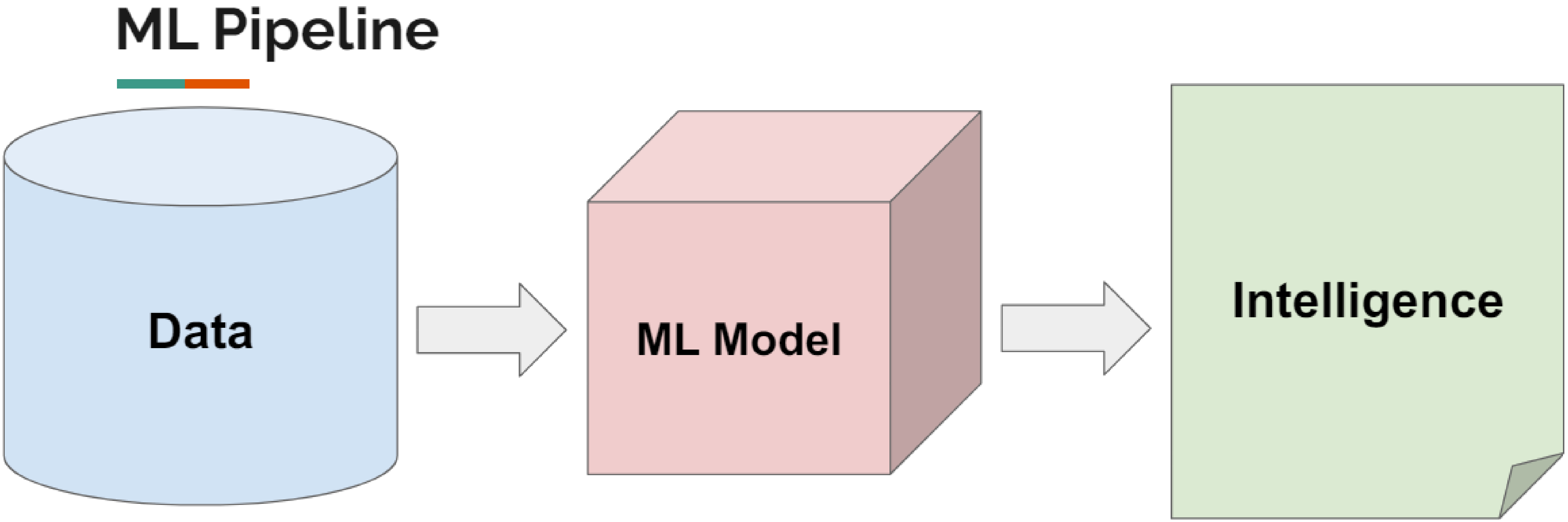
Naïve Bayes

Rina BUOY

AMERICAN UNIVERSITY OF PHNOM PENH

STUDY LOCALLY. LIVE GLOBALLY.

Machine Learning



From **Wikipedia**: "Machine learning is the study of computer algorithms that improve automatically through experience."

Spam Filter

- In real life, you may have seen a lot of spam emails like this.
- Building a good spam filter helps protect users from potential scams, unnecessary advertising, or malware links.

Evaluation

Training Set

Email	Label
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham

Test Set

Email	Label
You buy viagra!	Spam
You need viagra sir.	Spam
I hope you are healthy.	Ham
...	...
...	...

We "**train**" our spam filter on the training set, and **evaluate** performance using a test set (data that is unseen by the spam filter initially). This gives an unbiased estimate of performance.

Spam Filter Task

Training Set

Email	**Label**
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham

Predict whether this email is spam or ham:

You buy Viagra!

