CS 481

Artificial Intelligence Language Understanding

February 23, 2023

Announcements / Reminders

- **Please follow the Week 07 To Do List instructions**
- **PA #01 due on ~~Monday (02/20/23) at 11:59 PM CST~~ Thursday (02/23/23) at 11:59 PM CST**
- **Written Assignment #02 due on Thursday (03/02/23) at 11:59 PM CST**

- **Exam dates:**
  - **Midterm: 03/02/2023 during Thursday lecture time**
  - **Final: 04/27/2023 during Thursday lecture time**

Plan for Today

- **Naïve Bayes classifier**

Bag of Words: Document Vector

Pre-defined Vocabulary:							
Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	...	Word N

Document A Non-binary Vector [0-word absent | >0-word count]:							
6	0	2	3	1	0	. . .	4

Document B Non-binary Vector [0-word absent | >0-word count]:							
4	2	0	0	5	0	. . .	1

Document C Non-binary Vector [0-word absent | >0-word count]:							
0	0	3	0	0	7	. . .	0

Document vectors can be used to compare documents.

Bag of Words: Classification

$$\textbf{category} = h\left(\begin{array}{|c|}\hline 6 \\ \hline 5 \\ \hline 4 \\ \hline 3 \\ \hline 3 \\ \hline 2 \\ \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline \ldots \\ \hline \end{array}\right)$$

Learned Classifier model (hypothesis)

Bayes' Rule

$$P(y \mid x) = \frac{P(x \mid y) * P(y)}{P(x)}$$

$$P(Category \mid Document) = \frac{P(Document \mid Category) * P(Category)}{P(Document)}$$

$$P(Category \mid Instance) = \frac{P(Instance \mid Category) * P(Category)}{P(Instance)}$$

$$P(Category \mid Sample) = \frac{P(Sample \mid Category) * P(Category)}{P(Sample)}$$

Classification: Conditional Probability

$$\boldsymbol{P}(y \mid \boldsymbol{x}) = \frac{\boldsymbol{P}(\boldsymbol{x} \mid y) * \boldsymbol{P}(y)}{\boldsymbol{P}(\boldsymbol{x})}$$

$\mathbf{x} = x_1, x_2, \ldots, x_N$, **so:**

How to calculate?

$$\boldsymbol{P}(y \mid x_1 \wedge x_2 \wedge \ldots \wedge x_N) = \frac{\boxed{\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y)} * \boldsymbol{P}(y)}{\boxed{\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N)}}$$

constant

Naive Bayes Assumption

$\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N \wedge y) =$
$\boldsymbol{P}(x_1 \mid x_2 \wedge \ldots \wedge x_N \wedge y) * \boldsymbol{P}(x_2 \wedge \ldots \wedge x_N \wedge y) =$
$\boldsymbol{P}(x_1 \mid x_2 \wedge \ldots \wedge x_N \wedge y) * \boldsymbol{P}(x_2 \mid x_3 \wedge \ldots \wedge x_N \wedge y) * \boldsymbol{P}(x_3 \wedge \ldots \wedge x_N \wedge y) =$
$\boldsymbol{P}(x_1 \mid x_2 \wedge \ldots \wedge x_N \wedge y) * \boldsymbol{P}(x_2 \mid x_3 \wedge \ldots \wedge x_N \wedge y) * \boldsymbol{P}(x_3 \mid x_4 \wedge \ldots \wedge x_N \wedge y) * \boldsymbol{P}(x_3 \wedge \ldots \wedge x_N \wedge y) =$
$\ldots$
$\boldsymbol{P}(x_1 \mid x_2 \wedge \ldots \wedge x_N \wedge y) * \boldsymbol{P}(x_2 \mid x_3 \wedge \ldots \wedge x_N \wedge y) * \ldots * \boldsymbol{P}(x_N \mid y) * \boldsymbol{P}(y)$

Now let's assume that all events $x_1, x_2, \ldots, x_N$ are mutually independent (not true in reality) and conditionally independent given y → Naive Bayes assumption.

Under this assumption:

$$\boldsymbol{P}(x_i \mid x_{i+1} \wedge \ldots \wedge x_N \wedge y) = \boldsymbol{P}(x_i \mid y)$$

Naive Bayes Assumption

Under Naive Bayes assumption:

$$P(x_1 \wedge x_2 \wedge ... \wedge x_N \wedge y) =$$
$$P(x_1 \mid x_2 \wedge ... \wedge x_N \wedge y) * P(x_2 \wedge ... \wedge x_N \wedge y) =$$
$$P(x_1 \mid x_2 \wedge ... \wedge x_N \wedge y) * P(x_2 \mid x_3 \wedge ... \wedge x_N \wedge y) * P(x_3 \wedge ... \wedge x_N \wedge y) =$$
$$P(x_1 \mid x_2 \wedge ... \wedge x_N \wedge y) * P(x_2 \mid x_3 \wedge ... \wedge x_N \wedge y) * P(x_3 \mid x_4 \wedge ... \wedge x_N \wedge y) * P(x_3 \wedge ... \wedge x_N \wedge y) =$$
$$...$$
$$P(x_1 \mid x_2 \wedge ... \wedge x_N \wedge y) * P(x_2 \mid x_3 \wedge ... \wedge x_N \wedge y) * ... * P(x_N \mid y) * P(y)$$

becomes:

$$P(x_1 \wedge x_2 \wedge ... \wedge x_N \wedge y) =$$
$$P(x_1 \mid y) * P(x_2 \mid y) * P(x_3 \mid y) * ... * P(x_{N-1} \mid y) * P(x_N \mid y) * P(y) =$$
$$P(y) * [P(x_1 \mid y) * P(x_2 \mid y) * P(x_3 \mid y) * ... * P(x_{N-1} \mid y) * P(x_N \mid y)] =$$
$$P(y) * \prod_{i=1}^{N} P(x_i \mid y)$$

Naive Bayes Classifier

Under Naive Bayes assumption:

$$y_{\boldsymbol{MAP}} \propto \underset{y \in Y}{argmax} \left(\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y) * \boldsymbol{P}(y)\right)$$

becomes:

$$y_{\boldsymbol{MAP}} \propto \underset{y \in Y}{argmax} \left(\boldsymbol{P}(y) * \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{P}(x_i \mid y)\right)$$

MAP: Maximum a posteriori (corresponds to the most likely class).

Naive Bayes Classifier

Under Naive Bayes assumption:

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(\boxed{P(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y)} * P(y)\right)$$

becomes:

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(P(y) * \boxed{\prod_{i=1}^{N} P(x_i \mid y)} \right)$$

MAP: Maximum a posteriori (corresponds to the most likely class).

Naive Bayes Classifier

Under Naive Bayes assumption:

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(P(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y) * P(y)\right)$$

becomes:

How to calculate?

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(P(y) * \boxed{\prod_{i=1}^{N} P(x_i \mid y)}\right)$$

MAP: Maximum a posteriori (corresponds to the most likely class).

Text Classification: Supervised ML

Input:

- a document $\boldsymbol{x}$
- a fixed set of classes $Y = \{y_1, y_2,\ldots, y_J\}$
- a training set of N hand-labeled documents $(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_N,y_N)$

Output:

- a learned classifier $h:\boldsymbol{x} \rightarrow y$ ($y = h(\boldsymbol{x})$)

Text Classification: Classifier

$$\text{category/class} = h(\mathbf{document})$$

Learned Classifier model (hypothesis)

Text Classification: Classifier

$$y = \color{red}{h}(\mathbf{x})$$

Learned Classifier model (hypothesis)

Text Classification: Supervised ML

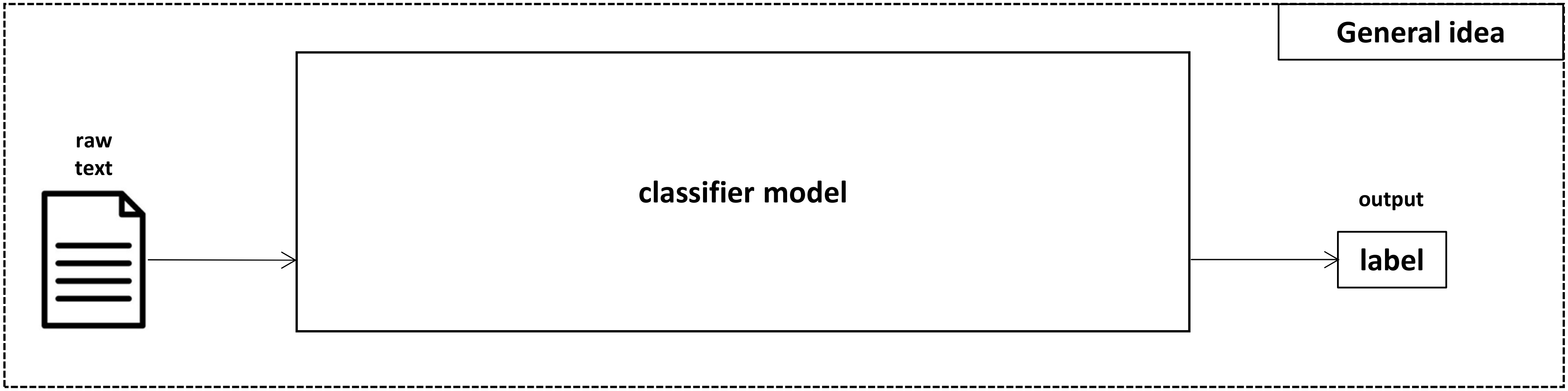
Input:

