# Naïve Bayesians

Back to Basics Series

27 Feb 2021

## Goal

# Developing the Bayesian muscle to solve a wide range of problems

# Naïve Bayesian Philosophy

## Intuitive (Visual) Understanding of the Bayesian Reasoning

## Ability to model real world problems in a Bayesian Setting

Starting from Simple Probabilistic modelling

Adapting it in a a Bayesian setting
And moving towards ML models

## Fluency in the Calculus of Bayesian Stats & ML model

# Season 2: Back to Basics

| | |
| --- | --- |
| Ep 1 | Bayes Theorem |
| Ep 2 – Ep 3 | Problems with Binomial Likelihoods |
| Ep 4 | Disease Detection |
| Ep 5 | Naive Bayes Classification |
| Ep 6 | Gaussian Naive Bayes Classification |
| Ep 7 | German Tank Problem |
| Ep 8 | Waiting Times (Continuous Distributions) |

# Back to Basics

| | | Canonical Problem | Applications |
|---|---|---|---|
| Ep 1 | Bayes Theorem | There are 2 boxes from which cookies can be taken from. Box A and Box B. Box A contains 10 chocolate cookies, Box B contains 5 ginger cookies. Given that you get a chocolate cookie which box was it taken from? | The Shy Librarian Problem<br>Naive Bayes algorithm |
| Ep 2<br>Ep 3 | Problems with Binomial Likelihoods | You have 2 coins C1 and C2. p(heads for C1) = .7 & P(heads for C2) = 0.6<br>You flip the coin 10 times. What is the probability that the given coin you picked is C1 given you have 7 heads and 3 tails? | A/B Testing |
| Ep 4 | Disease Detection | A particular disease affects 1% of the population. There is an imperfect test for this disease: The test gives a positive result for 90% of people who have the disease, and 5% of the people who are disease-free. Given a positive test result – what is the probability of having the disease? | COVID Tests (PCR & Antibody)!<br>Fraud Detection |
| Ep 5 | Naive Bayes Classification | Given these words occur in this text what's the probability it's spam? | Any Classification Problem |
| Ep 6 | Gaussian Naive Bayes Classification | Given the weights and heights of basketball players, what's the probability that person a is a basketball player given weight = w and height = h? | |

# Back to Basics

| | | Canonical Problem | Applications |
|---|---|---|---|
| Ep 7 | German Tank Problem | Suppose tanks were given a serial number based on the order in which they were manufactured. Given that you've observed a tank with serial number "10", how many tanks were actually manufactured in total? | ? |
| Ep 8 | Waiting Times (Continuous Distributions) | Suppose you need to gather 10 patients for a trial. Each signup happens at time t_i (i=1, 10). How long do you have to wait after it took you 3 weeks to accrue 2 signups? | Planning Trials<br>Estimating Queues |

## Bayes Rule

Posterior Likelihood Prior

$$P(\theta_i \mid D) = \frac{P(D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

## *Recap* | Canonical Problems

Given the words “Dear Friend” occur in this email what's the probability it's spam?

$$P(\ S \mid \text{Dear Friend}\ )$$

Given the size of an email is 1.8 MB & the time to read it is 2 seconds what's the probability it's spam?

$$P(\ S \mid 1.8\text{MB},\ 2\text{sec}\ )$$

| N | S |
| --- | --- |
| Normal | Spam |

## Canonical Problem

Given the words “Dear Friend” in an email and its size is 1.8 MB what's the probability it's spam?

$$P(\ S \mid \text{Dear Friend}\ , 1.8\text{MB})$$

N

Normal

S

Spam

## Canonical Problem

Given the words “Dear Friend” in an email and its size is 1.8 MB what's the probability it's spam?

