

Мультиномиальный Байесовский классификатор.

Вероятность принадлежности к классу рассчитывается как

$$\hat{P}(c) = \frac{N_c + 1}{N + |c|}$$

где N_c - число документов в классе c , N - общее число документов, $|c|$ - общее число классов.

Условная вероятность слова в документе —

$$\hat{P}(w|c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|}$$

где $\text{count}(c)$ - число слов в документах класса c , $\text{count}(w, c)$ - число слова w в документах класса c , $|V|$ - мощность словаря по всем документам.

Тогда вероятность того, что документ d относится к классу c , равна

$$\hat{P}(c|d) \propto \hat{P}(c) \cdot \prod_{i=1}^k \hat{P}(w_i|c)$$

Рассмотрим 2 простых документа:

"собака зверь собака зверь"

"машина машина машина зверь"

У первого пусть будет тег "животные", у второго - "авто". Поучимся на них. Рассчитаем вероятности:

$$p(\text{зверь}|\text{"животные"}) = \frac{2+1}{4+3} = \frac{3}{7} = 0.42857$$

$$p(\text{зверь}|\text{"авто"}) = \frac{1+1}{4+3} = \frac{2}{7} = 0.2857$$

$$p(\text{машина}|\text{"животные"}) = \frac{0+1}{4+3} = \frac{1}{7} = 0.142857$$

$$p(\text{машина}|\text{"авто"}) = \frac{3+1}{4+3} = \frac{4}{7} = 0.5714$$

$$p(\text{собака}|\text{"животные"}) = \frac{2+1}{4+3} = \frac{3}{7} = 0.42857$$

$$p(\text{собака}|\text{"авто"}) = \frac{0+1}{4+3} = \frac{1}{7} = 0.142857$$

$$p(\text{"авто"}) = \frac{1}{2} = 0.5$$

$$p(\text{"животные"}) = \frac{1}{2} = 0.5$$

Теперь рассмотрим текст "зверь". К какому классу он принадлежит?

Посчитаем $p(\text{"животные"}|\text{"зверь"}) = p(\text{"животные"}) \cdot p(\text{"зверь"}|\text{"животные"}) = 0.5 \cdot 0.42857 = 0.21428571428$

$p(\text{"авто"}|\text{"зверь"}) = p(\text{"авто"}) \cdot p(\text{"зверь"}|\text{"авто"}) = 0.5 \cdot 0.2857 = 0.14285714285$

Нормируем вероятности:

$$p(\text{"животные"}|\text{"зверь"}) = \frac{0.21428571428}{0.21428571428+0.14285714285} = 0.6$$

$$p(\text{"авто"}|\text{"зверь"}) = \frac{0.14285714285}{0.21428571428+0.14285714285} = 0.4$$

Теперь рассмотрим текст "зверь зверь". Снова посчитаем вероятности:

$$p(\text{"животные"}|\text{"зверь зверь"}) = p(\text{"животные"}) \cdot p(\text{"зверь"}|\text{"животные"})^2 = 0.5 \cdot 0.42857^2 = 0.09183673469$$

$$p(\text{"авто"}|\text{"зверь зверь"}) = p(\text{"авто"}) \cdot p(\text{"зверь"}|\text{"авто"})^2 = 0.5 \cdot 0.2857^2 = 0.04081632653$$

Нормируем вероятности:

$$p(\text{"животные"}|\text{"зверь зверь"}) = \frac{0.09183673469}{0.04081632653+0.09183673469} = 0.6923$$

$$p(\text{"авто"}|\text{"зверь зверь"}) = \frac{0.04081632653}{0.04081632653+0.09183673469} = 0.30769$$

Теперь рассмотрим текст "собака машина". Снова посчитаем вероятности:

$$p(\text{"животные"}|\text{"собака машина"}) = p(\text{"животные"}) \cdot p(\text{"собака"}|\text{"животные"}) \cdot$$

$$p(\text{"машина"}|\text{"животные"}) = 0.5 \cdot 0.42857^2 = 0.03061224489$$

$$p(\text{"авто"}|\text{"собака машина"}) = p(\text{"авто"}) \cdot p(\text{"собака"}|\text{"авто"}) \cdot p(\text{"машина"}|\text{"авто"}) = 0.5 \cdot 0.2857^2 = 0.04081632653$$

Нормируем вероятности:

$$p(\text{"животные"}|\text{"собака машина"}) = \frac{0.03061224489}{0.04081632653+0.03061224489} = 0.42857142851$$

$$p(\text{"авто"}|\text{"собака машина"}) = \frac{0.04081632653}{0.04081632653+0.03061224489} = 0.57142857148$$

Вроде пока всё сходится с Weka, надо потестить побольше.

Более полно про классификатор можно почитать в Andrew McCallum, Kamal Nigam: A Comparison of Event Models for Naive Bayes Text Classification. In: AAAI-98 Workshop on 'Learning for Text Categorization', 1998.