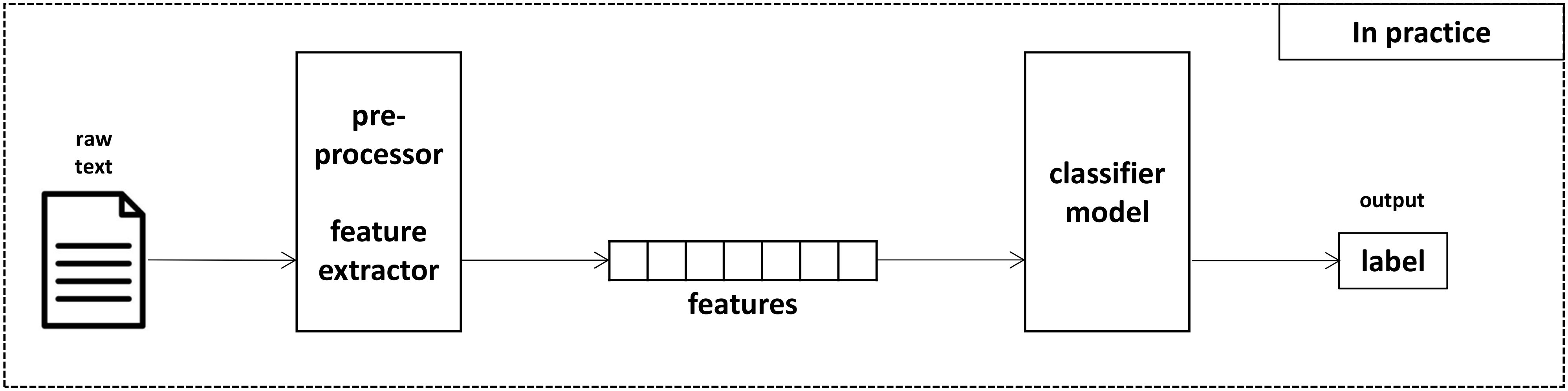
- a document $\boldsymbol{x}$
- a fixed set of classes $Y = \{y_1, y_2, \ldots, y_J\}$
- a **training set** of N hand-labeled documents $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$

Output:

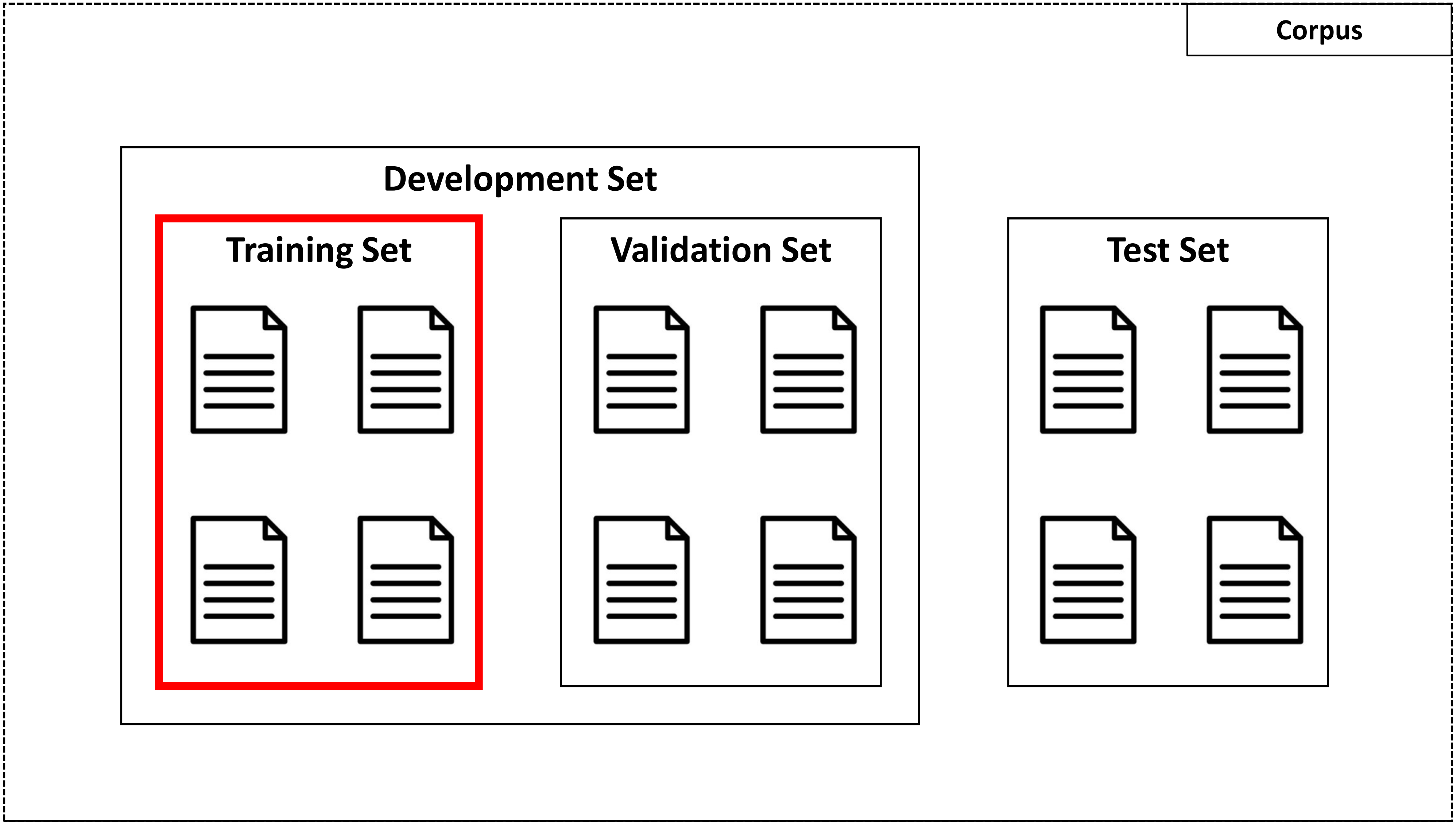
- a learned classifier $h: \boldsymbol{x} \rightarrow y$ ($y = h(\boldsymbol{x})$)

