Классификация: как фильтр узнаёт спам

Можно ли узнать рекламу по длине сообщения? Соберём фильтр, передвинем его границу и увидим, кому достаются ошибки.

Вы ждёте сообщение об отмене занятия. Телефон показывает рекламу розыгрыша, а нужное уведомление находится в спаме. Фильтр ошибся дважды, но последствия разные: рекламу можно закрыть, об отмене занятия можно узнать уже у запертой двери.

Соберём маленький фильтр и проследим каждое его решение. Он будет смотреть только на длину SMS. Это намеренно грубое правило: на нём легко увидеть, как настройка помогает одним сообщениям и мешает другим.

1. Сначала попробуем сами

Прочитайте: «Напиши, когда будешь готов» и «Вы выиграли приз, позвоните за наградой». Первое сообщение похоже на обычную переписку, второе вызывает подозрение. Вы использовали смысл текста и жизненный опыт. Нашему фильтру этот опыт пока недоступен.

Его задача называется **классификацией**: для каждого объекта выбрать один из заранее заданных классов. Объект у нас один SMS, классы «обычное» и «спам». Мы договорились, что ответ «спам» отправляет сообщение в отдельную папку. Для другого продукта можно придумать проверку человеком или промежуточную папку, но здесь будем выбирать из двух.

Попробуем правило «длинные сообщения считать спамом». Оно ещё недостаточно точное: для одного человека 50 символов много, для другого мало. Компьютеру нужна граница, чтобы одно и то же сообщение каждый раз получало один ответ.

> Какое действие фильтра решает задачу классификации?
>
> - ○ Выбирает для сообщения папку «Входящие» или «Спам»
> - ○ Переводит сообщение на русский язык
> - ○ Показывает сообщения по времени получения
> - ○ Считает символы в сообщении

2. Проведём границу на длине 60

Ниже шесть настоящих сообщений из SMS Spam Collection. Это специально выбранные примеры, на которых видно столкновение классов. Их пропорция не описывает весь набор или вашу переписку. В колонке «Содержание» дан короткий русский пересказ; длину считаем по полному английскому оригиналу.

Ранее

В прошлом сюжете мы превратили сообщения в строки таблицы. Длина описывает SMS, а метка человека сообщает, считалось ли оно спамом. Теперь попробуем принять решение по одному признаку.

← Данные и признаки: что машина знает о мире

← Карта задач ИИ

Пример

В настоящем фильтре SpamAssassin правила начисляют письму баллы, а порог определяет решение. Здесь вместо суммы баллов возьмём длину: её можно пересчитать самостоятельно. Документация настройки required_score.

Подумай

Друг прислал длинный адрес с указаниями, как найти вход. Какая информация поможет вам понять сообщение? Какая из неё останется фильтру, если мы передадим ему только длину?

№	Содержание	Символов	Метка в наборе
1	Напиши, когда будешь готов	31	обычное
2	У вас новое сообщение, позвоните по номеру	42	спам
3	Завтра ведь нет занятий?	30	обычное
4	Вы выиграли деньги или приз, позвоните	69	спам
5	Мой номер в Лутоне, позвони, если будешь рядом	51	обычное
6	Вам досталась камера, позвоните для получения	105	спам

Ссылки

SMS Spam Collection в UCI. Для этих примеров взяты строки оригинального файла 371, 5384, 1353, 577, 264 и 3291. Пробелы и знаки препинания входят в длину; метка и разделяющая табуляция не входят. Полные оригиналы открываются в интерактиве ниже.

Назначим порог 60: **если в сообщении 60 символов или больше, отправляем его в спам**. Остальные оставляем во входящих. Равенство тоже оговорено: сообщение длиной ровно 60 попадёт в спам.

Применим правило к строкам. Сообщения 4 и 6 достаточно длинные; фильтр отправит их в спам. Сообщения 1, 2, 3 и 5 останутся во входящих. Сравним результат с метками: всё верно, кроме сообщения 2. Короткая реклама прошла через фильтр.



