Теоретическая часть

1. Наивный байесовский классификатор, модель Бернулли

Для каждого документа d_i вектор $B_i = \langle B_{i_t} \rangle = \langle B_{i_1}, \dots B_{i_t}, \dots B_{i_{|V|}} \rangle$ показывает встречается ли терм t из словаря V в документе d_i.

a) Вероятность встретить i-е слово в документе j-го класса – число документов с этим словом к числу всех документов класса:

$$P(v_t|c_j) = \frac{\sum_{d_i : d_i \in c_j} B_{i_t}}{\left|\{d_i : d_i \in c_j\}\right|} = \frac{\sum_{i=1}^{|D|} B_{i_t} P(c_j|d_i)}{\sum_{i=1}^{|D|} P(c_j|d_i)}, \text{где } P(c_j|d_i) = \{0, 1\}$$

Учитывая сглаживание Лапласа, получим:

$$P(v_t|c_j) = \frac{1 + \sum_{d_i : d_i \in c_j} B_{i_t}}{2 + \left|\{d_i : d_i \in c_j\}\right|} = \frac{1 + \sum_{i=1}^{|D|} B_{i_t} P(c_j|d_i)}{2 + \sum_{i=1}^{|D|} P(c_j|d_i)}, \text{где } P(c_j|d_i) = \{0, 1\}$$

b) Вероятность сгенерировать документ = подбрасывание монет, соответствующих словам – вычислим ее из функции распределения Бернулли:

$$P(d_i|c_j) = \prod_{t=1}^{|V|} P(v_t|c_j)^{B_{i_t}} \left(1 - P(v_t|c_j)\right)^{(1 - B_{i_t})} = \prod_{t=1}^{|V|} B_{i_t} P(v_t|c_j) + (1 - B_{i_t}) \left(1 - P(v_t|c_j)\right)$$

c) Для вывода вероятности принадлежности документа к классу воспользуемся формулой Байеса:

$$P(c_j|d_i) = \frac{P(c_j) P(d_i|c_j)}{P(d_i)} = \left\{ \begin{matrix} P(d_i)\text{одинаково} \\ \text{для всех классов} \end{matrix} \right\} = P(c_j) P(d_i|c_j)$$

d) В качестве предсказанного класса выберем самый вероятный:

$$\text{c} = argmax_{c \in C} P(c|d) = argmax_{c \in C} P(c) P(d|c), \text{где } P(c) = \frac{N_c}{N} - \text{априорная вероятность класса } c, N_c - \text{число документов класса } c, N - \text{число всех документов}$$

2. Мультиномиальный наивный байесовский классификатор

Для каждого документа d_i вектор $N_i = \langle N_{i_t} \rangle = \langle N_{i_1}, \dots N_{i_t}, \dots N_{i_{|V|}} \rangle$ показывает число вхождений терма t из словаря V в документ d_i.

a) Вероятность встретить i-е слово в документе j-го класса – число вхождений этого слова в класс к числу вхождений всех слов в класс:

$$P(v_t|c_j) = \frac{\sum_{d_i:d_i\in c_j} N_{i_t}}{\sum_{d_i:d_i\in c_j} |N_i|} = \frac{\sum_{i=1}^{|D|} B_{it} P(c_j|d_i)}{\sum_{i=1}^{|D|} P(c_j|d_i)}, \text{где } P(c_j|d_i) = \{0, 1\}$$

Учитывая сглаживание Лапласа, получим:

$$P(v_t|c_j) = \frac{1 + \sum_{d_i:d_i\in c_j} N_{i_t}}{\sum_{d_i:d_i\in c_j} 1 + |N_i|} = \frac{1 + \sum_{d_i:d_i\in c_j} N_{i_t}}{|V| + \sum_{d_i:d_i\in c_j} |N_i|}$$

$$= \frac{1 + \sum_{i=1}^{|D|} B_{it} P(c_j|d_i)}{|V| + \sum_{i=1}^{|D|} P(c_j|d_i)}, \text{где } P(c_j|d_i) = \{0, 1\}$$

b) Вероятность сгенерировать документ = подбрасывание кубика с гранями, соответствующими словам – вычислим ее из функции мультиномиального распределения:

$$P(d_i|c_j) = \frac{\left(\sum_{t=1}^{|V|} N_{i_t}\right)!}{\prod_{t=1}^{|V|} N_{it}!} \prod_{t=1}^{|V|} P(w_t|c_j)^{N_{it}} = \frac{|N_i|!}{\prod_{t=1}^{|V|} N_{it}!} \prod_{t=1}^{|V|} P(w_t|c_j)^{N_{it}}$$

c) Для вывода вероятности принадлежности документа к классу перейдем в логарифмическое пространство:

$$\log\left(P(c_j|d_i)\right) = \log\left(P(c_j)P(d_i|c_j)\right) = \log(P(c_j) + \log\left(P(d_i|c_j)\right)$$

$$= \log(P(c_j) + \sum_{t=1}^{|V|} N_{it} \log(P(w_t|c_j))$$

d) В качестве предсказанного класса выберем самый вероятный:

$$\text{c} = argmax_{c\in C} \log(P(c|d))$$

$$= argmax_{c\in C} \log(P(c)) + \log(P(d|c))$$