$$P(\ S \mid \text{Dear Friend}\ , 1.8\text{MB})$$

N

Normal

S

Spam

# You collect some data: Word frequency & Email sizes

## 8 Normal Emails

| | |
|---|---|
| 180 | 976 |
| 190 | 1280 |
| 256 | 1500 |
| 780 | 1798 |

KB

Dear: 8
Friend 5
Lunch: 3
Money: 1

## 4 Spam Emails

| |
|---|
| 980 |
| 1850 |
| 1950 |
| 2000 |

KB

Dear: 2
Friend: 1
Lunch: 0
Money: 4

# You collect some data: Word frequency & Email sizes

## 8 Normal Emails

$\mu = 870$ KB

$\sigma = 628$ KB

KB

| | |
|---|---|
| Dear: | 8 |
| Friend | 5 |
| Lunch: | 3 |
| Money: | 1 |

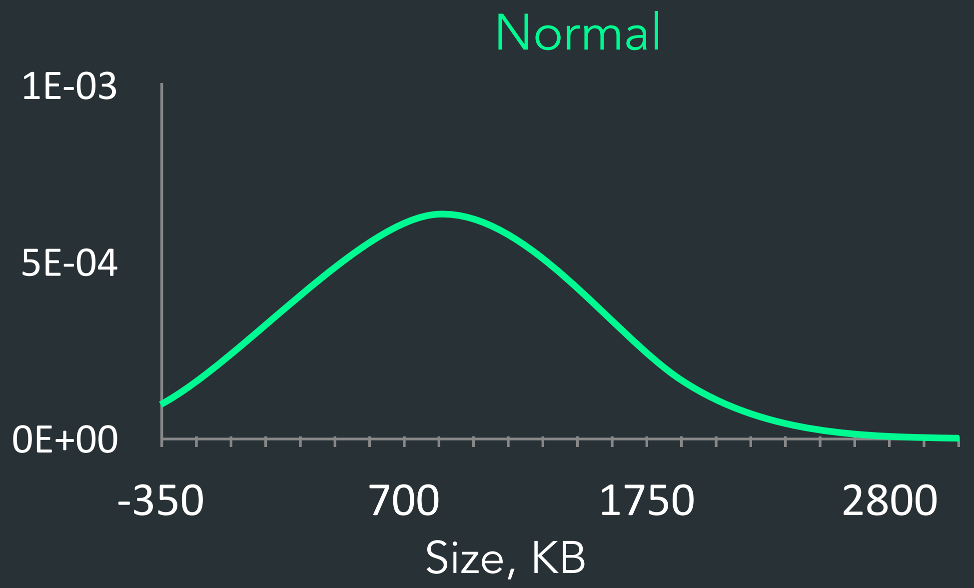
## 4 Spam Emails

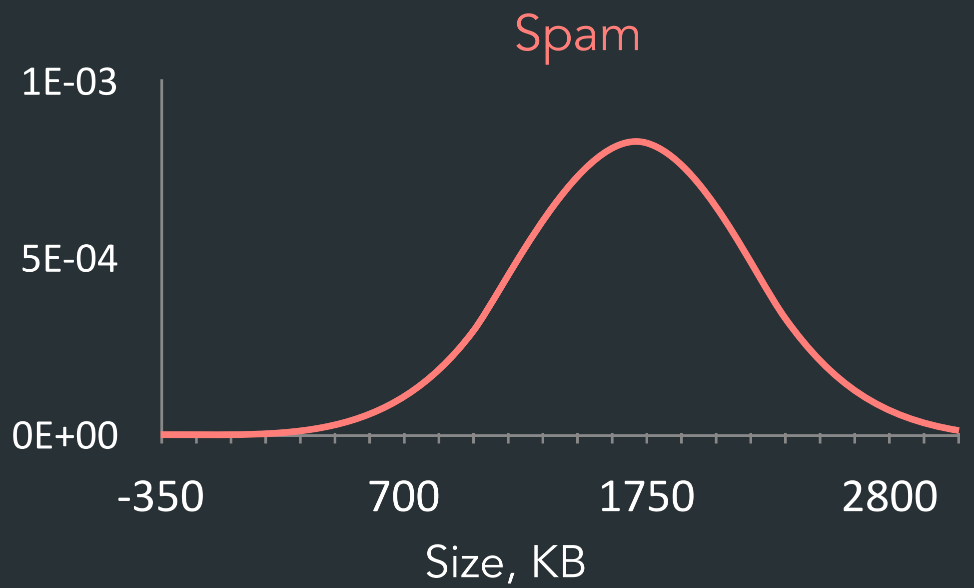
$\mu = 1697$ KB

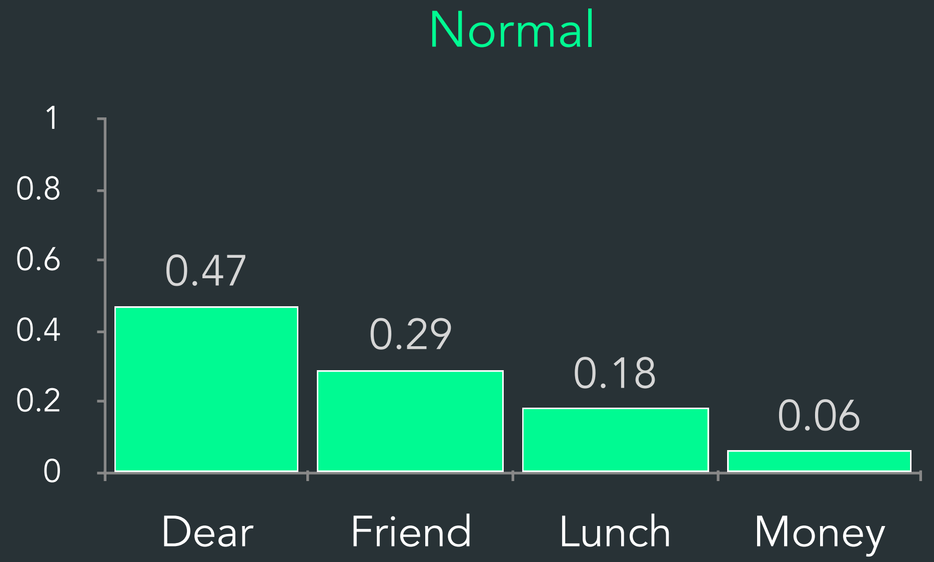
$\sigma = 481$ KB

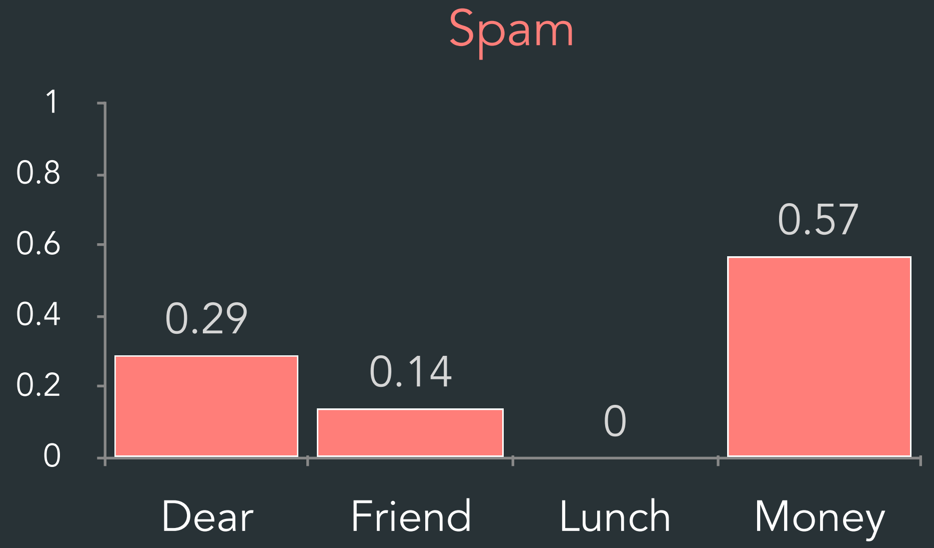
| | |
|---|---|
| Dear: | 2 |
| Friend: | 1 |
| Lunch: | 0 |
| Money: | 4 |