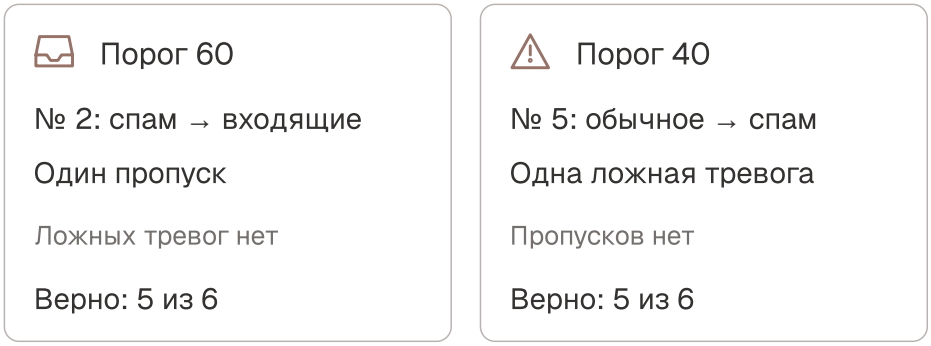
Три действия фильтра: прочитать текст, посчитать его длину, сравнить с порогом. Метка человека нужна позже, чтобы проверить результат.

> Фильтр отправляет в спам SMS длиной 60 символов и больше, остальные — во входящие. Обычные SMS имеют длины 20 и 75 символов, спам — 60 и 90. Сколько из этих четырёх сообщений фильтр распределит верно?

3. Два способа ошибиться

Пропуск случается, когда спам остаётся во входящих. Так произошло с сообщением № 2. **Ложная тревога** случается, когда обычное сообщение отправляется в спам. Пока при пороге 60 её нет.

Теперь снизим порог до 40. Сообщение № 2 попадёт в спам, как мы и хотели. Вместе с ним туда отправится сообщение № 5: 51 символ тоже больше 40. По метке это обычная переписка. Мы исправили пропуск и получили ложную тревогу.



Два порога дают одинаковое число ошибок, но ошибаются на разных SMS. Это полный разбор двух состояний интерактива для печатной версии.

В обоих случаях верно распределены пять сообщений из шести. Одно число «пять верных» скрыло различие: при первом пороге читатель увидит рекламу, при втором может не увидеть сообщение знакомого.

> Обычное SMS длиной 51 символ фильтр отправил в спам по правилу «40 символов и больше → спам». Что произошло?
>
> - ○ Метка сообщения автоматически сменилась с «обычное» на «спам»
> - ○ Ложная тревога: обычное сообщение признали спамом
> - ○ Пропуск: спам оставили во входящих
> - ○ Верное решение: длинные сообщения по определению спам

Заметка

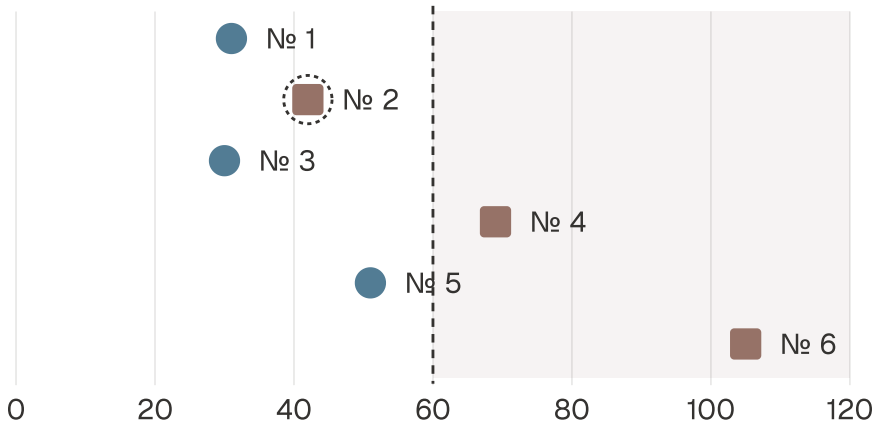
Метка в наборе служит ответом для проверки. Само пороговое правило её не читает: иначе получился бы ответ по подсказке. В реальной работе разметка тоже может ошибаться; здесь мы принимаем опубликованные метки как эталон упражнения.

