Данные и признаки: что машина знает о мире

Как убрать рассылки и не потерять сообщение от друга? Соберём данные для фильтра и посмотрим, что меняется вместе с вопросом.

Вы договорились встретиться с другом после занятий и ждёте, куда он предложит пойти. Телефон показывает несколько новых сообщений. Среди них — реклама, уведомления и личная переписка. Хочется убрать лишнее, но сообщение друга потерять нельзя. Поручим программе простую работу: отделять спам от обычных сообщений.

С чего начать? Кажется, спам легко узнать по заглавным буквам, восклицаниям и телефонным номерам. Проверим эту догадку на настоящем SMS. По-русски его смысл такой: «Мой номер в Лутоне... Позвони, если будешь рядом!». А вот оригинал из набора SMS Spam Collection:

> *MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H**

Здесь есть все три подозрительных признака. Отправили бы вы это сообщение в спам?

Как сообщение размечено в наборе
Как обычное: человек сообщает свой номер и предлагает позвонить.
Правило про заглавные, телефон и восклицание спрятало бы личное сообщение вместе с рекламой.

Чтобы придумать правило получше, сначала разберёмся, какую информацию дать программе. Все примеры этой главы будут возвращать нас к одному вопросу: что нужно знать о сообщении, чтобы принять нужное нам решение?

Ранее

В первой главе мы меняли вопрос к музыкальной рекомендации и проверяли скрытую оценку. Теперь соберём входные данные для другой знакомой задачи: разобрать сообщения на телефоне.

← Карта задач ИИ

Пример

Настоящие данные. Английские SMS и их метки взяты из SMS Spam Collection. Русский текст рядом объясняет смысл. Признаки считаются по оригиналу, включая его пробелы и опечатки.

1. Строка, столбец, ответ

Начнём с трёх SMS. Номер позволяет найти оригинал в файле. Длина включает буквы, пробелы и знаки препинания.

Строка источника	Оригинал SMS	Символов	Цифр	Разметка
371	Cool, text me when you're ready	31	0	обычное
5384	You have 1 new message. Call 0207-083-6089	42	12	спам
264	MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*	51	10	обычное

Первое означает «Напиши, когда будешь готов». Второе предлагает позвонить, чтобы получить новое сообщение. По числу цифр они различаются. Но у обычного SMS про Лутон цифр тоже много: человек передаёт номер телефона.

Объект здесь одно SMS: для него нужен один ответ. **Признаки** описывают доступные сведения об объекте, например число цифр. **Разметка** содержит уже известные ответы для учебных примеров. Её держат отдельно от признаков: у нового SMS ответа ещё нет.



Один объект превращается в строку признаков. Метку берём отдельно из размеченного набора.

Заметка
Одна строка зависит от вопроса. Если предсказываем спам, строкой будет SMS. Если хотим понять, кто из отправителей рассылает спам, строкой станет отправитель; число его сообщений за прошедший день может быть признаком.

> Фильтр должен при получении SMS определить, обычное оно или спам. В учебных примерах есть тексты и ответы проверяющих. Как разделить сведения для обучения?
>
> - ○ Объект: одно SMS; признаки: длина и число цифр; разметка: обычное или спам
> - ○ Объект: одно SMS; признаки: правильная метка; разметка: длина
> - ○ Объект: одна цифра; признаки: все сообщения; разметка: телефон
> - ○ Объект: одно SMS; признаки: жалоба на него завтра; разметка: число цифр

2. Что исчезает при превращении в числа

Мы решили передать фильтру четыре простых признака: длину SMS, количество цифр, процент заглавных среди букв и наличие «!». Для последнего записываем 1, если знак есть, и 0, если его нет. У всех сообщений счётчики работают одинаково. Слова и их порядок в такую строку не попали.

Перед экспериментом сделайте прогноз: справится ли правило «есть цифры → спам» с уведомлением о новом сообщении и SMS про Лутон? Откройте оба оригинала и проверьте свой прогноз. Затем выключите этот признак и сравните, что осталось.

Соберите строку признаков

Короткий спам

You have **1** new message. Call **0207-083-6089**

Смысл по-русски: У вас одно новое сообщение. Позвоните по номеру 0207-083-6089.

Символов	Цифр	Заглавных, % букв	Есть «!» (1/0)
42	12	10	0

> Разметка источника: спам. Этот ответ не вычисляется счётчиками.

Строка 5384 оригинального файла · 4 из 4 признаков включено. Длина включает пробелы и знаки; перевод в подсчёте не участвует.

Обычное SMS с цифрами

MY NO. IN LUTON **0125698789** RING ME IF UR AROUND! H*

Смысл по-русски: Мой номер в Лутоне... Позвони, если будешь рядом!

