CS 585

Natural Language Processing

February 12, 2024

Announcements / Reminders

- Please follow the Week 05 To Do List instructions (if you haven't already
- Written Assignment #03: posted
- Programming Assignment #02 will be posted soon
- Midterm Exam: 02/26/2024
  - Section 02 – Make arrangements with Mr. Charles Scott

Plan for Today

- **Text Classification**

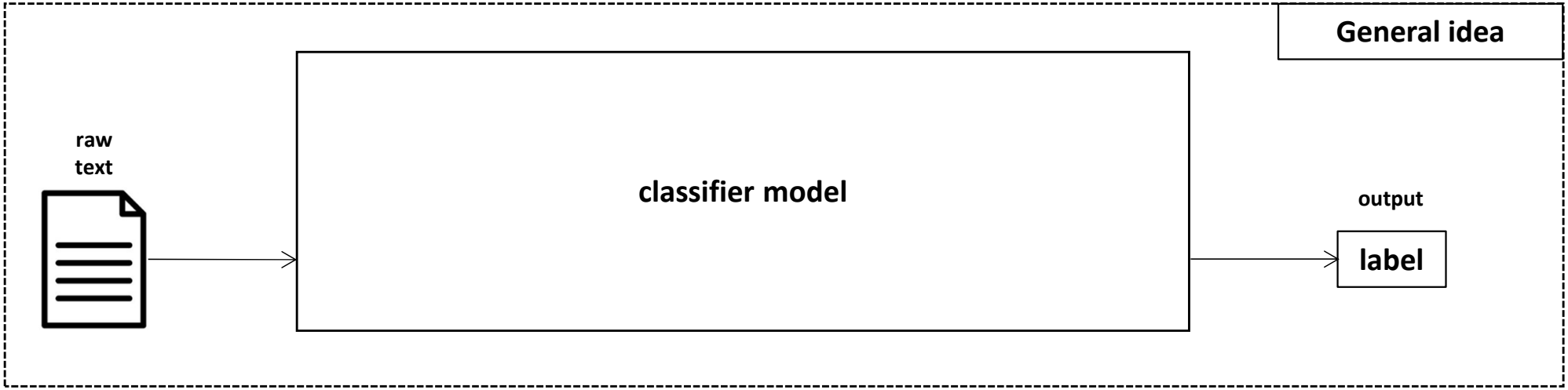
Text Classification: Definition

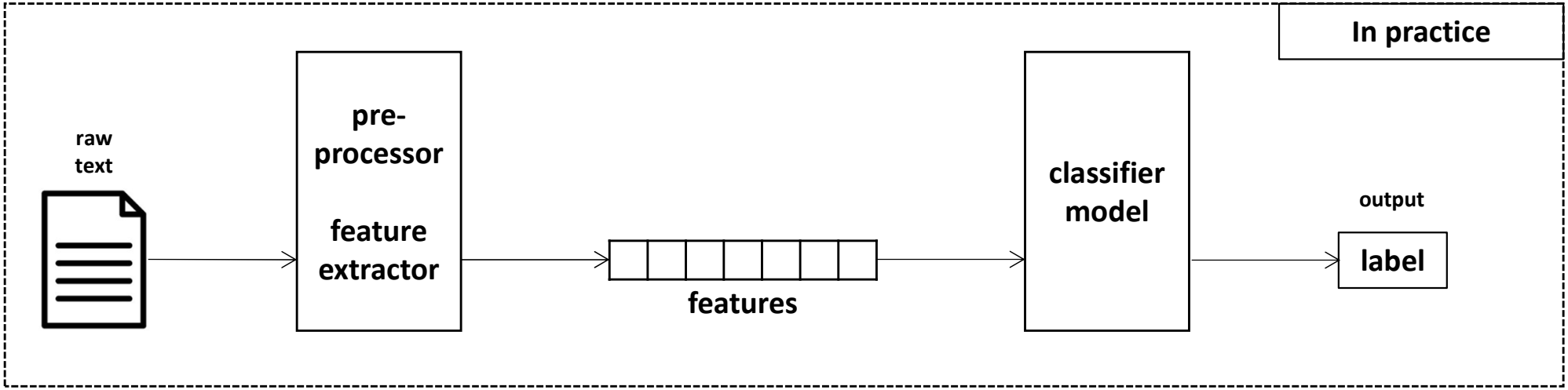
Input:

- a document d
- a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

Output: a predicted class $c \in C$

Text Classification: the Idea

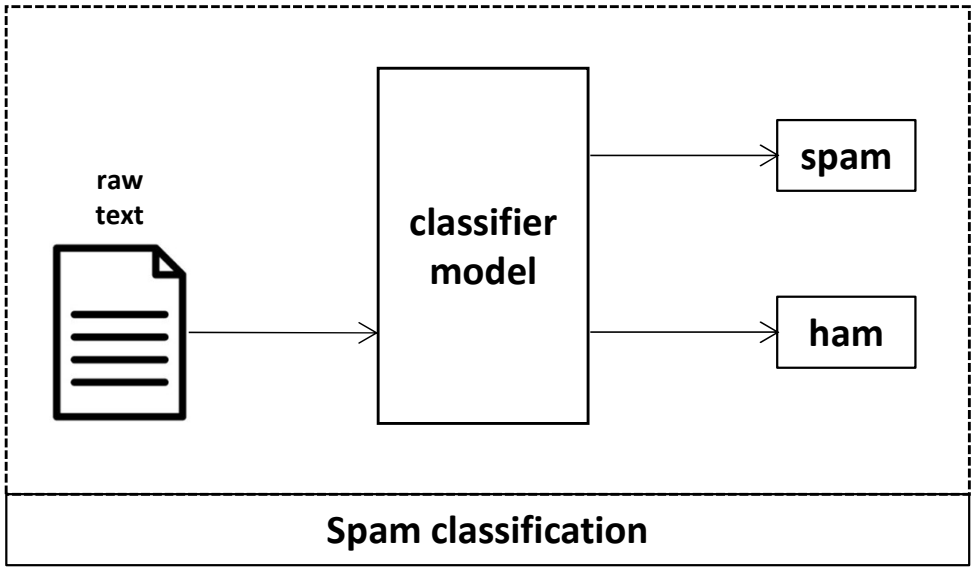




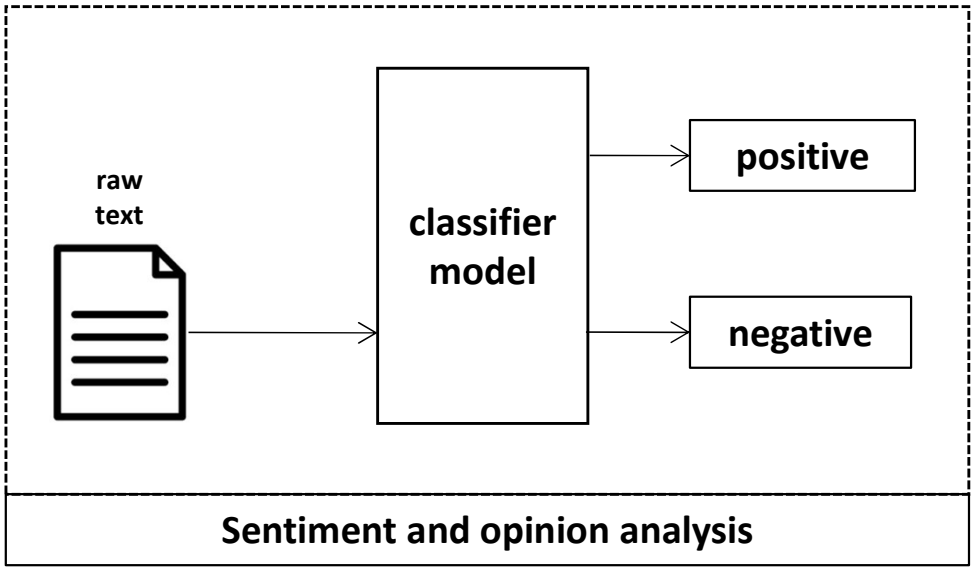
Text Classification: Applications

- Sentiment / opinion analysis
- Spam detection
- Gender identification
- Authorship identification
- Language identification
- Assigning subject categories, topics, or genres
- …

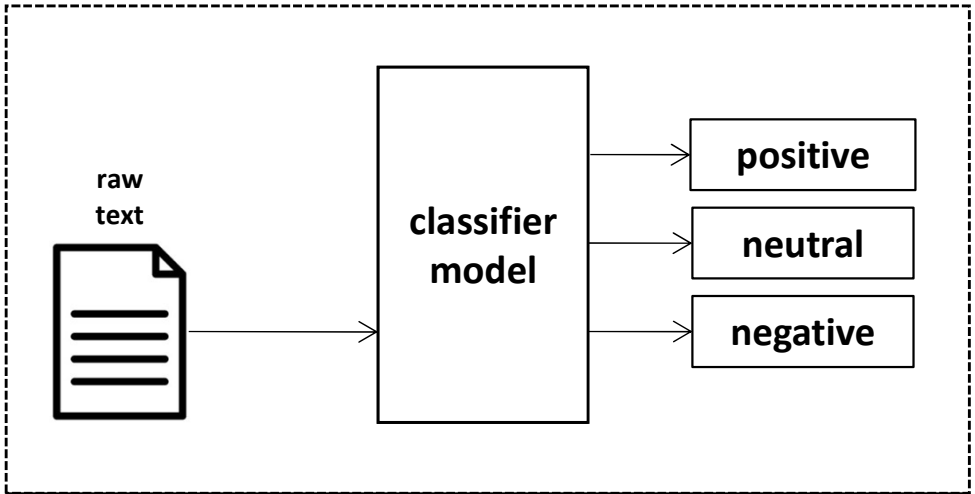
Text Classification: Applications



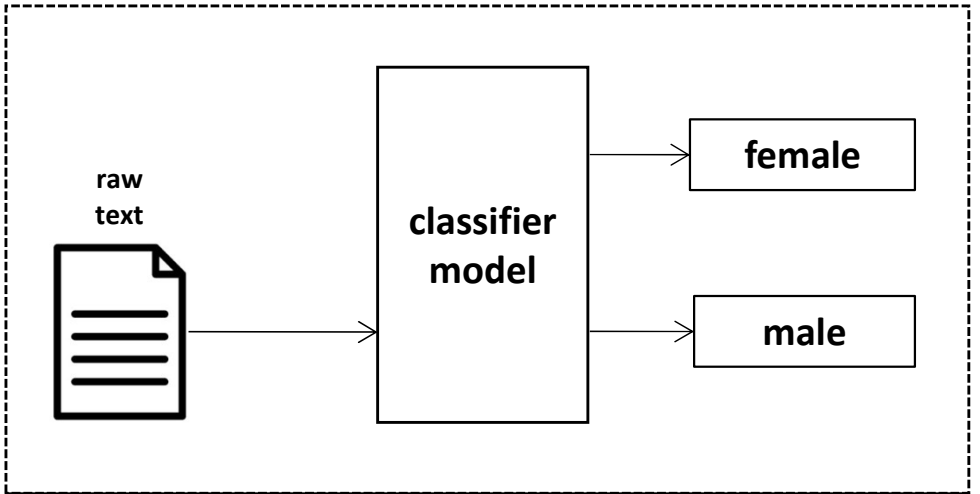
Spam classification



Sentiment and opinion analysis



Sentiment and opinion analysis



Gender identification / Attribution

Text Classification: Rule-Based

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “you have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Text Classification: Supervised ML

