

# Языковые модели

## 1. Предсказание следующего слова

Одной из задач обработки текстов на естественном языке выступает задача предсказания следующего слова по предшествующим. Примерами таких задач являются задача предсказания следующего слова при наборе текста в смартфоне или дополнение поисковых запросов.

Данная задача может быть сведена к оценке вероятностей встретить каждое из возможных слов после имеющегося. Соответствующая языковая модель примет вид:

$$w^* = \operatorname{argmax}_{w_k \in V} P(w_k | w_{k-1}), \quad (1)$$

$$P(w_k | w_{k-1}) = \frac{P(w_k) P(w_{k-1} | w_k)}{P(w_{k-1})}, \quad (2)$$

где  $V$  – множество всех возможных слов, а  $P(w_k | w_{k-1})$  – вероятность встретить слово  $w_k$  после  $w_{k-1}$ .

Для удобства полученную модель можно представить в следующем виде:

$$w^* = \operatorname{argmax}_{w_k \in V} [\log P(w_k) + \log P(w_{k-1} | w_k)]. \quad (3)$$

Иногда по последнему слову оказывается невозможным определить следующее, поскольку в нем не учитывается контекст. Например, если последнее слово является союзом, то приемлемые варианты следующего слова определяет и предшествующее данному союзу слово. Оценка вероятности в данном случае проводится на основе  $m$  последних слов. Таким образом, модель (1)-(2) примет вид:

$$w^* = \operatorname{argmax}_{w_k \in V} P(w_k | w_{k-1}, w_{k-2}, \dots, w_{k-m}), \quad (4)$$

$$P(w_k | w_{k-1}, w_{k-2}, \dots, w_{k-m}) = ((P(w_k) P(w_{k-1}, w_{k-2}, \dots, w_{k-m} | w_k)) / P(w_{k-1}, w_{k-2}, \dots, w_{k-m})). \quad (5)$$

Предположим, что текущее слово  $w_k$  зависит только от того, какие слова встретились перед ним и не зависит от того, в каком порядке они встретились. Тогда

$$w^* = \operatorname{argmax}_{w_k \in V} \left[ P(w_k) \prod_{i=1}^m P(w_{k-i} | w_k) \right] = \operatorname{argmax}_{w_k \in V} \left[ \log P(w_k) + \sum_{i=1}^m \log P(w_{k-i} | w_k) \right]. \quad (6)$$

Полученная языковая модель также используется для решения задачи классификации документов.

### 1.1. Наивный байесовский классификатор

Пусть стоит задача определить категорию новостей по их тексту. Тогда оценки вероятностей принадлежности новости к каждой  $c \in C$  категории могут быть найдены по формуле:

$$c^* = \operatorname{argmax}_{c \in C} \frac{P(c) P(d | c)}{P(d)}. \quad (7)$$

Здесь  $P(c)$  – вероятность встретить новость данной категории, рассчитываемая по формуле:

$$P(c) = \frac{N_c}{N}, \quad (8)$$

где  $N_c$  – количество новостей в категории  $c$ ,  $N$  – общее число новостей;

$$P(d | c) = P(w_1 | c) P(w_2 | c) \dots P(w_m | c) = \prod_{i=1}^m P(w_i | c), \quad (9)$$

где  $w_1, \dots, w_m$  – слова, встретившиеся в новости  $d$  (в предположении, что все слова независимы), а

$$P(w_i | c) = \frac{v_{ic} + 1}{\sum_{i' \in V} (v_{i'c} + 1)} = \frac{v_{ic} + 1}{|V| + \sum_{i' \in V} v_{i'c}} \quad (10)$$

после применения сглаживания Лапласа для устранения проблемы неизвестных слов;

$P(d)$  – оценка вероятности встретить новость, состоящую из данного набора слов. Поскольку  $P(d)$  не оказывает влияния на результат, то итоговая модель классификации может быть представлена в виде:

$$c^* = \operatorname{argmax}_{c \in C} \left[ \log P(c) + \sum_{i=1}^m \log P(w_i | c) \right]. \quad (11)$$

## 1.2. Оценка условных вероятностей в языковой модели на основе словаря

Условные вероятности в формуле (3) могут быть оценены непосредственно по имеющемуся словарю слов. Для этого необходимо рассмотреть имеющиеся в словаре биграммы, тогда:

$$P(w_{k-1} | w_k) = \frac{v_{(w_{k-1}, w_k)}}{\sum_{(w_i, w_j) \in V} v_{(w_i, w_j)}}, \quad (12)$$

где  $v_{(w_{k-1}, w_k)}$  – частота встречаемости словосочетания  $(w_{k-1}, w_k)$ .

## 1.3. Подход на основе машинного обучения

Задачу предсказания следующего слова по предыдущим можно представить как задачу многоклассовой классификации, где  $y = 1$ , если данное слово следует за предыдущим, и  $y = 0$  в противном случае. Обучающую выборку в данном легко получить с помощью скользящего окна.

Рассмотрим строку «*Success is the ability to go from failure to failure without losing your enthusiasm*» и пройдем по ней окном размера 3. Таким образом наберется обучающая выборка, где первые два слова являются признаками, а последний столбец – целевой переменной.

Success	is	the
is	the	ability
the	ability	to
ability	to	go
to	go	from
go	from	failure
from	failure	to
failure	to	failure
to	failure	without
failure	without	losing
without	losing	your
losing	your	enthusiasm

Представить слова обучающей выборки в виде вектора можно с помощью метода One-hot-encoding. Число уникальных слов в представленной строке равно 12. Тогда для первого объекта на вход модели будет подаваться следующий вектор:

$$\{1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\},$$

а соответствующее значение целевой переменной:

{0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0}.

Таким образом проводится оценка вероятностей каждого возможного варианта. Для эффективного обучения такой модели требуется большой объем данных.