Наивный байесовский классификатор

В первом задании вы будете создавать наивный байесовский классифи катор. Нашей целью будет написать функцию, принимающую email (как строку) в качестве параметра и классифицирующую её как либо спам, ли бо не спам. В её основе будет лежать модель, которая *обучается* на наборе образцов.

1 Подготовка email-ов

Обработка текста зависит от конкретной задачи. Для построения байесов ского классификатора мы будем обращаться с ними как просто с набором слов в нижнем регистре. Т. е. каждый текст должен быть переведён в мно жество слов (порядок слов не имеет значения) так, чтобы:

1. Любые знаки пунктуации были просто проигнорированы;
2. Все слова были бы переведены в нижний регистр;
3. Разумеется, дублирующиеся слова учитывались бы только один раз.

Например, письмо с текстом “Купите наш товар!” представляется как мно жество: {купите, наш, товар}.

2 Правило принятия решения

Для принятия решения нам потребуется научиться вычислять следующие величины:

$$P(\text{спам} \mid \{\text{купите, наш, товар}\}) \text{ и } P(\text{не спам} \mid \{\text{купите, наш, товар}\}).$$

Здесь под $P(\text{спам} \mid \{\text{купите, наш, товар}\})$ мы понимаем вероятность собы тия “письмо является спамом”, если событие “в письме встречаются слова купите, наш, товар” выполняется. Такой способ записи событий (словами в фигурных скобках) является неформальным, но удобным для изложения.

Сумма этих вероятностей равна 1 (убедитесь, что понимаете почему) и мы будем предсказывать спам, если

$$P(\text{спам} \mid \{\text{купите, наш, товар}\}) > 0.5,$$

и не спам при меньшем значении. В случае ничьей вы можете сделать про извольный выбор.

2.1 Но как их вычислять?

Вычислить эти значения нам поможет *теорема Байеса*:

$$P(\text{спам} \mid \{\text{купите, наш, товар}\}) = \frac{P(\{\text{купите, наш, товар}\} \mid \text{спам})}{P(\{\text{купите, наш, товар}\})}$$

$$= P(\{\text{купите, наш, товар}\} \mid \text{спам})$$

$$P(\{\text{купите, наш, товар}\} \mid \text{спам})P(\text{спам}) + P(\{\text{купите, наш, товар}\} \mid \text{не спам})P(\text{не спам})\cdot$$

Формула выглядит сложнее, но входящие в неё вероятности проще вы числить. Так вероятность спама можно оценить отношением количества спамовых писем к количеству вообще всех писем:

$$P(\text{спам}) = \frac{\#\text{ спам–писем}}{\#\text{ всех писем}}\cdot$$

И наоборот:

$$\text{писем } \#\text{ всех писем}\cdot$$

$$P(\text{не спам}) = \#\text{ обычных}$$

Можно попробовать аналогично вычислять и следующие вероятности: $P(\{\text{купите, наш, товар}\} \mid \text{спам})$

$$= \#\text{ спам–писем со словами “купите”, “наш”, “товар”}$$

$$\#\text{ спам–писем,}$$

но в случае реальных писем, вероятность того, что все слова некоторого письма встретятся в каком-то другом письме, может оказаться слишком мала, поэтому такое определение оказывается неудачным.

2.1.1 Условная независимость слов

Чтобы вычислять вероятность появления набора слов, используется *наив ное предположение* об их *условной независимости*. Это означает, что если известен класс письма Y , слова в нём считаются независимыми друг от друга. Формально: события $A_1, \ldots, A_n$ называются условно независимыми при условии B, если

$$P(A_1 \ldots A_n \mid B) = \prod_{k=1}^{n} P(A_k \mid B).$$

Применяя это предположение к нашей задаче, получаем, что для любого класса Y :

$$P(\{w_1, w_2, \ldots, w_n\} \mid Y) \approx \prod_{k=1}^{n} P(w_k \mid Y).$$

Например, для набора {купите, наш, товар} имеем:

$$P(\{\text{купите, наш, товар}\} \mid \text{спам}) = P(\text{купите} \mid \text{спам})P(\text{наш} \mid \text{спам})P(\text{товар} \mid \text{спам}).$$

Таким образом, вместо маловероятного события совместного появления всех слов мы оперируем произведением более простых вероятностей для отдельных слов.