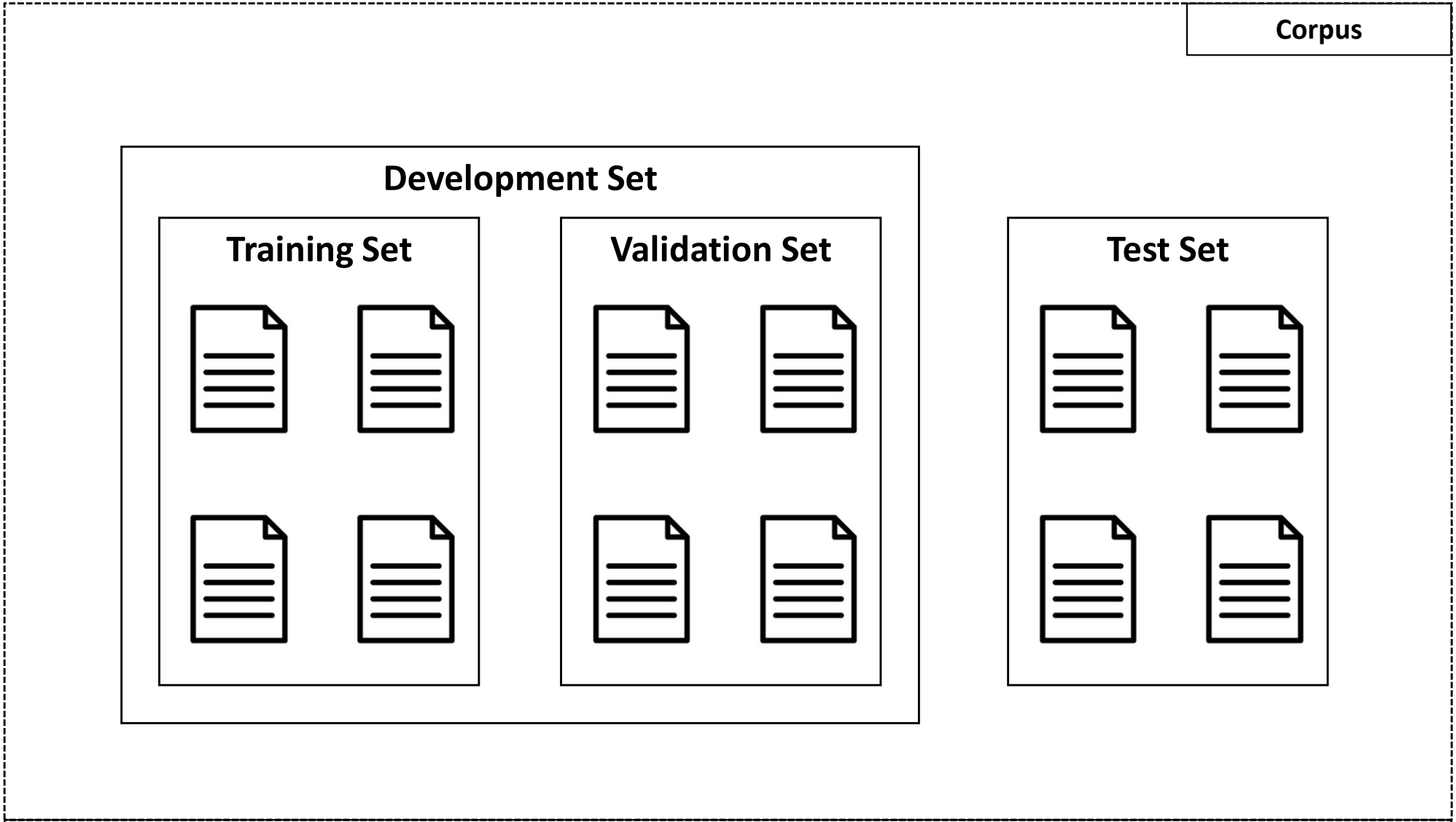
Input:

- a document d
- a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
- a training set of m hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$

Output:

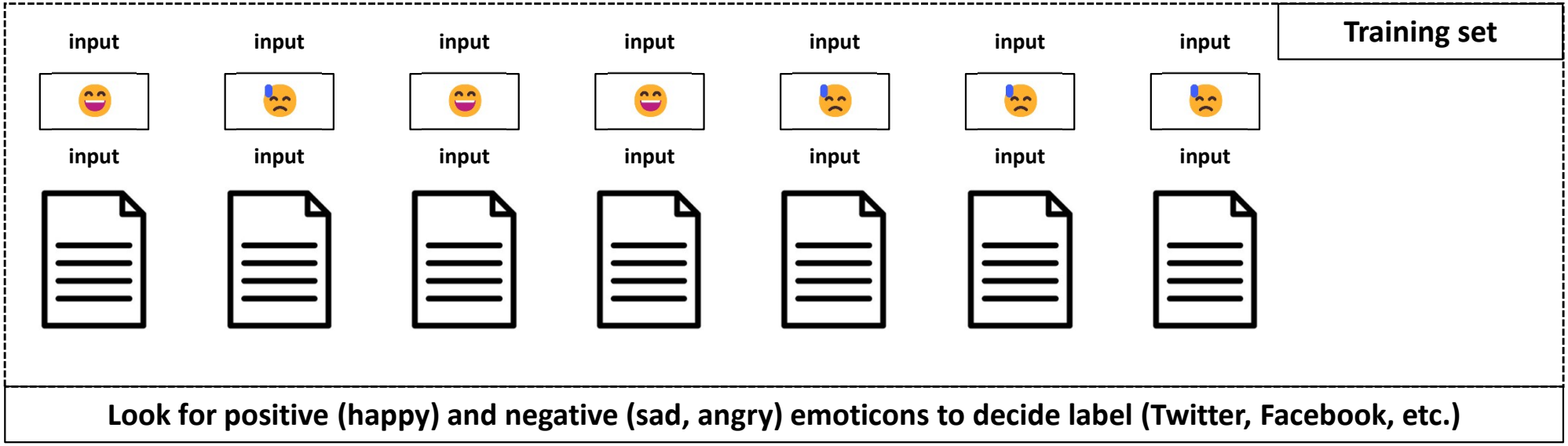
- a learned classifier $\gamma: d \rightarrow c$

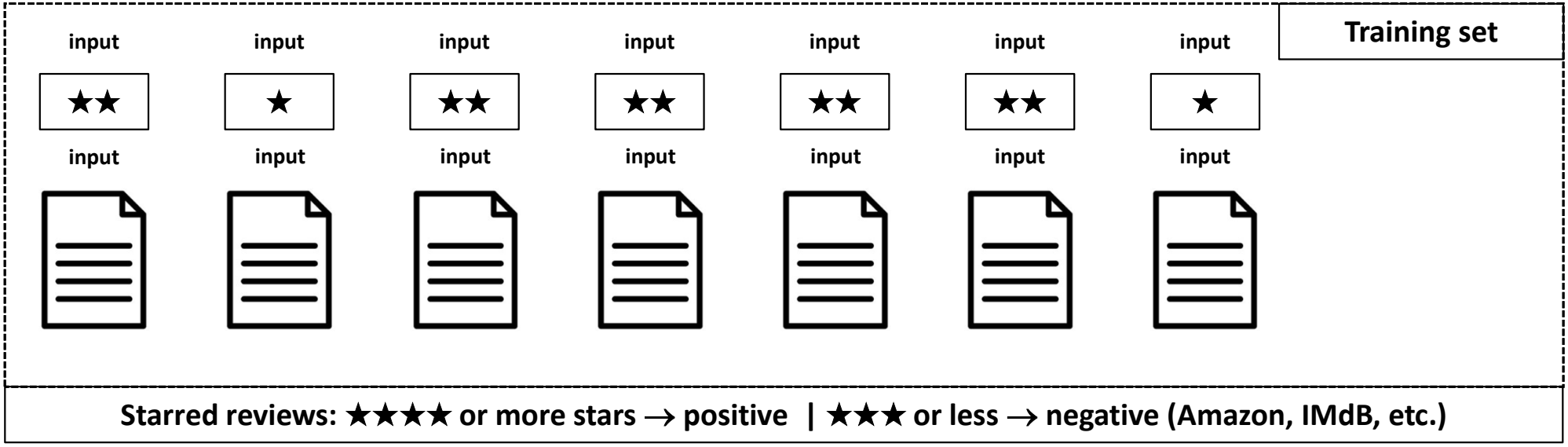
Corpus: Training / Validation / Test



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

Text Training Set (Auto) Labeling

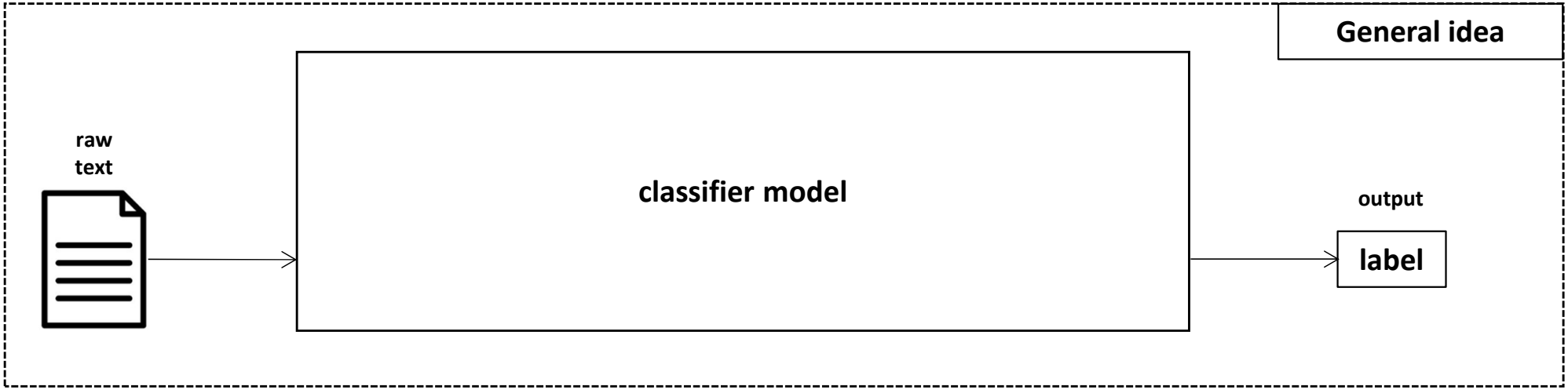


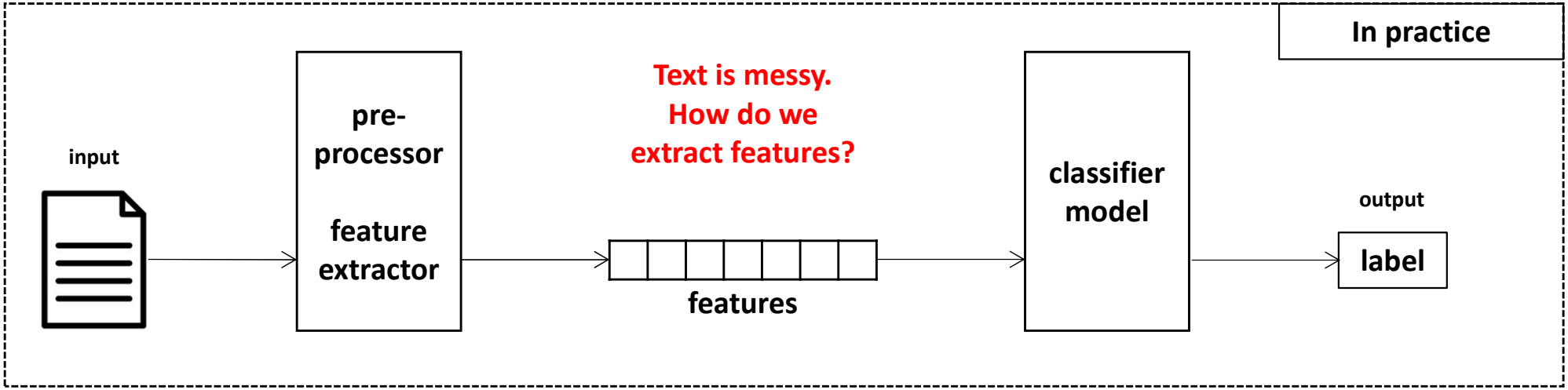


Text Classification: Supervised ML

- Various Machine Learning supervised learning classifier approaches can be employed:
  - Naïve Bayes
  - Logistic regression
  - Neural networks
  - k-Nearest Neighbors
  - etc.

