Наивный байесовский классификатор

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

1 Подготовка email-ов

Как именно стоит работать с текстом зависит от каждой конкретной задачи. Для построения байесовского классификатора мы будем обращаться с ними как просто с набором слов в нижнем регистре. Т.е. каждый текст должен быть переведён в тип данных `set` 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!" должно быть представлено в виде следующего словаря: `{'купите', 'наш', 'товар'}`.

2 Правило принятия решения

Для принятия решения нам потребуется научиться вычислять следующие величины:

$$\mathbb{P}(\text{спам}|\{'\text{купите}','\text{наш}','\text{товар}'\}) \quad \text{и} \quad \mathbb{P}(\text{не спам}|\{'\text{купите}','\text{наш}','\text{товар}'\}).$$

Сумма обоих величин должна быть равна 1 (подумайте почему это так) и мы будем предсказывать **спам**, если $\mathbb{P}(\text{спам}|\{'\text{купите}','\text{наш}','\text{товар}'\}) > 0.5$ и **не спам** иначе. В случае ничьей вы можете сделать произвольный выбор.

2.1 Но как их вычислять?

Вычислить эти значения нам поможет *теорема Байеса*:

$$\mathbb{P}(\text{спам}|\{'\text{купите}','\text{наш}','\text{товар}'\}) = \frac{\mathbb{P}(\{'\text{купите}','\text{наш}','\text{товар}'\}|\text{спам})}{\mathbb{P}(\{'\text{купите}','\text{наш}','\text{товар}'\})}$$

$$= \frac{\mathbb{P}(\{'\text{купите}','\text{наш}','\text{товар}'\}|\text{спам})}{\mathbb{P}(\{'\text{купите}','\text{наш}','\text{товар}'\}|\text{спам})\mathbb{P}(\text{спам}) + \mathbb{P}(\{'\text{купите}','\text{наш}','\text{товар}'\}|\text{не спам})\mathbb{P}(\text{не спам})}.$$

Кажется что формула стала только хуже, но участвующие в ней вероятности гораздо проще находить. Так вероятность спама можно оценить отношением количества спамовых писем к количеству вообще всех писем:

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

И наоборот:

$$\mathbb{P}(\text{не спам}) = \frac{\#\text{ обычных писем}}{\#\text{ всех писем}}.$$

Можно попробовать аналогично вычислять и следующие вероятности:

$$\mathbb{P}(\{'\text{купите}','\text{наш}','\text{товар}'\}|\text{спам}) = \frac{\#\text{ спам–писем со словами 'купите', 'наш', 'товар'}}{\#\text{ спам–писем}},$$

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.

Поэтому мы выдвигаем *наивное предположение* о том, что слова *условно независимы* при условии метки **спам** или **не спам**:

$$\mathbb{P}(\{'\text{купите}','\text{наш}','\text{товар}'\}|\text{спам})$$

$$= \mathbb{P}(\{'\text{купите}'|\text{спам})\mathbb{P}(\{'\text{наш}'|\text{спам})\mathbb{P}(\{'\text{товар}'|\text{спам}),$$

$$\mathbb{P}(\{'\text{купите}',' \text{наш}',' \text{товар}'\}|\text{не спам})$$

$$= \mathbb{P}(\{'\text{купите}'|\text{не спам})\mathbb{P}(\{'\text{наш}'|\text{не спам})\mathbb{P}(\{'\text{товар}'|\text{не спам}).$$

Таким образом от вас требуется отдельно оценить вероятности вида

$$\mathbb{P}(\{'\text{купите}'\}|\text{спам}) = \frac{\#\text{ спам–писем со словом 'купите'}}{\#\text{ спам–писем}}$$

и затем перемножить их для всех входящих в классифицируемое письмо слов.

2.1.1 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 мы для каждого слова мы притворяемся, что у нас есть ещё два письма: одно содержащее это слово и одно без такого слова. Т.е. финальная формула для вычисления такой вероятности выглядит так:

$$\mathbb{P}(\{'\text{купите}'\}|\text{спам}) = \frac{\#\text{ спам–писем со словом 'купите'} + 1}{\#\text{ спам–писем} + 2}.$$

2.1.2 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:

$$\mathbb{P}(\text{спам}|\{'w_1, w_2, \ldots, w_n\}) =$$

$$= \frac{\mathbb{P}(\text{спам})\prod_{k=1}^{n}\mathbb{P}(w_k|\text{спам})}{\mathbb{P}(\text{спам})\prod_{k=1}^{n}\mathbb{P}(w_k|\text{спам}) + \mathbb{P}(\text{не спам})\prod_{k=1}^{n}\mathbb{P}(w_k|\text{не спам})}.$$

Но в результате произведения вероятностей мы можем получить очень маленькие числа в числителе и знаменателе, приводя нас к неправильному ответу. Но так как числитель является одним из слагаемых знаменателя, то чтобы утверждать что некоторое письмо - **спам** нам достаточно проверить следующее неравенство (убедитесь что понимаете почему):

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

После чего потери погрешности в результате перемножения вероятностей можно избежать просто взятием логарифма:

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

3 Где брать письма для обучения и проверки классификатора

В качестве возможных источников для обучающего набора писем предлагаются базы данных Enrom Spam или PU Corpora, но вы можете использовать любой другой источник для обучающего набора писем. На крайний случай вы можете *сгенерировать* все письма.

Рекомендуется использовать 80% ваших писем для *обучения* модели, т.е. для вычисления описанных выше вероятностей и проверять её точность на оставшихся 20%. Используйте самою простую меру точности:

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