Лекция 5. Классификация текстовых документов

Курс «Методы машинного обучения»
С.Ю. Папулин (papulin.study@yandex.ru)

СОДЕРЖАНИЕ

1. Преобразование текстовых документов в числовой вектор

- Вхождение слова:

$$\mathrm{to}(t,d) = \begin{cases} 1, \text{если } t \text{ в документе } d \\ 0 \text{ иначе} \end{cases}$$

- Частота слова (Term Frequency – TF):

$$\mathrm{tf}(t,d) = n_t,$$

где t – терм; d – документ; n_t – количество термов t в документе d.

Нормализованная частота слов:

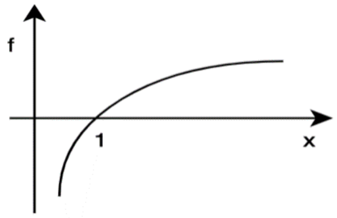
$$\mathrm{tf}_N(t,d) = \frac{n_t}{N_d},$$

где N_d – количество термов в документе d.

- Обратная частота документа (Inverse Document Frequency – IDF):

$$\mathrm{idf}(t,D) = \log\frac{N_D}{\mathrm{df}(t)},$$

где D – коллекция документов; N_D – количество документов в коллекции; $\mathrm{df}(t)$ – количество документов, в которых встречается терм t.



Сглаженная IDF:

$$\mathrm{idf}_s(t, D) = \log\frac{N_D + 1}{1 + \mathrm{df}(t)}$$

- TF-IDF:

$$\text{tf-idf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

Из множества документов D строится словарь термов V. $|V| = N$ – количество уникальных термов в коллекции документов D.

Пусть $d \in D$, то документ можно представить в виде бинарного вектора:

$$b(d) = (b_1, b_2, \dots, b_N),$$

где b_i есть $\mathrm{to}(t_i, d)$ для некоторого терма t_i, $b_i \in \{0,1\}$

Вектор частоты слов:

$$\mathrm{tf}(d) = (x_1, x_2, \dots, x_N),$$

где x_i есть $\mathrm{tf}(t_i, d)$ для некоторого терма t_i, $x_i \in \mathbb{R}_{0+}$

Вектор TF-IDF:

$$\text{tf-idf}(d, D) = (x_1, x_2, \dots, x_N),$$

где x_i есть $\text{tf-idf}(t_i, d, D)$ для некоторого терма t_i, $x_i \in \mathbb{R}_{0+}$

2. Наивный байесовский классификатор

2.1. Формула Байеса

Вероятность отнесения некоторого документа d классу с:

$$p(C = \mathrm{c}|D = d) = \frac{p(C = c) \cdot p(D = d|C = c)}{p(D = d)} = \frac{p(c) \cdot p(d|c)}{p(d)}.$$

Так как вероятность $P(d)$ одинакова для всех документов, то можно записать

$$p(\mathrm{c}|d) \propto p(c) \cdot p(d|c)$$

Задача найти максимальную вероятности из всех классов:

$$\hat{y} = \underset{c}{\operatorname{argmax}}\, p(\mathrm{c}|d)$$

2.2. Модель Бернулли

$$d \to b = (b_1, \dots, b_N),$$

где

$$b_i \in \{0,1\}.$$

Для одного документа:

$$p(d|c) \sim p(b|c) = \prod_{j=1}^{N} \left[p(t_j|c)^{b_j} \cdot \left(1 - p(t_j|c)\right)^{(1-b_j)} \right]$$

Обучение

Для коллекции документов:

$$p(c|D) \propto \prod_{i=1}^{N_D} p(c_i) p(d_i|c_i) = \prod_{i=1}^{N_D} p(c_i) \prod_{j=1}^{N} \left[p(t_i|c_i)^{b_{ij}} \cdot \left(1 - p(t_j|c_i)\right)^{(1-b_{ij})} \right]$$

где c_i – класс документа d_i.

