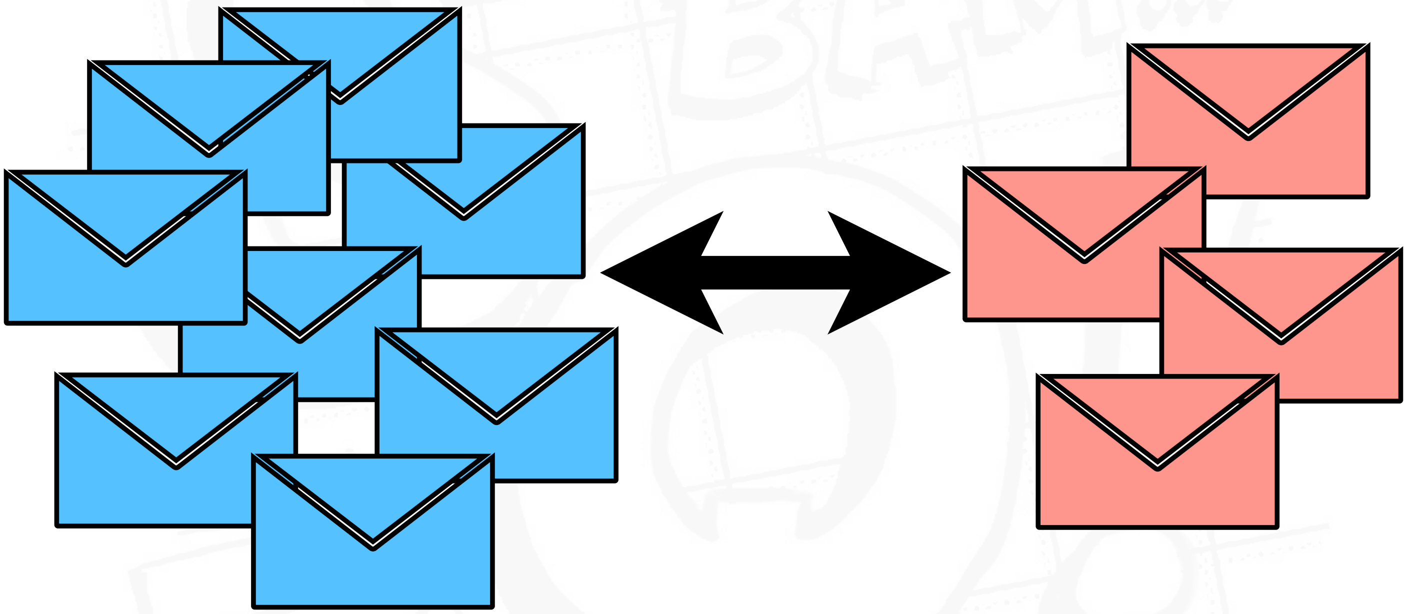
StatQuest!!!

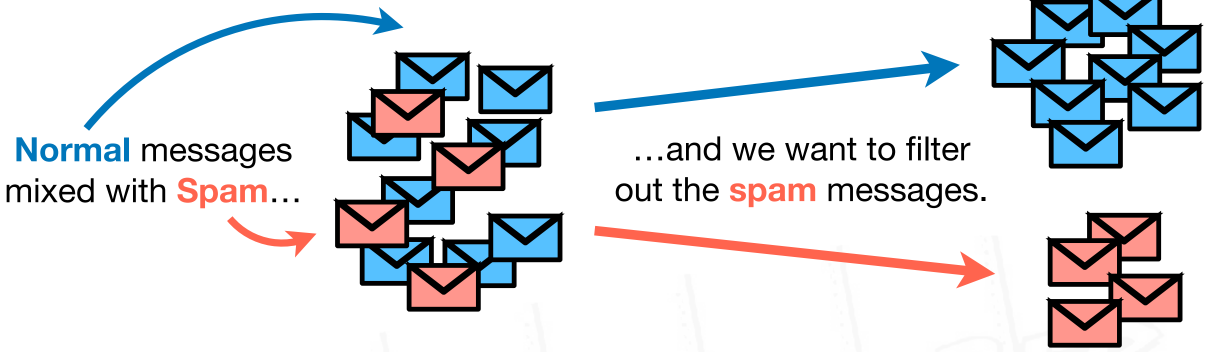
(Multinomial)

Naive Bayes



Study Guide!!!

