Что такое машинное обучение (МО)?

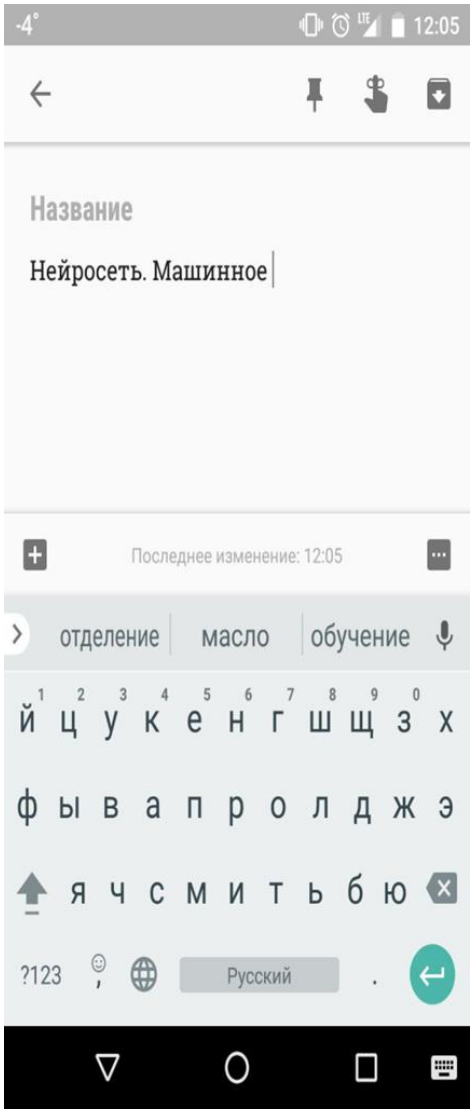
Лекция 1

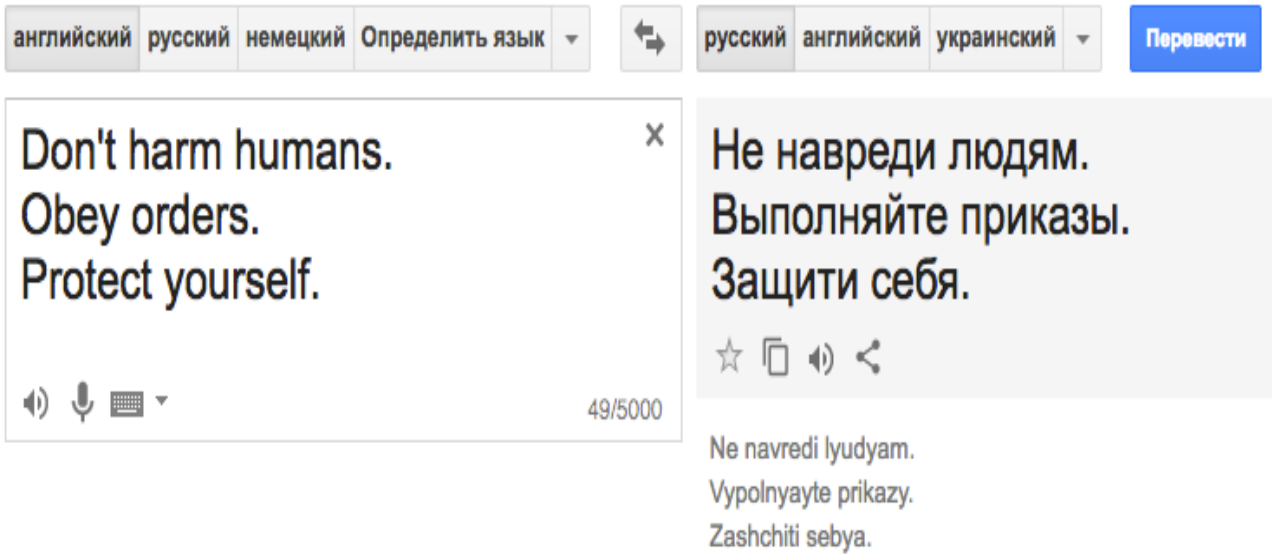
Меня зовут: Шевляков Артём

Где можно встретить модели машинного обучения (искусственного интеллекта, распознавания образов)?