Emails as word collections

Email	**Set of Words in the Email**
SUBJECT: Top Secret Business Venture Dear Sir. First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret…	{top, secret, business, venture, dear, sir, first, I, must, solicit, your, confidence, in, this, transaction, is, by, virtue, of, its, nature, as, being, utterly, confidencial, and}
Hello hello hello there.	{hello, there}
You buy Viagra!	{you, buy, viagra}

For simplicity, we will

- Ignore Duplicate Words
- Ignore Punctuation
- Ignore Casing

Idea

Compute and Compare:

$$\mathbb{P}(\text{spam} \mid \text{"You buy Viagra!"}) \qquad \mathbb{P}(\text{ham} \mid \text{"You buy Viagra!"})$$

Then predict whichever is larger! Can we get away with just computing one of them?

Equivalently, note that these add to 1, so we can just compute $\mathbb{P}(\text{spam} \mid \text{"You buy Viagra!"})$

and if it is greater than 0.5, then we predict **spam**.

Otherwise, we predict **ham**.

<u>Note:</u> We resolve the tie in favor of **ham**.

Naive Bayes Classifier - The bayes part

Bayes Theorem:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$$

Apply it to our example:

$$\mathbb{P}(\text{spam} \mid \text{"You buy Viagra!"}) = \frac{\mathbb{P}(\text{"You buy Viagra!"} \mid \text{spam})\,\mathbb{P}\,(\text{spam})}{\mathbb{P}(\text{"You buy Viagra!"})}$$

Naive Bayes Classifier - What we Calculate

$$\mathbb{P}(\text{spam} \mid \text{"You buy Viagra!"}) = \frac{\mathbb{P}(\text{"You buy Viagra!"} \mid \text{spam})\ \mathbb{P}(\text{spam})}{\mathbb{P}(\text{"You buy Viagra!"})}$$

$$= \frac{\mathbb{P}(\{\text{"you"},\text{"buy"},\text{"viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam})}{\mathbb{P}(\{\text{"you"},\text{"buy"},\text{"viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam}) + \mathbb{P}(\{\text{"you"},\text{"buy"},\text{"viagra"}\} \mid \text{ham})\ \mathbb{P}(\text{ham})}$$

[LTP]

$$\mathbb{P}(\text{spam}) = \frac{\text{total spam emails (in training set)}}{\text{total emails (in training set)}} \qquad \mathbb{P}(\text{ham}) = \frac{\text{total ham emails (in training set)}}{\text{total emails (in training set)}}$$

(our approximation for these probabilities, based on the training set)

Naive Bayes Classifier - The naive part

It is somewhat unlikely that we have the email ”You buy Viagra!” in our training data. (In this case we don’t!)

We <u>naively</u> assume that words are conditionally independent from each other, given the label (In reality, they aren’t):

$$\mathbb{P}(\{\text{“you”},\ \text{“buy”},\ \text{“viagra”}\} \mid \text{spam})$$
$$\approx \mathbb{P}(\text{“you”} \mid \text{spam})\mathbb{P}(\text{“buy”} \mid \text{spam})\mathbb{P}(\text{“viagra”} \mid \text{spam})$$

Then we estimate for example that

$$\mathbb{P}(\text{“you”} \mid \text{spam}) = \frac{\text{number of spam emails containing “you” (in training set)}}{\text{number of spam emails (in training set)}}$$

Why is this Naive?

Consider for example the following two emails:

“!!!Lunch free for You!!!!!”

Spam

“You free for lunch?”

Ham

One shortfalling of our model is that it will make the same prediction for these since they have the same set of words!

$$\mathbb{P}\,(\text{spam} \mid \text{“You buy Viagra”})$$

Example

$$= \frac{\mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam})}{\mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam}) + \mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{ham})\ \mathbb{P}(\text{ham})}$$

$$= \frac{\mathbb{P}(\text{“you”} \mid \text{spam})\mathbb{P}(\text{“buy”} \mid \text{spam})\mathbb{P}(\text{“viagra”} \mid \text{spam})\mathbb{P}(\text{spam})}{\mathbb{P}(\text{“you”} \mid \text{spam})\mathbb{P}(\text{“buy”} \mid \text{spam})\mathbb{P}(\text{“viagra”} \mid \text{spam})\mathbb{P}(\text{spam}) + \mathbb{P}(\text{“you”} \mid \text{ham})\mathbb{P}(\text{“buy”} \mid \text{ham})\mathbb{P}(\text{“viagra”} \mid \text{ham})\mathbb{P}(\text{ham})}$$