Символов	Цифр	Заглавных, % букв	Есть «!» (1/0)
51	10	100	1

> Разметка источника: обычное сообщение. Этот ответ не вычисляется счётчиками.

Строка 264 оригинального файла · 4 из 4 признаков включено. Длина включает пробелы и знаки; перевод в подсчёте не участвует.

Один признак: обычное SMS с «!»

MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*

Смысл по-русски: Мой номер в Лутоне... Позвони, если будешь рядом!

Есть «!» (1/0)
1

Разметка источника: обычное сообщение. Этот ответ не вычисляется счётчиками.

Строка 264 оригинального файла · 1 из 4 признаков включено. Длина включает пробелы и знаки; перевод в подсчёте не участвует.

Тот же признак: спам с «!»

Camera - You are awarded a SiPix Digital Camera! call 09061221066 fromm landline. Delivery within 28 days

Смысл по-русски: Вам присуждена цифровая камера. Позвоните с городского телефона. Доставка в течение 28 дней.

Есть «!» (1/0)
1

Разметка источника: спам. Этот ответ не вычисляется счётчиками.

Строка 3291 оригинального файла · 1 из 4 признаков включено. Длина включает пробелы и знаки; перевод в подсчёте не участвует.

Пропуск текста

Текст не загрузился

Это учебный сбой. В оригинале текст есть.

Символов	Цифр	Заглавных, % букв	Есть «!» (1/0)
нет данных	нет данных	нет данных	нет данных

Разметка источника: обычное сообщение. Этот ответ не вычисляется счётчиками.

Строка 264 оригинального файла · 4 из 4 признаков включено. Пропуск не равен нулю: длина неизвестна.

Выберите SMS, включайте признаки и открывайте разметку. Цифры подсвечиваются в оригинале. Кнопка пропуска имитирует сбой загрузки; это учебное изменение, в исходном наборе текст сохранён.

Интерактивный эксперимент — в электронной версии урока.

В уведомлении 12 цифр, в SMS про Лутон 10. Правило «есть цифры → спам» назовёт спамом оба и ошибётся на обычном SMS.

Если выключить все признаки, фильтр потеряет любые различия между сообщениями. Если оставить только «есть восклицательный знак», обычное SMS про Лутон и реклама камеры получат одинаковое значение: 1. Какое бы правило мы ни придумали на одном этом признаке, оно даст им одинаковый ответ и ошибётся хотя бы на одном.

Подумай

Можно ли спасти такой фильтр, добавив ещё тысячу примеров, но оставив единственный признак «есть !»? Примеры помогут узнать частоты, однако эта пара останется неразличимой.

Теперь нажмите «Смоделировать пропуск текста». Длина стала неизвестной. Ноль означал бы, что текст получен и пуст. Подмена пропуска нулём сообщила бы фильтру факт, которого у нас нет.

При этом метка источника может сохраниться. Такую строку можно учесть при подсчёте меток «спам», но вычислить длину по ней нельзя. Полезность строки зависит от того, какой вопрос мы задаём.

> Сколько цифр содержит настоящее SMS «MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*»?

Профи. **Числовое представление и его ограничения**

Обозначим текст через s, выбранные признаки через x, метку через y. Мы используем отображение $s \mapsto x = (\text{длина}, \text{цифры}, \text{доля заглавных}, \text{есть !})$. Доля заглавных считается только среди букв; если букв нет, в этой лаборатории договоримся записывать ноль.

Модель получает x и предсказывает y. Совпадение x не доказывает, что совпадают исходные тексты. Поэтому удалять строки как дубликаты лишь по совпадению нескольких признаков нельзя.

Посчитать цифры в SMS

Цель: Исправьте счётчик, чтобы получить 10 цифр

```
text = "MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*"
digits = len(text)  # TODO: сейчас считаются все символы
print(digits)
```

3. Меняем вопрос — меняем нужные ответы

История

Сборник опубликован исследователями в 2011 году. Часть спама извлекли из жалоб на британском форуме Grumbletext; значительную часть обычных сообщений получили из добровольно переданных SMS корпуса Национального университета Сингапура. Это смесь нескольких коллекций.

В учебном наборе к каждому тексту приложена метка `ham` («обычное») или `spam`. Она пришла из источника данных. Мы не получили её подсчётом заглавных. Во время обучения модель ищет связь признаков с этими ответами; при проверке сравнивают её решения с сохранёнными метками.

Разметка требует договорённости. Учебный пример: магазин прислал скидку человеку, который согласился на рассылку, а другому отправил тот же текст без согласия. Если спам определён как нежелательная рассылка, ответы могут различаться. Самого текста для решения недостаточно: нужно знать о согласии. В SMS Spam Collection таких сведений нет.