Text Classification: Feature Extraction





Bag of Words: the Idea

Elon Musk @elonmusk · Feb 18

So amazing that we decompress from a single cell to over 30 trillion cells

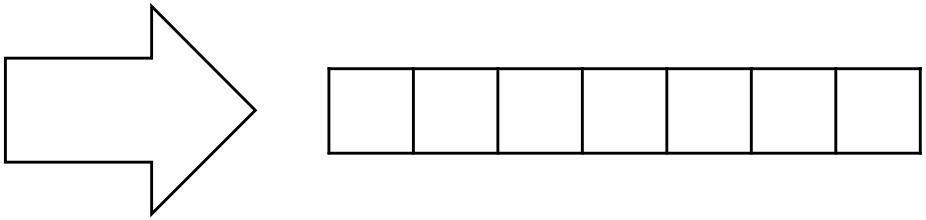
14.7K 15.2K 190.2K

By Amy Bizzarri 1st March 2022

Obtained from the autumnal flowering of the strawberry tree on the island of Sardinia, corbezzolo honey isn't sweet and has a history that dates back more than 2,000 years.

Corbezzolo honey tricks the palate. Instead of the sweetness one would expect, this extremely rare honey, born in the mountains of the Italian island of Sardinia, is surprisingly bitter, with notes of leather, liquorice and smoke. Nomadic beekeepers have been setting up beehives in the region to collect this aromatic treat – derived from the white, bell-shaped flowers of the wild strawberry tree – for more than 2,000 years.

Statesman, lawyer and philosopher Marcus Tullius Cicero (106-43 BCE) mentioned the honey in his defence of a Roman citizen accused of murder in Nora, Sardinia. "*Omne quod Sardinia fert, homines et res, mala est! Etiam mel quod ea insula abundat, amarum est!* (Everything that the island of Sardinia produces, men and things, is bad!)," he exclaimed. "Even the honey, abundant on that island, is bitter!"



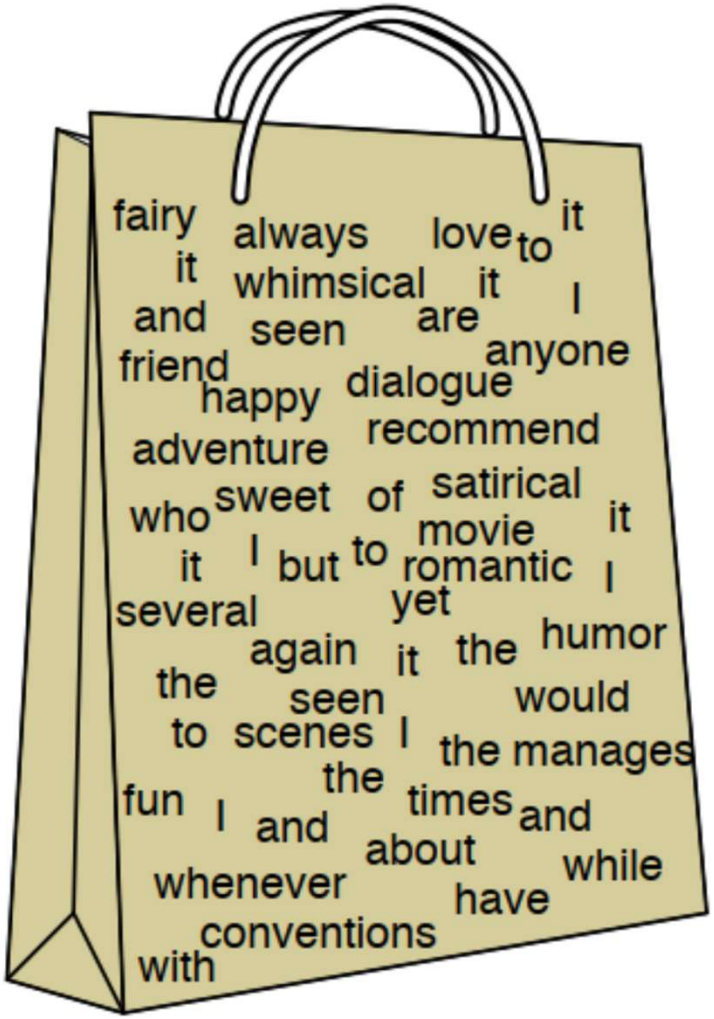
FIXED size

Feature vector

Bag of Words: the Idea

Some document:

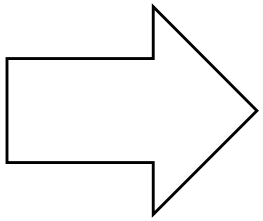
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



Bag of words assumption: word/token position does not matter.

Bag of Words: the Idea

Some document:
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

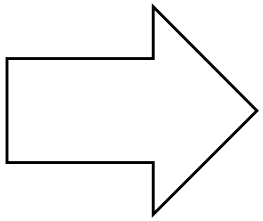


Word:	Frequency:
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
whimsical	1
times	1
....	...

Bag of words assumption: word/token position does not matter.

Bag of Words: Document Vector

Some document:
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



Word:	Frequency:
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
whimsical	1
times	1
....	...

vector

Bag of Words: Document Vector

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

Document A Binary Vector [0-word absent | 1-word present]:							
1	0	1	1	1	0	. . .	1

Document B Binary Vector [0-word absent | 1-word present]:							
1	1	0	0	1	0	. . .	1

Document C Binary Vector [0-word absent | 1-word present]:							
0	0	1	0	0	1	. . .	0

Document vectors can be used to compare documents.

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: Document Vector

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

Document A Binary Vector [0-word absent | 1-word present]:							
1	0	1	1	1	0	. . .	1

Document B Binary Vector [0-word absent | 1-word present]:							
1	1	0	0	1	0	. . .	1

Document C Binary Vector [0-word absent | 1-word present]:							
0	0	1	0	0	1	. . .	0

Document vectors can be used to compare documents.

Document Vector = Feature Vector

Pre-defined Features:							
Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	...	Feature N

Document A Binary Vector [0-word absent | 1-word present]:							
1	0	1	1	1	0	. . .	1

Document B Binary Vector [0-word absent | 1-word present]:							
1	1	0	0	1	0	. . .	1

Document C Binary Vector [0-word absent | 1-word present]:							
0	0	1	0	0	1	. . .	0

Document vectors can be used to compare documents.

Bag of Words: Document Vector

Pre-defined Vocabulary:							
she	want	to	walk	drive	fly	there	or

"She wants to walk there today": Binary Document Vector							
1	1	1	1	0	0	1	0

"She wants to drive there today": Binary Document Vector							
1	1	1	0	1	0	1	0

"She wants to fly or drive there today": Binary Document Vector							
1	1	1	0	1	1	1	1

Note: sentences lemmatized and lowercased.

Bag of Bigrams: Document Vector

Pre-defined Bigrams:							
w1, w2	w2, w3	w3, w4	w4, w5	w5, w6	w6,w7	...	wN-1,wN

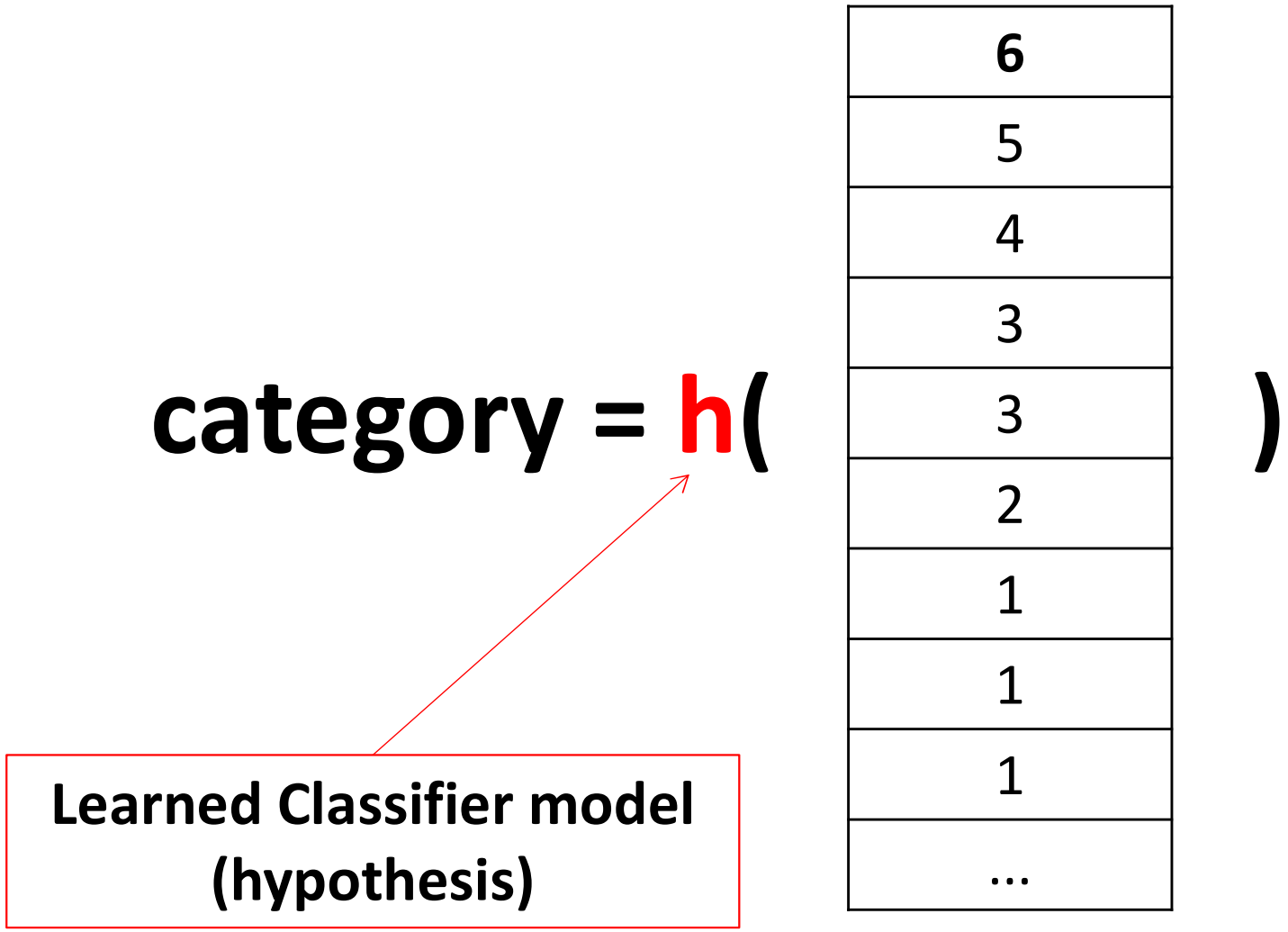
Document A Binary Vector [0-word absent | 1-word present]:							
1	0	1	1	1	0	. . .	1

Document B Binary Vector [0-word absent | 1-word present]:							
1	1	0	0	1	0	. . .	1

Document C Binary Vector [0-word absent | 1-word present]:							
0	0	1	0	0	1	. . .	0

Document vectors can be used to compare documents.

