Четвертая технологическая революция строится на вездесущем и мобильном Интернете, искусственном интеллекте и машинном обучении.

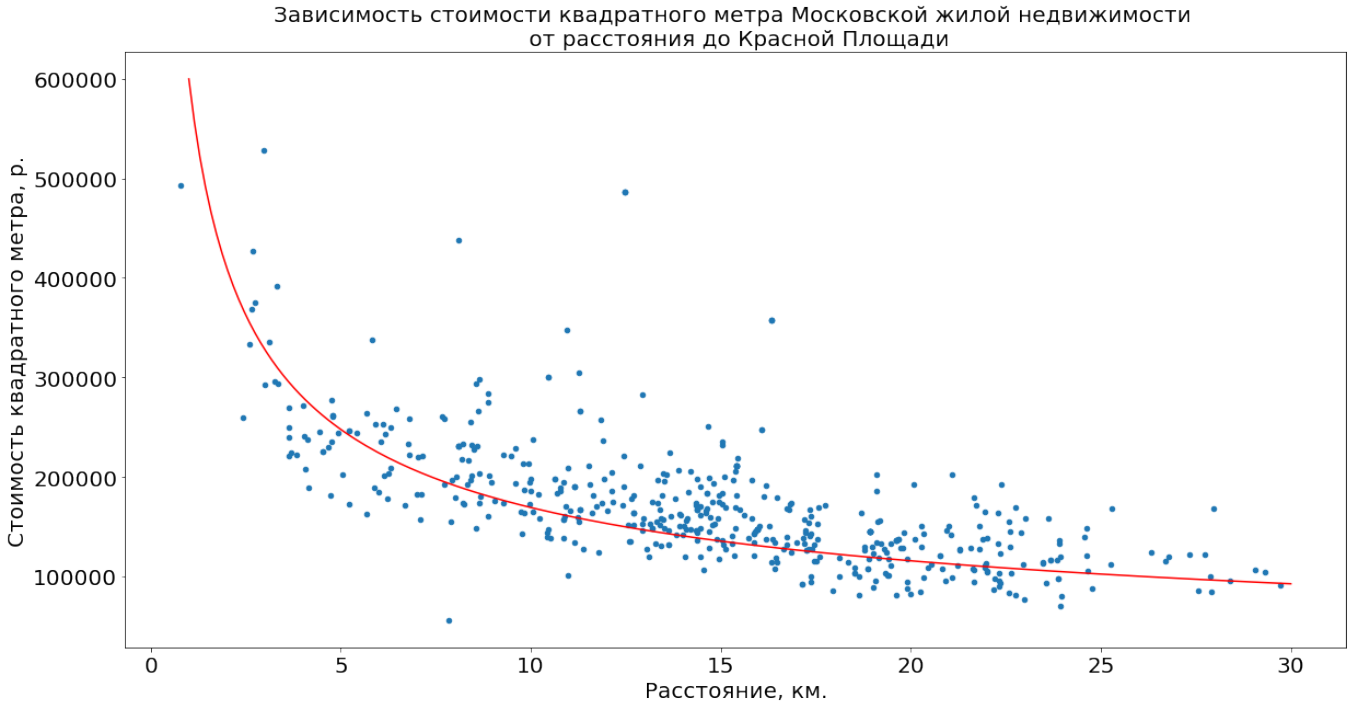
Президент Всемирного экономического форума

Клаус Мартин Шваб

Нации уделяющие значительное внимание научно-исследовательским и опытно-конструкторским работам в области машинного обучения займут лидирующие позиции в автоматизации будущего. Основными сферами применения достижений исследовательских разработок станут:

- Цифровая и распределенная экономика
- Автоматизация и сокращение издержек
- Автономный транспорт и роботизация
- Оптимизация логистики и цепей поставок
- Оптимизация энергетических сетей
- Автоматизация банковских услуг
- Автоматизация юридических услуг
- Мониторинг сельского хозяйства
- Персональная медицина
- Персональные образовательные траектории
- Автономные системы вооружений.

“**Preparing for the future of artificial intelligence**”, Доклад начальника штаба президента США по национальной науке и технике, Вашингтон, 2016.