Text Classification: the Idea



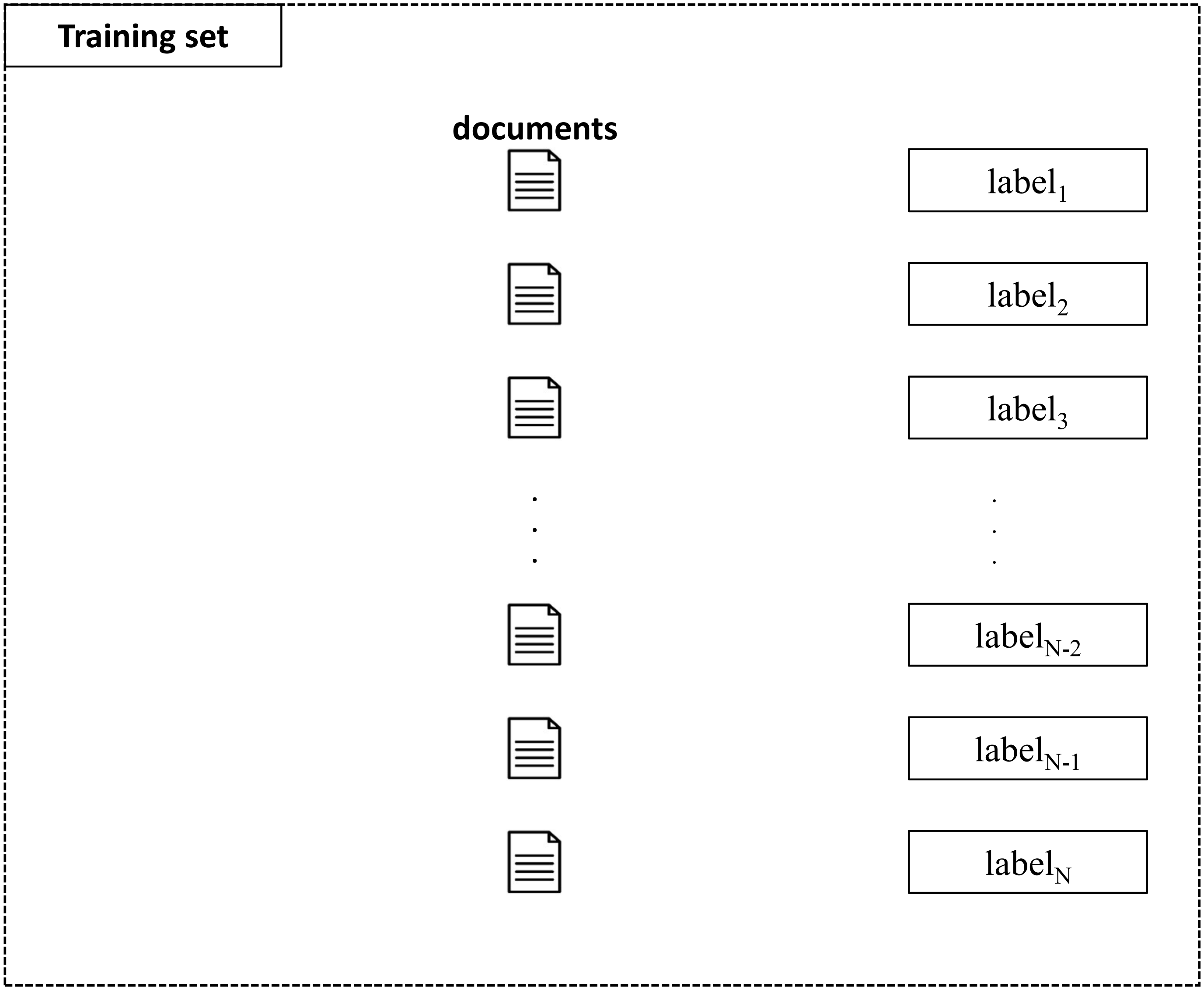


Corpus: Training / Validation / Test

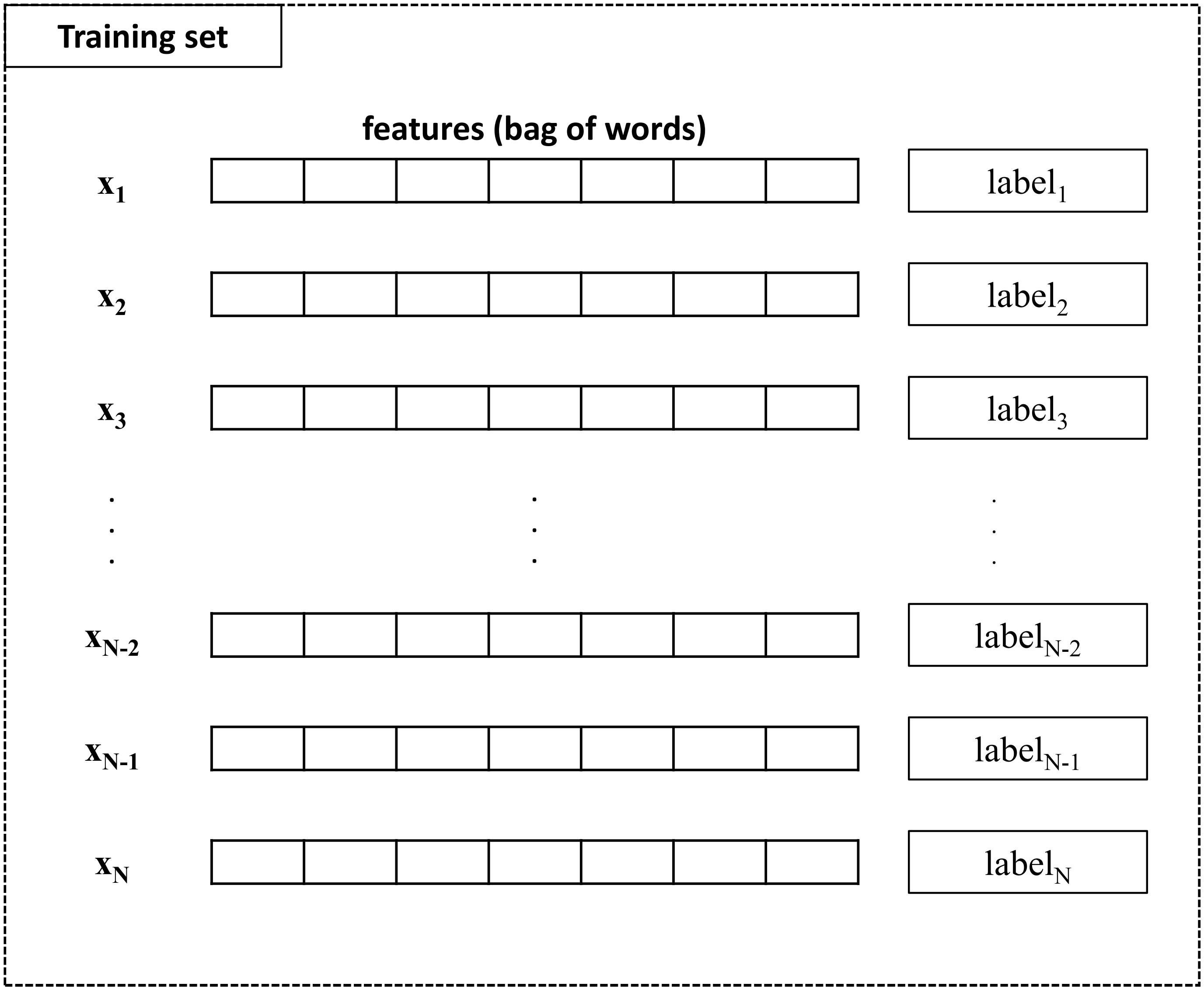


Typical training / validation / test set split for a text corpora

Text Classification: Training Set

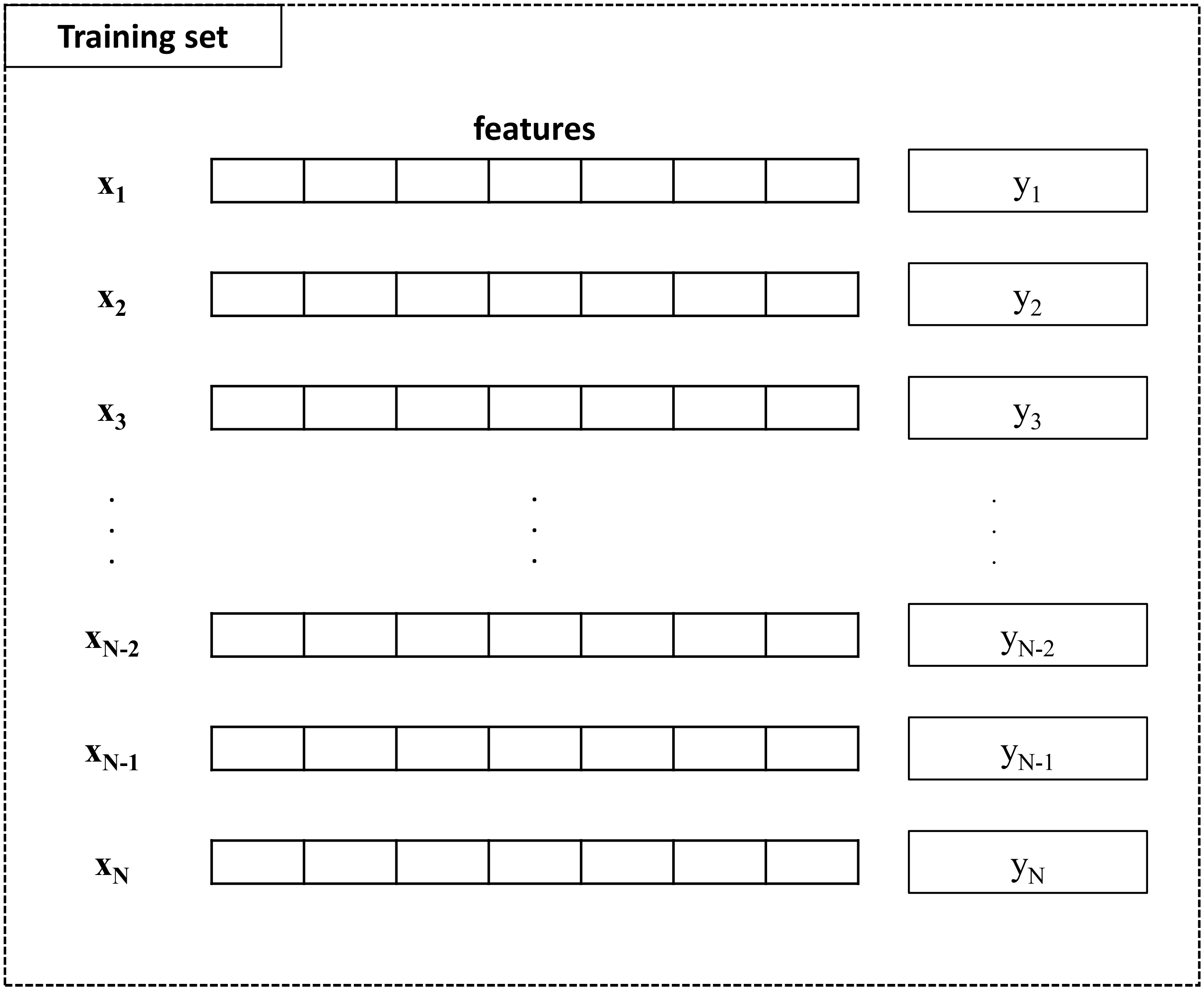


Text Classification: Training Set



$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**)

Text Classification: Training Set



$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Training Set

Training set

Vocabulary V

	word1	rolex	word3	replica	word5	word6	word7	
$\mathbf{x}_1$	0	0	1	0	1	1	1	y_1=HAM
$\mathbf{x}_2$	1	0	1	1	0	1	1	y_2=HAM
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM
⋮				⋮				⋮
$\mathbf{x}_{N-2}$	1	1	1	1	0	1	1	y_{N-2}=HAM
$\mathbf{x}_{N-1}$	1	1	0	1	0	0	1	y_{N-1}=SPAM
$\mathbf{x}_N$	1	0	0	1	0	0	1	y_N=HAM

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \dots, y_{N-2}, y_{N-1}, y_N$ - labels

Text Classification: Bag of Words

Bag of words document representation (feature vector)

$$\text{category/class} = h\left(\begin{array}{|c|}\hline 6 \\ \hline 5 \\ \hline 4 \\ \hline 3 \\ \hline 3 \\ \hline 2 \\ \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline 2 \\ \hline\end{array}\right)$$

Learned Classifier model (hypothesis)

Text Classification: Bag of Words

Bag of words binary document representation (feature vector)

$$\text{category/class} = h\left(\begin{array}{|c|}\hline 1\\\hline 1\\\hline 0\\\hline 1\\\hline 0\\\hline 0\\\hline 1\\\hline 1\\\hline 1\\\hline 0\\\hline\end{array}\right)$$

Learned Classifier model (hypothesis)

Text Classification: Bag of Words

Bag of words document representation (feature vector)

$$\text{category/class} = h(\boxed{}\boxed{}\boxed{}\boxed{}\boxed{}\boxed{}\boxed{})$$

Learned Classifier model (hypothesis)

Spam Detection: Learning

Training set

Vocabulary V

	word1	rolex	word3	replica	word5	word6	word7	
$\mathbf{x}_1$	0	0	1	0	1	1	1	y_1=HAM
$\mathbf{x}_2$	1	0	1	1	0	1	1	y_2=HAM
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM
. . .				. . .				. . .
$\mathbf{x}_{N-2}$	1	1	1	1	0	1	1	y_{N-2}=HAM
$\mathbf{x}_{N-1}$	1	1	0	1	0	0	1	y_{N-1}=SPAM
$\mathbf{x}_N$	1	0	0	1	0	0	1	y_N=HAM

Learning

Naive Bayes Classifier:

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(\boldsymbol{P}(y) * \prod_{i=1}^{N} \boldsymbol{P}(x_i \mid y) \right)$$

Probability estimates (Maximium Likelihood estimation):

$$P(y_k) = \frac{N_{samples\ labeled\ y_k}}{N}$$

$$P(x_i \mid y_k) = \frac{count(x_i, y_k)}{\sum_{x \in V} count(x, y_k)}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Learning

Training set

Vocabulary V

	word1	rolex	word3	replica	word5	word6	word7	label
$\mathbf{x}_1$	0	0	1	0	1	1	1	y_1=HAM
$\mathbf{x}_2$	1	0	1	1	0	1	1	y_2=HAM
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM
$\mathbf{x}_4$	1	1	1	1	0	0	0	y_4=HAM
$\mathbf{x}_5$	1	1	1	1	0	1	1	y_5=HAM
$\mathbf{x}_6$	1	1	0	1	0	0	1	y_6=SPAM
$\mathbf{x}_7$	1	0	0	1	0	0	1	y_7=HAM

Learning

Naive Bayes Classifier:

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(\boldsymbol{P}(y) * \prod_{i=1}^{N} \boldsymbol{P}(x_i \mid y) \right)$$

Probability estimates (Maximium Likelihood estimation):

$$P(y = HAM) = \frac{N_{samples\ labeled\ HAM}}{N} = \frac{5}{7}$$

$$P(y = SPAM) = \frac{N_{samples\ labeled\ SPAM}}{N} = \frac{2}{7}$$

$$P(x_i = rolex \mid y = SPAM) = \\ = \frac{count(x_i = rolex, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{2}{8}$$

and so on...

