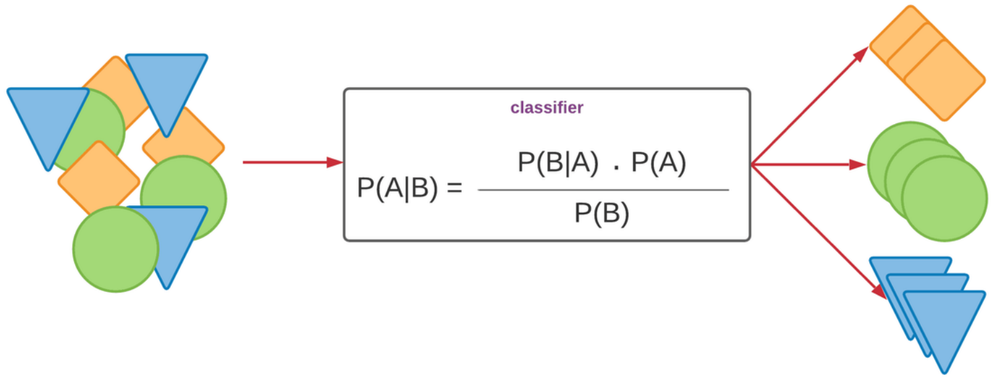
Mutinomial Naive Bayes

Multinomial Naive Bayes



- The multinomial model captures word frequency information in documents
- Capturing frequency information of tokenization can help classification
- Require a small number of training sets but effective classification results
- The more applicable algorithm in text classification
- The most popular among naïve bayes classifiers

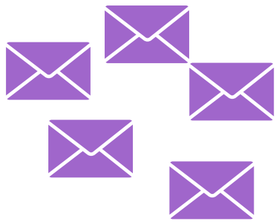
Conditional probalities

vector of words from message

$$P(Spam|x) = \frac{p(x|Spam) * P(Spam)}{P(x)}$$

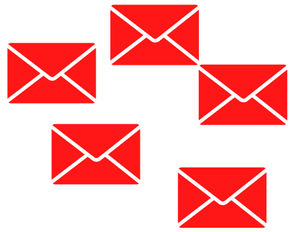
x=[w1,w2,w3,w4,...,wn]

Conditional probalities



Word	Count
Dear	9
Friend	6
Promotion	1
Free	1
Total	**17**

p(Dear| normal) = 9/17 =0.52
p(Friend| normal) = 6/17 =0.35
p(promotion| normal)= 1/17 =0.29
p(free| normal) = 1/17 =0.29



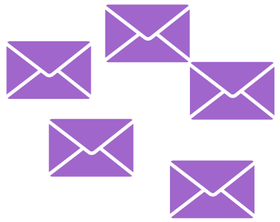
Word	Count
Dear	1
Friend	1
Promotion	1
Free	4
Total	7

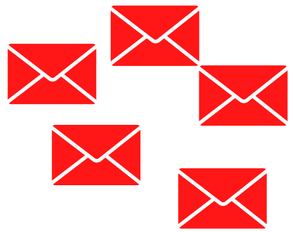
p(Dear| spam) = 1/7 =0.14
p(Friend| spam) = 1/7 =0.14
p(promotion| spam)= 1/7 =0.14
p(free| spam) = 4/7 =0.57

likelihoods

Conditional probalities

17 messages → 12 training data

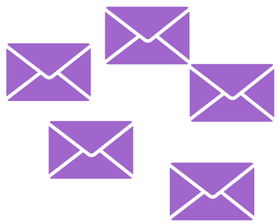
8 → 

4 → 

p(normal)= $\frac{8}{12}$ =0.67

p(spam)= $\frac{4}{12}$ =0.33

Conditional probalities

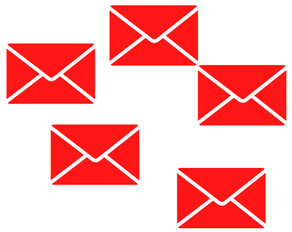


p(Dear| normal) = 8/17 =0.47
p(Friend| normal) = 5/17 =0.29
p(promotion| normal)= 3/17 =0.18
p(free| normal) = 1/17 =0.29

p(normal)=0.67 p(normal) * p(Dear|normal) * p(Friend|normal)

Dear Friend

0.67 * 0.47 * 0.29= 0.091



p(Dear| spam) = 1/7 =0.14
p(Friend| spam) = 1/7 =0.14
p(promotion| spam)= 1/7 =0.14
p(free| spam) = 4/7 =0.57

p(spam)=0.33 p(spam) * p(Dear|spam) * p(Friend|spam)

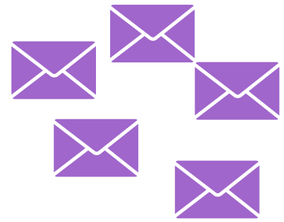
0.33 * 0.14 * 0.14= 0.006

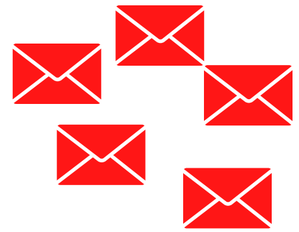
Conditional probalities

0.67 * 0.47 * 0.29= 0.091

0.33 * 0.14 * 0.14= 0.006

Dear Friend





Conditional probalities

p(Dear| normal) = 8/17 =0.47
p(Friend| normal) = 5/17 =0.29
p(promotion| normal)= 3/17 =0.18
p(free| normal) = 1/17 =0.29

Promotion Free Free

p(Dear| spam) = 1/7 =0.14
p(Friend| spam) = 1/7 =0.14
p(promotion| spam)= 1/7 =0.14
p(free| spam) = 4/7 =0.57

p(normal) * p(promotion|normal) * p(free|normal) * p(free|normal)

0.18* 0.29*0.29= 0.015

<

p(spam) * p(promotion|spam) * p(free|spam) * p(free|spam)

0.14*0.57*0.57= 0.045

Promotion Free Free