Bag of Words: Classification

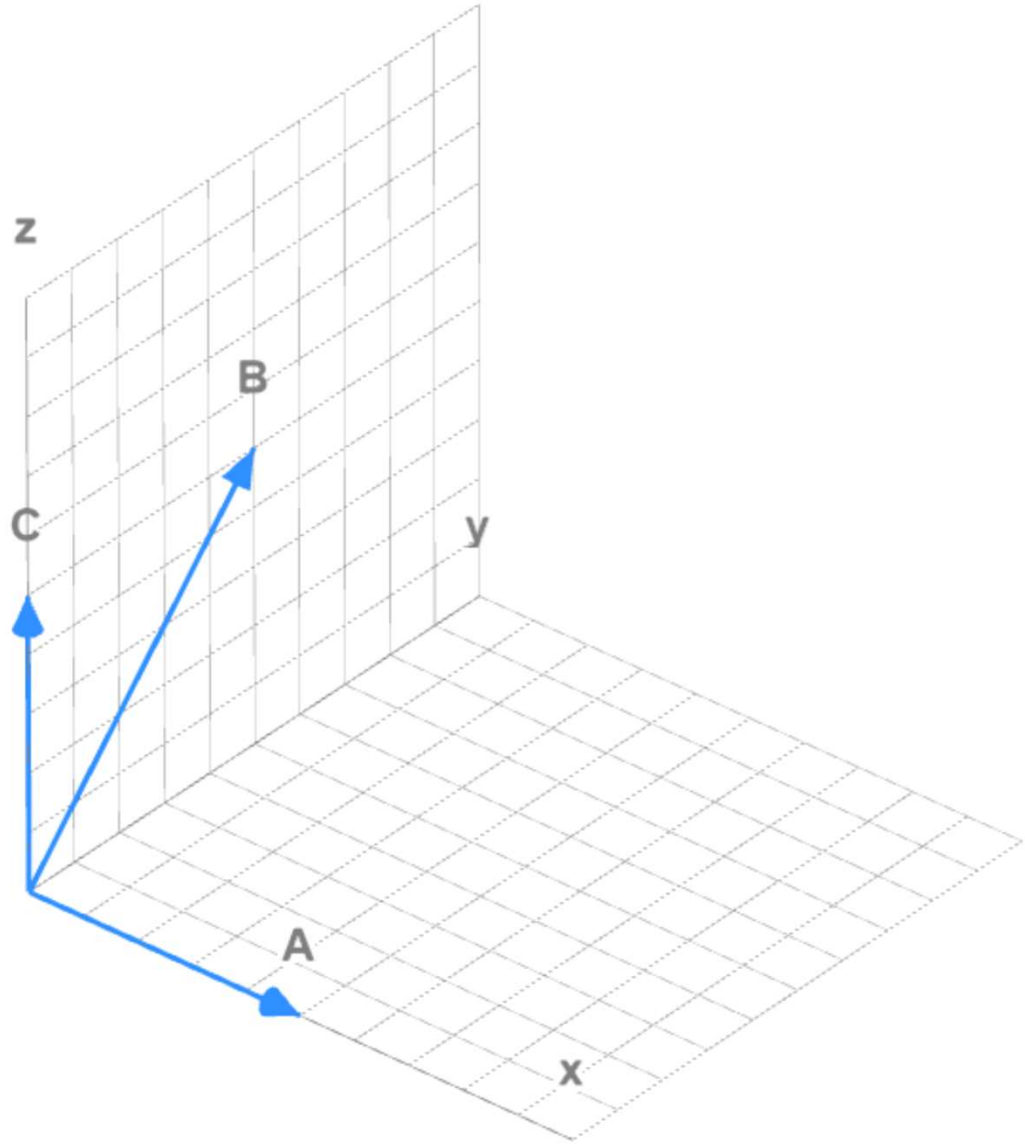


Similar Documents

=

Similar Structure

Document Vectors in Vector Space



Note: vector space can be N-dimensional (N - feature vector length).

How similar are two documents?

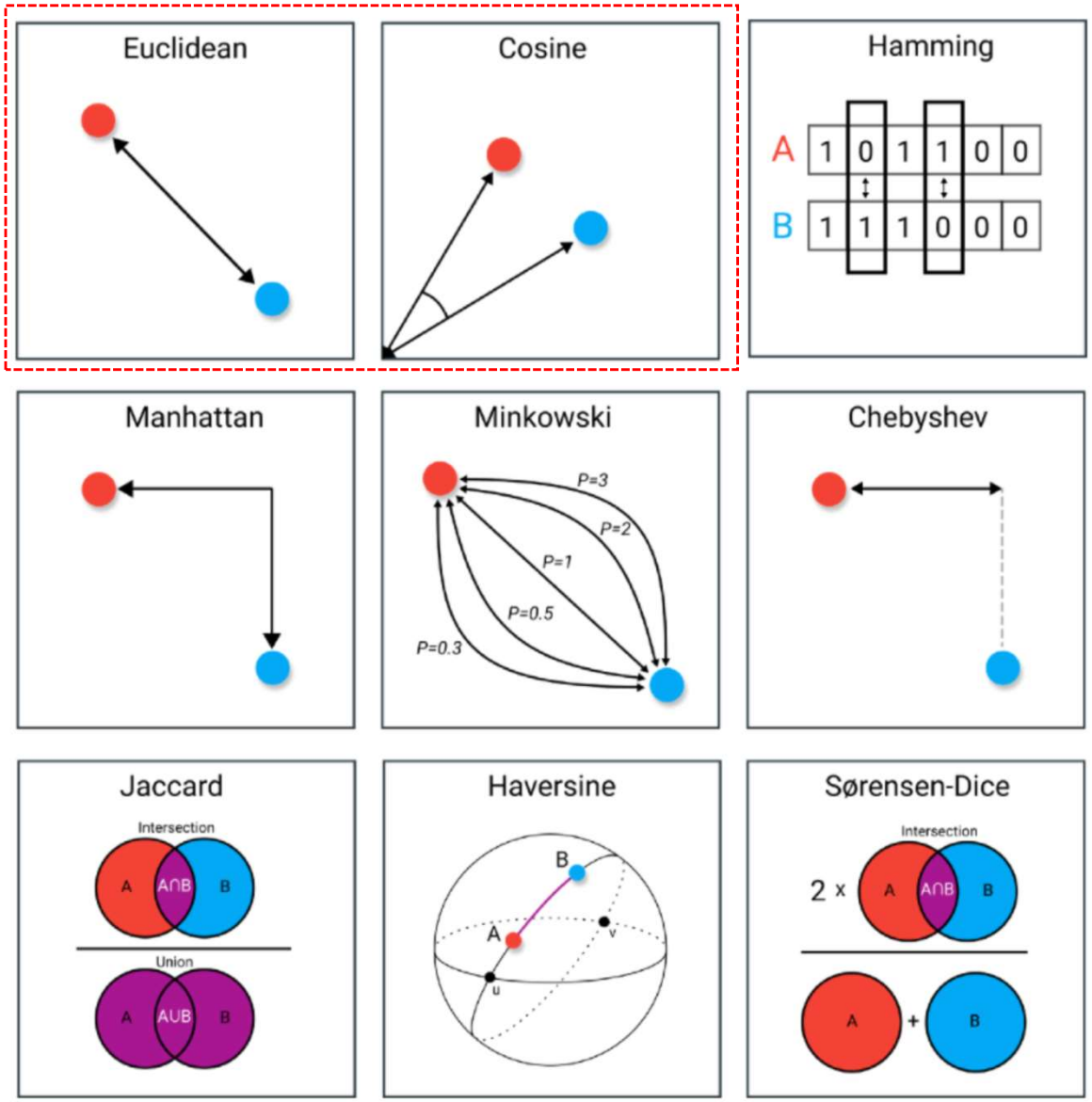
=

How similar are their structures?

=

How close (in a vector space) are points defined by their document vectors

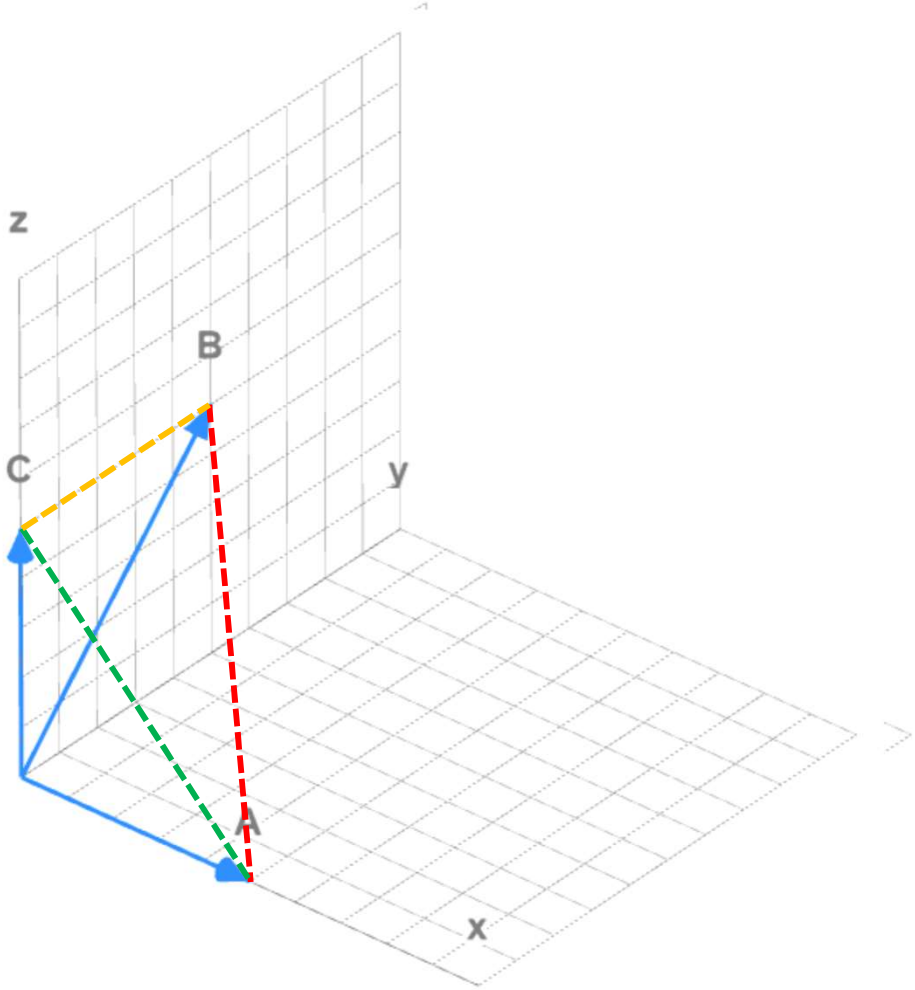
Distance Measures



Source: https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

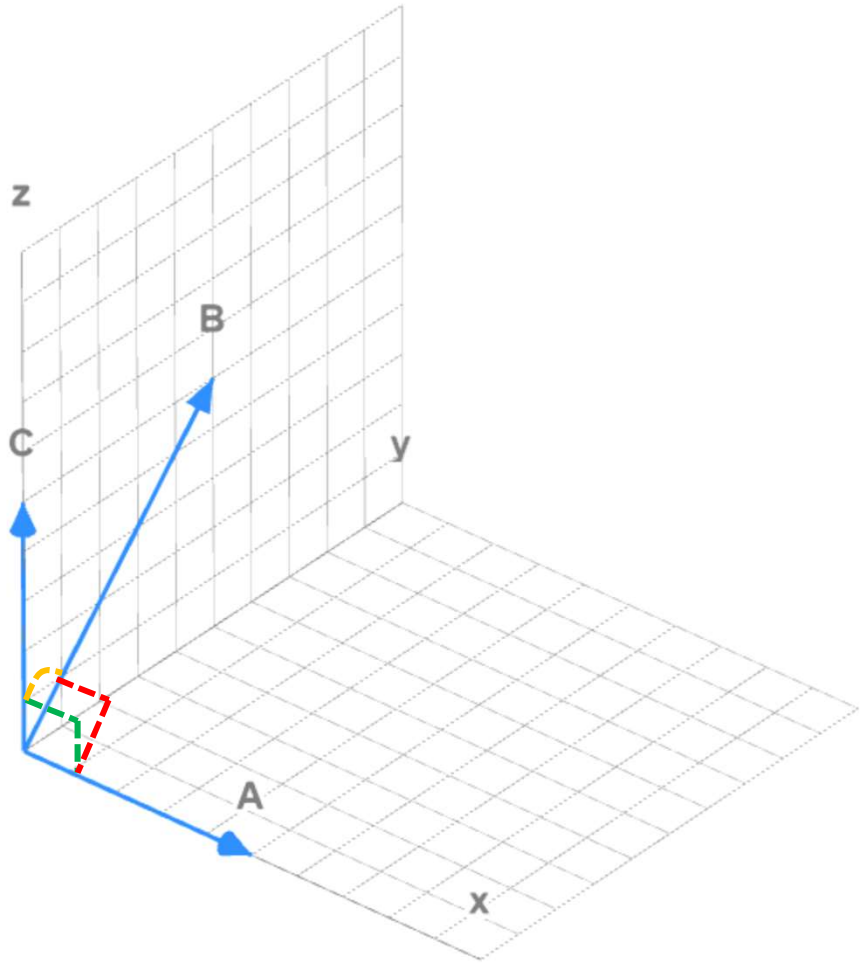
Distance Measures

Euclidean distance



$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2)}$$

Cosine similarity



$$D(x,y) = cos(\theta) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

Bag of Words: Limitations

- Word locations ignored
- Semantics ignored
  - similar / synonymous words could become distinct features

Pre-defined Vocabulary (features):							
Word 1	soccer	Word 3	Word 4	football	Word 6	…	Word N

  - similar sentences will have different vectors

buy	old	desktop	purchase	used	PC	…	Word N
"Buy old desktop" Vector [0-word absent | 1-word present]:							
1	1	1	0	0	0	. . .	0
"Purchase used PC" vector [0-word absent | 1-word present]:							
0	0	0	1	1	1	. . .	0

- New / unknown words | vocabulary range

Text Classification: Definition

Input:

- a document x
- a fixed set of classes $Y = \{y_1, y_2, \ldots, y_J\}$

Output: a predicted class $y \in Y$

Classification: Key Question

Given a document (email, tweet, etc.):

which category / class does it belong to?