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Learning

Training set

Vocabulary $\mathbf{V}$

	word1	rolex	word3	replica	word5	word6	word7	
$\mathbf{x}_1$	0	0	1	0	1	1	1	y_1=HAM
$\mathbf{x}_2$	1	0	1	1	0	1	1	y_2=HAM
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM
. . .				. . .				. . .
$\mathbf{x}_{N-2}$	1	1	1	1	0	1	1	y_{N-2}=HAM
$\mathbf{x}_{N-1}$	1	1	0	1	0	0	1	y_{N-1}=SPAM
$\mathbf{x}_N$	1	0	0	1	0	0	1	y_N=HAM

Learning

Naive Bayes Classifier:

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(\boldsymbol{P}(y) * \prod_{i=1}^{N} \boldsymbol{P}(x_i \mid y) \right)$$

Probability estimates:

$$P(y_k) = \frac{N_{samples\ labeled\ y_k}}{N}$$

or

- **equiprobable (all classes have equal probability)**

$$P(y = HAM) = P(y = SPAM) = 0.5$$

- **can be determined by experts in the area**

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Classifier

$$y_{MAP} = \underset{y \in Y}{argmax}\left(\boldsymbol{P}(y \mid \boldsymbol{x})\right) = \underset{y \in Y}{argmax}\left(\frac{\boldsymbol{P}(\boldsymbol{x} \mid y) * \boldsymbol{P}(y)}{\boldsymbol{P}(\boldsymbol{x})}\right)$$

$$\mathbf{x} = x_1, x_2, \ldots, x_N, \textbf{ so:}$$

$$y_{MAP} = \underset{y \in Y}{argmax}\left(\frac{\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y) * \boldsymbol{P}(y)}{\boxed{\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N)}}\right)$$

constant | we can drop

$$y_{MAP} \propto \underset{y \in Y}{argmax}\left(\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y) * \boldsymbol{P}(y)\right)$$

MAP: Maximum a posteriori (corresponds to the most likely class).

Naive Bayes Classifier: Assumptions

- **All events (words) $x_1, x_2, ..., x_N$ are mutually independent**
  - **Bag-of-words representation: the order of the words in a document d makes no difference (repetitions do)**
- **All events (words) $x_1, x_2, ..., x_N$ are conditionally independent given y (category / class)**
  - **words appear independently of each other, given the document category / class y (e.g. if you see word "*car*", the word "*drive*" is no more likely to appear than if you saw "*dog*")**

Naive Bayes Classifier

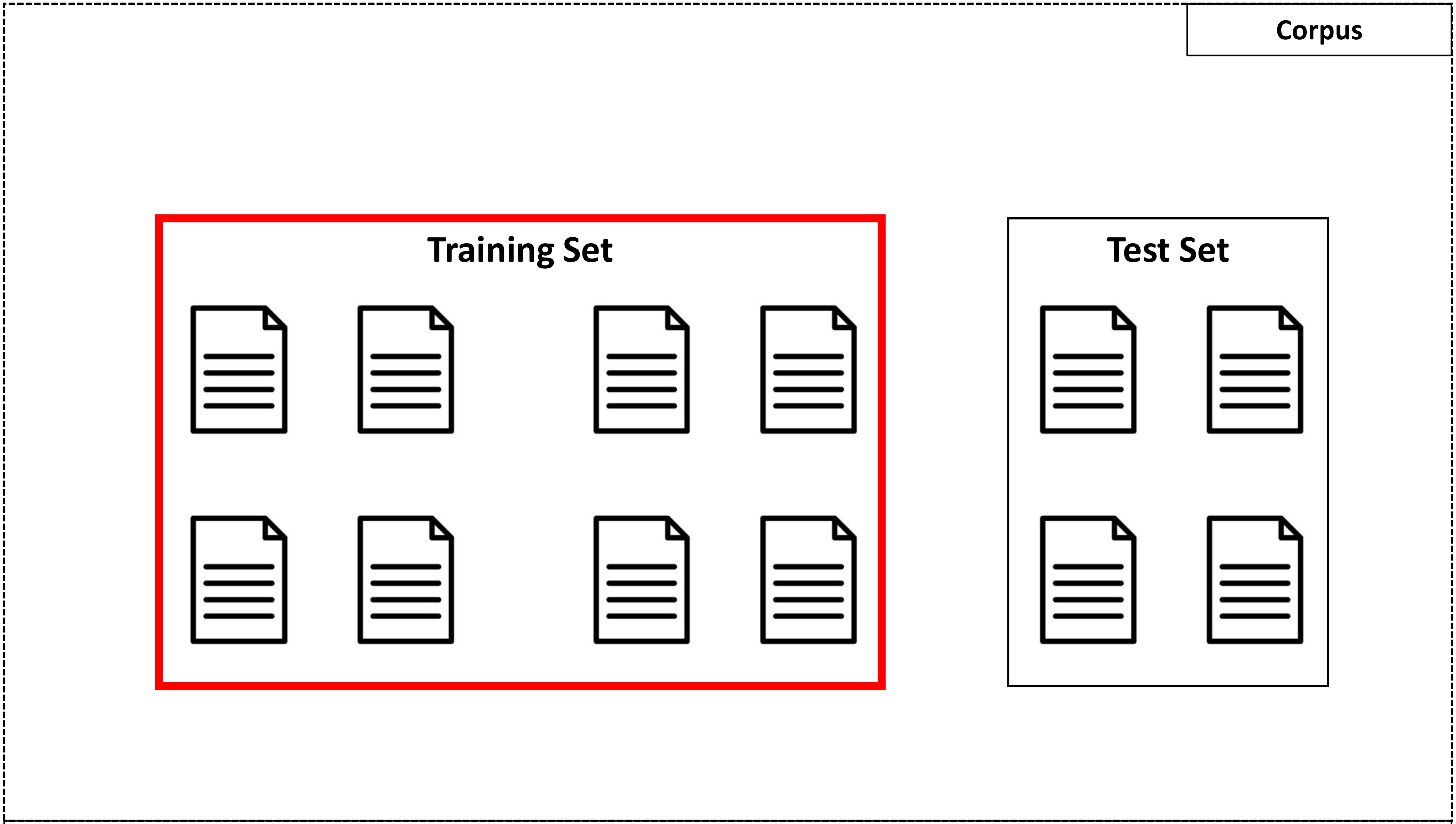
category/class = h(**document**)

Finding model / hypothesis h $\rightarrow$ **Finding probabilities for** y_{MAP}

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(\boxed{P(y)} * \prod_{i=1}^{N} \boxed{P(x_i \mid y)} \right)$$

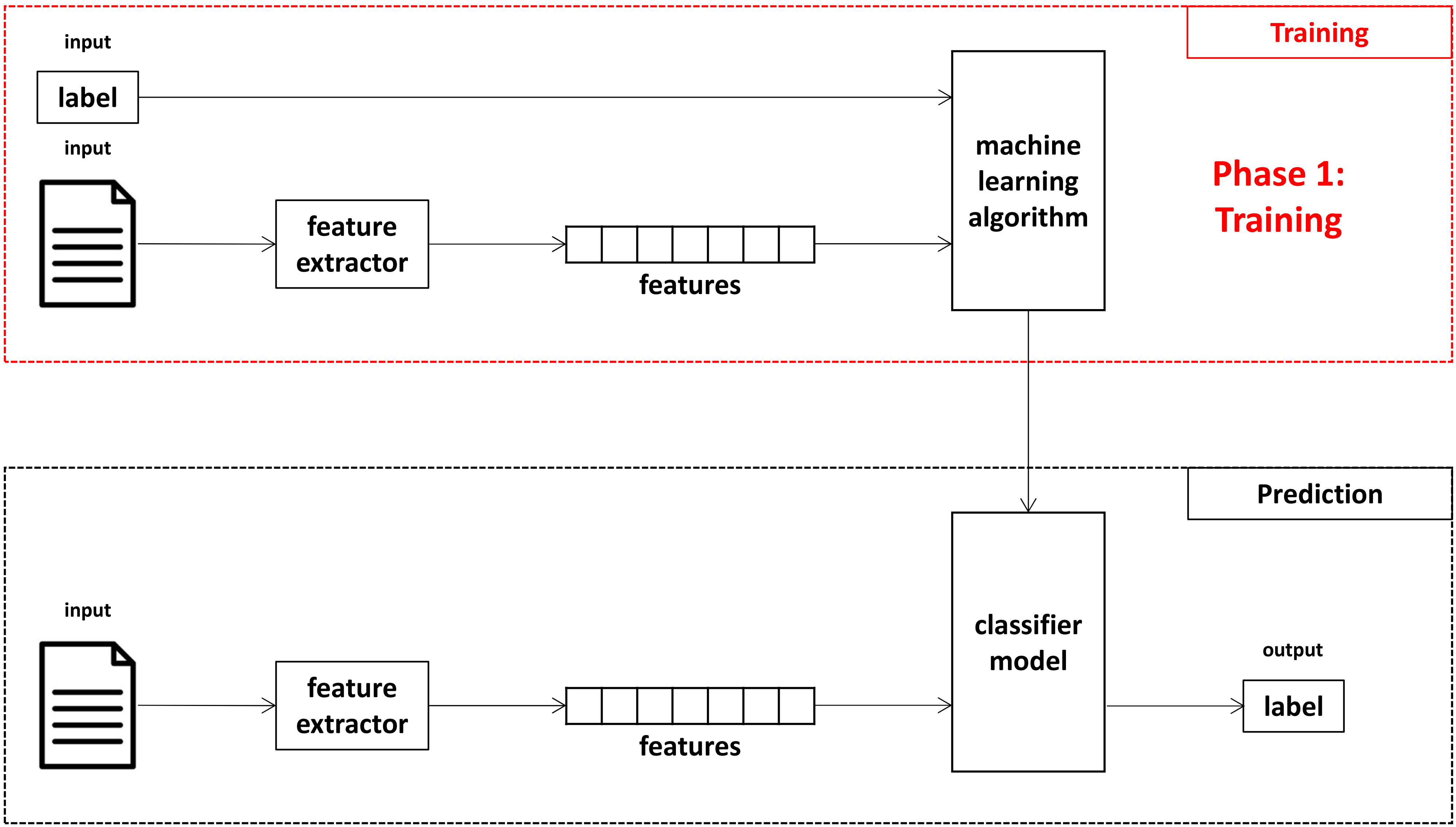
MAP: Maximum a posteriori (corresponds to the most likely class).

