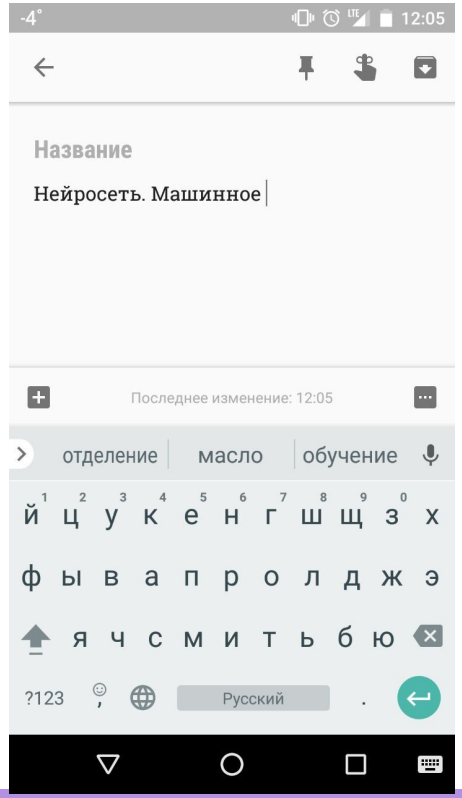
Задачи машинного обучения

Лекция 1

7bits

Рекомендательные системы

Работа с текстом



Горячие Продажа Оригинал CHEERSON Мини CX-STARS RC Micro Квадрокоптер 2.4 Г 4CH 6-осевой Карман Беспилотный БПЛА 3d-ролл Вертолеты ПРОТИВ FQ777-124

английский | русский | немецкий | Определить язык

русский | английский | украинский | Перевести

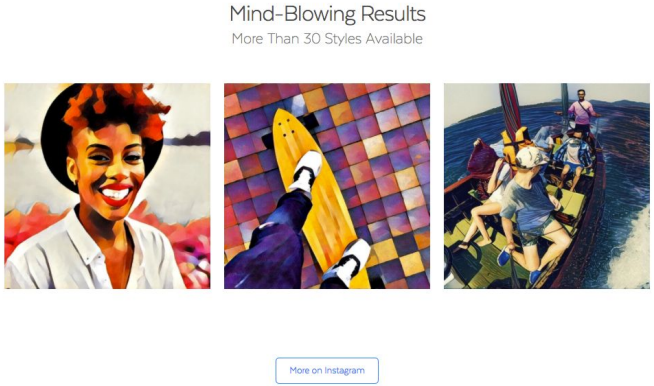
Don't harm humans.
Obey orders.
Protect yourself.

49/5000

Не навреди людям.
Выполняйте приказы.
Защити себя.

Ne navredi lyudyam.
Vypolnyayte prikazy.
Zashchiti sebya.

Обработка изображений



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).
- Число n называется объемом выборки.

Разные задачи ML

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

Постановка задачи

- X – пространство объектов
- x – объекты из пространства
- Y – пространство ответов
- y: X -> Y – целевая функция
- y = y(x) – ответ

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

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.

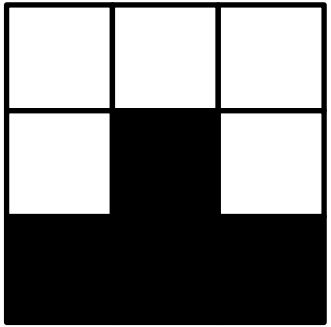
Вот вам упражнение

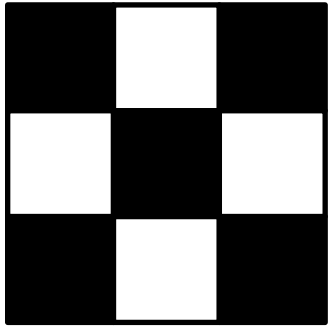
Назовите тип каждого признака для этой таблицы