Classification: Key Question

Given a document (email, tweet, etc.):

which category / class is the best (predicted) match for this document?

Classification: Key Question

Given a document (email, tweet, etc.):

which category / class is the most probable (= lowest error) for this document?

Classification: Key Question

Given a document (email, tweet, etc.):

which category / class has the highest

$$P(y = \text{class} \mid \mathbf{x} = \quad)?$$

Classification: Key Question

Which category / class has the highest

$$P(y = \text{class}_1 \mid \mathbf{x} = \text{  }) = \text{???}$$

$$P(y = \text{class}_2 \mid \mathbf{x} = \text{  }) = \text{???}$$

...

$$P(y = \text{class}_j \mid \mid \mathbf{x} = \text{  }) = \text{???}$$

Calculate all probabilities ...

Classification: Key Question

Which category / class has the highest

$$P(y = \text{class}_1 \mid \mathbf{x} = \text{  }) = \mathbf{0.1}$$

$$P(y = \text{class}_2 \mid \mathbf{x} = \text{  }) = \mathbf{0.3}$$

…

$$P(y = \text{class}_j \mid \mathbf{x} = \text{  }) = \mathbf{0.2}$$

… and pick the maximum $P()$.

Classification: Key Question

Which category / class has the highest

$$P(y = \text{class}_1 \mid \mathbf{x} = \text{  }) = \mathbf{0.1}$$

$$P(y = \text{class}_2 \mid \mathbf{x} = \text{  }) = \mathbf{0.3}$$

...

$$P(y = \text{class}_j \mid \mathbf{x} = \text{  }) = \mathbf{0.2}$$

Corresponding class → most probable.

Classification: Key Question

Which category / class has the highest

$$P(y = \text{class}_1 \mid \mathbf{x} = \text{  }) = \text{???}$$

$$P(y = \text{class}_2 \mid \mathbf{x} = \text{  }) = \text{???}$$

...

$$P(y = \text{class}_j \mid \mid \mathbf{x} = \text{  }) = \text{???}$$

Calculate all probabilities ... but how?

Bayes' Rule

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

Bayes' Rule: Another Interpretation

Another way to think about Bayes' rule: it allows us to update the hypothesis H in light of some new data/evidence e.

$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

$$P(Hypothesis \mid evidence) = \frac{P(evidence \mid Hypothesis) * P(Hypothesis)}{P(evidence)}$$

where:

- P(H) - **probability of the Hypothesis H being true BEFORE we see new data/evidence e (prior probability)**
- P(H | e) - **probability of the Hypothesis H being true AFTER we see new data/evidence e (posterior probability)**
- P(e | H) - **probability of new data/evidence e being true under the Hypothesis H (likelihood)**
- P(e) - **probability of new data/evidence e being true under ANY hypothesis (normalizing constant)**

Bayes' Rule: Another Interpretation

Another way to think about Bayes' rule: it allows us to update the hypothesis H in light of some new data/evidence e.

$$P(H \mid e) = \frac{P(e \mid H) * P(H)}{P(e)}$$

$$P(Hypothesis \mid evidence) = \frac{P(evidence \mid Hypothesis) * P(Hypothesis)}{P(evidence)}$$

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

$$P(class \mid document) = \frac{P(document \mid class) * P(class)}{P(document)}$$

$$P(y \mid x_1, x_1, \ldots, x_N) = \frac{P(x_1, x_1, \ldots, x_N \mid y) * P(y)}{P(x_1, x_1, \ldots, x_N)}$$

for example:

$$P(y = y_k \mid x_1 = 1, x_1 = 3, \ldots, x_N = 0) = \frac{P(x_1 = 1, x_1 = 3, \ldots, x_N = 0 \mid y = y_k) * P(y = y_k)}{P(x_1 = 1, x_1 = 3, \ldots, x_N = 0)}$$

Bayes' Rule

$$posterior = \frac{likelihood * prior}{evidence}$$

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

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

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

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

Classification: Conditional Probability

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

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

How to calculate?

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

constant

Classifier

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

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

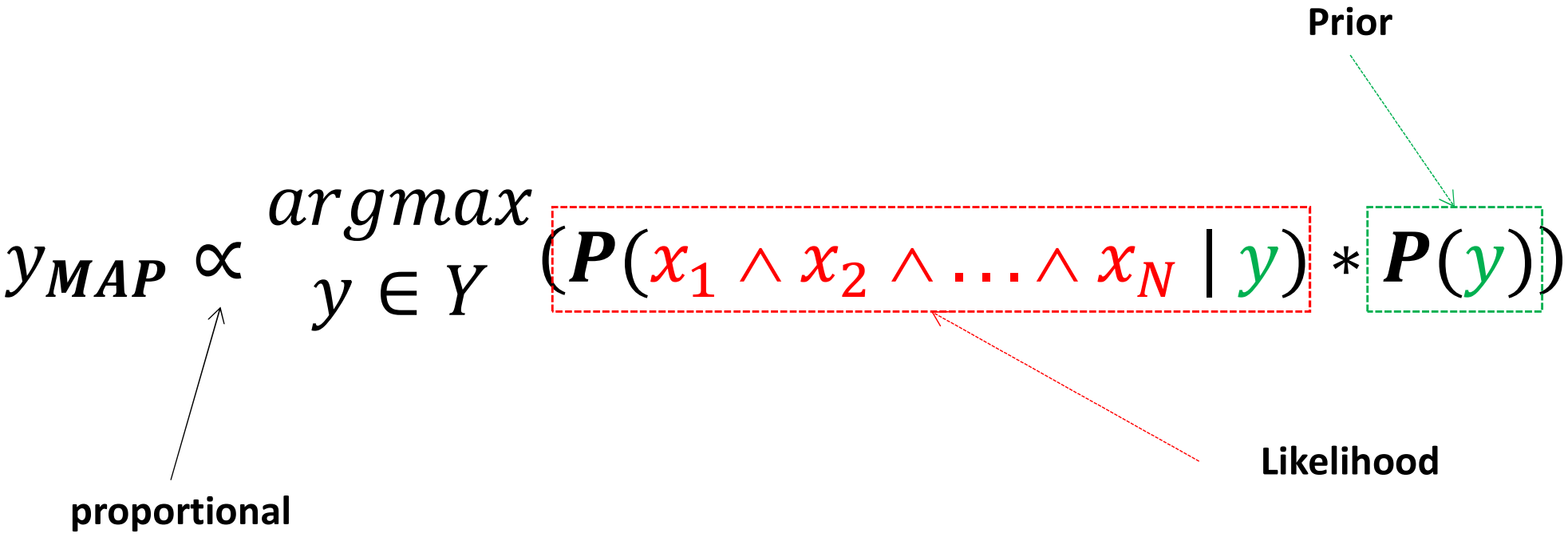
$$y_{MAP} = \underset{y \in Y}{argmax} \left(\frac{\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y) * \boldsymbol{P}(y)}{\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N)}\right)$$

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).

Classifier



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

Classifier

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

How to calculate?

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

Classifier

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

How to calculate?

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

Conditional Probability (Product Rule)

$$P(A \wedge B) = P(A \mid B) * P(B)$$

so:

$$P(A \mid B) * P(B) = P(A \wedge B)$$

Conditional Probability (Product Rule)

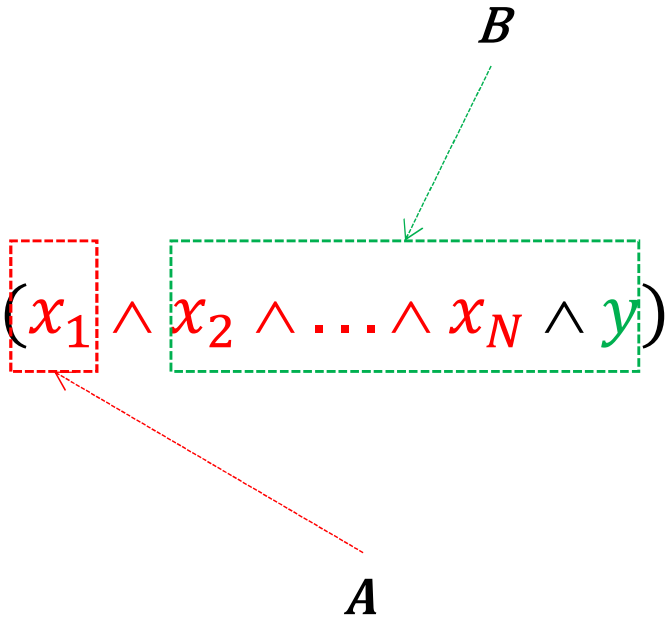
$$\boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y) * \boldsymbol{P}(y) = \boldsymbol{P}(x_1 \wedge x_2 \wedge \ldots \wedge x_N \wedge y)$$

so:

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

Conditional Probability (Product Rule)

$$P(x_1 \wedge x_2 \wedge \ldots \wedge x_N \mid y) * P(y) = P(x_1 \wedge x_2 \wedge \ldots \wedge x_N \wedge y)$$



and

$$P(A \wedge B) = P(A \mid B) * P(B)$$

so:

$$P(x_1 \wedge x_2 \wedge \ldots \wedge x_N \wedge y) = P(x_1 \mid x_2 \wedge \ldots \wedge x_N \wedge y) * P(x_2 \wedge \ldots \wedge x_N \wedge y)$$

Chain Rule

Conditional probabilities can be used to decompose conjunctions using the chain rule. For any events $f_1, f_2, \ldots, f_n$:

$$P(f_1 \wedge f_2 \wedge \ldots \wedge f_n) =$$
$$P(f_1) *$$
$$P(f_2 \mid f_1) *$$
$$P(f_3 \mid f_1 \wedge f_2) *$$
$$\ldots$$
$$P(f_n \mid f_1 \wedge \ldots \wedge f_{n-1}) =$$
$$= \prod_{i=1}^{n} P(f_i \mid f_1 \wedge \ldots \wedge f_{i-1})$$

Expansion

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

Independence

Assume that the knowledge of the truth of one proposition Y, does not affect the agent's belief in another proposition, X, in the context of other propositions Z. We say that X is independent of Y given Z.

Conditional Independence

Random variable X is conditionally independent of random variable Y given Z if for all $x \in Dx$, for all $y \in Dy$, and for all $z \in Dz$, such that

$$P(Y = y \wedge Z = z) > 0 \text{ and } P(Y = y' \wedge Z = z) > 0$$

$$P(X = x \mid Y = y \wedge Z = z) = P(X = x \mid Y = y' \wedge Z = z)$$

In other words, given a value of Z, knowing Y's value DOES NOT affect your belief in the value of X.

Conditional Independence

The following four statements are equivalent as long as conditional probabilities:

1. X **is conditionally independent of** Y **given** Z
2. Y **is conditionally independent of** X **given** Z
3. $P(X \mid Y, Z) = P(X \mid Z)$
4. $P(X, Y \mid Z) = P(X \mid Z) * P(Y \mid Z)$

Naive Bayes Assumption

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

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 $y \rightarrow$ Naive Bayes assumption.