Введение

Машинное обучение применяется всюду, где собираются данные. По сути машинное обучение, это наука о том, как восстановить неизвестную зависимость по экспериментальным данным, как провести функцию $y^*(x)$ по точкам $\{x_1, \dots, x_l\} \in X$. На рисунке выше приведена экспериментальная зависимость стоимости квадратного метра Московской жилой недвижимости в зависимости от одного фактора – расстояния до Красной Площади. 500 синих точек на графике – 500 московских квартир. Красная кривая – восстановленная зависимость. Математические методы решающие подобные задачи известны не одну сотню лет.

Машинное обучение – развитие этих методов для случая, когда точки принадлежат многомерным пространствам, а ответы принимают не только числовые значения. Задача определения стоимости квадратного метра квартиры в зависимости от целого набора её признаков (года постройки дома, района города, этажа, площади квартиры и отдельных комнат, качества ремонта и т.д.) – задача машинного обучения. Иными словами, методы машинного обучения позволяют по набору известных признаков некоторого объекта предсказать значение

некоторого неизвестного признака этого объекта. Методы машинного обучения позволяют построить алгоритм оценки стоимости квадратного метра, «цифрового риелтора», который «осмотрев» квартиру выносит вердикт о её стоимости.

Различные особенности множества признаков и множества допустимых значений восстанавливаемой функции привели к появлению массы методов машинного обучения.

Основные понятия и определения.

Попытаемся формализовать постановку задачи машинного обучения.

Пусть нам задано множество *объектов* X, множество *допустимых ответов* Y и существует *целевая функция* (target function) $y^*: Y \to X$, значения которой $y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $\{x_1, \dots, x_l\} \in X$. Совокупность пар $X^l = (x_i, y_i)_{i=1}^l$ называется *обучающей выборкой* (training sample).

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

Решающая функция должна допускать эффективную компьютерную реализацию; по этой причине её называют *алгоритмом.*

Объекты и признаки

Признак (feature) f объекта x - это результат измерения некоторой характеристики объекта. Формально признаком называется отображение $f: X \to D_f$, где $D_f -$ множество допустимых значений признака. В зависимости от природы множества D_f признаки делятся на несколько типов:

Если $D_f = \{0, 1\}$, то $f -$ *бинарный* признак, например наличие балкона у квартиры;

Если $D_f -$ конечное множество, то $f -$ категориальный признак, например район города (Центральный, Сокольники, Тушинский …);

Если $D_f = \mathbb{R}$, то $f -$ вещественный признак, например жилая площадь квартиры, количество комнат, год постройки дома

Пусть имеется набор признаков $f_1, \dots, f_n$, тогда вектор составленный из значений признаков ($f_1(x), \dots, f_n(x)$) называют признаковым описанием объекта $x \in X$. Совокупность признаковых описаний всех объектов выборки X^l, записанную в виде таблицы размера $\ell \times$ n, названются *матрицей объектов-признаков:*

$$F = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_l) & \cdots & f_n(x_l) \end{pmatrix}$$

Например, если предположить, что $f_1 -$ наличие балкона, $f_2 -$ площадь квартиры, $f_3 -$ количество комнат, $f_4 -$ расстояние до Красной Площади, то некоторые три квартиры могли бы быть описаны следующей матрицей объектов-признаков.

$$\begin{pmatrix} 1 & 50{,}2 & 2 & 15{,}8 \\ 0 & 36{,}4 & 1 & 12{,}3 \\ 1 & 78{,}1 & 3 & 22{,}1 \end{pmatrix}$$

Ответы и типы задач.

В зависимости от природы множества допустимых ответов Y задачи машинного обучения делятся на следующие типы. Если $Y = \{1, \dots, M\}$, то это задача классификации (classification) на M непересекающихся классов. В этом случае всё множество объектов X разбивается на классы, и алгоритм a(x) должен давать ответ на вопрос «какому классу принадлежит x?». В некоторых приложениях классы называют образами и говорят о задаче распознавания образов (pattern recognition). Если $Y = \{0,1\}^M$, то это задача классификации на M пересекающихся классов (к примеру классификация новостей по темам: новость «Путин провел двустороннюю встречу с представителями ЦК КНДР» может быть отнесена как к теме «Президент России», так и к теме «Китай»). В простейшем случае эта задача сводится к решению M независимых задач классификации с двумя непересекающимися классами. Если $Y = \mathbb{R}$, то это задача восстановления регрессии (regression estimation). Задачи прогнозирования (forecasting), когда x ∈ X — описание прошлого поведения объекта x, y ∈ Y — описание некоторых характеристик его будущего поведения, являются частными случаями классификации или восстановления регрессии.