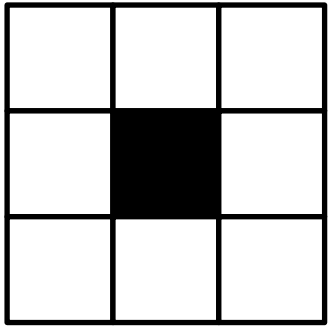
Студент	Пол	Рост	Вес	Место на олимпиаде
Иванов	1	172	107	3
Запеканка	1	185	64	4
Ватрушкина	0	168	61	2
Ололоева	0	201	85	1

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

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





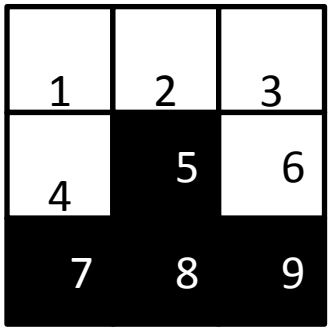


Допустим, есть черно-белые рисунки, состоящие из 9 пикселов.

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?

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, что ровно половина из элементов pi 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 500тыс руб.

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

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)

sP=0.

Для P=(-2,-1,0,1,2)

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

$$s_P = \sqrt{\frac{1}{5-1}[(-2-0)^2 + (-1-0)^2 + (0-0)^2 + (1-0)^2 + (2-0)^2}} = \sqrt{\frac{10}{4}} = 1.58$$

Выводы:

Отклонение всегда неотрицательно

Отклонение выборки равно нулю, если… (додумайте сами).

Чем больше величина отклонения, тем сильнее разброс значений выборки вокруг среднего значения.

Проблема!

В таблице могут быть незаполненные ячейки, или же содержимое некоторых ячеек заведомо некорректное и противоречит здравому смыслу

Студент	Пол	Рост	Вес	Место на олимпиаде
Иванов	1	172	107	3
Запеканка	?	185	64	-4
Ватрушкина	0	168	666	2
Ололоева	0	?	85	1

Что нужно делать?

- Удалить объект (т.е. строку).
- Удалить столбец, если в нём очень много пропусков.
- Заменить значение в ячейке на среднее (медиану, моду…) из значений столбца.
- Например, для столбца-признака (0,1,2,4,4,5,?) пропущенное значение можно заменить на среднее 2.67, медиану 3, моду 4.

А что делать с номинальными признаками?

Есть признак «пол человека» (0,0,0,1,1,?). На что заменить пропущенное значение???

Ваши предложения?

Еще подход: восстановление данных с помощью метрики

А что такое метрика? Это обобщение понятия расстояния из геометрии, но метрика – это расстояние…

1. не обязательно между двумя точками;
2. не обязано вычисляться по формуле из учебника геометрии.

Известные метрики

Если даны два набора

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

то расстояние между P и Q может быть найдено, например, такими способами:

1. как в учебнике геометрии (евклидова метрика)

$$\rho(P, Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

Известные метрики

2. Метрика Манхеттен

$$\rho(P,Q) = |p_1 - q_1| + |p_2 - q_2| + \ldots + |p_n - q_n|$$

3. max-метрика

$$\rho(P,Q) = \max\{|p_1 - q_1|, |p_2 - q_2|, \ldots, |p_n - q_n|\}$$

В общем, метрик очень много, вы можете придумать свои собственные.

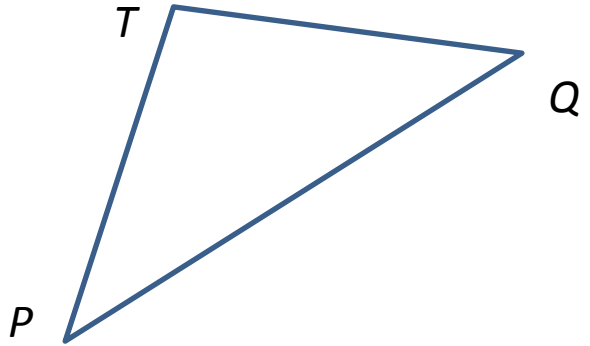
Свойства метрики

1. $\rho(P,P) = 0$(в принципе, логично)
2. $\rho(P,Q) = \rho(Q,P)$(разумно). Но можете ли привести пример из жизни, когда расстояние от P до Q не равно расстоянию от Q до P?
3. $\rho(P,Q) \leq \rho(P,T) + \rho(T,Q)$ (неравенство треугольника).