Under this assumption:

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

Naive Bayes Assumption

Under Naive Bayes assumption:

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

becomes:

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

Naive Bayes Classifier

Under Naive Bayes assumption:

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

becomes:

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

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

Naive Bayes Classifier

Under Naive Bayes assumption:

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

becomes:

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

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

Naive Bayes Classifier

Under Naive Bayes assumption:

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

becomes:

How to calculate?

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

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

Text Classification: Supervised ML

Input:

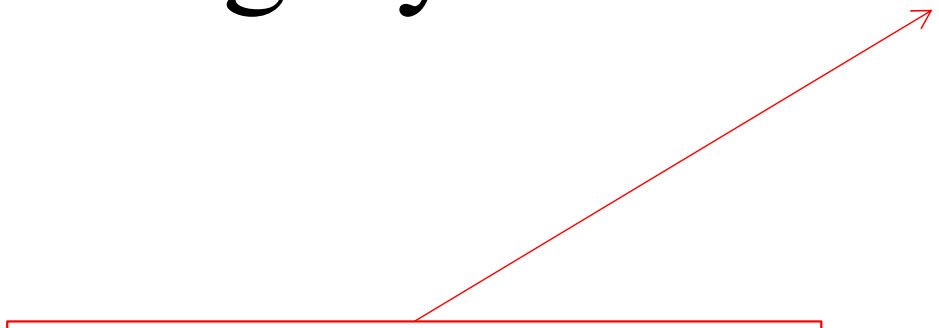
- a document $\boldsymbol{x}$
- a fixed set of classes $Y = \{y_1, y_2, \ldots, y_J\}$
- a training set of N hand-labeled documents $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$

Output:

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

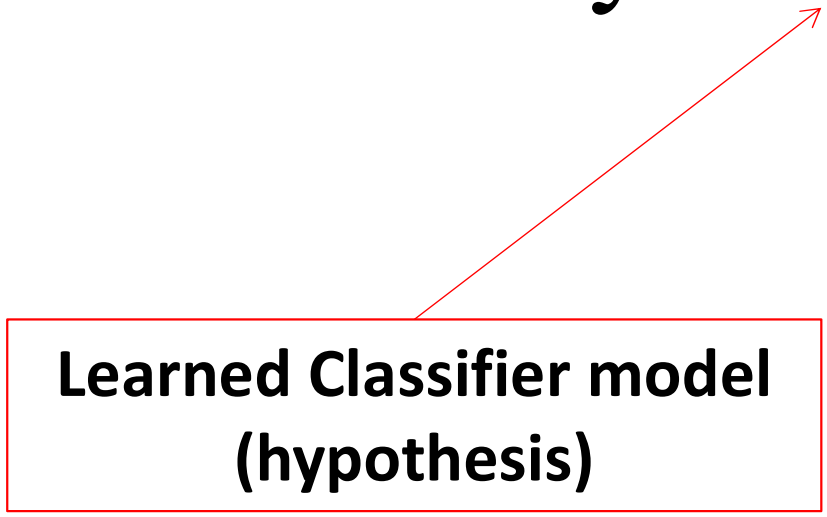
Text Classification: Classifier

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



Text Classification: Classifier

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



Text Classification: Supervised ML

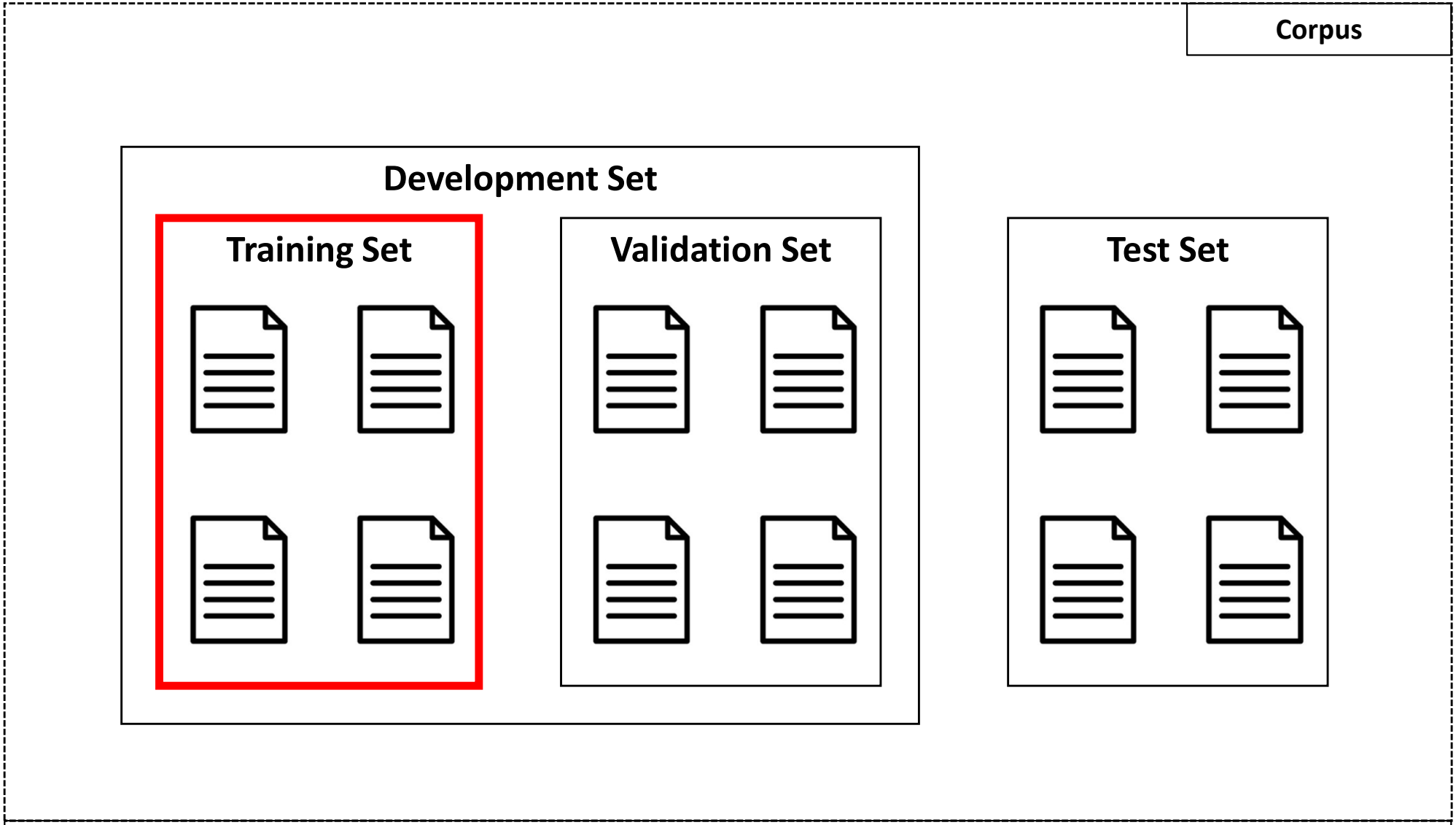
Input:

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

Output:

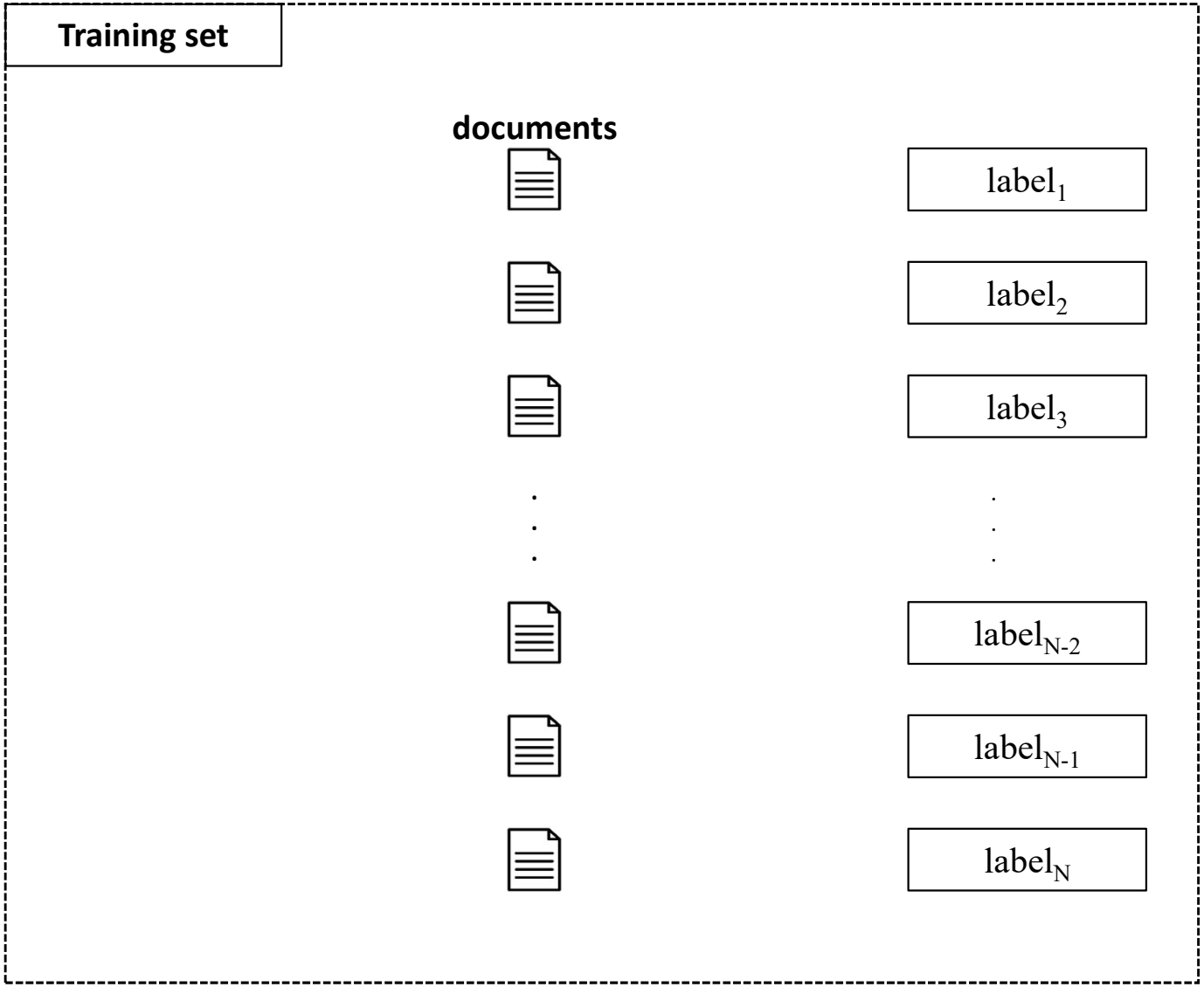
- a learned classifier $h: \mathbf{x} \rightarrow y$ ($y = h(\mathbf{x})$)