The Problem



The Solution - A Naive Bayes Classifier

If we get this message:

Dear Friend

We multiply the **Prior** probability the message is **Normal**…

…by the probabilities of seeing the words **Dear** and **Friend**, given that it's a **Normal Message**…

$$p(\ N\ ) \times p(\ \textbf{Dear} \mid N\ ) \times p(\ \textbf{Friend} \mid N\ )$$

…and compare that to the **Prior** probability the message is **Spam**…

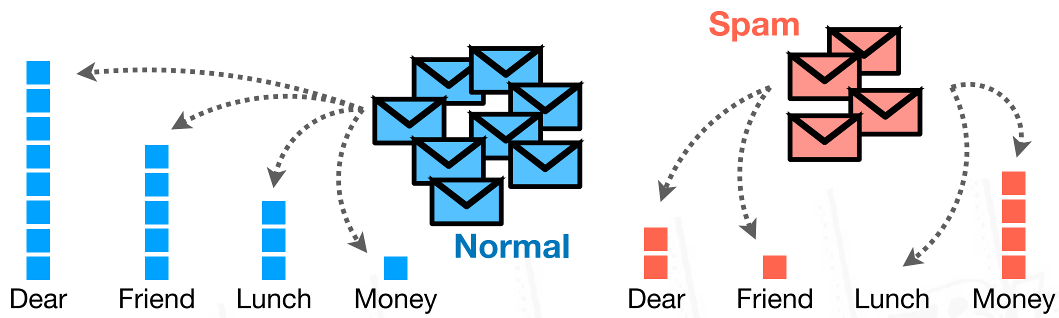
…multiplied by the probabilities of seeing the words **Dear** and **Friend**, given that it's **Spam**.

$$p(\ S\ ) \times p(\ \textbf{Dear} \mid S\ ) \times p(\ \textbf{Friend} \mid S\ )$$

Whichever classification has the highest probability is the final classification.

NOTES:

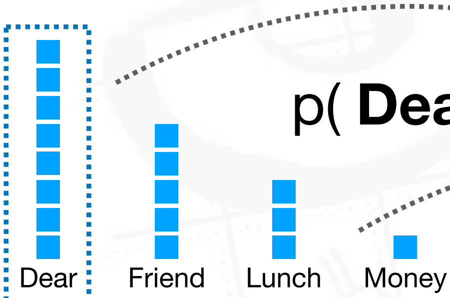
Step 1) Make histograms for all words



NOTE: The word **Lunch** did not appear in the **Spam**. This will cause problems that we will see and fix later.

Step 2a) Calculate conditional probabilities for **Normal**, **N**

For example, the probability that the word **Dear** occurs given that it is in a **Normal** message is the number of times **Dear** occurred in **Normal** messages, **8**…



$$p(\,\textbf{Dear} \mid \textbf{N}\,) = \frac{8}{17} = \boxed{0.47}$$

…divided by **17**, the total number of words in **Normal** messages.

$\boxed{p(\,\textbf{Dear} \mid \textbf{N}\,) = 0.47}$

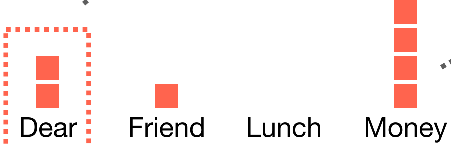
p(**Friend** | **N**) = 0.29

p(**Lunch** | **N**) = 0.18

p(**Money** | **N**) = 0.06

Step 2b) Calculate conditional probabilities for **Spam**, **S**

For example, the probability that the word **Dear** occurs given that it is in **Spam** is the number of times **Dear** occurred in **Spam**, **2**…



$$p(\,\textbf{Dear} \mid \textbf{S}\,) = \frac{2}{7} = \boxed{0.29}$$

…divided by **7**, the total number of words in **Spam** messages.

$\boxed{p(\,\textbf{Dear} \mid \textbf{S}\,) = 0.29}$

p(**Friend** | **S**) = 0.14

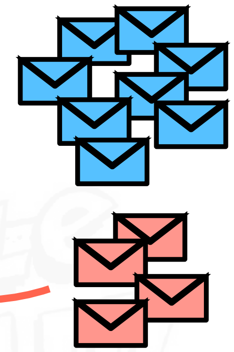
p(**Lunch** | **S**) = 0.00

p(**Money** | **S**) = 0.57

Step 3a) Calculate prior probability for Normal, p(N)

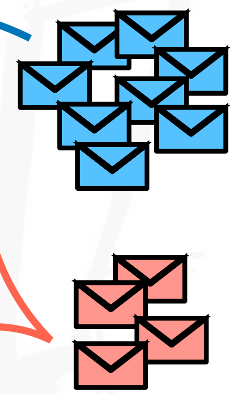
NOTE: The **Prior Probabilities** can be set to any probabilities we want, but a common guess is estimated from the training data like so:

$$p(\,\mathbf{N}\,) = \frac{\textbf{\# of Normal Messages}}{\textbf{Total \# of Messages}} = \frac{8}{8 + 4} \boxed{= 0.67}$$



Step 3b) Calculate prior probability for Spam, p(S)

$$p(\,\mathbf{S}\,) = \frac{\textbf{\# of Spam Messages}}{\textbf{Total \# of Messages}} = \frac{4}{8 + 4} \boxed{= 0.33}$$



NOTE: The reason **Naive Bayes** is ***naive***, is that it does not take word order or phrasing into account.