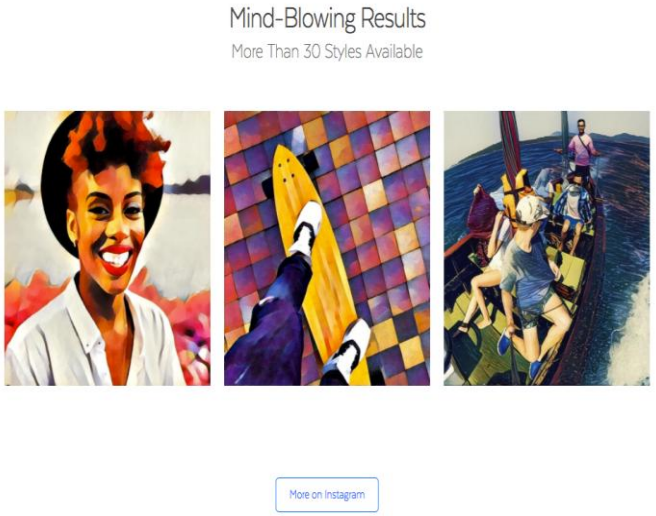
Рекомендательные системы

Работа с текстом





Обработка изображений



Другие задачи машинного обучения

- Предсказание стоимости жилья
- Кредитный скоринг
- Прогнозирование спроса на товары
- Медицинская диагностика
- Ранжирование поисковой выдачи
- Поиски аномалий в данных

Короче: Машинное обучение – это поиск скрытых закономерностей в данных.

Методы машинного обучения используют результаты мат. дисциплин:

- математическая статистика
- методы оптимизации
- идеи из геометрии и линейной алгебры

Основные задачи машинного обучения

1. Восстановление пропущенных или поврежденных данных.
2. Поиск выбросов (outlier detection). Есть множество объектов M. Найти в нем все аномальные объекты.
3. Поиск новизны (novelty detection). Есть множество объектов M. Определить, является ли объект $A \notin M$ похожим на объекты из M или нет?

Основные задачи машинного обучения

4. Кластеризация (clustering). Дано множество объектов. Их нужно разбить на несколько групп (кластеров), состоящих из похожих друг на друга объектов.

5. Предсказание (prediction). Есть множество объектов M с известными значениями признака Y. Найти значение признака Y для нового объекта $A \notin M$.

Насколько эффективно МО:

Проблемы:

- некомпетентность заказчиков;
- МО как фетиш;
- эффект от внедрения МО можно заметить не сразу;
- Кто будет нести ответственность за ошибки модели МО?
- если модель МО дает правильный ответ, то не совпадение ли это?
- сколько случаев совпадения с правильным ответом является ли признаком присутствия искусственного интеллекта?

Поэтому при работе в области МО нужно знать:

- основные идеи алгоритмов МО; что они могут делать, а что они сделать не в состоянии;
- какие операции при работе с данными не поддаются автоматизации, насколько результат МО зависит от «человеческого фактора»;
- методы обмана заказчика (это должны знать и исполнители и заказчики).

Представление данных для МО

Данные

Данные будем представлять в виде таблицы

Объекты	Признак 1	Признак 2	…	Признак m
A_1	P_{11}	P_{12}	…	P_{1m}
A_2	P_{21}	P_{22}	…	P_{2m}
…	…	…	…	…
A_n	P_{n1}	P_{n2}	…	P_{nm}

Здесь *n* объектов, у каждого их которых имеется *m* признаков (фич).

С помощью P(A) будем обозначать значение признака P для объекта A.

Число *n* называется объёмом выборки.

Например

Студент	Пол	Рост	Вес	Место на олимпиаде
Иванов	1	172	107	3
Запеканка	1	185	64	4
Ватрушкина	0	168	61	2
Ололоева	0	201	85	1

Значения признаков имеют разную природу.

Рост (вес) – это вещественные числа (количественный признак).

Пол – это бинарный признак (надолго ли?)

Место на олимпиаде – это порядковый признак.