4. Двигаем порог и выбираем последствия

Перед запуском предскажите результат: если поднять порог с 60 до 80, какое сообщение сменит папку? Затем проверьте себя. Квадрат обозначает спам по метке набора, круг обозначает обычное сообщение. Кольцо выделяет ошибку фильтра; нажатие на номер открывает оригинал и перевод.

Какие SMS сменят папку?

Порог 60



Длина в символах. Круг: обычное по метке; квадрат: спам по метке. Кольцо: ошибка. Справа от линии и на ней → в спам. Высота точки нужна только для разделения сообщений.

Порог 60: ложных тревог 0, пропусков 1. Верно 5 из 6.

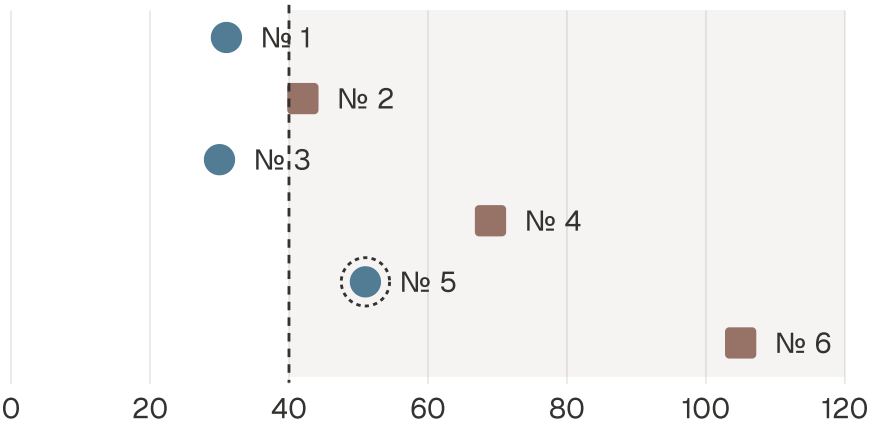
SMS 2 · 42 символов · строка источника 5384

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

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

Метка: спам. Правило: во входящие. Пропуск.

Порог 40



Длина в символах. Круг: обычное по метке; квадрат: спам по метке. Кольцо: ошибка. Справа от линии и на ней → в спам. Высота точки нужна только для разделения сообщений.

Порог 40: ложных тревог 1, пропусков 0. Верно 5 из 6.

SMS 2 · 42 символов · строка источника 5384

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

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

Метка: спам. Правило: в спам. Верное решение.

Справа от порога и на нём правило отправляет сообщения в спам. Форма показывает метку набора; кольцо и подпись показывают ошибку. Начальная настройка: порог 60, один пропуск, ни одной ложной тревоги.

Интерактивный эксперимент — в электронной версии урока.

При пороге 80 сообщение № 4 длиной 69 вернётся во входящие. Реклама № 2 уже была там. Получилось два пропуска. Обычные сообщения при таком сдвиге не могут впервые попасть в спам: мы только сузили область, которую фильтр считает подозрительной.

Сравните теперь 40 и 60. Для ящика с важными приглашениями владелец может предпочесть 60: он согласен вручную удалять короткую рекламу, чтобы не прятать обычное сообщение. Если каждое подозрительное SMS отправляют сотруднику на проверку, порог 40 может быть удобнее: сотрудник увидит весь спам из наших примеров и проверит одно лишнее сообщение.

Пример

«Ошибка дороже» не обязательно значит денежную сумму. Это может быть пропущенное приглашение, лишняя минута проверки или риск поверить мошеннику. Для выбора порога сначала обсуждают действие после решения фильтра.