# You collect some data: Word frequency & Email sizes

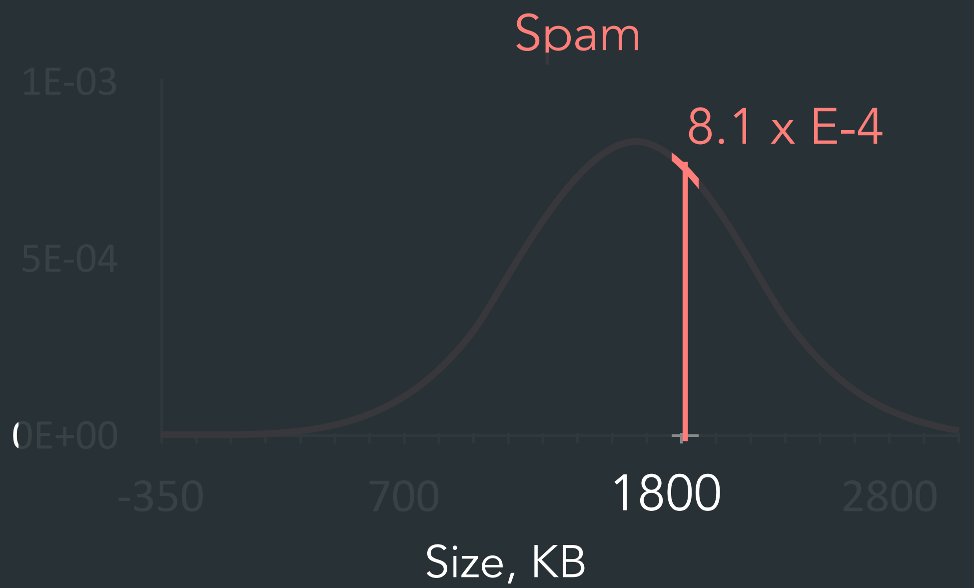


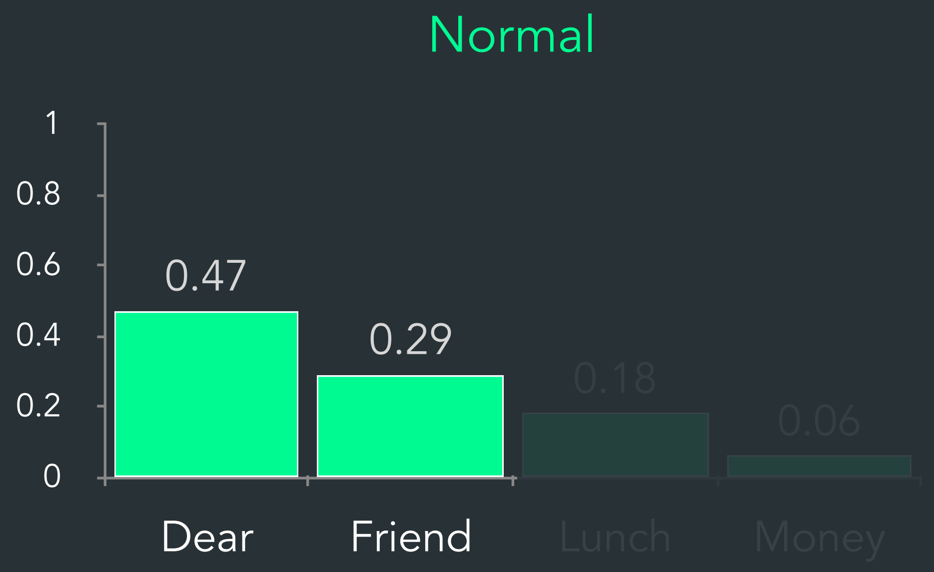