Метка тоже может быть ошибочной или спорной. Подозрительный случай стоит перечитать, записать причину сомнения и проверить по принятому правилу. Незаметно менять метки под ответы собственной модели нельзя: тогда она будет проверять сама себя.

А если хочется не только убирать спам? Те же сообщения можно использовать в других задачах, но прежних меток уже не хватит.

Что хотим узнать?	Какой нужен ответ?	Что должно быть в примерах?
Это спам?	Обычное / спам — классификация	Метка для каждого сообщения
Через сколько минут получатель ответит?	Число минут — регрессия	Время получения и ответа; отдельно отмечаем отсутствие ответа
Что показать первым?	Порядок сообщений — ранжирование	Предпочтения получателя: какое сообщение важнее другого

У SMS про встречу и у банковского уведомления может быть одна метка «обычное», но разная важность для вас сейчас. Поэтому метка «не спам» не заменяет ответ на вопрос «показать первым». В нашем настоящем наборе есть тексты и метки спама. Времени ответа и личных приоритетов нет: их пришлось бы собирать отдельно.

Мы хотим предсказывать время ответа, но в таблице есть только тексты и метки «обычное / спам». Чего не хватает для обучения такому прогнозу?

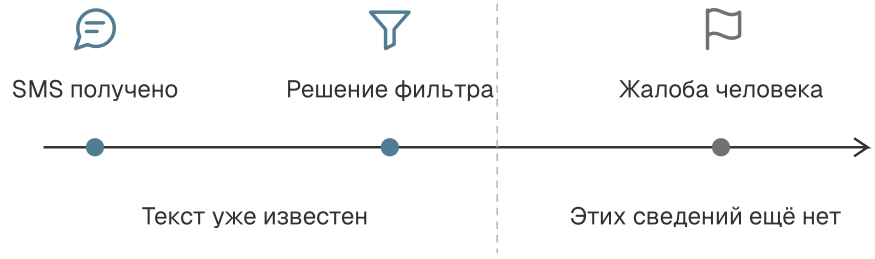
- ○ Ещё одной копии столбца «спам»
- ○ Более длинного названия модели
- ○ Наблюдений о времени получения и ответа, с отметкой случаев без ответа
- ○ Перевода слова «обычное» в число 1

В исходном файле 5574 сообщения, из них 747 помечены как спам. Сколько процентов сообщений помечено как спам? Округлите до десятых.

4. Подсказка из будущего

Вернёмся к фильтру, который должен сохранить сообщение друга. Представьте выгрузку: рядом с SMS записано, нажал ли получатель «Пожаловаться на спам». На старых сообщениях это кажется прекрасным признаком. Но фильтр должен решить, куда положить SMS, до того, как человек его прочитает. Жалобы в этот момент ещё нет.

Такую ошибку называют **утечкой ответа**: модель учится использовать сведения, которых не будет при настоящем решении. Проверять доступность признака нужно по времени, когда фильтр срабатывает.



Жалоба на это SMS возникает после решения фильтра. Включать её в признаки при получении сообщения нельзя.

Заметка

История жалоб на отправителя, накопленная до нового SMS, может быть допустимым признаком. Жалоба на само новое сообщение появляется позже. Похожие названия столбцов скрывают разный порядок событий.

Фильтр принимает решение сразу при получении SMS. Какой признак доступен ему в этот момент и не содержит сведений из будущего?

- ○ Итог проверки этого SMS модератором завтра
- ○ Удалит ли получатель это SMS вечером
- ○ Жалоба получателя на это SMS через час
- ○ Число жалоб на отправителя, накопленных до получения этого SMS

5. Паспорт: откуда примеры и кому они подходят

Соберём паспорт нашего набора: что именно попало в файл и для каких решений этих данных достаточно.

Вопрос	Паспорт нашего набора
Что за объект?	Одно англоязычное SMS
Какой ответ?	Метка обычное / спам из исходной коллекции
Сколько?	5574 строки, 747 со спамом
Откуда?	Несколько публичных коллекций, включая Grumbletext и NUS
Что сделали в уроке?	Выбрали восемь примеров вручную для сравнения, оригиналы сохранили; обучение и проверку модели здесь не проводим
Повторы?	5171 различная пара «текст, метка»; остальные 403 строки повторяют такие пары
Где осторожнее?	Другой язык, новые рассылки, другая аудитория

Ссылки

Страница UCI содержит описание источников и лицензию CC BY 4.0. Архив с оригинальным файлом позволяет пересчитать строки. Восемь примеров лаборатории подобраны для объяснения и не передают частоты всей коллекции.

