Bayes' Theorem.

Conditional Probability.

Independent Events

Conditionally Independent Events.

Bayes' Theorem

naive Bayes algorithm.

Independent Events

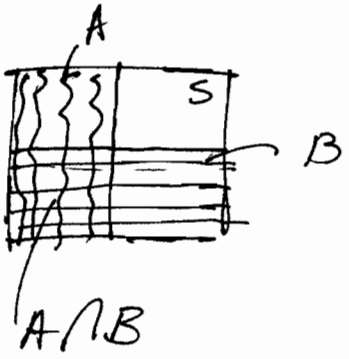
Events A, B are independent if

$$P(A \cap B) = P(A) \cdot P(B)$$

or

$$P(B|A) = P(B)$$

Venn diagram

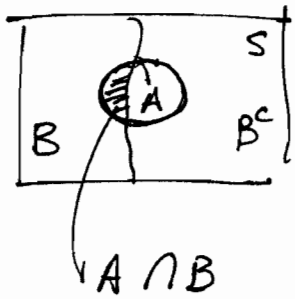


Conditional Probability.

The probability of B given A

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Venn diagram.



S = Sample Space

A is the reduced Sample Space.

So $P(A \cap B) = P(B|A) \cdot P(A)$

or $P(A \cap B) = P(A|B) \cdot P(B)$

note $P(A \cap B) = P(B \cap A)$

In the application of the naive Bayes algorithm, presented in the book, independence would mean all words in the corpus are independent. This is not what is being assumed in naive Bayes. What is assumed is conditional independence. That is the words are independent given the class, spam or ham.

In the 1ST Edition the denominator has an error.

$$P(W_1, \neg W_2, \neg W_3, W_4) \neq$$

$$P(W_1) \cdot P(\neg W_2) \cdot P(\neg W_3) \cdot P(W_4)$$

Conditionally Independent Events

Events B_1, B_2 are conditionally independent given A if

$$P(B_1, B_2 \mid A) = P(B_1 \mid A) \cdot P(B_2 \mid A)$$

which does not imply independence

$$P(B_1, B_2) = P(B_1) \cdot P(B_2)$$

In the naive Bayes application to the SMS data, conditional independence means all words are independent given the class.

So the words in the "spam" dictionary are assumed independent. And the words in the "ham" dictionary are assumed independent. Both are conditional on knowing the class, ham or spam.

But all the words in the combined dictionary are not independent

Bayes' Theorem.

posterior

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{\overbrace{P(A|B)}^{\text{likelihood}} \cdot \overbrace{P(B)}^{\text{prior}}}{P(A)}$$

$$= \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)}$$

Since the denominator is constant, i.e., a number.

$$P(B|A) \propto P(A|B) \cdot P(B)$$

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}.$$

Naive Bayes algorithm.

The naive Bayes learner is trained by constructing a likelihood table for the appearance of the words. For the book example W_1, W_2, W_3, W_4

As a new message is received, the posterior probability is calculated to determine whether the new message is more likely spam or ham, given the likelihood of the words found in the message text.

$$P(\text{Spam} \mid W_1, \neg W_2, \neg W_3, W_4)$$

$$\propto P(W_1 \mid \text{Spam}) \cdot P(\neg W_2 \mid \text{spam}) \cdot P(\neg W_3 \mid \text{Spam}) \cdot P(W_4 \mid \text{Spam}) \cdot P(\text{Spam})$$

$$\propto \left[\left(\frac{4}{20}\right)\cdot\left(\frac{10}{20}\right)\cdot\left(\frac{20}{20}\right)\cdot\left(\frac{12}{20}\right)\right]\cdot\left(\frac{20}{100}\right)$$

$$= .012$$

$$P(\text{ham} \mid W_1, \neg W_2, \neg W_3, W_4)$$

$$\propto P(W_1 \mid \text{ham}) \cdot P(\neg W_2 \mid \text{ham}) \cdot P(\neg W_3 \mid \text{ham}) \cdot P(W_4 \mid \text{ham}) \cdot P(\text{ham})$$

$$\propto \left[\left(\frac{1}{80}\right) \cdot \left(\frac{66}{80}\right) \cdot \left(\frac{71}{80}\right) \cdot \left(\frac{23}{80}\right) \right] \left(\frac{80}{100}\right)$$

$$= .002$$

Because $.012 / .002 = 6$, there is six times more likely the message is spam than ham.