Corpus: Training / Test



Typical training / test set split for a text corpora [I will ignore validation set for the sake of an example]

Supervised Learning with ML



Spam Detection: Training Set

Training set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	Label	Text
$\mathbf{x}_1$	1	1	1	0	1	0	0	y_1=HAM	I own rolex watch
$\mathbf{x}_2$	1	0	1	0	1	0	0	y_2=HAM	I own watch
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM	buy cheap rolex replica
$\mathbf{x}_4$	1	0	1	0	0	0	0	y_4=HAM	I own
$\mathbf{x}_5$	1	0	1	1	0	0	1	y_5=HAM	I own cheap replica
$\mathbf{x}_6$	0	1	0	1	0	0	1	y_6=SPAM	cheap rolex replica
$\mathbf{x}_7$	1	0	0	0	1	0	0	y_7=HAM	I watch

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Learning

Training set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_1$	1	1	1	0	1	0	0	y_1=HAM
$\mathbf{x}_2$	1	0	1	0	1	0	0	y_2=HAM
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM
$\mathbf{x}_4$	1	0	1	0	0	0	0	y_4=HAM
$\mathbf{x}_5$	1	0	1	1	0	0	1	y_5=HAM
$\mathbf{x}_6$	0	1	0	1	0	0	1	y_6=SPAM
$\mathbf{x}_7$	1	0	0	0	1	0	0	y_7=HAM

Learning

Probability estimates:

Naive Bayes Classifier:

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(\boldsymbol{P}(y) * \prod_{i=1}^{N} \boldsymbol{P}(x_i \mid y) \right)$$

Probability estimates (Maximum Likelihood estimation):

$$P(y = HAM) = \frac{N_{samples\ labeled\ HAM}}{N}$$

$$P(y = SPAM) = \frac{N_{samples\ labeled\ SPAM}}{N}$$

$$P(x_i = word \mid y = CLASS) = \frac{count(x_i = word, y = CLASS)}{\sum_{x \in V} count(x, y = CLASS)}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Learning

Training set

Vocabulary $\mathbf{V}$

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_1$	1	1	1	0	1	0	0	y_1=HAM
$\mathbf{x}_2$	1	0	1	0	1	0	0	y_2=HAM
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM
$\mathbf{x}_4$	1	0	1	0	0	0	0	y_4=HAM
$\mathbf{x}_5$	1	0	1	1	0	0	1	y_5=HAM
$\mathbf{x}_6$	0	1	0	1	0	0	1	y_6=SPAM
$\mathbf{x}_7$	1	0	0	0	1	0	0	y_7=HAM

Learning

$$P(y = HAM) = \frac{N_{samples\ labeled\ HAM}}{N} = \frac{5}{7}$$

$$P(y = SPAM) = \frac{N_{samples\ labeled\ SPAM}}{N} = \frac{2}{7}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Learning

Training set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x_1}$	1	1	1	0	1	0	0	y_1=HAM
$\mathbf{x_2}$	1	0	1	0	1	0	0	y_2=HAM
$\mathbf{x_3}$	0	1	0	1	0	1	1	y_3=SPAM
$\mathbf{x_4}$	1	0	1	0	0	0	0	y_4=HAM
$\mathbf{x_5}$	1	0	1	1	0	0	1	y_5=HAM
$\mathbf{x_6}$	0	1	0	1	0	0	1	y_6=SPAM
$\mathbf{x_7}$	1	0	0	0	1	0	0	y_7=HAM

Learning

$$P(x_i = I \mid y = HAM) = \frac{count(x_i = I, y = HAM)}{\sum_{x \in V} count(x, y = HAM)} = \frac{5}{15}$$

$$P(x_i = rolex \mid y = HAM) = \frac{count(x_i = rolex, y = HAM)}{\sum_{x \in V} count(x, y = HAM)} = \frac{1}{15}$$

$$P(x_i = own \mid y = HAM) = \frac{count(x_i = own, y = HAM)}{\sum_{x \in V} count(x, y = HAM)} = \frac{4}{15}$$

$$P(x_i = replica \mid y = HAM) = \frac{count(x_i = replica, y = HAM)}{\sum_{x \in V} count(x, y = HAM)} = \frac{1}{15}$$

$$P(x_i = watch \mid y = HAM) = \frac{count(x_i = watch, y = HAM)}{\sum_{x \in V} count(x, y = HAM)} = \frac{3}{15}$$

$$P(x_i = buy \mid y = HAM) = \frac{count(x_i = buy, y = HAM)}{\sum_{x \in V} count(x, y = HAM)} = \frac{0}{15}$$

$$P(x_i = cheap \mid y = HAM) = \frac{count(x_i = cheap, y = HAM)}{\sum_{x \in V} count(x, y = HAM)} = \frac{1}{15}$$

$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_{N-2}}, \mathbf{x_{N-1}}, \mathbf{x_N}$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Learning

Training set

Vocabulary $\mathbf{V}$

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_1$	1	1	1	0	1	0	0	y_1=HAM
$\mathbf{x}_2$	1	0	1	0	1	0	0	y_2=HAM
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM
$\mathbf{x}_4$	1	0	1	0	0	0	0	y_4=HAM
$\mathbf{x}_5$	1	0	1	1	0	0	1	y_5=HAM
$\mathbf{x}_6$	0	1	0	1	0	0	1	y_6=SPAM
$\mathbf{x}_7$	1	0	0	0	1	0	0	y_7=HAM

Learning

$$P(x_i = I \mid y = SPAM) = \frac{count(x_i = I, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{0}{7}$$

$$P(x_i = rolex \mid y = SPAM) = \frac{count(x_i = rolex, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{2}{7}$$

$$P(x_i = own \mid y = SPAM) = \frac{count(x_i = own, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{0}{7}$$

$$P(x_i = replica \mid y = SPAM) = \frac{count(x_i = replica, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{2}{7}$$

$$P(x_i = watch \mid y = SPAM) = \frac{count(x_i = watch, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{0}{7}$$

$$P(x_i = buy \mid y = SPAM) = \frac{count(x_i = buy, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{1}{7}$$

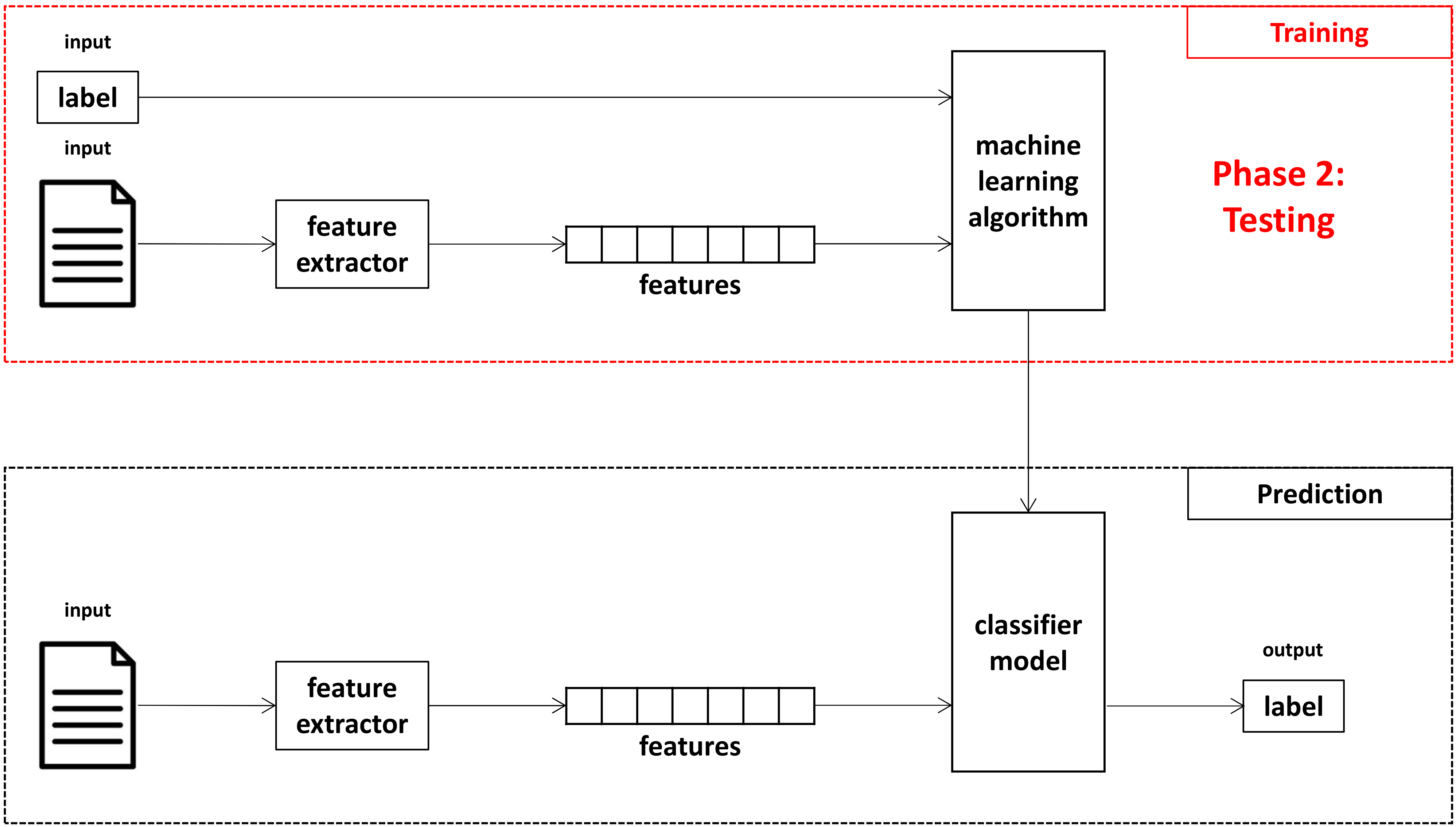
$$P(x_i = cheap \mid y = SPAM) = \frac{count(x_i = cheap, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{2}{7}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Corpus: Training / Test

