режиме on line ценно само по себе.

В-четвертых, можно провести анализ текстовой информации и, например, понять тональность отзыва о компании.

И наконец, в-пятых, можно найти новые зависимости в данных и на их основе прийти к новым знаниям о предмете анализа.

Data Science — зачем она нужна?

Спасаем Мир

Эпидемия Эбола в 2014 г. унесла более 11 000 жизней, и каждый день приносил новые смерти. Для data scientist задача по анализу данных и построению модели стала вызовом. И в 2014 г. the Leiden Centre of Data Science (LCDS) принял этот вызов. В результате разработана комплексная модель симуляции лихорадки Эбола, включающая диагностику распространения и испытание лекарств. Врачи, используя эту модель, остановили эпидемию Эбола.

Немного о модели

Стандартную модель симуляции распространения эпидемии дополнили картой, составленной на основе SMS-сообщений, звонков и другой активности с мобильных телефонов и добавили в нее все источники масс-медиа. На основе данной карты построили модель по пересечению и вычленению реальных данных. Использование данных с мобильных телефонов позволило установить направление распространения эпидемии и уже на основе этих данных получить оптимальные места для развертывания медицинских центров. Комплексная сеть данных (мобильные, масс-медиа и правительственные данные) и социальная сеть контактов дали возможность спрогнозировать скорость и направление развития эпидемии.

Модели, полученные из анализа данных пациентов, позволяют оценивать эффективность лекарств и проводить быстро множество тестов.

Познаем вселенную

The Center for Computational Astrophysics разрабатывает новый фреймворк (каркас программной системы), который предназначен для анализа астрономических данных. Он используется для построения модели Вселенной и оценки космологических констант. В XXI веке на основе нейронных сетей разработана 3D-модель Вселенной, в которой учтено распространение темной материи и есть возможность предсказания космологических констант.

Контрольные вопросы

1. Что является основной целью науки о данных?
2. Приведите примеры задач, которые можно решать с помощью науки о данных.

Основы обучения с учителем

Основные понятия

Итак, во введении мы рассмотрели области применения машинного обучения и его возможности на примерах. Результаты применения поражают воображение. Теперь настала пора разобраться в деталях и понять, как же это становится возможным.

В машинном обучении выделяют 2 основных подхода — обучение с учителем и обучение без учителя. В этой главе рассмотрим первый подход — с учителем.

Начнем с жизненной ситуации. У девушки — день рождения, и Пете нужно подарить ей цветы. Известно, что она не любит экзотические цветы, но какие нравятся — не известно. Петя подошел к проблеме с точки зрения машинного обучения и собрал данные о том, какие цветы больше всего любят девушки, и выбрал девушек близкого к имениннице возраста и внешности. Оказалось, что в предпочтениях лидируют два самых популярных цветка — роза и гербера. Причем 80 % девушек отдают предпочтение розам, а 20 % — герберам.

Теперь разберемся, что же сделал Петя. Все девушки, о которых Петя собрал информацию о предпочтениях в цветах, являются обучающей выборкой. Параметры, по которым Петя выбирал девушек, а именно возраст и цвет волос, являются признаками или факторами выборки. Информация о каждой отдельной девушке (цвет волос, возраст и любимый цветок) является объектом выборки. Причем цвет волос и возраст являются параметрами объекта, которые обычно обозначаются как x_1 и x_2,

а любимый цветок — ответ на данном объекте выборки (обозначается как y_i, где i — номер девушки в списке). Информация о возрастах и цвете волос всех девушек в выборке является пространством объектов X. Информация о любимых цветках всех девушек из выборки является пространством ответов Y.

Далее Петя должен найти зависимость между цветом волос и возрастом девушек и любимым цветком, то есть Петя должен найти коэффициенты w_1 и w_2 в уравнении $y = w_1 x_1 + w_2 x_2$. Зная их, он подставит возраст и цвет волос именинницы и получит ответ. Уравнение, которое определит Петя, — это алгоритм обучения. Строго говоря, алгоритм обучения $a(X)$ — функция перехода из пространства X в пространство Y.

Однако, возможно, Петя ошибется и ответ получит неверный, поэтому всегда нужно иметь инструмент оценки качества алгоритма. Таким инструментом является функционал ошибки. Функционал ошибки $Q(a, X)$ — ошибка алгоритма a на выборке X.

Обучение с учителем имеет два основных применения: классификация и регрессия. Классификация дает возможность предсказать по значениям признаков, к какому классу будет относиться новый объект. Кстати, пример с Петей — это типичная задача бинарной классификации. Регрессионный анализ дает возможность предсказать значение целевой переменной (ответа) по имеющейся модели, построенной на обучающей выборке.