Фильтр отправляет в спам сообщения длиной не меньше порога, остальные — во входящие. Порог снизили; сообщения и их настоящие метки не меняли. Что можно обещать?

- ○ Ложных тревог станет меньше, поскольку порог стал меньше
- ○ Общее число ошибок обязательно уменьшится
- ○ Все обычные сообщения останутся во входящих
- ○ Уже пойманный спам останется пойманным; ложных тревог может стать больше

5. А что будет с новым сообщением?

Мы назначали пороги вручную и сразу видели ошибки. Обучения сложной модели здесь не было. Даже очень удачная настройка на этих шести строках ещё не обещает удачи на остальных SMS.

Зафиксируйте порог 60. Представьте новое сообщение длиной 90 символов. Правило отправит его в спам, даже если это подробные указания от друга. Мы можем уверенно предсказать действие правила; правильность действия узнаем только после проверки содержания или метки.

На наших шести строках даже идеальный выбор порога не уберёт все ошибки: спам длиной 42 расположен левее обычного сообщения длиной 51. Чтобы поймать первый, правило вынуждено захватить второе. Одного сдвига границы недостаточно. Следующий шаг разработки: попробовать другие признаки, например слова сообщения, а затем проверить новую версию на отложенных примерах.

История

Набор SMS Spam Collection опубликован в UCI в 2012 году и объединяет сообщения из нескольких источников. Наш эксперимент показывает механизм на исторических примерах; о качестве фильтра для сегодняшней переписки по нему судить нельзя. Происхождение описано в паспорте UCI.

Фильтр отправляет в спам SMS длиной 60 символов и больше. Новое сообщение длиной 90 символов ушло в спам; его настоящая метка неизвестна. Какой вывод уже можно сделать?

- ○ Порог 60 подходит для любой переписки
- ○ Сообщение действительно рекламное
- ○ Правило выполнено; чтобы оценить правильность, нужна проверка сообщения
- ○ Фильтр обучился на новых данных

Проверка за пределами шести примеров

Оставим порог 60 и возьмём ещё два настоящих SMS из той же коллекции. Здесь указан смысл по-русски; длины посчитаны по английским оригиналам. До раскрытия меток решите, куда фильтр положит каждое сообщение.

Строка источника	Смысл сообщения	Символов
81	Извини, перезвоню позже	22

Строка источника	Смысл сообщения	Символов
5040	Спасибо за разговор в субботу. Ты мне дорог, я ценю тебя как брата и пример для себя	133

Проверка двух новых решений
Оба SMS помечены в источнике как обычные. Первое останется во входящих — верно. Второе уйдёт в спам — ложная тревога. Оно длинное, потому что человек подробно благодарит близкого. Эти два сообщения не входили в шесть строк, на которых мы сравнивали пороги. Двух примеров мало для общей оценки, но достаточно, чтобы увидеть ещё одну конкретную ошибку.

Есть и хитрый способ получить много верных ответов: вообще ничего не отправлять в спам. В полном файле 4827 обычных сообщений и 747 спам-сообщений. Такое правило даст **4827 верных ответов из 5574**, примерно 87 %, — и не поймает ни одного спама. Большая доля верных ответов ещё не означает, что фильтр выполняет свою задачу.

Профи. Посчитать ошибки и подобрать порог кодом

Обозначим длину через x, порог через t, спам числом 1, обычное сообщение числом 0. Правило: $f_t(x) = 1$ при $x \ge t$ и $f_t(x) = 0$ иначе.

Четыре результата часто записывают коротко: TP — спам найден, FN — спам пропущен, FP — ложная тревога, TN — обычное сообщение оставлено во входящих. При пороге 60 получаем TP = 2, FN = 1, FP = 0, TN = 3.

$$\text{доля верных} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{5}{6}.$$

Если цена ложной тревоги c_{FP}, а пропуска c_{FN}, суммарная цена ошибок равна $c_{FP} FP + c_{FN} FN$. Эти цены задают люди; измеренные длины их не определяют.