Алгоритмы и их обучение.

Одним из простейших алгоритмов машинного обучения являются *линейные алгоритмы* с набором параметров $\theta = (\theta_1, \dots, \theta_n)\ \epsilon\ \Theta = \mathbb{R}^n$:

$$a(x, \theta) = \sum_{j=1}^{n} \theta_j f_j(x)$$

Процесс подбора оптимальных параметров θ по обучающий выборке X^l, позволяющих решить поставленную задачу с наилучшим качеством, называют *обучением* (training, learning) алгоритма $a\ \epsilon\ A$.

В случае задачи предсказания стоимости квадратного метра жилья на основе его признакового описания линейный алгоритм будет выглядеть так:

$$\theta_1 f_1(x) + \theta_2 f_2(x) + \theta_3 f_3(x) + \theta_4 f_4(x) = \text{стоимость\_квадратного\_метра}$$

Итак, в задачах машинного обучения чётко различаются два этапа:

- этапе *обучения* в процессе которого по выборке X^l строит алгоритм a.
- этапе *применения (*prediction*)* для новых объектов x алгоритм a выдает ответы $y = a(x)$.

Этап обучения как правило сводится к поиску параметров модели, доставляющих оптимальное значение заданному функционалу качества.

Функционал качества

Функция потерь (loss function) – это неотрицательная функция $\mathcal{L}(a, x)$, характеризующая величину ошибки алгоритма a на объекте x. Если $\mathcal{L}(a, x) = 0$, то ответ $\mathrm{a}(x)$ называется корректным.

Функционал качества алгоритма a на выборке X^l:

$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} \mathcal{L}(a, x_i)$$

Функционал Q называют также функционалом *средних потерь*. Пример функционала средних потерь с квадратичной функцией ошибки, который может применяться в задаче оценки стоимости квадратного метра:

$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} (y_i - a(x_i, \theta))^2$$

Классический *метод обучения*, называемый *минимизацией средних потерь*, заключается в том, чтобы найти набор параметров θ алгоритм a, доставляющий минимальное значение функционалу Q на заданной обучающей выборке X^l:

$$\mu(X^l) = \arg\underbrace{min}_{\theta}\ Q(a, X^l)$$

Вопросы по прочитанному.

1. Что такое объект?

- То, для чего нужно делать предсказания.
- То, с помощью чего измеряется качество предсказаний.
- То, что необходимо предсказывать.

2. Что такое признаки?

- То, с помощью чего измеряется качество предсказаний.
- То, с помощью чего описываются ответы.
- То, с помощью чего описываются объекты.

3. Что такое алгоритм?

- Функция, которая принимает на вход объекты и выдаёт подсчитанные для них признаки.
- Функция, которая принимает на вход предсказания на обучающей выборке и выдаёт оценку качества этих предсказаний.
- Функция, которая принимает на вход объект и выдаёт предсказанный ответ.

4. Выберите верные утверждения про признаки.

- [x] Набор значений признаков на объекте представляет собой вектор определённой размерности.

Правильно
Полный набор признаков для объекта удобно задавать с помощью вектора, размерность которого совпадает с числом признаков.

- [ ] Признаки задаются только вещественными числами.

правильно, этот вариант не должен быть выбран

- [x] Предсказания для объектов делаются на основе значений признаков.

Правильно
Алгоритмы принимают на вход именно признаковые описания объектов.

- [x] Признаки могут иметь только значения 0 или 1.

Правильно, этот вариант не должен быть выбран

5. Выберите вещественные признаки из списка

- [x] Количество детей

Правильно

- [ ] Наличие у клиента банка военного билета

правильно, этот вариант не должен быть выбран

- [ ] Город, в котором прописан клиент

правильно, этот вариант не должен быть выбран

- [x] Температура воздуха

Правильно

- [x] Год рождения

Правильно

6. Выберите признаки, которые могут рассматриваться только как категориальные (и не могут рассматриваться как бинарные или вещественные).