А есть ситуации, когда

это неравенство

не выполняется?



Восстановление данных с помощью метрики

Пусть у объекта A значение признака P не корректно. Как восстановить P для A?

Выкинем из таблицы столбец с признаком P (мы предполагаем, что остальные ячейки в таблице нормальные).

Найдем расстояния (с помощью некоторой метрики) от строки A до остальных объектов таблицы. Получим числа

$$\rho(A, A_1), \rho(A, A_2), \ldots, \rho(A, A_n)$$

Восстановление данных с помощью метрики

3. Пусть значения признака P для объектов A1,A2,…,An равны P(A1), P(A2),…, P(An). Тогда восстановленное значение признака P для A равно… (щас прервёмся на примерчик)

Пример

Объекты	P1	P2	P3	P4	P
A1	3	4	5	3	4
A2	5	5	5	4	3
A3	4	3	3	2	5
A	5	4	3	3	?

Если пропуск заменить на среднее или медиану по столбцу (здесь они равны друг другу), то нужно писать 4.

Попытаемся заполнить пропуск с помощью различных метрик.

Объекты	P1	P2	P3	P4	P
A1	3	4	5	3	4
A2	5	5	5	4	3
A3	4	3	3	2	5
A	5	4	3	3	?

Вид метрики	От A до A1	От A до A2	От A до A3
Евклид	2,83	2,45	1,73
Манхеттен	4	4	3
Макс	2	2	1

Кто догадается, что с этими числами делать дальше?

Ключевая идея: признак P для объекта A должен быть близок к значению признака P у близких к A объектов.

Нужно взять их лин. комбинацию со значениями признака P

То есть по евклидовой метрике надо так:

Вид метрики	От A до A1	От A до A2	От A до A3
Евклид	2,83	2,45	1,73
Манхеттен	4	4	3
Макс	2	2	1

$$\frac{1}{\frac{1}{2.83}+\frac{1}{2.45}+\frac{1}{1.73}}\left(\frac{4}{2.83}+\frac{3}{2.45}+\frac{5}{1.73}\right)=4.15$$

Объекты	P
A1	4
A2	3
A3	5
A	?

Аналогично дается ответ с помощью других метрик

Вид метрики	От A до A1	От A до A2	От A до A3
Евклид	2,83	2,45	1,73
Манхеттен	4	4	3
Макс	2	2	1

Например,

макс-метрика даёт:

Объекты	P
A1	4
A2	3
A3	5
A	?

$$\frac{1}{1/2 + 1/2 + 1/1}(4/2 + 3/2 + 5/1) = 4.25$$

Умная формула для восстановления данных с помощью метрики

$$P(A) = \frac{1}{\sum_{i=1}^{n} \frac{1}{\rho(A, A_i)}} \left(\sum_{j=1}^{n} \frac{P(A_j)}{\rho(A, A_j)} \right)$$

Восстановление данных с помощью других столбцов

(Более подробно идеи этой темы будут обсуждаться в теме «Отбор признаков» и «Линейная регрессия»)

В отличие от предыдущего метода тут есть трудность: значения в разных столбцах могут иметь разный масштаб.

	Грудь	Талия	Бёдра	Рост	Вес
Ж1	99	56	91	160	58
Ж2	89	58	89	157	48
Ж3	91	64	91	165	54
Ж4	91	51	91	170	54
Ж5	86	56	84	157	44
Ж6	97	53	86	175	56
Ж7	?	51	91	165	54

Восстановление данных с помощью других столбцов

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.

Формула для КК

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

столбцы (строки) из таблицы.

Тогда КК считается по формуле

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

Коэффициенты корреляции:

r(Грудь,Талия)= -0,22

r(Грудь,Бедра)= 0,34

r(Грудь,Рост)= 0,46

r(Грудь,Вес)= 0,91

Смысл в том, признак с большим (по модулю) КК имеет большее влияние на размер груди.

