First ingredient we need: Bayes Theorem

\[ P(A \text{ and } B) = P(B \text{ and } A), \quad P(A \text{ and } B) = P(A|B)P(B) \]
\[ P(A|B)P(B) = P(B|A)P(A) \]
\[ P(A|B) = \frac{P(B|A)P(A)}{P(B)} \]

That's it!

Bayes then lets us "flip around" conditional probabilities.

In the text classification it is easier to maximize \( Pr(d|c) \) than \( Pr(c|d) \).

\[
\text{MAP} = \arg \max_c \frac{Pr(c|d)}{Pr(c)}
\]
\[
= \arg \max_c \frac{Pr(d|c)Pr(c)}{Pr(d)}
\]

Now, to find a way to calculate \( \text{MAP} \) we need to make some simplifications.

I. The document we want to classify is constant, this means that \( Pr(d) \) is the same regardless of \( c \Rightarrow \) it won't change what \( \text{MAP} \) is:

\[
\text{MAP} = \arg \max_c Pr(d|c)Pr(c)
\]

Next, the document \( d \) is a collection of words:

\[ Pr(d|c) = Pr(w_1, w_2, \ldots, w_n | c) \]

This is not the bag of words (BOW) model.

Assume Bag-of-Words \( \Rightarrow \) position/order of words doesn't matter.
Assume conditional independence \( \Rightarrow \) probabilities of different words appearing together are independent given the document class.

\[ Pr(d|c) = Pr(w_1, \ldots, w_n | c) = Pr(w_1 | c) \cdot Pr(w_2 | c) \cdots Pr(w_n | c) \]

Note these last two assumptions are certainly not true for a real text!
Put these together and you have constructed a (text) classifier called **Naive Bayes**

\[
C_{\text{MAP}} = \arg\max_C \Pr(d|c) \Pr(c)
\]

\[
\downarrow
\]

\[
C_{\text{NB}} = \arg\max_C \Pr(c) \prod_{i=1}^{\hat{P}} \Pr(w_i|c)
\]

**Learning Naive Bayes**

How to compute these probabilities...

Training corpus \(N_{\text{doc}}\) documents, each labeled \(w/c = \text{spam or ham}\).

Estimate Probabilities:

\[
\hat{P}(c) = \frac{\# \text{docs labeled } c}{N_{\text{docs}}}
\]

\[
\hat{P}(w_i|c) = \frac{\text{count}(w_i,c)}{\sum_j \text{count}(w_j,c)}
\]

Problem! What if, we use this \(\hat{P}\) estimator and then, when we attempt to classify a new document we see a new word we have never seen before?

\rightarrow \text{word will have a count of zero} \rightarrow \hat{P}(w_i|c) = 0

plug into \(C_{\text{NB}}\) and it becomes zero:

\[
C_{\text{NB}} = \arg\max_C \hat{P}(c) \prod_{i=1}^{\hat{P}} \hat{P}(w_i|c)
\]

The fix is **maddness** \(\rightarrow\) Laplace (or additive) smoothing!

\[
\hat{\Pr}(w_i|c) = \frac{\text{count}(w_i,c) + 1}{\sum_j \text{count}(w_j,c) + 1} = \frac{\text{count}(w_i,c) + 1}{\sum_{j=1}^{n} \text{count}(w_j,c) + n}
\]