Типы признаков

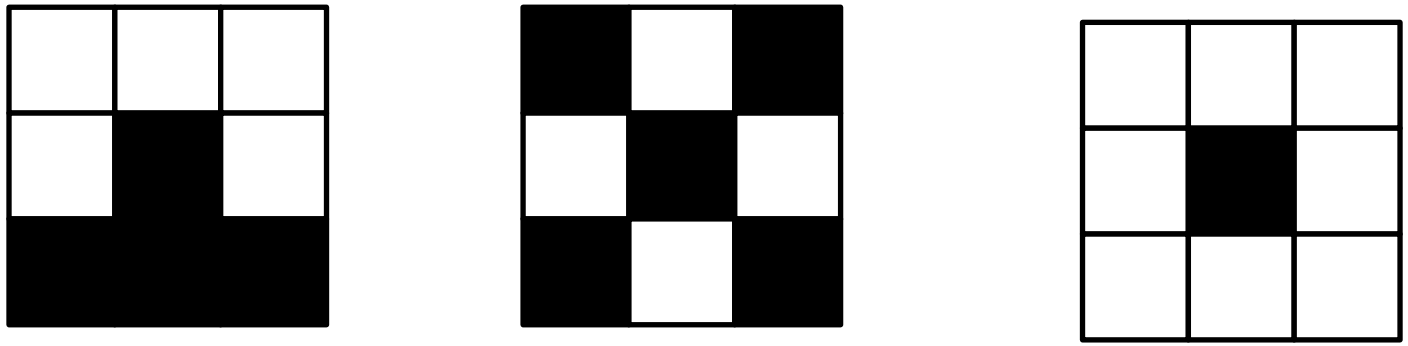
Признак называется:

- количественным (числовым), если область его значений – вещественные числа (и сам признак имеет числовую природу)
- порядковым – если признак задает порядок на объектах.
- номинальным (категориальным) – если признак не имеет числовой природы и (как правило) число его возможных значений конечно. В частности, бинарный признак – это номинальный признак с 2-мя возможными значениями.

Вот вам упражнение

Какие еще числовые, порядковые, бинарные признаки существуют у человека? А слабо найти номинальный, но не бинарный признак у человека?

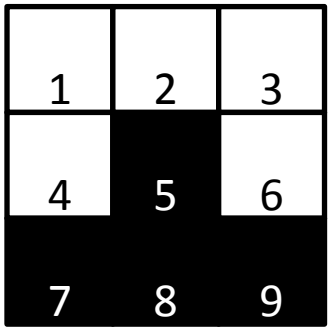
А как представить в виде таблицы набор картинок, аудио-файлов, текстов и т.п.?



- Допустим, есть черно-белые рисунки, состоящие из 9 пикселов.
- Нумеруем пиксели. Получаем 9 бинарных (так как рисунки черно-белые) признаков.

Превращение рисунка в строку из таблицы

Рисунок	пр1	пр2	пр3	пр4	пр5	пр6	пр7	пр8	пр9
Рис1	0	0	0	0	1	0	1	1	1
Рис2	1	0	1	0	1	0	1	0	1
Рис3	0	0	0	0	1	0	0	0	0



Вопрос: какие недостатки у такого представления рисунков?

Важное ограничение

Все объекты должны иметь одинаковое количество признаков (чтобы их можно было запихнуть в одну таблицу). Выполнить это требование не всегда легко:

1. Если объект А имеет более сложную структуру (более богатую историю) чем В.
2. Если объекты А,В – картинки, то они должны иметь одинаковый формат и размер.
3. Что делать если объекты имеют совершенно разную структуру (форму)? Например, это актуально для земельных участков.

Характеристики признаков

Поизучаем признаки по отдельности

Пусть *P* – столбец со значениями числового признака P из нашей таблицы:

$$P = (p_1, p_2, \dots, p_n)$$

Что можно посчитать для *P*?

1. Минимальное и максимальное значение.
2. Среднее значение

$$\bar{p} = \frac{p_1 + p_2 + \cdots + p_n}{n}$$

Что еще можно подсчитать для признака?

3. Медиану - такое число h_P, что ровно половина из элементов p_i больше него, а другая половина меньше него.

Медиана – это не то же самое, что и среднее:

Для (3,5,5,9,11) среднее значение 4.6, а медиана 5.

Для выборок чётного объёма медиана не определена однозначно.

Где тут медиана: (1,3,4,5)?

Но на практике в качестве медианы берут среднее арифметическое между двумя «центральными» значениями.

То есть в примере выше медиана равна (3+4)/2=3.5

Совет: чтобы быстро вычислить медиану, нужно мысленно упорядочить массив по возрастанию и взять число из середины.