Corpus: Training / Validation / Test

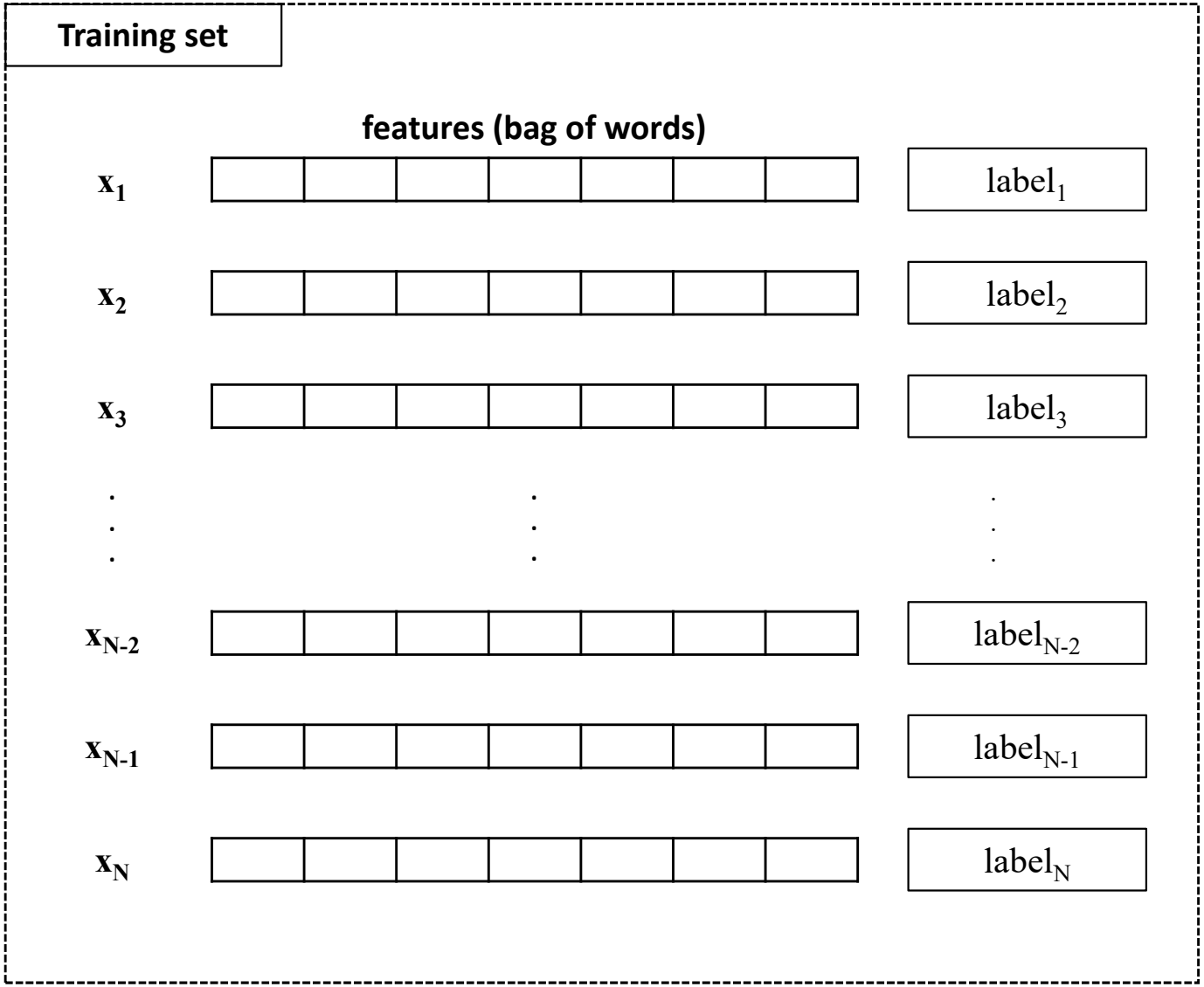


Typical training / validation / test set split for a text corpora

Text Classification: Training Set

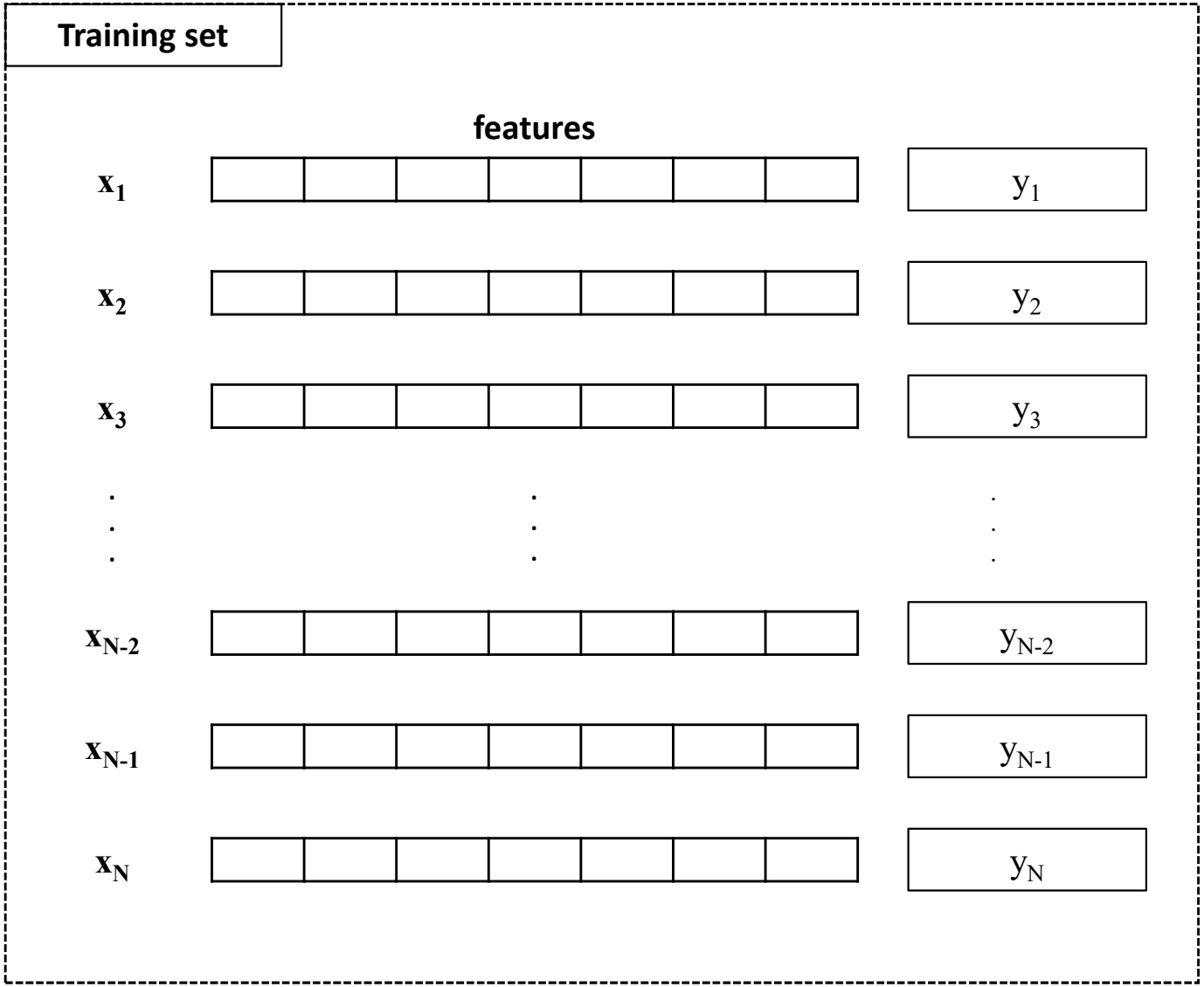


Text Classification: Training Set



$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - 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