Typical training / test set split for a text corpora [I will ignore validation set for the sake of an example]

Supervised Learning with ML



Spam Detection: Test Set

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(P(y) * \prod_{i=1}^{N} P(x_i \mid y) \right)$$

category/class = h(**document**)

Learned h()

$$P(y = HAM) = \frac{5}{7} \quad P(y = SPAM) = \frac{2}{7}$$

$$P(x_i = I \mid y = HAM) = \frac{5}{15}$$

$$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$$

$$P(x_i = own \mid y = HAM) = \frac{4}{15}$$

$$P(x_i = replica \mid y = HAM) = \frac{1}{15}$$

$$P(x_i = watch \mid y = HAM) = \frac{3}{15}$$

$$P(x_i = buy \mid y = HAM) = \frac{0}{15}$$

$$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$$

$$P(x_i = I \mid y = SPAM) = \frac{0}{7}$$

$$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$$

$$P(x_i = own \mid y = SPAM) = \frac{0}{7}$$

$$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$$

$$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$$

$$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$$

$$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1
$\mathbf{x}_9$	1	1	1	0	1	0	1
$\mathbf{x}_{10}$	0	0	0	1	0	0	0

$\mathbf{x}_8$: buy cheap rolex replica rolex — y_8=SPAM

$\mathbf{x}_9$: I own cheap rolex watch — y_9=HAM

$\mathbf{x}_{10}$: replica — y_{10}=HAM

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(\boldsymbol{P}(y) * \prod_{i=1}^{N} \boldsymbol{P}(x_i \mid y) \right)$$

category/class = h($\mathbf{x}_8$)

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax}\left(P(y) * \prod_{i=1}^{N} P(x_i \mid y)\right)$$

category/class = h(**buy cheap rolex replica rolex**)

Learned h()

$P(y = HAM) = \frac{5}{7}$ $P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax}\left(P(y) * \prod_{i=1}^{N} P(x_i \mid y)\right)$$

category/class = h(| 0 | 1 | 0 | 1 | 0 | 1 | 1 |)

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary **V**

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1

buy cheap rolex replica rolex

y_8=SPAM

$\mathbf{x}_9$	1	1	1	0	1	0	1

I own cheap rolex watch

y_9=HAM

$\mathbf{x}_{10}$	0	0	0	1	0	0	0

replica

y_{10}=HAM

Testing

Which probability is higher? Which y maximizes P()?

$$P(y = HAM \mid \boldsymbol{x_8}) \propto P(y = HAM) * \prod_{i=1}^{5} P(x_i \mid y = HAM)$$

$$P(y = SPAM \mid \boldsymbol{x_8}) \propto P(y = SPAM) * \prod_{i=1}^{5} P(x_i \mid y = SPAM)$$

Learned h()

$P(y = HAM) = \frac{5}{7}$ $P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1

buy cheap rolex replica rolex

y_8=SPAM

$\mathbf{x}_9$	1	1	1	0	1	0	1

I own cheap rolex watch

y_9=HAM

$\mathbf{x}_{10}$	0	0	0	1	0	0	0

replica

y_{10}=HAM

Testing

Which probability is higher? Which y maximizes P()?

$$\boldsymbol{P}(y = HAM \mid \boldsymbol{x}_8) \propto \boldsymbol{P}(y = HAM) * \prod_{i=1}^{5} \boldsymbol{P}(x_i \mid y = HAM) =$$

$$\boldsymbol{P}(y = HAM) * \boldsymbol{P}(x_1 = buy \mid y = HAM) *$$
$$\boldsymbol{P}(x_2 = cheap \mid y = HAM) * \boldsymbol{P}(x_3 = rolex \mid y = HAM) *$$
$$\boldsymbol{P}(x_4 = replica \mid y = HAM) * \boldsymbol{P}(x_5 = rolex \mid y = HAM)$$

Learned h()

$$P(y = HAM) = \frac{5}{7} \quad P(y = SPAM) = \frac{2}{7}$$
$$P(x_i = I \mid y = HAM) = \frac{5}{15}$$
$$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$$
$$P(x_i = own \mid y = HAM) = \frac{4}{15}$$
$$P(x_i = replica \mid y = HAM) = \frac{1}{15}$$
$$P(x_i = watch \mid y = HAM) = \frac{3}{15}$$
$$P(x_i = buy \mid y = HAM) = \frac{0}{15}$$
$$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$$
$$P(x_i = I \mid y = SPAM) = \frac{0}{7}$$
$$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$$
$$P(x_i = own \mid y = SPAM) = \frac{0}{7}$$
$$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$$
$$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$$
$$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$$
$$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$P(y = HAM \mid \mathbf{x}_8) \propto P(y = HAM) * \prod_{i=1}^{5} P(x_i \mid y = HAM) =$$

$$\frac{5}{7} * \frac{0}{15} * \frac{1}{15} * \frac{1}{15} * \frac{1}{15} * \frac{1}{15} = 0$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1
$\mathbf{x}_9$	1	1	1	0	1	0	1
$\mathbf{x}_{10}$	0	0	0	1	0	0	0

$\mathbf{x}_8$: buy cheap rolex replica rolex — y_8=SPAM

$\mathbf{x}_9$: I own cheap rolex watch — y_9=HAM

$\mathbf{x}_{10}$: replica — y_{10}=HAM

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x}_8) \propto \boldsymbol{P}(y = SPAM) * \prod_{i=1}^{5} \boldsymbol{P}(x_i \mid y = SPAM) =$$

$$\boldsymbol{P}(y = SPAM) * \boldsymbol{P}(x_1 = buy \mid y = SPAM) *$$
$$\boldsymbol{P}(x_2 = cheap \mid y = SPAM) * \boldsymbol{P}(x_3 = rolex \mid y = SPAM) *$$
$$\boldsymbol{P}(x_4 = replica \mid y = SPAM) * \boldsymbol{P}(x_5 = rolex \mid y = SPAM)$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary **V**

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

Which probability is higher? Which y maximizes P()?

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x}_8) \propto \boldsymbol{P}(y = SPAM) * \prod_{i=1}^{5} \boldsymbol{P}(x_i \mid y = SPAM) =$$

$$\frac{2}{7} * \frac{1}{7} * \frac{2}{7} * \frac{2}{7} * \frac{2}{7} * \frac{2}{7} \approx \mathbf{0.00027}$$

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1

buy cheap rolex replica rolex

y_8=SPAM

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_9$	1	1	1	0	1	0	1

I own cheap rolex watch

y_9=HAM

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_{10}$	0	0	0	1	0	0	0

replica

y_{10}=HAM

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$\boldsymbol{P}(y = HAM \mid \boldsymbol{x_8}) = 0$$

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x_8}) \approx 0.00027$$

For document $\mathbf{x}_8$ $y = SPAM$ maximizes P(). Class = $SPAM$.

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1
$\mathbf{x}_9$	1	1	1	0	1	0	1
$\mathbf{x}_{10}$	0	0	0	1	0	0	0

$\mathbf{x}_8$: buy cheap rolex replica rolex — y_8=SPAM

$\mathbf{x}_9$: I own cheap rolex watch — y_9=HAM

$\mathbf{x}_{10}$: replica — y_{10}=HAM

Testing

Which probability is higher? Which y maximizes P()?

For document $\mathbf{x}_8$ $y = SPAM$ maximizes P(). Class = $SPAM$.

Correct classification!

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1

buy cheap rolex replica rolex

y_8=SPAM

$\mathbf{x}_9$	1	1	1	0	1	0	1

I own cheap rolex watch

y_9=HAM

$\mathbf{x}_{10}$	0	0	0	1	0	0	0

replica

y_{10}=HAM

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax}\left(P(y) * \prod_{i=1}^{N} P(x_i \mid y)\right)$$

category/class = h($\mathbf{x}_9$)

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM

$\mathbf{x}_8$: buy cheap rolex replica rolex

$\mathbf{x}_9$: I own cheap rolex watch

$\mathbf{x}_{10}$: replica

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(P(y) * \prod_{i=1}^{N} P(x_i \mid y) \right)$$

category/class = h(**I own cheap rolex watch**)

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax}\left(P(y) * \prod_{i=1}^{N} P(x_i \mid y)\right)$$

category/class = h(| 1 | 1 | 1 | 0 | 1 | 0 | 1 |)

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary $\mathbf{V}$

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$\boldsymbol{P}(y = HAM \mid \boldsymbol{x_9}) \propto \boldsymbol{P}(y = HAM) * \prod_{i=1}^{1} \boldsymbol{P}(x_i \mid y = HAM)$$

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x_9}) \propto \boldsymbol{P}(y = SPAM) * \prod_{i=1}^{1} \boldsymbol{P}(x_i \mid y = SPAM)$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

Which probability is higher? Which y maximizes P()?

$$P(y = HAM \mid \mathbf{x}_9) \propto P(y = HAM) * \prod_{i=1}^{5} P(x_i \mid y = HAM) =$$

$$P(y = HAM) * P(x_1 = I \mid y = HAM) *$$
$$P(x_2 = own \mid y = HAM) * P(x_3 = cheap \mid y = HAM) *$$
$$P(x_4 = rolex \mid y = HAM) * P(x_5 = watch \mid y = HAM)$$

Learned h()

$$P(y = HAM) = \frac{5}{7} \quad P(y = SPAM) = \frac{2}{7}$$
$$P(x_i = I \mid y = HAM) = \frac{5}{15}$$
$$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$$
$$P(x_i = own \mid y = HAM) = \frac{4}{15}$$
$$P(x_i = replica \mid y = HAM) = \frac{1}{15}$$
$$P(x_i = watch \mid y = HAM) = \frac{3}{15}$$
$$P(x_i = buy \mid y = HAM) = \frac{0}{15}$$
$$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$$
$$P(x_i = I \mid y = SPAM) = \frac{0}{7}$$
$$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$$
$$P(x_i = own \mid y = SPAM) = \frac{0}{7}$$
$$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$$
$$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$$
$$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$$
$$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

Which probability is higher? Which y maximizes P()?

$$P(y = HAM \mid \mathbf{x}_9) \propto P(y = HAM) * \prod_{i=1}^{5} P(x_i \mid y = HAM) =$$

$$\frac{5}{7} * \frac{5}{15} * \frac{4}{15} * \frac{1}{15} * \frac{1}{15} * \frac{3}{15} \approx 0.000056$$

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$P(y = SPAM \mid \mathbf{x}_9) \propto P(y = SPAM) * \prod_{i=1}^{5} P(x_i \mid y = SPAM) =$$

$$P(y = SPAM) * P(x_1 = I \mid y = SPAM) *$$
$$P(x_2 = own \mid y = SPAM) * P(x_3 = cheap \mid y = SPAM) *$$
$$P(x_4 = rolex \mid y = SPAM) * P(x_5 = watch \mid y = SPAM)$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary **V**

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x}_9) \propto \boldsymbol{P}(y = SPAM) * \prod_{i=1}^{5} \boldsymbol{P}(x_i \mid y = SPAM) =$$

$$\frac{2}{7} * \frac{0}{7} * \frac{0}{7} * \frac{2}{7} * \frac{2}{7} * \frac{0}{7} = 0$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

Which probability is higher? Which y maximizes P()?

$$\boldsymbol{P}(y = HAM \mid \boldsymbol{x}_9) \approx 0.000056$$

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x}_9) = 0$$

For document $\mathbf{x}_8$ $y = HAM$ maximizes P(). Class = HAM.