Медиана VS среднее

- Значение медианы не так сильно (как среднее) зависит от попадания в выборку аномально больших и аномально малых значений признака.

Медиана VS среднее

Теперь усекаете, почему официальная пропаганда всегда употребляет понятия:

- «Средняя зарплата по стране»
- «Средняя продолжительность жизни»

а не

- «Медианная зарплата по стране»
- «Медиана продолжительности жизни»?

Задачка на медиану и среднее

Вот по телеку недавно сказали, что средний «объем задолженности жителя РФ по ипотечным кредитам» равен 100тыс руб (для этого они взяли суммарную ипотечную задолженность по стране и поделили на число всех жителей).

Вопрос: а чему равно медианное значение величины в кавычках? Дайте точный ответ.

Симметричные выборки

Если медиана и среднее близки друг к другу (не как в задаче про ипотеку), то выборка называется симметричной.

Важность симметричных выборок: для них проще искать аномалии.

А что значит: «близки друг к другу»? (См. ниже)

Мода

Мода - значение, которое встречается наиболее часто в выборке. Например, модой здесь (2,0,1,1,3,2,3,2) будет 2.

Мода не всегда определена однозначно.

!!! Мода (в отличие от среднего и от медианы) имеет смысл и для номинальных признаков.

Отклонение

Среднее и медиана не достаточны для адекватного описания выборки. Например, выборки (0,0,0,0,0) и (-2,-1,0,1,2) имеют одинаковые средние и медианы. Однако во второй выборке значения чаще отклоняются от среднего.

Отклонение (полное название: среднее квадратическое отклонение) считается по формуле:

$$s_P = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(p_i - \bar{p})^2}$$

Пример

Для P=(0,0,0,0,0)

s_P=0.

$$s_P = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(p_i - \bar{p})^2}$$

Для P=(-2,-1,0,1,2)

$$s_P = \sqrt{\frac{1}{5-1}[(-2-0)^2 + (-1-0)^2 + (0-0)^2 + (1-0)^2 + (2-0)^2} = \sqrt{\frac{10}{4}}$$

$$= 1.58$$

Выводы про отклонение:

- Отклонение всегда неотрицательно
- Отклонение значений признака P равно нулю, если…
- Чем больше величина отклонения, тем сильнее разброс значений выборки вокруг среднего значения.

Выводы про отклонение:

- Отклонение всегда неотрицательно
- Отклонение значений признака P равно нулю, если P состоит из одинаковых значений.
- Чем больше величина отклонения, тем сильнее разброс значений выборки вокруг среднего значения.

Симметричные выборки (дежавю)

Если медиана h_P и среднее $\bar{p}$ близки друг к другу (не как в задаче про ипотеку), то выборка называется симметричной.

А что значит: близки друг к другу? На практике считают, что если выполнено неравенство

$$|h_P - \bar{p}| \leq 3\sqrt{s_P^2/n}$$

Основная литература курса:

- Т. Сегаран «Программируем коллективный разум»
- Лекции М.Воронцова http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_%28курс_лекций%2C_К.В.Воронцов%29
- Видео с его лекций www.youtube.com/watch?v=qLBkB4sMztk&list=PLJOzdkh8T5kp99tGTEFjH_b9zqEQiiBtC
- Хабрахабр, особенно блог ODS habrahabr.ru/company/ods/
- Блог А.Дьяконова https://alexanderdyakonov.wordpress.com/
- Материалы компании 7bits https://github.com/7bits/ml-course-7bits

Коэффициент корреляции

Нужна величина, которая показывает, как значения одного признака определяют значения другого признака. Эта величина должна иметь смысл и для признаков с разными единицами измерения.

В статистике для таких задач используют коэффициент корреляции (КК).

