Naive Bayes

Naive Bayes to probabilistyczny algorytm klasyfikacji tekstu, oparty na regule Bayesa i założeniu niezależności cech. W filtrowaniu opartym na treści oblicza prawdopodobieństwo przynależności dokumentu do klasy na podstawie rozkładów słów (Wzór 8):

$$P(c_j|d_i;\hat{\theta}) = \frac{P(c_j|\hat{\theta})P(d_i|c_j;\hat{\theta})}{P(d_i|\hat{\theta})} \tag{1}$$

W modelu Bernoulliego dokumenty są reprezentowane binarnie (obecność słów), a prawdopodobieństwo obliczane jest jako (Wzór 9):

$$P(d_i|c_j;\theta) = \prod_{t=1}^{|V|} (B_{it}P(w_t|c_j;\theta) + (1 - B_{it})(1 - P(w_t|c_j;\theta))) \tag{2}$$

Estymacja $P(w_t|c_j;\theta)$ w tym modelu to (Wzór 10):

$$P(w_t|c_j;\theta) = \frac{1 + \sum_{i=1}^{|D|} B_{it}P(c_j|d_i)}{2 + \sum_{i=1}^{|D|} P(c_j|d_i)} \tag{3}$$

W modelu multinomialnym uwzględnia się częstotliwość słów (Wzór 11):

$$P(d_i|c_j;\theta) = P(|d_i|)\prod_{t=1}^{|d_i|} P(w_t|c_j;\theta)^{N_{it}} \tag{4}$$

Estymacja $P(w_t|c_j;\theta)$ to (Wzór 12):

$$P(w_t|c_j;\theta) = \frac{1 + \sum_{i=1}^{|D|} N_{it}P(c_j|d_i)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N_{is}P(c_j|d_i)} \tag{5}$$

Model multinomialny jest skuteczniejszy przy dużych słownikach. Mimo uproszczenia, Naive Bayes działa dobrze w systemach rekomendacyjnych, np. Syskill Webert [1].

Gdzie:
$P(c_j|d_i;\hat{\theta})$ – prawdopodobieństwo klasy c_j dla dokumentu d_i,
$P(c_j|\hat{\theta})$, $P(d_i|c_j;\hat{\theta})$, $P(d_i|\hat{\theta})$ – prawdopodobieństwa a priori i warunkowe,
B_{it} – obecność słowa w_t w d_i (0 lub 1),
$P(w_t|c_j;\theta)$ – prawdopodobieństwo słowa w_t w klasie c_j,
N_{it} – liczba wystąpień w_t w d_i,
$|V|$ – liczba słów w słowniku,
$|D|$ – liczba dokumentów,
$|d_i|$ – długość dokumentu d_i,
θ, $\hat{\theta}$ – parametry i ich estymacje.

References

[1] Źródło o Syskill Webert.