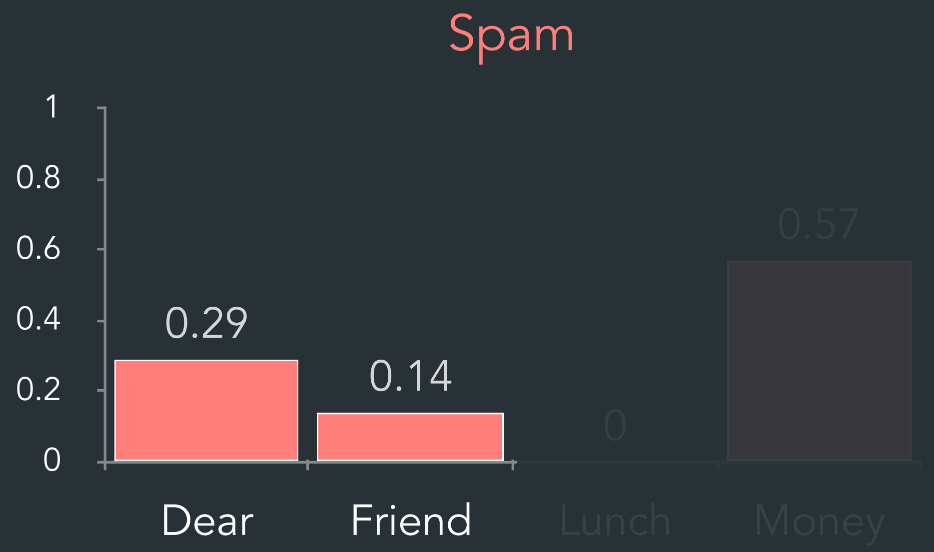



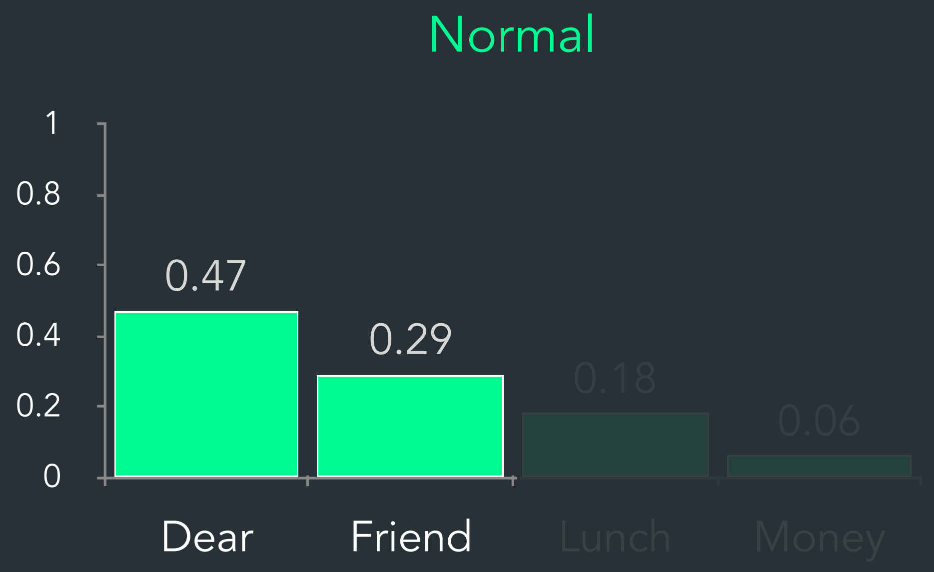
# Given the words “Dear Friend” in an email and its size is 1.8 MB