Пример зависимости между столбцами

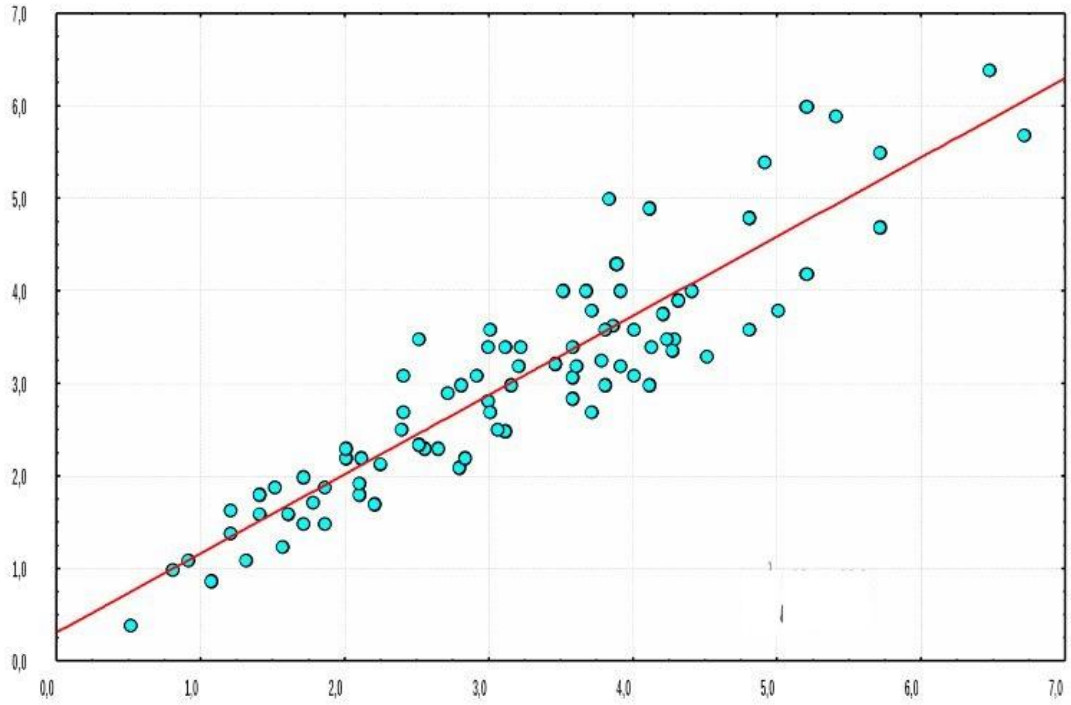
Здесь сильная зависимость между признаком *P1* и признаками *P3,P4,P5.*

P1	P2	P3	P4	P5
0	1	0	10	4
1	0	100	11	3
2	3	200	12	2
3	2	300	13	1

Так как их пары их значений ложатся на прямую.

Коэффициент корреляции

Более точно: КК – это показатель того, как значения признаков ложатся на прямую



Формула для КК

Пусть $P = (p_1, p_2, \dots, p_n) \quad Q = (q_1, q_2, \dots, q_n)$

признаки (столбцы из таблицы).

Тогда КК считается по формуле

$$r(P,Q) = \frac{\sum_{i=1}^{n} p_i q_i - n\bar{p}\bar{q}}{(n-1)s_P s_Q}$$

Свойства КК

Коэффициент корреляции (КК) – это число из отрезка [–1,1], которое имеет следующий смысл:

1. Если КК=0 (или близок к нему), то очевидной зависимости между признаками *P,Q* нет.
2. Если КК>0, то бОльшим значениям признака *P*, как правило, соответствуют бОльшие значения признака Q.
3. Если КК<0, то бОльшим значениям признака *P*, как правило, соответствуют меньшие значения признака Q.
4. Чем ближе значение КК к единице, тем сильнее зависимость между признаками *P,Q.*
5. Если модуль КК равен 1, то между признаками *P,Q* существует линейная зависимость.

Задача на понимание

Однажды я попросил, чтобы студенты ответили на 2 вопроса анкеты «ваш год рождения» и «ваш возраст». Из их ответов я сформировал таблицу, в которой был столбец P=«год рождения студента» и Q=«возраст студента».

Вопрос 1: оцените (приближенно) КК $r(P, Q)$.

Задача на понимание

Однажды я попросил, чтобы студенты ответили на 2 вопроса анкеты «ваш год рождения» и «ваш возраст». Из их ответов я сформировал таблицу, в которой был столбец P=«год рождения студента» и Q=«возраст студента».

Вопрос 1: оцените (приближенно) КК $r(P,Q)$.

Вопрос 2: как зависит $r(P,Q)$ от месяца, в котором проводится опрос (я не шучу)?