Goal: $P(\text{msg is spam} \mid \text{msg content})$

Use Bayes' rule:

$$P(\text{msg is spam} \mid \text{msg content}) = \frac{P(\text{msg content} \mid \text{spam}) \cdot P(\text{spam})}{P(\text{msg content})}$$

$$P(\text{msg content} \mid \text{spam}) = P(\text{word 1} \mid \text{spam}) \cdot P(\text{word 2} \mid \text{spam}) \cdots P(\text{word N} \mid \text{spam})$$

$$P(\text{msg content} \mid \text{ham}) = P(\text{word 1} \mid \text{ham}) \cdot P(\text{word 2} \mid \text{ham}) \cdots P(\text{word N} \mid \text{ham})$$

The classification depends on the likelihood ratio:

$$\frac{P(\text{spam} \mid \text{msg content})}{P(\text{ham} \mid \text{msg content})}$$

$\Downarrow$

ratio $= 1$ (spam & ham equally likely)

ratio > 1 (spam is more likely)

ratio < 1 (ham is more likely)

$$\frac{P(\text{spam} \mid \text{msg content})}{P(\text{ham} \mid \text{msg content})} = \frac{P(\text{word 1} \mid \text{spam})}{P(\text{word 1} \mid \text{ham})} \cdot \frac{P(\text{word 2} \mid \text{spam})}{P(\text{word 2} \mid \text{ham})} \cdots \cdots \frac{P(\text{spam})}{P(\text{ham})}$$

$$\log\left[\frac{P(\text{spam} \mid \text{msg content})}{P(\text{ham} \mid \text{msg content})}\right] = \log\left[\frac{P(\text{word 1} \mid \text{spam})}{P(\text{word 1} \mid \text{ham})}\right] + \log\left[\frac{P(\text{word 2} \mid \text{spam})}{P(\text{word 2} \mid \text{ham})}\right] + \ldots +$$

Implementation

$$P(\text{word is present} \mid \text{spam}) \approx \frac{\#\text{ of spam messages with word} + 1/2}{\#\text{ of spam messages} + 1/2}$$

$$P(\text{word is absent} \mid \text{spam}) \approx \frac{\#\text{ of spam messages without the word} + 1/2}{\#\text{ of spam messages} + 1/2}$$

$$\sum_{\substack{\text{words}\\ \text{in}\\ \text{message}}} \log P(\text{word present} \mid \text{spam}) - \log(\text{word present} \mid \text{ham}) +$$

$$+ \sum_{\substack{\text{words}\\ \text{not in}\\ \text{message}}} \log(\text{word absent} \mid \text{spam}) - \log(\text{word absent} \mid \text{ham}) +$$

$$+ \log P(\text{spam}) - \log P(\text{ham})$$

$$LLR = \log\left(\prod_{\substack{\text{words}\\ \text{in}\\ \text{message}}} \frac{P(\text{word present} \mid \text{spam})}{P(\text{word present} \mid \text{ham})}\right) + \log\left(\prod_{\substack{\text{words}\\ \text{not}\\ \text{in}\\ \text{message}}} \frac{P(\text{word absent} \mid \text{spam})}{P(\text{word absent} \mid \text{ham})}\right)$$