Email	Label
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham

$$\mathbb{P}(\text{spam}) = \frac{3}{5} \qquad \mathbb{P}(\text{ham}) = \frac{2}{5}$$

$$\mathbb{P}(\text{"you"} \mid \text{spam}) = \frac{1}{3} \quad \mathbb{P}(\text{"you"} \mid \text{ham}) = \frac{1}{2}$$

$$\mathbb{P}(\text{"buy"} \mid \text{spam}) = \frac{1}{3} \quad \mathbb{P}(\text{“buy”} \mid \text{ham}) = 0$$

$$\mathbb{P}(\text{"viagra"} \mid \text{spam}) = 1 \quad \mathbb{P}(\text{"viagra"} \mid \text{ham}) = \frac{1}{2}$$

$$\mathbb{P}\left(\text{spam} \mid \text{“You buy Viagra”}\right)$$

Example

$$= \frac{\mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam})}{\mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam}) + \mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{ham})\ \mathbb{P}(\text{ham})}$$

$$= \frac{\mathbb{P}(\text{“you”} \mid \text{spam})\mathbb{P}(\text{“buy”} \mid \text{spam})\mathbb{P}(\text{“viagra”} \mid \text{spam})\mathbb{P}(\text{spam})}{\mathbb{P}(\text{“you”} \mid \text{spam})\mathbb{P}(\text{“buy”} \mid \text{spam})\mathbb{P}(\text{“viagra”} \mid \text{spam})\mathbb{P}(\text{spam}) + \cancel{\mathbb{P}(\text{“you”} \mid \text{ham})\mathbb{P}(\text{“buy”} \mid \text{ham})\mathbb{P}(\text{“viagra”} \mid \text{ham})\mathbb{P}(\text{ham})}}$$

= 1 (Marked as spam since no ham email contained “buy”)

Email	Label
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham

$$\mathbb{P}(\text{spam}) = \frac{3}{5} \qquad \mathbb{P}(\text{ham}) = \frac{2}{5}$$

$$\mathbb{P}(\text{"you"} \mid \text{spam}) = \frac{1}{3} \qquad \mathbb{P}(\text{"you"} \mid \text{ham}) = \frac{1}{2}$$

$$\mathbb{P}(\text{"buy"} \mid \text{spam}) = \frac{1}{3} \qquad \mathbb{P}(\text{“buy”} \mid \text{ham}) = 0$$

$$\mathbb{P}(\text{"viagra"} \mid \text{spam}) = 1 \qquad \mathbb{P}(\text{"viagra"} \mid \text{ham}) = \frac{1}{2}$$

What happen if we got a 0?

P(ham | “You buy Viagra!”) = 0 since P(“buy”| ham) = 0, since no ham email in our training data contained the word **‘buy’**.

But does that mean we will never encounter a ham email with word **‘buy’**?

Laplace smoothing

Pretend in spam emails (training set):

- We saw one extra spam email **with** word w_i
- We saw one extra spam email **without** word w_i

$$\mathbb{P}(w_i \mid \text{spam}) = \frac{\left|\text{total spam emails (training set) containing } w_i\right| + 1}{\left|\text{total spam emails (training set)}\right| + 2}$$