- [x] Тарифный план клиента мобильного оператора

Правильно

Тарифных планов много, и вряд ли все из них можно сравнивать.

- [ ] Наличие у клиента банка военного билета

правильно, этот вариант не должен быть выбран

- [x] Цвет автомобиля

Правильно

Цветов бывает больше двух.

- [x] Тип дома — кирпичный, блочный, панельный и т.д.

Правильно

Тут конечное число вариантов и уже приведено больше двух.

- [ ] Скорость интернет-соединения

правильно, этот вариант не должен быть выбран

- [ ] Год рождения

правильно, этот вариант не должен быть выбран

7. Выберите из списка задачи классификации.

- [ ] Определение возраста человека по его активности в сети.

правильно, этот вариант не должен быть выбран

- [ ] Поиск групп схожих пользователей мобильного оператора.

правильно, этот вариант не должен быть выбран

- [x] Предсказание тарифного плана, который клиент захочет себе подключить.

Правильно

Возможных тарифных планов конечное число, поэтому задача относится к задачам многоклассовой классификации.

- [ ] Определение стоимости одежды по фотографии.

правильно, этот вариант не должен быть выбран

8. Выберите верные утверждения про функционал качества

- ( ) Функционал качества определяет уровень шума в признаках.
- ( ) Функционал качества позволяет определить, насколько данный алгоритм подходит для решения задачи на конкретной выборке.

Вопросы на общую математическую подготовку

1. Найдите производную $\frac{\partial a(x,y)}{\partial x}$ при $a(x,y) = \sin(xy)\, e^x$:

- $x \cos(xy)\, e^x$
- $x \cos(xy)\, e^x + \sin(xy)\, e^x$
- $x\, cos(xy) + e^x$
- $\mathrm{x} \cos(y) + \sin(xy)\, e^x$

2. Сколько решений у следующей системы линейных уравнений?

$$2x=3$$

$$2x=4$$

- Ни одного

Правильно

Нет такого x, чтобы он одновременно был равен 1.5 и 2.

- Одно
- Конечное число, больше единицы
- Бесконечно много

3. Что такое ранг матрицы?

☑ Максимальное число линейно независимых строк.

Правильно

Это утверждение верное.

☑ Максимальное число линейно независимых столбцов.

Правильно

Это утверждение верное.

☐ Максимальное число различных элементов.

правильно, этот вариант не должен быть выбран

☐ Число решений системы линейных уравнений с данной матрицей коэффициентов и нулевой правой частью.

правильно, этот вариант не должен быть выбран

4. Предположим, что в некоторой популяции до 60 лет доживает 50%, а до 80 лет — 20%. Какова вероятность (от 0 до 1), что случайно выбранный шестидесятилетний представитель популяции доживёт до восьмидесяти? Запишите ответ с точностью до одного знака после десятичной точки.

0.4

Правильный ответ По формуле условной вероятности $0.2/0.5 = 0.4$.

5. В супермаркете 60% яблок из Турции и 40% яблок из Индии. 10% турецких и 15% индийских яблок — червивые. Какова вероятность, что яблоко, купленное в этом магазине, окажется червивым? Запишите ответ от 0 до 1 с двумя знаками после десятичной точки.

0.12

Правильный ответ

Пусть событие A происходит, если яблоко из Турции, а событие B — если яблоко червивое. Воспользуемся формулой полной вероятности:

$P(B)=P(B|A)P(A)+P(B|\sim A)P(\sim A)=0.1\cdot 0.6+0.15\cdot 0.4=0.12$

6. Журнал "COSMOPOLITAN" читает 6.4% целевой аудитории вашего продукта, а журнал "Караван историй" — 3.7%. Если вы разместите рекламу в обоих журналах, что можно сказать о доле p целевой аудитории, которая увидит рекламу хотя бы один раз?

- $p\leq 0.037$
- $0.037\leq p\leq 0.064$
- $0.037\leq p\leq 0.101$
- $0.064\leq p\leq 0.101$