Повтор текста не всегда ошибка: одинаковое «Позвоню позже» могут отправить разные люди. Но если копия уже встречалась при обучении, проверка на ней слабее проверки на новом сообщении. Для честного экзамена группы одинаковых текстов держат вместе при разделении. Совпадение длины и числа цифр само по себе повтором не считается.

Ещё одна ловушка: привычные слова могут выдавать источник коллекции. Фильтр рискует научиться отличать студенческую переписку одного региона от рекламных жалоб другого. Поэтому успех на этих SMS не доказывает, что он справится с русскоязычной рассылкой вашей школы.

> Набор SMS Spam Collection содержит англоязычные SMS из нескольких публичных коллекций. Школе нужен фильтр русскоязычных сообщений. Какой вывод о применимости набора обоснован?
>
> - ○ На нём можно разобрать механизм, а качество для школы нужно проверять на подходящих школьных сообщениях
> - ○ Если использовать только длину, отдельная проверка больше не нужна
> - ○ Достаточно скопировать набор десять раз, чтобы охватить школьную переписку
> - ○ Любой фильтр с этого набора готов для школы

Профи. Повторы, пропуски и аккуратный подсчёт

Если 20 строк не содержат текста, 15 не содержат метки, а у 5 отсутствует и то и другое, удалить придётся $20 + 15 - 5 = 30$ строк. Пять строк вошли в обе группы; второй раз их удалять нельзя. Это отдельный учебный пример с пропусками, не описание оригинального SMS Spam Collection.

> В учебной выгрузке 450 строк. Без текста 49, без метки 36, а без обоих 8. Сколько останется после удаления строк с любым пропуском?

Эпик. **Предел любой модели с выбранными признаками**

Пусть пять сообщений получили один и тот же набор признаков. Три размечены как обычные, два как спам. Детерминированная модель, использующая только эти признаки, обязана дать всем пяти один ответ. Назвав их обычными, она ошибётся дважды; назвав спамом, трижды. Две ошибки неизбежны.

Докажите общий результат. Разбейте набор на группы с одинаковыми признаками. Пусть в группе g количества двух меток равны a_g и b_g. Наименьшее число ошибок произвольной детерминированной функции на этом конечном наборе:

$$E_{\min} = \sum_{g} \min(a_g, b_g).$$

Нижняя оценка получается внутри каждой группы. Она достижима: каждой группе назначаем её более частую метку. Добавление признака может разделить группу и снизить предел. Усложнение модели при прежних признаках эту неоднозначность не устранит. Это утверждение о конечном наборе; оно не обещает качества на новых объектах.

Объект: одно сообщение, о котором нужен ответ.

Признаки и разметка: первые описывают доступный вход; вторая хранит ответы учебных примеров.

Новый вопрос: спам, время ответа и личная важность требуют разных сведений для проверки.

Ограничения данных: одинаковые признаки могут скрывать разный смысл. Проверяем пропуски, повторы и доступность сведений в момент решения.

Домашняя работа

1. Три учебных сообщения: А — «Встречаемся у входа в 17:00!» от друга; Б — «Вы выиграли приз!» из нежелательной рассылки; В — «Завтра занятий нет» от одноклассника. Составьте таблицу: одно сообщение на строку, признаки «число цифр» и «есть !», отдельно ответ «обычное / спам».
2. Учебные SMS: А — «Встречаемся у входа в 17:00!» от друга; Б — «Вы выиграли приз!» из нежелательной рассылки; В — «Завтра занятий нет» от одноклассника. Примените к ним правило «если есть ! — спам, иначе — обычное». Где оно ошибётся? Объясните, почему по одному этому признаку нельзя различить А и Б.

3. В архиве SMS есть только тексты и метки «обычное / спам». Нужен прогноз времени от получения сообщения до вашего ответа. Какие сведения придётся добавить в учебные примеры? Объясните, почему настоящее время ответа можно использовать для проверки, но нельзя дать модели как признак до прогноза.

Следующий сюжет

Из одного признака построим правило фильтра: выберем границу по длине SMS и посмотрим, какие обычные сообщения оно спрячет, а какой спам пропустит.

Источники

- SMS Spam Collection, UCI (https://archive.ics.uci.edu/dataset/228/sms+spam+collection) — авторы Almeida и Hidalgo, происхождение, описание меток и лицензия CC BY 4.0.
- Оригинальный архив UCI (https://archive.ics.uci.edu/static/public/228/sms+spam+collection.zip) — файл SMSSpamCollection; проверен 21 сентября 2026 года. Номера строк лаборатории: 371, 5384, 1353, 577, 264, 3291, 5040, 81.
- Scikit-learn: ошибки при подготовке данных (https://scikit-learn.org/stable/common_pitfalls.html#data-leakage) — почему сведения из проверки и будущего не должны попадать в обучение.