Рассмотрим подробнее задачу регрессии. Допустим, у нас есть данные по продажам рожков мороженого и данные по температуре воздуха в дни продаж. Построим график, где по оси X отложим температуру воздуха, а по оси Y — количество проданных рожков мороженого, и отложим точки на нем. Далее, выведем зависимость количества проданных рожков мороженого от температуры воздуха и построим ее на этом же графике. Итак, сами данные о продажах нам уже не нужны. Теперь, зная температуру и пользуясь прямой зависимости, можно узнать, сколько рожков мороженого может быть продано.

Рассмотрим математическое описание линейного алгоритма модели. Алгоритм линейной модели $a(x)$ выражается в виде взвешенной суммы

$$a(x) = w_0 + \sum_{j=1}^{d} w_i x_j, \quad (1)$$

где w_0, w_i — веса модели; j — номер объекта в выборке; d — количество объектов в выборке; i — номер параметра объекта.

Перейдем к векторной форме записи выражения (1). Для этого внесем w_0 в сумму, добавив еще один признак, всегда равный 1:

$$a(x) = \sum_{j=1}^{d+1} w_j x_j = \langle W, X \rangle, \quad (2)$$

где W — вектор весовых коэффициентов; X — вектор признаков.

Вернемся к вышеописанному примеру модели, предсказывающей количество проданных рожков мороженого в зависимости от температуры воздуха. Алгоритм данной модели $a(t)$ запишется в виде следующей суммы:

$$a(t) = w_0 + \sum_{j=1}^{d} w_1 t_j, \quad (3)$$

где t — температура воздуха; w_0 и w_1 — веса модели; j — номер дня, для которого известна температура воздуха и количество проданных рожков мороженого; d — количество дней.

Можно выражение (3) представить также в векторной форме:

$$a(t) = \langle W, X \rangle.$$

Рассмотрим функционал ошибки, используемый в задаче регрессии. В качестве функционала ошибки в задаче регрессии используется среднеквадратичная ошибка. По сути своей она отображает квадрат разности между значением, полученным алгоритмом ($a(x)$) и ответом на обучающей выборке (y):

$$Q(a,x)=\frac{1}{d}\sum_{i=1}^{d}(a(x_i)-y_i)^2, \tag{4}$$

где y_i — ответ на i-м объекте обучающей выборки.

В векторной форме среднеквадратичная ошибка представляется в виде выражения

$$Q(a,x)=\frac{1}{d}\sum_{i=1}^{d}(\langle W,X\rangle-y_i)^2. \tag{5}$$

Теперь мы выяснили, как выглядит алгоритм и как оценить его ошибку. Тем не менее, как видно из формул (1)–(2), в алгоритме известны X и Y, а вот веса модели нужно найти. Процесс вычисления значений весов модели называется обучением модели. Естественно, нужно найти такие веса модели, при которых ошибка модели будет минимальной. Поэтому цель обучения модели — найти минимум функционала модели:

$$Q(x)\to\min.$$

Выражение (4) можно представить в полной форме:

$$Q(a,x)=\frac{1}{d}\sum_{i=1}^{d}(a(x)-y_i)^2\to\min. \tag{6}$$

Перепишем выражение (6) в матричной форме:

$$Q(W,X)=\frac{1}{d}\|XW-Y\|^2\to\min_w, \tag{7}$$

где X и Y — матрицы параметров и ответов соответственно:

$$X=\begin{pmatrix} x_{11} & x_{12} & x_{1n} \\ & & \\ x_{m1} & & x_{mn} \end{pmatrix},\quad Y=\begin{pmatrix} y_1 \\ \\ y_m \end{pmatrix}.$$

Матричное выражение (7) имеет аналитическое решение относительно вектора W

$$w=\left(X^TX\right)^{-1}X^Ty. \tag{8}$$

Однако, как видно из выражения (8), в аналитическом решении необходимо определить обратную матрицу, что не всегда возможно и очень долго. Поэтому на практике для расчета коэффициентов модели используется градиентный спуск.

Суть градиентного спуска заключается в пошаговом изменении значения весов и пересчете ошибки модели, причем изменение шага предпринимается в направлении антиградиента функции ошибки. В конечном итоге достигается минимум функции ошибки, что и дает искомый набор значений весов модели. Подробнее о методе градиентного спуска и его модификациях будет описано в следующей главе. А теперь рассмотрим метрики качества регрессионной модели.

Первая метрика, на которой и происходит обучение модели, — это уже рассмотренная выше среднеквадратичная ошибка. Обратите внимание на ее график, который принимает вид параболы (рис. 1). Парабола является гладкой, непрерывной функцией с 1 глобальным минимумом и поэтому может быть использована для дифференцирования.