Same for ham emails:

$$\mathbb{P}(w_i \mid \text{ham}) = \frac{\left|\text{total ham emails (training set) containing } w_i\right| + 1}{\left|\text{total ham emails (training set)}\right| + 2}$$

$$\mathbb{P}(\text{spam} \mid \text{"You buy Viagra"})$$

Example

$$= \frac{\mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam})}{\mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{spam})\ \mathbb{P}(\text{spam}) + \mathbb{P}(\{\text{"you","buy","viagra"}\} \mid \text{ham})\ \mathbb{P}(\text{ham})}$$

$$= \frac{\mathbb{P}(\text{"you"} \mid \text{spam})\mathbb{P}(\text{"buy"} \mid \text{spam})\mathbb{P}(\text{"viagra"} \mid \text{spam})\mathbb{P}(\text{spam})}{\mathbb{P}(\text{"you"} \mid \text{spam})\mathbb{P}(\text{"buy"} \mid \text{spam})\mathbb{P}(\text{"viagra"} \mid \text{spam})\mathbb{P}(\text{spam}) + \mathbb{P}(\text{"you"} \mid \text{ham})\mathbb{P}(\text{"buy"} \mid \text{ham})\mathbb{P}(\text{"viagra"} \mid \text{ham})\mathbb{P}(\text{ham})}$$

$$= \frac{\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}}{\frac{2}{5} \cdot \frac{2}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} + \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{2}{5}} \approx 0.7544$$

Email	Label
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham

$$\mathbb{P}(\text{spam}) = \frac{3}{5} \qquad \mathbb{P}(\text{ham}) = \frac{2}{5}$$

$$\mathbb{P}(\text{"you"} \mid \text{spam}) = \frac{1+1}{3+2} = \frac{2}{5} \qquad \mathbb{P}(\text{"you"} \mid \text{ham}) = \frac{1+1}{2+2} = \frac{1}{2}$$

$$\mathbb{P}(\text{"buy"} \mid \text{spam}) = \frac{1+1}{3+2} = \frac{2}{5} \qquad \mathbb{P}(\text{"buy"} \mid \text{ham}) = \frac{0+1}{2+2} = \frac{1}{4}$$

$$\mathbb{P}(\text{"viagra"} \mid \text{spam}) = \frac{3+1}{3+2} = \frac{4}{5} \qquad \mathbb{P}(\text{"viagra"} \mid \text{ham}) = \frac{1+1}{2+2} = \frac{1}{2}$$

Underflow Prevention

- Multiplication of many probabilities, each of which will be between 0 and 1, can result in floating-point underflow. The product will be too small and will result in arithmetic underflow.

- Reminder: Log property:

$$\log(xy) = \log(x) + \log(y)$$

- Summing logs of probabilities is better than multiplying probabilities

$$\begin{aligned} \log\left(\prod_{i=1}^{n} p_i\right) = \log(p_1 p_2 \ldots p_n) &= \log(p_1) + \log(p_2) + \ldots + \log(p_n) \\ &= \sum_{i=1}^{n} \log(p_i) \end{aligned}$$

Applying underflow prevention

$$\mathbb{P}(\textbf{spam} \mid \{w_1, w_2, \ldots, w_n\}) \approx \frac{\mathbb{P}(\{w_1,w_2,\ldots,w_n\}\mid \textbf{spam})\ \mathbb{P}(\textbf{spam})}{\mathbb{P}(\{w_1,w_2,\ldots,w_n\}\mid \textbf{spam})\ \mathbb{P}(\textbf{spam})+\mathbb{P}(\{w_1,w_2,\ldots,w_n\}\mid \textbf{ham})\ \mathbb{P}(\textbf{ham})}$$

$$\mathbb{P}(\textbf{ham} \mid \{w_1, w_2, \ldots, w_n\}) \approx \frac{\mathbb{P}(\{w_1,w_2,\ldots,w_n\}\mid \textbf{ham})\ \mathbb{P}(\textbf{ham})}{\mathbb{P}(\{w_1,w_2,\ldots,w_n\}\mid \textbf{spam})\ \mathbb{P}(\textbf{spam})+\mathbb{P}(\{w_1,w_2,\ldots,w_n\}\mid \textbf{ham})\ \mathbb{P}(\textbf{ham})}$$

