Suppose I have 100 emails. 58 are spam, 42 are not spam.

\* 47 of those 58 emails that are spam have the word "buy"

\* 41 of those 58 emails that are spam have the word "win"

\* 3 of those 42 emails that aren't spam have the word "buy"

\* 5 of those 42 emails that aren't spam have the word "win"

A new email arrives that contains the word "win" but not the word "buy."  
  
Answer the following questions:

* Should this email be classified as spam or not spam?
* What is the posterior probability of the email being spam? Hint: Use normalization!

P(spam) = 58/100

P(~spam) = 42/100

P(buy|spam) = 47/58 -> 48/60

P(win|spam) = 41/58 -> 42/60  
  
P(buy|~spam) = 3/42 -> 4/44  
P(win|~spam) = 5/42 -> 6/44  
  
Naïve bayes:

HMAP = argmax[i] P(D|H)P(H) = argmax[H in spam/not spam] P(win, ~buy)P(H) =

= argmax[H in spam/not spam] P(win|H)P( ~buy|H)P(H)

Calc for spam:

P(win|spam)P( ~buy|spam)P(spam) = (42/60) \* (1-48/60) \* (58/100) = 0.0812

Calc for ~spam:

P(win|~spam)P( ~buy|~spam)P(~spam) = (6/44) \* (1-4/44) \* (42/100) = 0.0521

B/c .0812 > .0521, we can conclude the email is most likely spam.

Posterior probability of spam = 0.0812 / (0.0812 + 0.0521) =approx .6092

So there is roughly a 61% chance this email is spam.