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

Which probability is higher? Which y maximizes P()?

For document $\mathbf{x}_9$ $y = HAM$ maximizes P(). Class = HAM.

Correct classification!

Learned h()

$P(y = HAM) = \frac{5}{7}$ $P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \dots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(P(y) * \prod_{i=1}^{N} P(x_i \mid y) \right)$$

$$\text{category/class} = h(\mathbf{x}_{10})$$

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	label
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(P(y) * \prod_{i=1}^{N} P(x_i \mid y) \right)$$

category/class = h(**replica**)

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1
$\mathbf{x}_9$	1	1	1	0	1	0	1
$\mathbf{x}_{10}$	0	0	0	1	0	0	0

$\mathbf{x}_8$: buy cheap rolex replica rolex — y_8=SPAM

$\mathbf{x}_9$: I own cheap rolex watch — y_9=HAM

$\mathbf{x}_{10}$: replica — y_{10}=HAM

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(\boldsymbol{P}(y) * \prod_{i=1}^{N} \boldsymbol{P}(x_i \mid y) \right)$$

category/class = h(| 0 | 0 | 0 | 1 | 0 | 0 | 0 |)

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x_8}$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x_9}$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x_{10}}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing

Which probability is higher? Which y maximizes P()?

$$P(y = HAM \mid \mathbf{x_{10}}) \propto P(y = HAM) * \prod_{i=1}^{1} P(x_i \mid y = HAM)$$

$$P(y = SPAM \mid \mathbf{x_{10}}) \propto P(y = SPAM) * \prod_{i=1}^{1} P(x_i \mid y = SPAM)$$

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_{N-2}}, \mathbf{x_{N-1}}, \mathbf{x_N}$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$P(y = HAM \mid \mathbf{x}_{10}) \propto P(y = HAM) * \prod_{i=1}^{1} P(x_i \mid y = HAM) =$$

$$P(y = HAM) * P(x_1 = replica \mid y = HAM) = \frac{5}{7} * \frac{1}{15} \approx 0.048$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1
$\mathbf{x}_9$	1	1	1	0	1	0	1
$\mathbf{x}_{10}$	0	0	0	1	0	0	0

$\mathbf{x}_8$: buy cheap rolex replica rolex — y_8=SPAM

$\mathbf{x}_9$: I own cheap rolex watch — y_9=HAM

$\mathbf{x}_{10}$: replica — y_{10}=HAM

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x}_{10}) \propto \boldsymbol{P}(y = SPAM) * \prod_{i=1}^{1} \boldsymbol{P}(x_i \mid y = SPAM) =$$

$$\boldsymbol{P}(y = SPAM) * \boldsymbol{P}(x_1 = replica \mid y = SPAM) = \frac{2}{7} * \frac{2}{7} \approx 0.082$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary **V**

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Learned h()

$P(y = HAM) = \frac{5}{7}$ $\quad P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

$$P(y = HAM \mid \boldsymbol{x}_{10}) \approx 0.048$$

$$P(y = SPAM \mid \boldsymbol{x}_{10}) \approx 0.082$$

For document $\mathbf{x}_{10}$ $y = SPAM$ maximizes P(). Class = $SPAM$.

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Spam Detection: Testing Classifier

Test set

Vocabulary **V**

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_8$	0	1	0	1	0	1	1

buy cheap rolex replica rolex

y_8=SPAM

$\mathbf{x}_9$	1	1	1	0	1	0	1

I own cheap rolex watch

y_9=HAM

$\mathbf{x}_{10}$	0	0	0	1	0	0	0

replica

y_{10}=HAM

Learned h()

$P(y = HAM) = \frac{5}{7}$ $P(y = SPAM) = \frac{2}{7}$

$P(x_i = I \mid y = HAM) = \frac{5}{15}$

$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$

$P(x_i = own \mid y = HAM) = \frac{4}{15}$

$P(x_i = replica \mid y = HAM) = \frac{1}{15}$

$P(x_i = watch \mid y = HAM) = \frac{3}{15}$

$P(x_i = buy \mid y = HAM) = \frac{0}{15}$

$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$

$P(x_i = I \mid y = SPAM) = \frac{0}{7}$

$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$

$P(x_i = own \mid y = SPAM) = \frac{0}{7}$

$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$

$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$

$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$

$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$

Testing

Which probability is higher? Which y maximizes P()?

For document $\mathbf{x}_{10}$ $y = SPAM$ maximizes P(). Class = $SPAM$.

Incorrect classification! Misclassification.

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Classifier Evaluation: Confusion Matrix

		Predicted class		
		Positive	**Negative**	
Actual class	**Positive**	**True Positive (TP)**	**False Negative (FN)** **Type II Error**	**Sensitivity (Recall)** $\frac{TP}{TP+FN}$
	Negative	**False Positive (FP)** **Type I Error**	**True Negative (TN)**	**Specificity** $\frac{TN}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	**Negative Predictive Value** $\frac{TN}{TN+FN}$	**Accuracy** $\frac{TP+TN}{TP+TN+FP+FN}$

Classifier Evaluation: Confusion Matrix

		Predicted class		
		SPAM	HAM	
Actual class	SPAM	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity (Recall) $\frac{TP}{TP+FN}$
	HAM	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	Negative Predictive Value $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

Spam Detection: Evaluating Classifier

Test set

Vocabulary **V**

	I	rolex	own	replica	watch	buy	cheap	
$\mathbf{x}_8$	0	1	0	1	0	1	1	y_8=SPAM
	buy cheap rolex replica rolex							
$\mathbf{x}_9$	1	1	1	0	1	0	1	y_9=HAM
	I own cheap rolex watch							
$\mathbf{x}_{10}$	0	0	0	1	0	0	0	y_{10}=HAM
	replica							

Testing results

y_8=SPAM	correct
y_9=HAM	correct
y_{10}=SPAM	incorrect

Evaluation

y_8=SPAM	y_8=SPAM	true positive
y_9=HAM	y_9=HAM	true negative
y_{10}=HAM	y_{10}=SPAM	false positiive

No false negatives in this example.

Confusion matrix

	SPAM	HAM
SPAM	True Positive (TP)	False Negative (FN) Type II Error
HAM	False Positive (FP) Type I Error	True Negative (TN)