2.1.2 Сглаживание Лапласа

Но так как может оказаться даже, что отдельное слово больше нигде не встречалось, то следует пойти на шаг дальше и применить так называемое *сглаживание Лапласа*, при котором для каждого слова мы псчитаем, что к корпусу[1] добавлены два искусственных письма: одно со словом, другое без него. Т. е. финальная формула для вычисления такой вероятности выглядит так:

$$P(\text{купите} \mid \text{спам}) = \frac{\#(\text{спам–писем со словом “купите”}) + 1}{\#(\text{спам–писем}) + 2}.$$

2.1.3 Как избегать погрешностей

Пусть исследуемое письмо состоит из n слов $w_1, w_2, \ldots, w_n$. Тогда, согласно нашему определению, вероятность того, что это письмо спам, равна:

$$P(\text{спам} \mid \{w_1, w_2, \ldots, w_n\}) = \frac{P(\text{спам}) \prod_{k=1}^{n} P(w_k \mid \text{спам})}{P(\text{спам}) \prod_{k=1}^{n} P(w_k \mid \text{спам}) + P(\text{не спам}) \prod_{k=1}^{n} P(w_k \mid \text{не спам})}.$$

Но в результате произведения вероятностей мы можем получить очень маленькие числа в числителе и знаменателе, приводя нас к неправильному ответу. Так как числитель входит в знаменатель, достаточно проверить неравенство (убедитесь, что понимаете почему):

$$P(\text{спам}) \prod_{k=1}^{n} P(w_k \mid \text{спам}) > P(\text{не спам}) \prod_{k=1}^{n} P(w_k \mid \text{не спам}).$$

После чего потери точности в результате перемножения вероятностей можно избежать, просто взяв логарифм:

$$\log P(\text{спам}) + \sum_{k=1}^{n} \log P(w_k \mid \text{спам}) > \log P(\text{не спам}) + \sum_{k=1}^{n} \log P(w_k \mid \text{не спам}).$$

2.1.4 Почему 80/20 — это эвристика

Разделение датасета на обучающую и тестовую части в пропорции 80/20 — распространённая *практическая* эвристика, а не теоретическое правило. Наша цель здесь оставить достаточно данных для оценки параметров и при этом иметь независимую выборку для честной оценки качества. На маленьких наборах лучше использовать перекрёстную проверку (k-fold CV), например $k = 5$ или $k = 10$: данные делятся на k равных частей, обучение проводится k раз, каждый раз откладывая свою часть для теста, а метрика усредняется.

[1]Корпус — это совокупность всех текстов или писем, собранных для обучения и анализа модели.

3 Где брать письма для обучения и проверки классификатора

В качестве возможных источников для обучающего набора писем предлагаются базы данных Enron Spam или PU Corpora, но вы можете использовать любой другой источник для обучающего набора писем. В крайнем случае письма можно сгенерировать самостоятельно.

Рекомендуется использовать 80% писем для обучения модели (для вычисления описанных выше вероятностей) и проверять её работу на оставшихся 20%. Для оценки качества удобно начать с простой метрики точности:

$$\text{точность} = \frac{\#\text{ верно классифицированных писем}}{\#\text{ всех писем}}.$$

Однако одной точности недостаточно, так как она не показывает, какие ошибки чаще допускает модель. Поэтому дополнительно вычисляют *чувствительность* (sensitivity, полнота) и *специфичность* (specificity):

$$\text{чувствительность} = \frac{\#\text{ спамовых писем, верно классифицированных как спам}}{\#\text{ всех спамовых писем}},$$

$$\text{специфичность} = \frac{\#\text{ обычных писем, верно классифицированных как не спам}}{\#\text{ всех обычных писем}}.$$

Чувствительность показывает, насколько хорошо классификатор распо знаёт спам, а специфичность — насколько надёжно он не помечает обычные письма как спам.
Полезные ссылки:

• https://www2.aueb.gr/users/ion/data/enron-spam/

• http://www.aueb.gr/users/ion/data/PU123ACorpora.tar.gz 4