In other words, **Naive Bayes** would give the exact same probability to the phrase **I like pizza** as it would to the phrase **Pizza like I**…

…even though people frequently say **I like pizza** and almost never say **Pizza like I**.

Because keeping track of every phrase and word ordering would be impossible, **Naive Bayes** doesn't even try.

That said, **Naive Bayes** works well in practice, so keeping track of word order must not be super important.

4a) Calculate probability of seeing the words Dear Friend, given the message is Normal

The **Prior** probability the message is **Normal**…

…multiplied by the probabilities of seeing the words **Dear** and **Friend**, given that it's **Normal**.

$$p(\,N\,) \times p(\,\textbf{Dear} \mid N\,) \times p(\,\textbf{Friend} \mid N\,)$$

$p(\,N\,) = 0.67$ $\quad$ $p(\,\textbf{Dear} \mid N\,) = 0.47$ $\quad$ $p(\,\textbf{Friend} \mid N\,) = 0.29$

$$0.67 \times 0.47 \times 0.29 \boxed{= 0.09}$$

NOTE: This probability makes the *naive* assumption that **Dear** and **Friend** are not correlated.

In other words, this is not a realistic model (high bias), but it works in practice (low variance).

4b) Calculate probability of seeing the words Dear Friend, given the message is Spam

The **Prior** probability the message is **Spam**…

…multiplied by the probabilities of seeing the words **Dear** and **Friend**, given that it's **Spam**.

$$p(\,S\,) \times p(\,\textbf{Dear} \mid S\,) \times p(\,\textbf{Friend} \mid S\,)$$

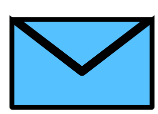
$p(\,S\,) = 0.33$ $\quad$ $p(\,\textbf{Dear} \mid S\,) = 0.29$ $\quad$ $p(\,\textbf{Friend} \mid S\,) = 0.14$

$$0.33 \times 0.29 \times 0.14 \boxed{= 0.01}$$

NOTE: In practice, these probabilities can get very small, so we calculate the **log()** of the probabilities to avoid underflow errors on the computer.

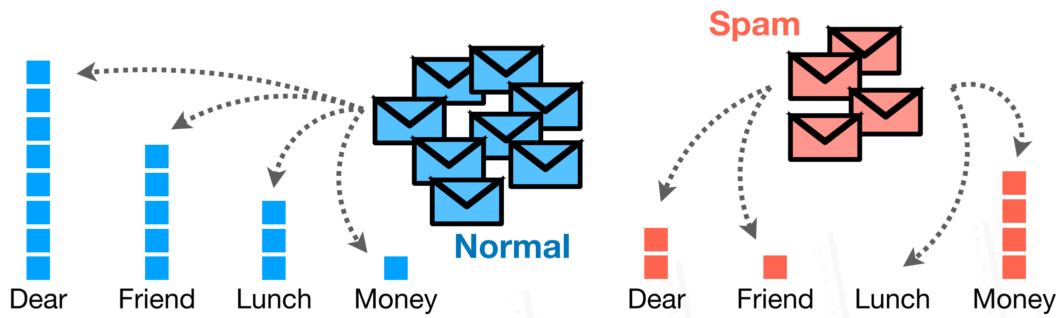
5) Classification

Because **Dear Friend** has a higher probability of being **Normal** (**0.09**) than **Spam** (**0.01**), we classify it as **Normal**.

Dear Friend 

BAM!!!

Dealing With Missing Data



Remember, the word **Lunch** did not occur in any of the **Spam**…

…and that means the probability of seeing **Lunch** in **Spam** = **0**.

p(**Dear** | **S**) = 0.29

p(**Friend** | **S**) = 0.14

$\boxed{\text{p(Lunch | S) = 0.00}}$

p(**Money** | **S**) = 0.57

This means that any message with the word **Lunch** in it will be classified as **Normal**, because the probability of being **Spam** = **0**.