Поймайте весь спам из шести примеров

Цель: Измените порог: добейтесь нуля пропусков и ровно одной ложной тревоги

```
lengths = [31, 42, 30, 69, 51, 105]
labels = [0, 1, 0, 1, 0, 1]
threshold = 60  # TODO: измените порог

def errors(t):
    predictions = [int(x >= t) for x in lengths]
    fp = sum(p == 1 and y == 0 for p, y in zip(predictions, labels))
    fn = sum(p == 0 and y == 1 for p, y in zip(predictions, labels))
    return fp, fn

fp, fn = errors(threshold)
print("Ложные тревоги:", fp, "Пропуски:", fn)
```

Эпик. **Сколько решений умеет одно пороговое правило?**

Пусть у $n \geq 2$ объектов попарно различные значения признака. Расположим их слева направо. Порог с фиксированным направлением «справа класс 1» может дать только $n + 1$ наборов ответов: всё единицы, затем один ноль слева, затем два нуля, вплоть до всех нулей. Всего же двоичных разметок 2^n.

Найдите разметку, которую такой порог не сможет выразить, и докажите это. Затем разрешите менять направление: считать классом 1 либо правую, либо левую часть оси. Сколько разных разметок получится теперь?

▾*Решение*

Возьмём две соседние точки: левая имеет метку 1, правая 0. Порог вида «справа класс 1», захватив левую точку, обязательно захватит и правую. Поэтому уже эта пара неразделима таким правилом.

При двух направлениях получается $2n$ разметок: два семейства по $n + 1$, но «все нули» и «все единицы» посчитаны дважды. Других совпадений нет: непостоянная разметка первого семейства начинается с нуля и заканчивается единицей, второго наоборот. При совпадающих значениях признака разметок может быть меньше. Подсчёт описывает возможности правила; гарантии на будущие данные из него не следует.

Где это встретится дальше

В маршруте первых четырёх уроков теперь переходим к уроку «Регрессия: предсказать число». Вместо категории ответом станет количество прокатов: сравнивать модели будем по величине промаха. К выбору класса вернёмся позже в уроке о ближайших соседях — там решение зависит от похожих примеров.

Правило: длина не меньше порога отправляет SMS в спам, меньшая длина оставляет во входящих.

Эксперимент: пороги 40 и 60 дали по пять верных решений, но ошиблись на разных сообщениях.

Ограничение: короткий спам и длинное обычное сообщение нельзя разделить таким порогом без ошибки.

Следующая проверка: зафиксировать настройку и оценить её на сообщениях, по которым порог не выбирали.

Домашняя работа

1. Возьмите учебный набор: обычные сообщения длиной 20, 55 и 90; спам длиной 45 и 80. Используйте правило «длина не меньше порога → спам». Сравните пороги 40 и 60 и объясните выбор для ящика с важными приглашениями.

2. Обычные SMS имеют длины 20, 55 и 90 символов, спам — 45 и 80. Фильтр отправляет в спам сообщения длиной 60 символов и больше, остальные — во входящие. Сколько спам-сообщений останется во входящих?

Источники

- SMS Spam Collection, UCI (https://archive.ics.uci.edu/dataset/228/sms+spam+collection): оригинальные сообщения, метки, происхождение и лицензия CC BY 4.0. Наши шесть примеров выбраны для объяснения, а не для оценки качества на всём наборе.
- Scikit-learn: настройка порога классификации (https://scikit-learn.org/stable/modules/classification_threshold.html): разделение предсказания и решения, выбор порога по последствиям ошибок.
- SpamAssassin: required_score (https://spamassassin.apache.org/full/3.4.x/doc/Mail_SpamAssassin_Conf.html): реальный пример решения по порогу баллов; наша учебная длина не имитирует баллы этого фильтра.