Осталось изобрести подходящую формулу…

Умная формула

Пусть P(A) - значение признака P объекта A.

$\bar{P}$ - среднее значение признака P.

Требуется определить P(A) по столбцам-признакам P1,P2,…,Pm

$$P(A) = \bar{P} + \frac{\sum_{j=1}^{n} r(P, P_i)(P_i(A) - \bar{P}_i)}{\sum_{i=1}^{m} |r(P, P_i)|}$$

Примечание: в формуле все величины вычисляются без учета строки объекта A

В нашем примере имеем

A	?	51	91	165	54

r(Грудь,Талия)= -0,22

r(Грудь,Бедра)= 0,34

r(Грудь,Рост)= 0,46

r(Грудь,Вес)= 0,91

	Среднее
Грудь	92,17
Талия	56,33
Бёдра	88,67
Рост	164
Вес	52,33

Формула дает

$$P(A) = 92.17 + \frac{1}{0.22 + 0.34 + 0.46 + 0.91} \cdot (-0.22(51 - 56.33) + 0.34(91 - 88{,}67) + 0.46(165 - 164) + 0.91(54 - 52.33))$$

$$= 92.17 + 0.52(1{,}17 + 0.79 + 0.46 + 1.52) = 94.22$$

Применение метрик и КК в рекомендательных системах (РС)

Формально РС - это таблица: строки соответствуют пользователям, столбцы – товарам, а в (i,j)-ячейке стоит оценка, которую поставил i-й пользователь j-му товару (в ячейке может быть пустой, если оценивания не было).

	Чупачупс	Доширак	Боярышник	ВАЗ-2101	Семечки
Вася	3	4	5	3	4
Петя	5	5	5	4	3
Маша	4	3	3	2	5
Саша	5	4	3	3	?

Предсказать оценку в РС – это фактически восстановить данные в таблице!!!

А выше было показано, как это делается с помощью метрики и КК.

!!! Только здесь нужно считать КК для строк (а не для столбцов как раньше).При использовании КК получится ответ 4.12

	Чупачупс	Доширак	Боярышник	ВАЗ-2101	Семечки
Вася	3	4	5	3	4
Петя	5	5	5	4	3
Маша	4	3	3	2	5
Саша	5	4	3	3	4.12

Восстановление данных с помощью моделей предсказания

(Более подробно идеи этой темы будут обсуждаться в теме «Модели классификации» и «Модели регрессии»)

Идея: если мы сумеем научить предсказывать значение признака P используя для этого все другие признаки, то автоматически будет решена проблема восстановления данных для признака P.
В этой таблице восстановление данных эквивалентно предсказанию места на олимпиаде по остальным данным студента

Студент	Пол	Рост	Вес	Место на олимпиаде
Иванов	1	172	107	?
Запеканка	1	185	64	?
Ватрушкина	0	168	61	?
Ополоева	0	201	85	?

Используемая литература

- Википедия (см. медиана, мода, отклонение…)
- Статья habrahabr.ru/company/infopulse/blog/283168/
- Обзорная статья: «Collaborative Filtering Recommender Systems» by M.Ekstrand, J.Riedl, J.Konstan
- Книга Т.Сегаран «Программируем коллективный разум»

Формат курса

- Лекция - про теорию
- Семинар - про приложение (python+scipy+sklearn)
- Домашнее задание в github
  - Заполнять пропуски в jupyter notebook
  - Проверять иногда ДЗ сокурсников
  - Финальный проект в виде ML-соревнования

Домашнее задание 1

1. Create @**gmail.com** mailbox
2. Create GitHub account
3. Email to teachers@7bits.it your Gmail and GitHub addresses
4. Install Git
5. Create and push something to GitHub

Домашнее задание 2

Следуйте инструкциям отсюда:

https://github.com/7bits/ml-course-7bits/blob/master/2017-fall/homeworks/homework-01.md

Полезные ссылки

- Задания этого курса https://github.com/7bits/ml-course-7bits

- Клуб машинного обучения в Омске vk@mlomsk
- Полезные материалы по ML&DS