We will output **spam** iff:

$$\mathbb{P}(\text{spam} \mid \{w_1, w_2, \ldots, w_n\}) > \mathbb{P}(\text{ham} \mid \{w_1, w_2, \ldots, w_n\})$$

$$\iff \mathbb{P}(\{w_1, w_2, \ldots, w_n\} \mid \text{spam})\mathbb{P}(\text{spam}) > \mathbb{P}(\{w_1, w_2, \ldots, w_n\} \mid \text{ham})\mathbb{P}(\text{ham})$$

$$\iff \mathbb{P}(w_1 \mid \text{spam})\mathbb{P}(w_2 \mid \text{spam}) \cdots \mathbb{P}(w_n \mid \text{spam})\mathbb{P}(\text{spam}) > \mathbb{P}(w_1 \mid \text{ham})\mathbb{P}(w_2 \mid \text{ham}) \cdots \mathbb{P}(w_n \mid \text{ham})\mathbb{P}(\text{ham})$$

Taking the log of two sides:

$$\iff \log(\mathbb{P}(\text{spam})) + \sum_{i=1}^{n} \log(\mathbb{P}(w_i \mid \text{spam}) > \log(\mathbb{P}(\text{ham})) + \sum_{i=1}^{n} \log(\mathbb{P}(w_i \mid \text{ham})$$

Summary: Naive Bayes Algorithm steps

1. **TRAINING**

   1.1. Compute the proportion of emails in the **training set** that is spam or ham:

$$\mathbb{P}(\text{spam}) = \frac{\text{total spam emails (in training set)}}{\text{total emails (in training set)}}$$

$$\mathbb{P}(\text{ham}) = \frac{\text{total ham emails (in training set)}}{\text{total emails (in training set)}}$$

   1.2. Iterate over the **training set**, for each unique word **x**, count:

- How many **spam emails** in the training set contain **x**
- How many **ham emails** in the training set contain **x**

Summary: Naive Bayes Algorithm steps

2. **TESTING**

Iterate over the **test set**, for each unlabelled email **D**:

- Create a set **S** of **n** unique words appearing in **D**: $\{w_1, w_2, \ldots, w_n\}$
- For each word w_i in set **S**, calculate:

$$\mathbb{P}(x \mid \text{spam}) = \frac{|\text{total spam emails (training set) containing } w_i| + 1}{|\text{total spam emails (training set)}| + 2}$$

$$\mathbb{P}(w_i \mid \text{ham}) = \frac{|\text{total ham emails (training set) containing } w_i| + 1}{|\text{total ham emails (training set)}| + 2}$$

  - Note: If word w_i doesn't appear in the training set, we still calculate the above probabilities, with:

$$|\text{total spam emails (training set) containing } w_i| = |\text{total ham emails (training set) containing } w_i| = 0$$