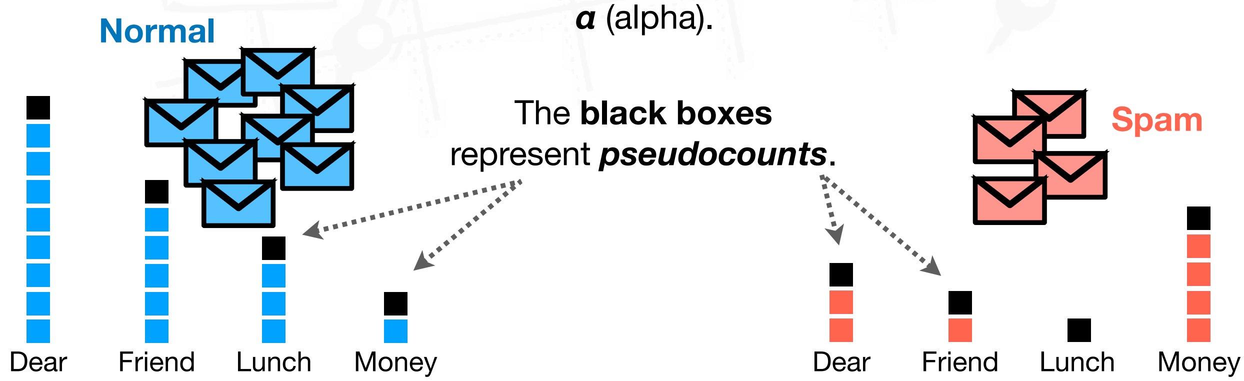
For example, the probability that this message is **Spam**: **Lunch Money Money Money Money**

$$\text{p(S)} \times \text{p(Lunch | S)} \times \text{p(Money | S)}^4$$

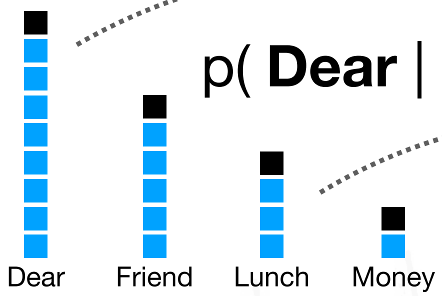
p(**S**) = 0.33   p(**Lunch** | **S**) = 0.00   p(**Money** | **S**) = 0.57

$$0.33 \times 0.00 \times 0.57^4 \boxed{= 0}$$

To solve this problem a ***pseudocount*** is added to each word. Usually that means adding **1** count to each word, but you can add any number by changing ***α*** (alpha).



Using Pseudocounts…



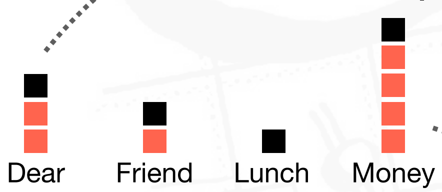
$$p(\textbf{Dear} \mid \textbf{N}) = \frac{8 + 1}{17 + 4} = \boxed{0.43}$$

$p(\textbf{Dear} \mid \textbf{N}) = 0.43$
$p(\textbf{Friend} \mid \textbf{N}) = 0.29$
$p(\textbf{Lunch} \mid \textbf{N}) = 0.19$
$p(\textbf{Money} \mid \textbf{N}) = 0.10$

$$p(\textbf{N}) \times p(\textbf{Lunch} \mid \textbf{N}) \times p(\textbf{Money} \mid \textbf{N})^4$$

$p(\textbf{N}) = 0.67$ $\quad$ $p(\textbf{Lunch} \mid \textbf{N}) = 0.19$ $\quad$ $p(\textbf{Money} \mid \textbf{N}) = 0.10$

$$0.67 \times 0.19 \times 0.10^4 \boxed{= 0.00001}$$



$$p(\textbf{Dear} \mid \textbf{S}) = \frac{2 + 1}{7 + 4} = \boxed{0.27}$$

$p(\textbf{Dear} \mid \textbf{S}) = 0.27$
$p(\textbf{Friend} \mid \textbf{S}) = 0.18$
$p(\textbf{Lunch} \mid \textbf{S}) = 0.09$
$p(\textbf{Money} \mid \textbf{S}) = 0.45$

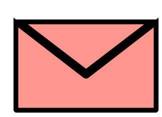
$$p(\textbf{S}) \times p(\textbf{Lunch} \mid \textbf{S}) \times p(\textbf{Money} \mid \textbf{S})^4$$

$p(\textbf{S}) = 0.33$ $\quad$ $p(\textbf{Lunch} \mid \textbf{S}) = 0.09$ $\quad$ $p(\textbf{Money} \mid \textbf{S}) = 0.45$

$$0.33 \times 0.09 \times 0.45^4 \boxed{= 0.00122}$$

Because **Lunch Money Money Money Money** has a higher probability of being Spam (**0.00122**) than Normal (**0.00005**), we classify it as Spam.

Lunch Money Money Money Money



SPAM!!!