Naive Bayes Model

$P(\text{cause}, \text{Effect}_1, \text{Effect}_2, .., \text{Effect}_n)$

$= P(\text{Effect}_1 | \text{cause})\ P(\text{Effect}_2 | \text{cause}) \cdots$

$P(\text{Effect}_2 | \text{cause})\ P(\text{cause})$

$= P(\text{cause}) \prod_{i=1}^{n} P(\text{Effect}_i | \text{cause})$

{ assuming $\text{Effect}_1, \ldots \text{Effect}_n$ are conditionally independent given cause

Text classification using Naive Baye's

Given a text, decide which of a predefined set of classes or categories it belongs to.

Cause: category or class variable

Effect: certain key words

* $P(\text{class} | \text{words})$

determine the posterior probability over categories, select the one with the highest posterior probability.

Newspaper articles classes: Sports, Weather

Document	words	class
1	Rain, Manchester, Score, League	S
2	Rain, warning	W
3	warning, Season, Flood	W
4	Rain, Season, Heavy	W
5	Score, League, warning	?

Rain	kw1	warning	kw5
Manchester	kw2	Flood	kw6
Score	kw3	Heavy	kw7
League	kw4	Season	kw8

$P(S \mid kw3, kw4, kw5) = ?$

$P(W \mid kw3, kw4, kw5) = ?$

$$P(kw_i \mid c) = \frac{count(kw_i, c) + 1}{\sum_{w \in V} (count(kw, c) + 1)}$$

$$= \frac{count(kw_i, c) + 1}{\left(\sum_{w \in V} count(kw, c)\right) + |V|}$$

$$P(s) = \frac{\#\text{ documents with class } s}{\#\text{ total documents}}$$ (4)

$$= \frac{1}{4}$$

$$P(w) = \frac{3}{4}$$

		Count	
		Sports	Weather
Rain	KW1	1	2
Manchester	KW2	1	0
score	KW3	1	0
League	KW4	1	0
warning	KW5	0	2
Flood	KW6	0	1
Heavy	KW7	0	1
Season	KW8	0	2
Total		4	8

$$P(\text{score}|s) = \frac{1+1}{4+8} = \frac{2}{12}$$

$$P(\text{league}|s) = \frac{1+1}{4+8} = \frac{2}{12}$$

$$P(\text{warning}|s) = \frac{0+1}{4+8} = \frac{1}{12}$$

$$P(\text{score}|w) = \frac{2+1}{8+8} = \frac{3}{16}$$

$$P(\text{league}|w) = \frac{0+1}{8+8} = \frac{1}{16}$$

$$P(\text{warning}|w) = \frac{2+1}{8+8} = \frac{3}{16}$$

$$P(s|kw3, kw4, kw5) = \frac{P(kw3, kw4, kw5|s)\,P(s)}{P(kw3, kw4, kw5)}$$

$$= \alpha\; P(s)\, P(kw3|s)\, P(kw4|s)\, P(kw5|s)$$

$$= \alpha\; \frac{1}{4} \cdot \frac{2}{12} \cdot \frac{2}{12} \cdot \frac{1}{12}$$

$$= \alpha\; \frac{1}{123} = \alpha\; \frac{1}{1,728} = \alpha\; 0.00057$$

$$P(w|kw3, kw4, kw5) = \frac{P(kw3, kw4, kw5|w)\; P(w)}{P(kw3, kw4, kw5)}$$

$$= \alpha\; P(w)\, P(kw3|w)\, P(kw4|w)\, P(kw5|w)$$

$$= \alpha\; \frac{3}{4} \cdot \frac{3}{16} \cdot \frac{1}{16} \cdot \frac{3}{16}$$

$$= \alpha\; \frac{9}{16,384} = \alpha\; 0.00054$$

$\therefore$ Document 5 belongs to class sports.