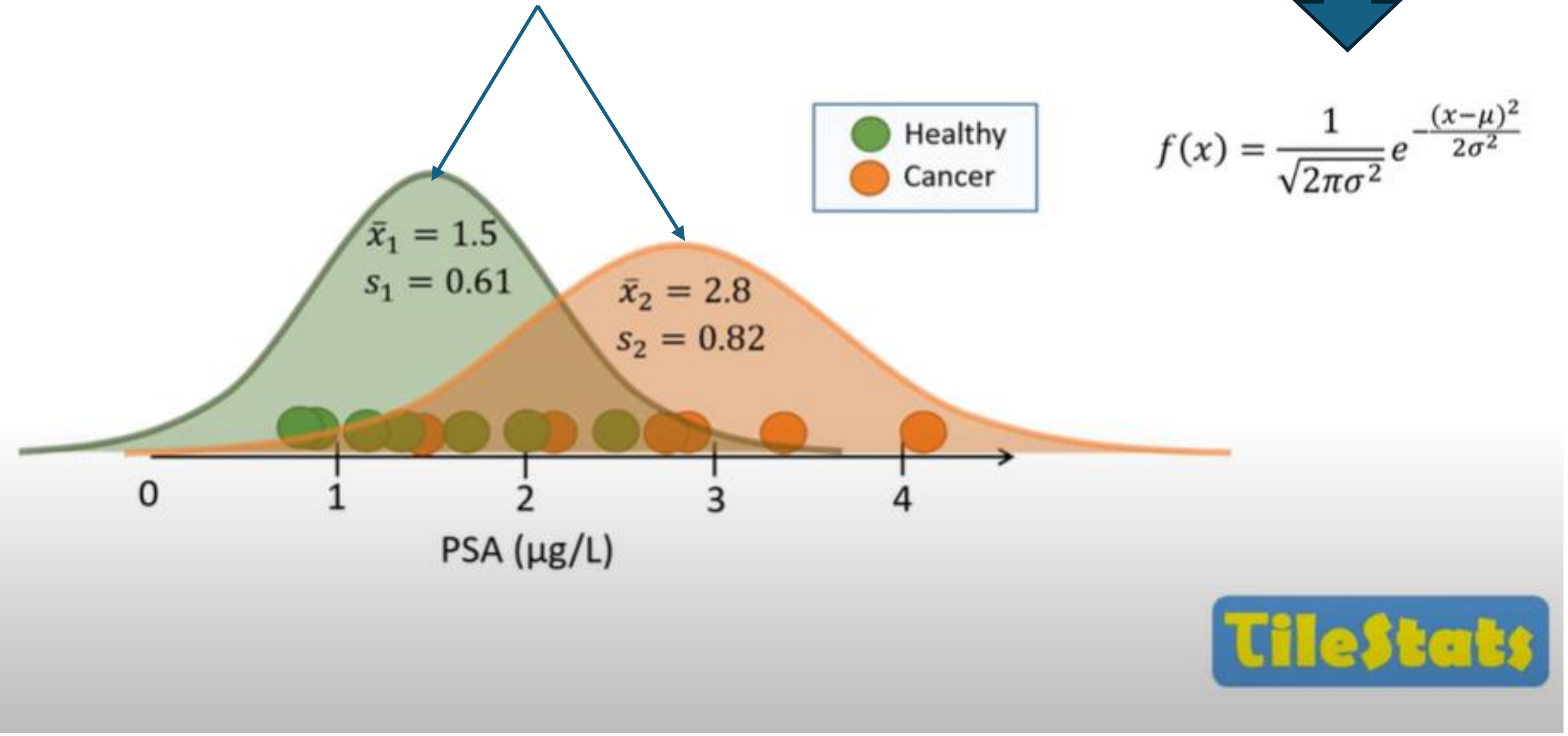
- if $\log(\mathbb{P}(\text{spam})) + \sum_{i=1}^{n} \log(\mathbb{P}(w_i \mid \text{spam})) > \log(\mathbb{P}(\text{ham})) + \sum_{i=1}^{n} \log(\mathbb{P}(w_i \mid \text{ham}))$

  Predict email D as **spam**
  Otherwise, predict email D as **ham**

Dealing with continuous features

Status	PSA
Cancer	4.1
Cancer	3.4
Cancer	2.9
Cancer	2.8
Cancer	2.7
Cancer	2.1
Cancer	1.6
Healthy	2.5
Healthy	2.0
Healthy	1.7
Healthy	1.4
Healthy	1.2
Healthy	0.9
Healthy	0.8