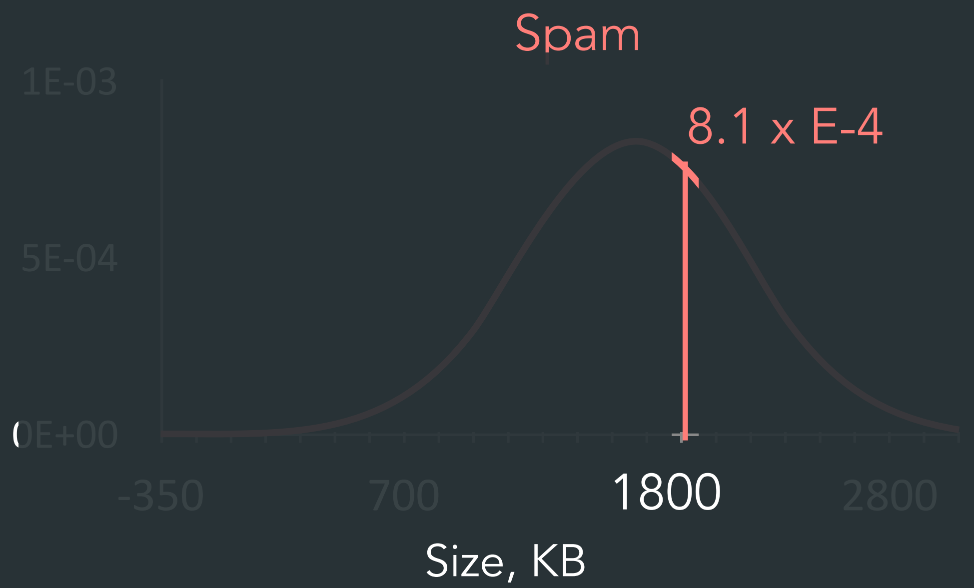



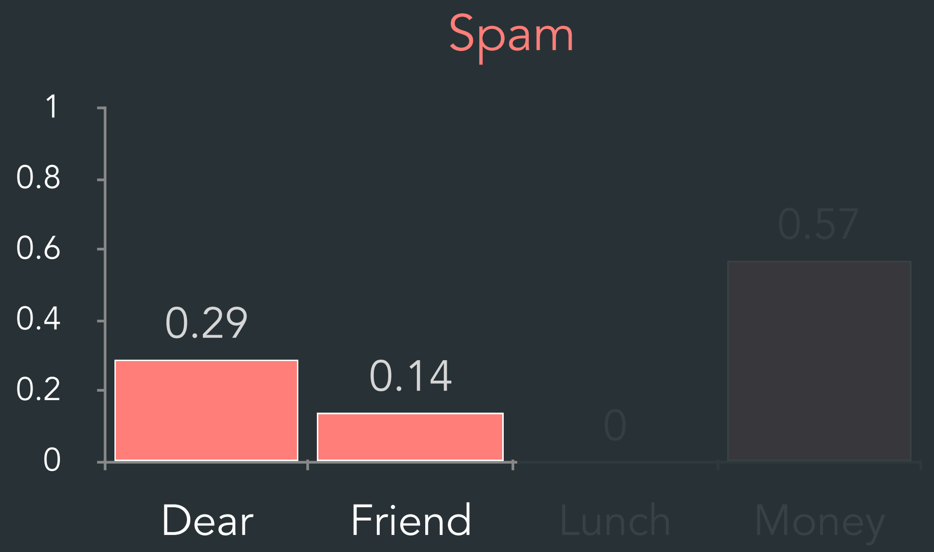



# Given the words “Dear Friend” in an email and its size is 1.8 MB

## Generalised Naive Bayes Solution

$$P(N \mid 1.8\text{MB}, 2\text{s}) \propto P(\text{Dear Friend}, 1.8\text{MB} \mid N)\, P(N)$$