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

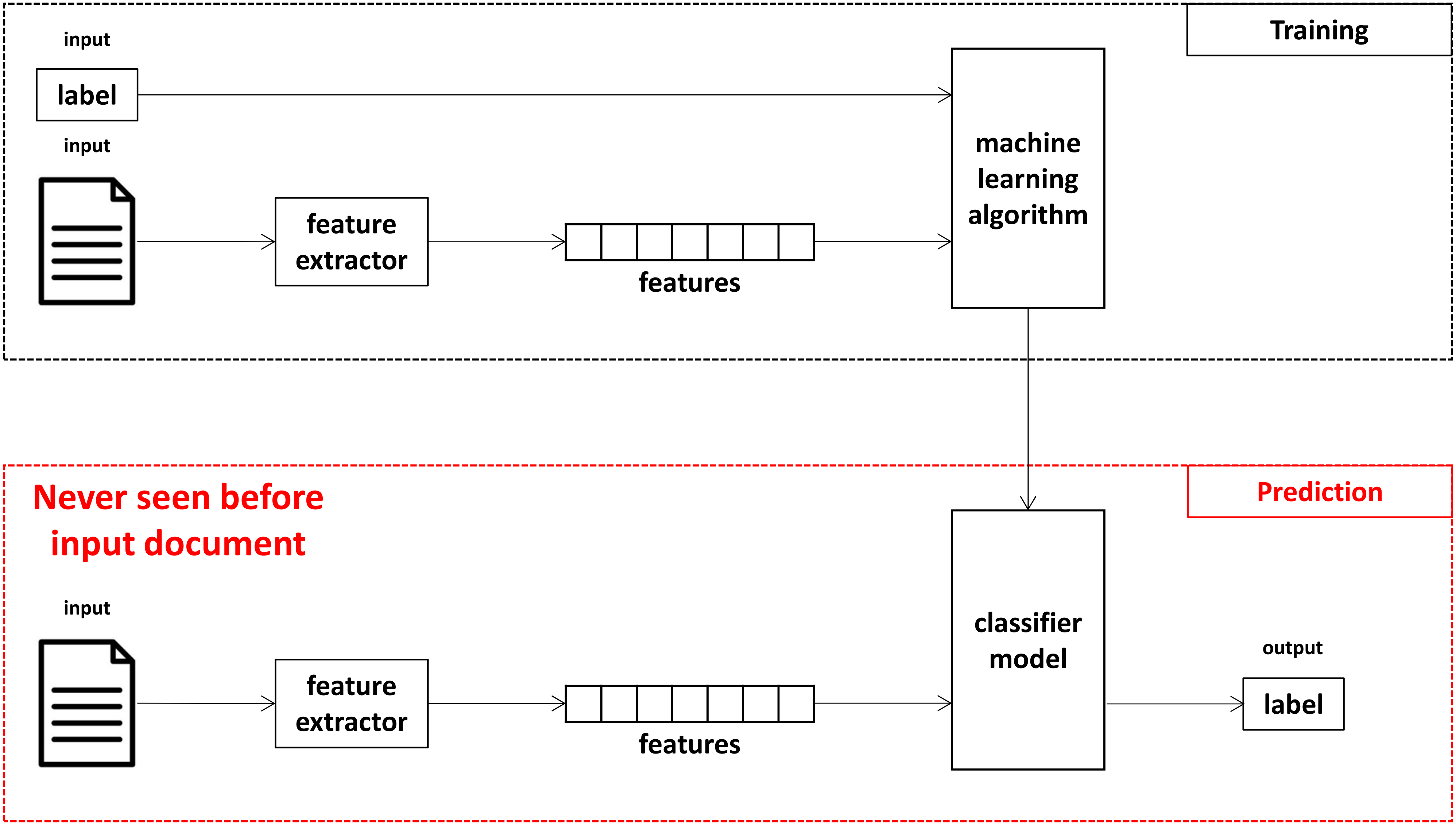
Classifier Evaluation: Confusion Matrix

		Predicted class		
		SPAM	HAM	
Actual class	SPAM	TP = 1 (x_8)	FN = 0	Sensitivity (Recall) $\frac{1}{1+0} = 1.0$
	HAM	FP = 1 (x_{10})	TN = 1 (x_9)	Specificity $\frac{1}{1+1} = 0.5$
		Precision $\frac{1}{1+1} = 0.5$	Negative Predictive Value $\frac{1}{1+0} = 1.0$	Accuracy $\frac{1+1}{1+1+1+0} = \frac{2}{3}$

Confusion Matrix Explained

- Accuracy (TP+TN)/(TP+TN+FP+FN):
  - Overall, how often is the classifier correct?
- Misclassification rate [Error Rate] (FP+FN)/(TP+TN+FP+FN):
  - Overall, how often is the classifier incorrect?
- Sensitivity [Recall | True Positive Rate] (TP)/(TP+FN):
  - When it's actually yes, how often does it predict yes?
- Specificity [True Negative Rate] (TN)/(TN+FP)
  - When it's actually no, how often does it predict no?
- Precision (TP)/(TP+FP)
  - When it predicts yes, how often is it correct?
- Negative Predictive Value (TN)/(TN+FN)
  - When it predicts no, how often is it correct?

Supervised Learning with ML



Spam Detection: Prediction

Unseen x

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap
$\mathbf{x}_?$	0	1	0	0	0	1	0

buy rolex

Label needs to be decided

$y_?$=????

Prediction

Which probability is higher? Which y maximizes P()?

$$P(y = HAM \mid x_?) \propto P(y = HAM) * \prod_{i=1}^{2} P(x_i \mid y = HAM) =$$

$$P(y = HAM) * P(x_1 = buy \mid y = HAM) * P(x_2 = rolex \mid y = HAM)$$

$$= \frac{5}{7} * \frac{0}{15} * \frac{1}{15} = 0$$

$$P(y = SPAM \mid x_?) \propto P(y = SPAM) * \prod_{i=1}^{2} P(x_i \mid y = SPAM) =$$

$$P(y = SPAM) * P(x_1 = buy \mid y = SPAM) * P(x_2 = rolex \mid y = SPAM)$$

$$= \frac{2}{7} * \frac{1}{7} * \frac{2}{7} \approx 0.012$$

For document $\mathbf{x}_?$ $y = SPAM$ maximizes P(). Class = $SPAM$

Learned h()

$$P(y = HAM) = \frac{5}{7} \quad P(y = SPAM) = \frac{2}{7}$$

$$P(x_i = I \mid y = HAM) = \frac{5}{15}$$

$$P(x_i = rolex \mid y = HAM) = \frac{1}{15}$$

$$P(x_i = own \mid y = HAM) = \frac{4}{15}$$

$$P(x_i = replica \mid y = HAM) = \frac{1}{15}$$

$$P(x_i = watch \mid y = HAM) = \frac{3}{15}$$

$$P(x_i = buy \mid y = HAM) = \frac{0}{15}$$

$$P(x_i = cheap \mid y = HAM) = \frac{1}{15}$$

$$P(x_i = I \mid y = SPAM) = \frac{0}{7}$$

$$P(x_i = rolex \mid y = SPAM) = \frac{2}{7}$$

$$P(x_i = own \mid y = SPAM) = \frac{0}{7}$$

$$P(x_i = replica \mid y = SPAM) = \frac{2}{7}$$

$$P(x_i = watch \mid y = SPAM) = \frac{0}{7}$$

$$P(x_i = buy \mid y = SPAM) = \frac{1}{7}$$

$$P(x_i = cheap \mid y = SPAM) = \frac{2}{7}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - feature vectors (in **bold**) | $y_1, y_2, y_3, \ldots, y_{N-2}, y_{N-1}, y_N$ - labels

Classifier Problems: Zero Counts

$$P(x_i = word \mid y = CLASS) = \frac{count(x_i = word, y = CLASS)}{\sum_{x \in V} count(x, y = CLASS)}$$

- **Unseen words:**
  - $count(x_i = word, y = CLASS)$ **can be zero**
- **Words NOT present in samples for one class (see: example):**
  - $count(x_i = word, y = CLASS)$ **can be zero**

- **Solution: smoothing (e.g. Laplace smoothing)**

$$P(x_i = word \mid y = CLASS) = \frac{count(x_i = word, y = CLASS) + \alpha}{\sum_{x \in V} count(x, y = CLASS) + \alpha * |V|}$$

where: α - pseudo-occurence (typically “add 1”), $|V|$ - vocabulary size

Classifier Problems: Underflow

$$P(y \mid x) \propto P(y) * \prod_{i=1}^{N} P(x_i \mid y)$$

- N can be large (100 and more):
  - long, “wordy”, documents
- some $P(x_i \mid y)$ can be very small (< 0.1)
  - the product $\prod_{i=1}^{N} P(x_i \mid y)$ may lead to underflow

- Solution: use logarithms

$$log(P(y \mid x)) \propto log(P(y)) + \sum_{i=1}^{N} log(P(x_i \mid y))$$

Naive Bayes Classifier

$$\text{category/class} = h(\textbf{document})$$

Finding model / hypothesis $h \rightarrow$ **Finding probabilities for** y_{MAP}

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(log\boxed{(P(y))} + \sum_{i=1}^{N} log\boxed{(P(x_i \mid y))} \right)$$

MAP: Maximum a posteriori (corresponds to the most likely class).

Naive Bayes Classifier

$$\text{category/class} = h(\textbf{document})$$

Finding model / hypothesis h $\rightarrow$ Finding probabilities for y_{MAP}

$$y_{MAP} \propto \underset{y \in Y}{argmax} \left(log\boxed{(P(y))} + \sum_{i=1}^{N} log\boxed{(P(x_i \mid y))} \right)$$

- **Taking log doesn't change the ranking of classes!**
  - **The class with highest probability also has highest log probability!**
- **It's a linear model:**
  - **Just a max of a sum of weights: a linear function of the inputs**
  - **So Naive Bayes is a linear classifier**

Naive Bayes: Training/Testing

function TRAIN NAIVE BAYES(D, C) **returns** log $P(c)$ and log $P(w|c)$

for each class $c \in C$ # Calculate $P(c)$ terms
  N_{doc} = number of documents in D
  N_c = number of documents from D in class c
  $logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$
  $V \leftarrow$ vocabulary of D
  $bigdoc[c] \leftarrow$ **append**(d) **for** d $\in$ D **with** class c
  for each word w in V # Calculate $P(w|c)$ terms
    $count(w,c) \leftarrow$ # of occurrences of w in $bigdoc[c]$
    $loglikelihood[w,c] \leftarrow \log \frac{count(w,c) + 1}{\sum_{w' \text{ in } V} (count(w',c) + 1)}$
return *logprior*, *loglikelihood*, V

function TEST NAIVE BAYES(*testdoc*, *logprior*, *loglikelihood*, C, V) **returns** best c

for each class $c \in C$
  $sum[c] \leftarrow logprior[c]$
  for each position i in *testdoc*
    $word \leftarrow testdoc[i]$
    if $word \in V$
      $sum[c] \leftarrow sum[c] + loglikelihood[word,c]$
return $\text{argmax}_c \, sum[c]$

Naive Bayes: Summary

- **Pros:**
  - **Very fast and easy-to-implement**
  - **Well-understood formally & experimentally**
    - **see "Naive (Bayes) at Forty", Lewis, ECML98**
- **Cons:**
  - **Seldom gives the very best performance (baseline)**
  - **"Probabilities" $P(y \mid x)$ are not accurate**
  - **Probabilities tend to be close to zero or one**

Naive Bayes: Stop Words

- Some systems ignore stop words
  - Stop words: very frequent words like the and a.
    - Sort the vocabulary by word frequency in training set
    - Call the top 10 or 50 words the stopword list.
    - Remove all stop words from both training and test sets
      - As if they were never there!
- But removing stop words doesn't usually help
  - So in practice most NB algorithms use all words and don't use stopword lists

Naive Bayes: Unknown Words

- What about unknown words
  - that appear in our test data
  - but not in our training data or vocabulary?
- We ignore them
  - Remove them from the test document!
  - Pretend they weren't there!
  - Don't include any probability for them at all!
- Why don't we build an unknown word model?
  - It doesn't help: knowing which class has more unknown words is not generally helpful!

Naive Bayes: More Than Two Classes

- Dealing with any-of or multivalue classification
  - A document can belong to 0, 1, or more than 1 classes.

- For each class $c \in C$
  - Build a classifier h_c to distinguish c from all other classes $c' \in C$
- Given test document d,
  - Evaluate it for membership in each class using each h_c
  - d belongs to any class for which h_c returns true

Naive Bayes: More Than Two Classes

- Dealing with one-of or multinomial classification
  - Classes are mutually exclusive: each document in exactly one class

- For each class $c \in C$
  - Build a classifier h_c to distinguish c from all other classes $c' \in C$
- Given test document d,
  - Evaluate it for membership in each class using each h_c
  - d belongs to the one class with maximum score