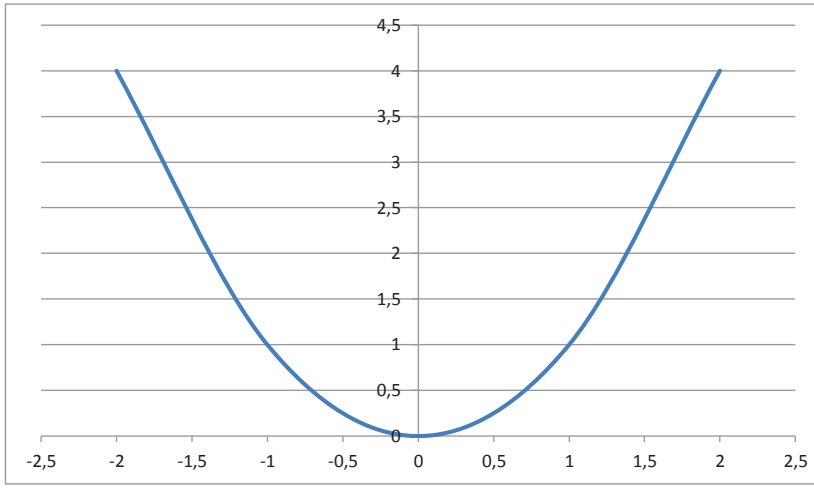

Рис. 1. График среднеквадратичной ошибки

Следующая метрика — средняя абсолютная ошибка ($MAE(a, X)$). Она отображает модуль разности между результатом, полученным алгоритмом модели и ответами на выборке:

$$MAE(a, X) = \frac{1}{d}\sum_{i=1}^{d}\left|a(x_i) - y_i\right|.$$

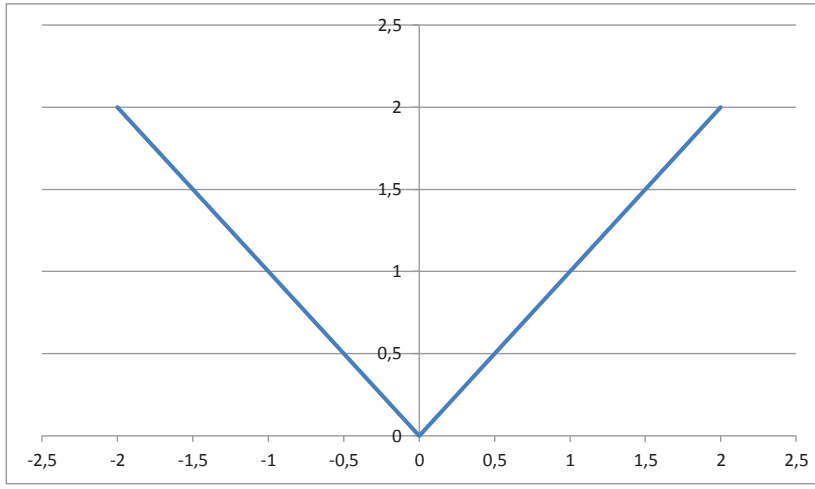

Рис. 2. График абсолютной ошибки

Как видно из рис. 2, абсолютную ошибку уже нельзя дифференцировать.

Следующая метрика качества регрессии — это коэффициент детерминации R^2:

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^{d}\left(a(x_i) - y_i\right)^2}{\sum_{i=1}^{d}\left(y_i - \bar{y}\right)},$$

где $\bar{y}$ — среднее значение ответов на выборке.

Коэффициент детерминации показывает, какую долю дисперсии ответов алгоритма модель может объяснить.

Контрольные вопросы

1. Запишите выражение для алгоритма линейного регрессора.
2. Что рассчитывается при обучении модели?
3. В чем смысл коэффициента детерминации?

Градиентный спуск

Итак, мы переходим к рассмотрению вопроса: как вычислять веса модели? Как уже было сказано ранее, для вычисления весов модели используется функционал ошибки, а именно среднеквадратичная ошибка. Значения весов модели, соответствующие точке минимума среднеквадратичной ошибки, являются искомыми весами модели.

Суть метода заключается в пошаговом приближении к минимуму функции среднеквадратичной ошибки. Сначала назначаются случайным образом значения весов модели, затем вычисляются градиенты для каждого веса модели. Значения весов изменяются в направлении антиградиента — вычисляется среднее значение ошибки на всей выборке. Сравнивается полученное значение со значением с предыдущего шага. Если изменение ошибки становится незначительным, значит, функция зашла в свой минимум и искомые значения весов найдены.

Рассмотрим несколько алгоритмов метода.

Пакетный метод градиентного спуска

Алгоритм метода пакетного градиентного спуска для модели с одним параметром при постоянном шаге

Сперва назначаются начальные значения весов модели w_0 и w_1.

Далее вычисляется градиент w_0:

$$\frac{\partial Q}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right), \tag{9}$$

где n — число объектов в выборке,