Для оценки вероятностей $p(c)$ и $p(t_j|c)$ используется метод максимального правдоподобия (MLE).

Оценка параметров:

- Оценка вероятности встретить документ класса с:

$$\hat{p}(\mathrm{c}) = \frac{N_\mathrm{c}}{N_D},$$

где N_c – количество документов класса с.

- Оценка вероятности встретить терм t_j в документах класса с:

$$\hat{p}(t_j|\mathrm{c}) = \frac{\mathrm{df}_\mathrm{c}(t_j)}{N_\mathrm{c}}$$

где $\mathrm{df}_\mathrm{c}(t_j)$ – количество документов, содержащих терм t_j.

Чтобы избежать нулевых вероятностей:

$$\hat{p}_{smooth}(t_j|c) = \frac{\mathrm{df}_c(t_j) + 1}{N_c + 2}$$

Предсказание

Для некоторого нового документа d_*:

$$\begin{aligned}\hat{y} = \underset{c}{\operatorname{argmax}}\, p(\mathrm{c}|d_*) = \underset{c}{\operatorname{argmax}}\, p(c)p(d_*|c) \\ = \underset{c}{\operatorname{argmax}} \left[\hat{p}(c) \cdot \prod_{j=1}^{N} \left[\hat{p}(t_j|c)^{b_{*j}} \cdot \left(1 - \hat{p}(t_j|c)\right)^{(1-b_{*j})}\right]\right]\end{aligned}$$

2.3. Мультиноминальная модель

$$d \to x = (x_1, \dots, x_N),$$

где

$$x_i \in \mathbb{N}_{0+} \text{ или } \mathbb{R}_{0+}$$

Для одного документа:

$$p(d|c) = \prod_{j=1}^{N} p(t_j|c)^{x_j}$$

Обучение

Для коллекции документов:

$$p(c|D) \propto \prod_{i=1}^{N_D} p(c_i)p(d_i|c_i) = \prod_{i=1}^{N_D} p(c_i) \prod_{j=1}^{N} p(t_j|c_i)^{x_{ij}}$$

Для оценки вероятностей $p(c)$ и $p(t_j|c)$ используется метод максимального правдоподобия (MLE).

Оценка параметров:

- Оценка вероятности встретить документ класса с:

$$\hat{p}(\mathrm{c}) = \frac{N_{\mathrm{c}}}{N_D},$$

где N_{c} – количество документов класса с.

- Оценка вероятности встретить терм t_j в документах класса с:

$$\hat{p}\left(t_j|\mathrm{c}\right) = \frac{\mathrm{tf}_c(t_j)}{\sum_{k=1}^{N}\mathrm{tf}_c(t_k)} = \frac{\mathrm{tf}_c(t_j)}{n_c}$$

где $\mathrm{tf}_{\mathrm{c}}\left(t_j\right)$ – количество терма t_j в документах класса с; n_c – количество термов в документах класса с.

Чтобы избежать нулевых вероятностей:

$$\hat{p}_{smooth}^{\alpha=1}(t_i|c) = \frac{1 + \mathrm{tf}_c(t_i)}{N + \sum_{k=1}^{N}\mathrm{tf}_c(t_k)}$$

$$\hat{p}_{smooth}^{\alpha}(t_i|c) = \frac{\alpha + \mathrm{tf}_c(t_i)}{\alpha N + \sum_{k=1}^{N}\mathrm{tf}_c(t_k)}$$

Предсказание

Для некоторого нового документа d_*:

$$\hat{y} = \underset{c}{\mathrm{argmax}}\, p(\mathrm{c}|d_*) = \underset{c}{\mathrm{argmax}}\, p(c)p(d_*|c) = \underset{c}{\mathrm{argmax}}\left[\hat{p}(c)\cdot\prod_{j=1}^{N}\hat{p}\left(t_j|c\right)^{x_{*j}}\right]$$