$$\approx P(\text{Dear} \mid N)\, P(\text{Friend} \mid N)\, P(1.8\text{MB} \mid N)\, P(N)$$

$$= (0.47) \times (0.29) \times (2.1\text{E-}4) \times (2/3)$$

$$= 1.9\text{E-}5$$

## Generalised Naive Bayes Solution

$$P(S \mid 1.8MB, 2s) \propto P(\text{Dear Friend}, 1.8MB \mid S)\, P(S)$$

$$\approx P(\text{Dear} \mid S)\, P(\text{Friend} \mid S)\, P(1.8MB \mid S)\, P(S)$$

$$= (0.29) \times (0.14) \times (8.1E\text{-}4) \times (1/3)$$

$$= 1.1E\text{-}5$$

## Generalised Naive Bayes Solution

$$P(S \mid 1.8MB, 2s) \propto 1.1E\text{-}5$$

$$P(N \mid 1.8MB, 2s) \propto 1.9E\text{-}5$$

$$P(S \mid 1.8MB, 2s) < P(N \mid 1.8MB, 2s)$$

N

Normal

# Generalised Naive Bayes Solution

## Bayes Rule

Posterior Likelihood Prior

$$P(\theta_i \mid D) = \frac{P(D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

## Bayes Rule

Posterior Likelihood Prior

$$P(\theta_i \mid D) = \frac{P(d_1, \ldots, d_D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

# Generalised Naive Bayes Solution

$$P(\theta_i \mid D) \approx \frac{P(d_1 \mid \theta_i)\, P(d_2 \mid \theta_i) \ldots P(d_D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

## Generalised Naive Bayes Solution

$$P(\theta_i \mid D) \approx \frac{P_1(d_1 \mid \theta_i)\, P_2(d_2 \mid \theta_i) \ldots P_D(d_D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P_j(D \mid \theta_j)\, P(\theta_j)}$$

Normalising Constant

## Generalised Naive Bayes Solution

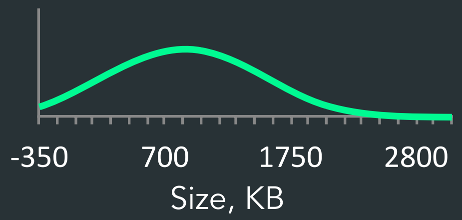
$$P(\theta_i \mid D) \approx \frac{P_1(d_1 \mid \theta_i)\, P_2(d_2 \mid \theta_i) \ldots P_D(d_D \mid \theta_i)\, P(\theta_i)}{\sum_{all\ j} P(D \mid \theta_j)\, P(\theta_j)}$$

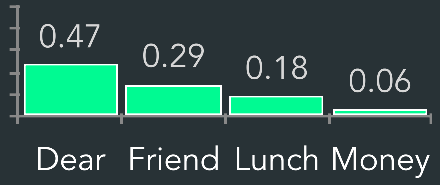
Normalising Constant

# Generalised Naive Bayes Solution

$P_1(d_1 \mid \theta)$ $P_2(d_2 \mid \theta)$

Normal



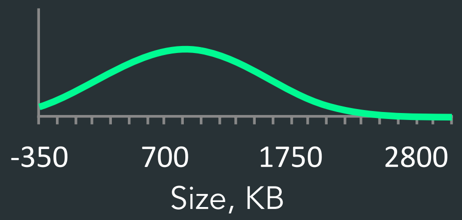



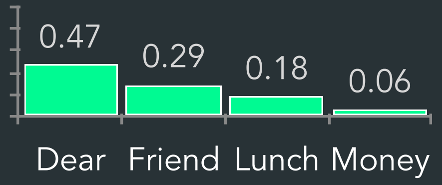

Spam


# Generalised Naive Bayes Solution

$P_1(d_1 \mid \theta)$ $P_2(d_2 \mid \theta)$

Normal




Spam



.
.
.

Important

# Generalised Naive Bayes Solution