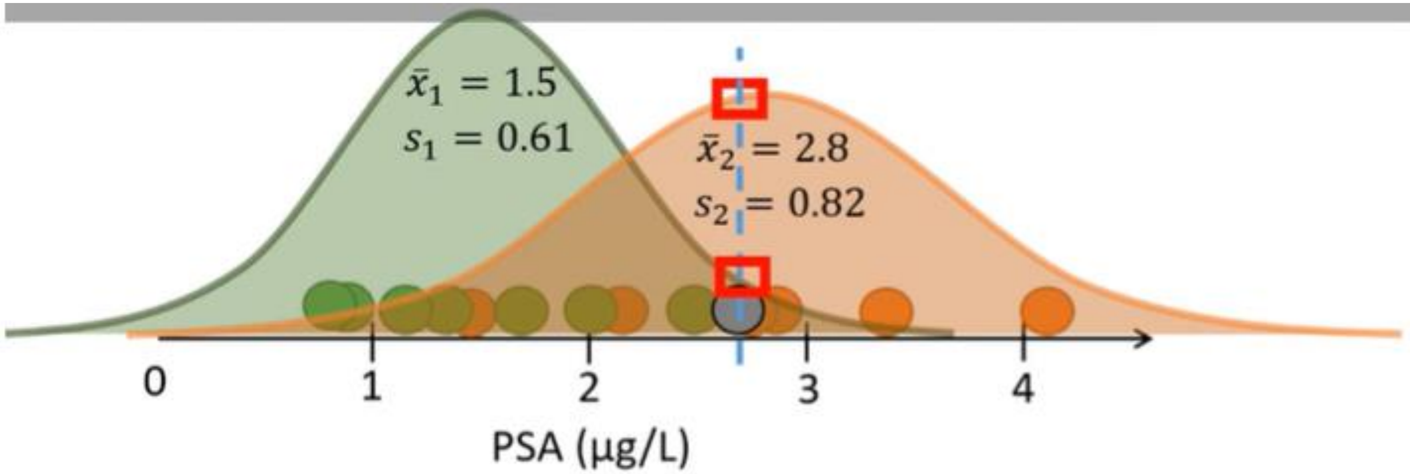
Gaussian Naïve Bayes



$\boldsymbol{P(PSA|Status)}$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian Naïve Bayes – one feature



$p(PSA = 2.6|Cancer) = 0.47$

$p(PSA = 2.6|Healthy) = 0.13$

$p(Cancer) = 0.5$

$p(Healthy) = 0.5$

$$\boxed{\begin{array}{l} posterior\ numerator(Healthy) = p(Healthy)p(PSA|Healthy) = 0.5 \cdot 0.13 = 0.065 \\ posterior\ numerator(Cancer) = p(Cancer)p(PSA|Cancer) = 0.5 \cdot 0.47 = 0.235 \end{array}}$$

Gaussian Naïve Bayes – two feature

Status	PSA	Age
Cancer	4.1	78
Cancer	3.4	70
Cancer	2.9	62
Cancer	2.8	66
Cancer	2.7	70
Cancer	2.1	65
Cancer	1.6	58
Healthy	2.5	68
Healthy	2.0	64
Healthy	1.7	62
Healthy	1.4	70
Healthy	1.2	72
Healthy	0.9	67
Healthy	0.8	59

$$PSA = 2.6$$
$$Age = 70$$

$\bar{x}_{PSA2} = 2.8$

$s_{PSA2} = 0.82$

$\bar{x}_{Age2} = 67$

$s_{Age2} = 6.45$

$$p(Age|Healthy) = \frac{1}{\sqrt{2\pi \cdot 4.58^2}} e^{-\frac{(70-66)^2}{2 \cdot 4.58^2}} = 0.059$$

$\bar{x}_{PSA1} = 1.5$

$s_{PSA1} = 0.61$

$\bar{x}_{Age1} = 66$

$s_{Age1} = 4.58$

$$posterior\ numerator(Healthy) = p(Healthy)p(PSA|Healthy)p(Age|Healthy)$$
$$= 0.5 \cdot 0.13 \cdot 0.059 = 0.004$$

$$posterior\ numerator(Cancer) = p(Cancer)p(PSA|Cancer)p(Age|Cancer)$$
$$= 0.5 \cdot 0.47 \cdot 0.055 = 0.013$$