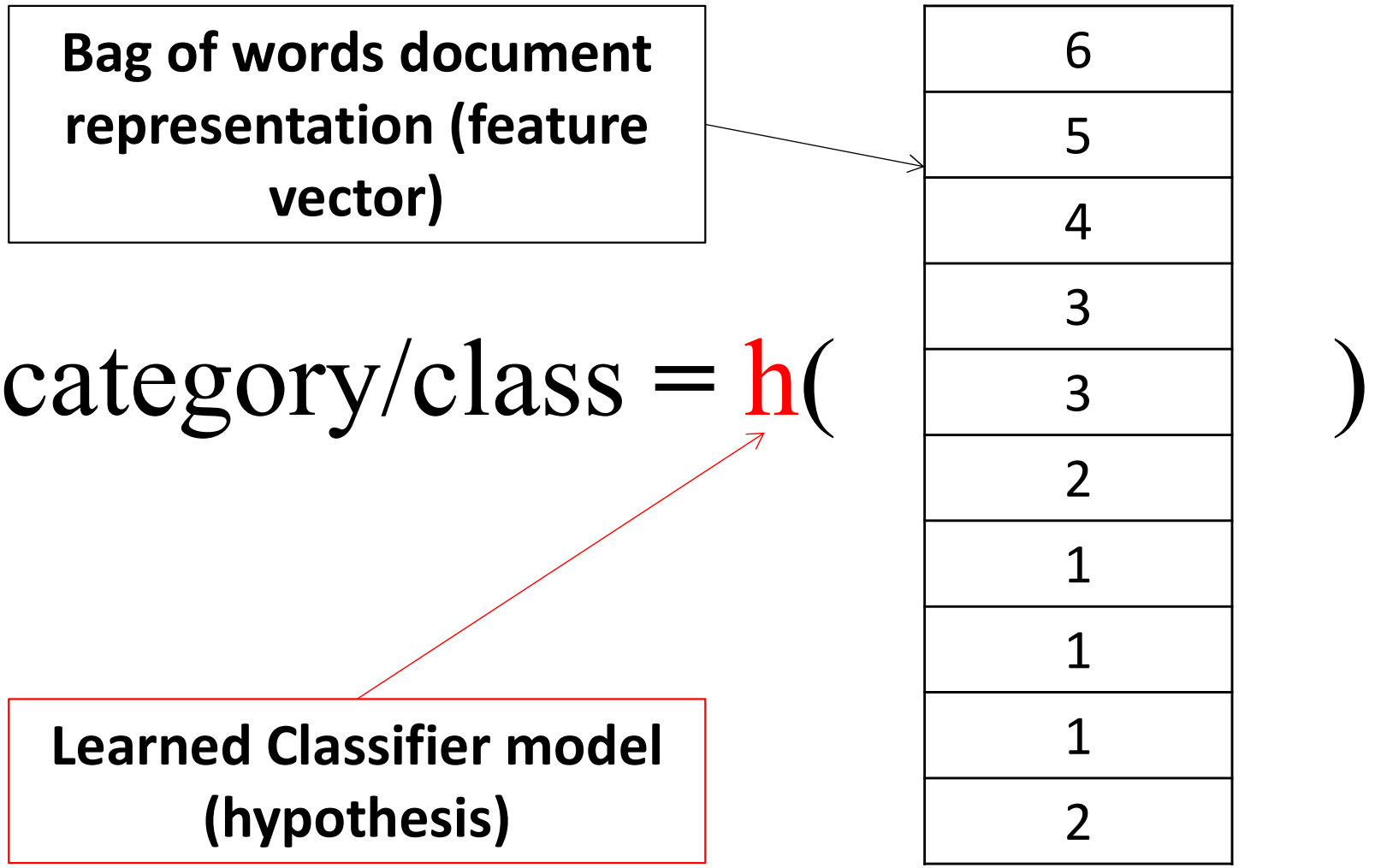
Training set

Vocabulary V

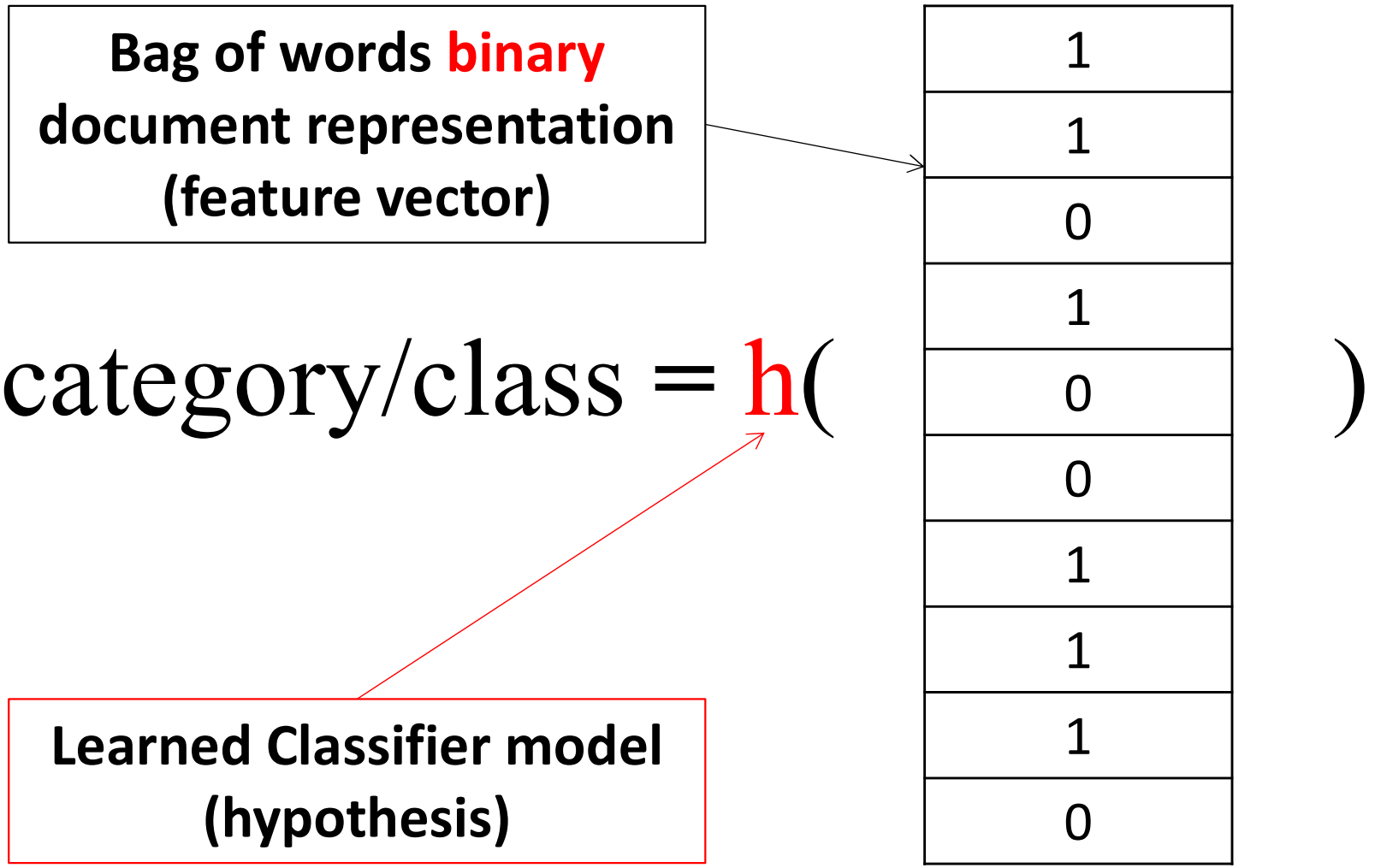
	word1	rolex	word3	replica	word5	word6	word7	
$\mathbf{x}_1$	0	0	1	0	1	1	1	y_1=HAM
$\mathbf{x}_2$	1	0	1	1	0	1	1	y_2=HAM
$\mathbf{x}_3$	0	1	0	1	0	1	1	y_3=SPAM
. . .	. . .			. . .				. . .
$\mathbf{x}_{N-2}$	1	1	1	1	0	1	1	y_{N-2}=HAM
$\mathbf{x}_{N-1}$	1	1	0	1	0	0	1	y_{N-1}=SPAM
$\mathbf{x}_N$	1	0	0	1	0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

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

Text Classification: Bag of Words

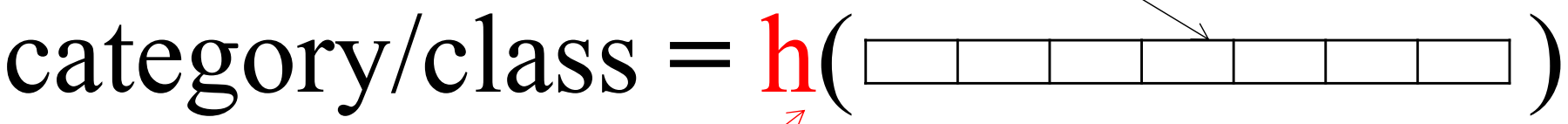


Text Classification: Bag of Words



Text Classification: Bag of Words

Bag of words document representation (feature vector)



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	
$\mathbf{x}_1$	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 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

Building a Naïve Bayes Classifier

Classifier

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

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

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

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) $x_1, x_2, \ldots, x_N$ are mutually independent**
  - **Bag-of-words representation: the order of the words in a document d makes no difference (repetitions do)**
- **All events (words) $x_1, x_2, \ldots, x_N$ 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

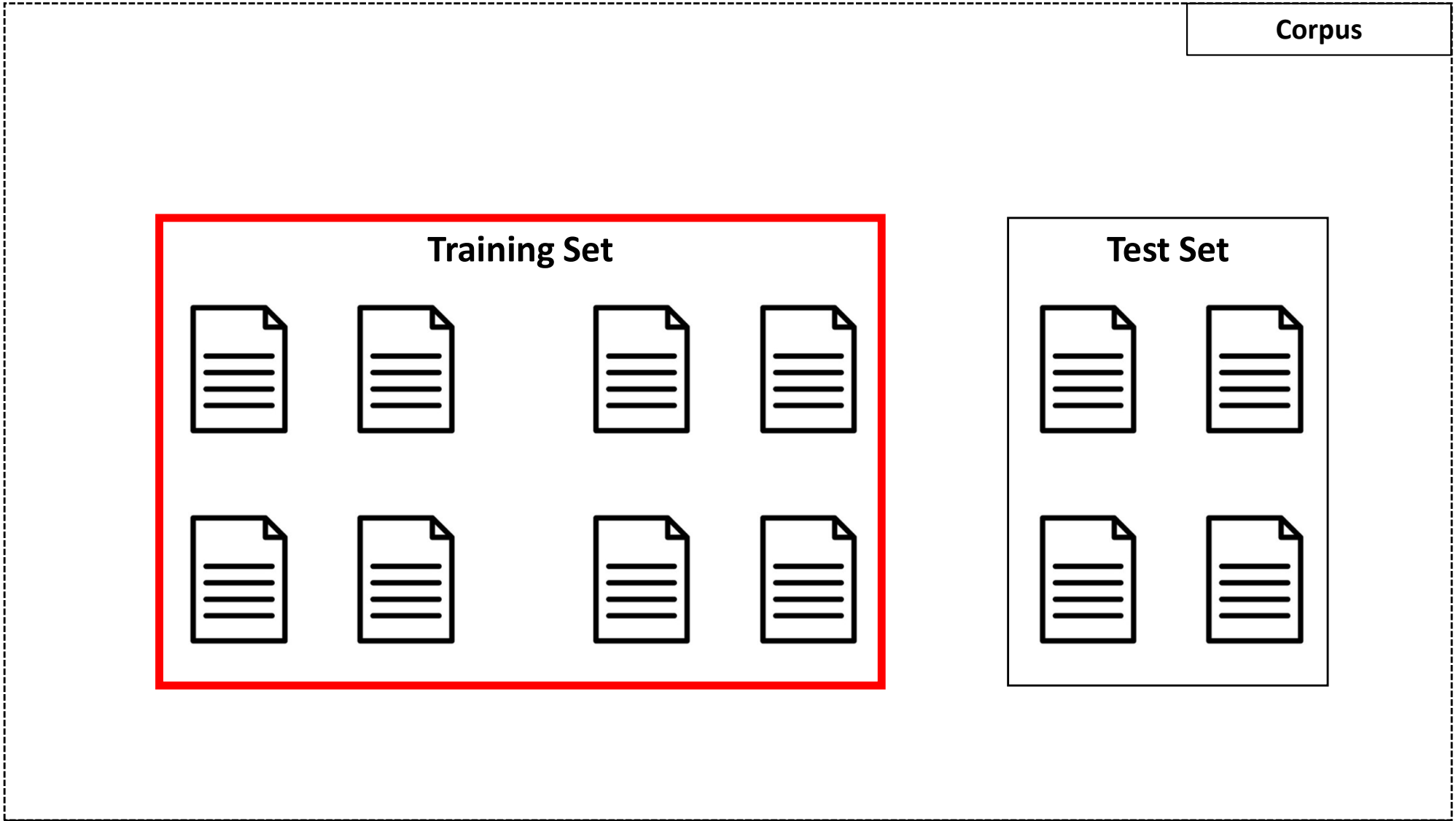
$$\text{category/class} = \text{h}(\textbf{document})$$

Finding model / hypothesis h → Finding probabilities for y_{MAP}

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

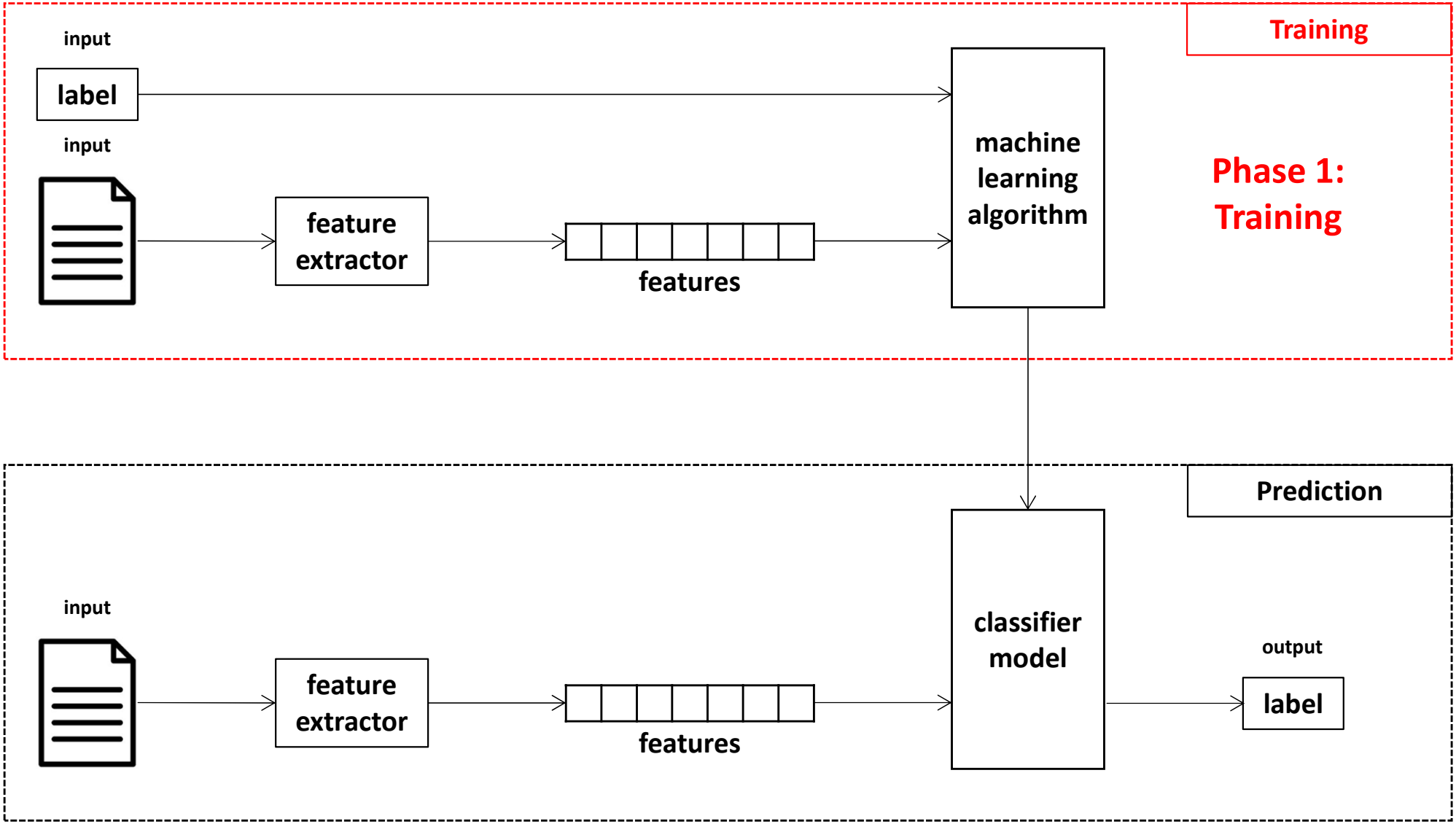
MAP: Maximum a posteriori (corresponds to the most likely class).

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	
$\mathbf{x}_1$	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 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

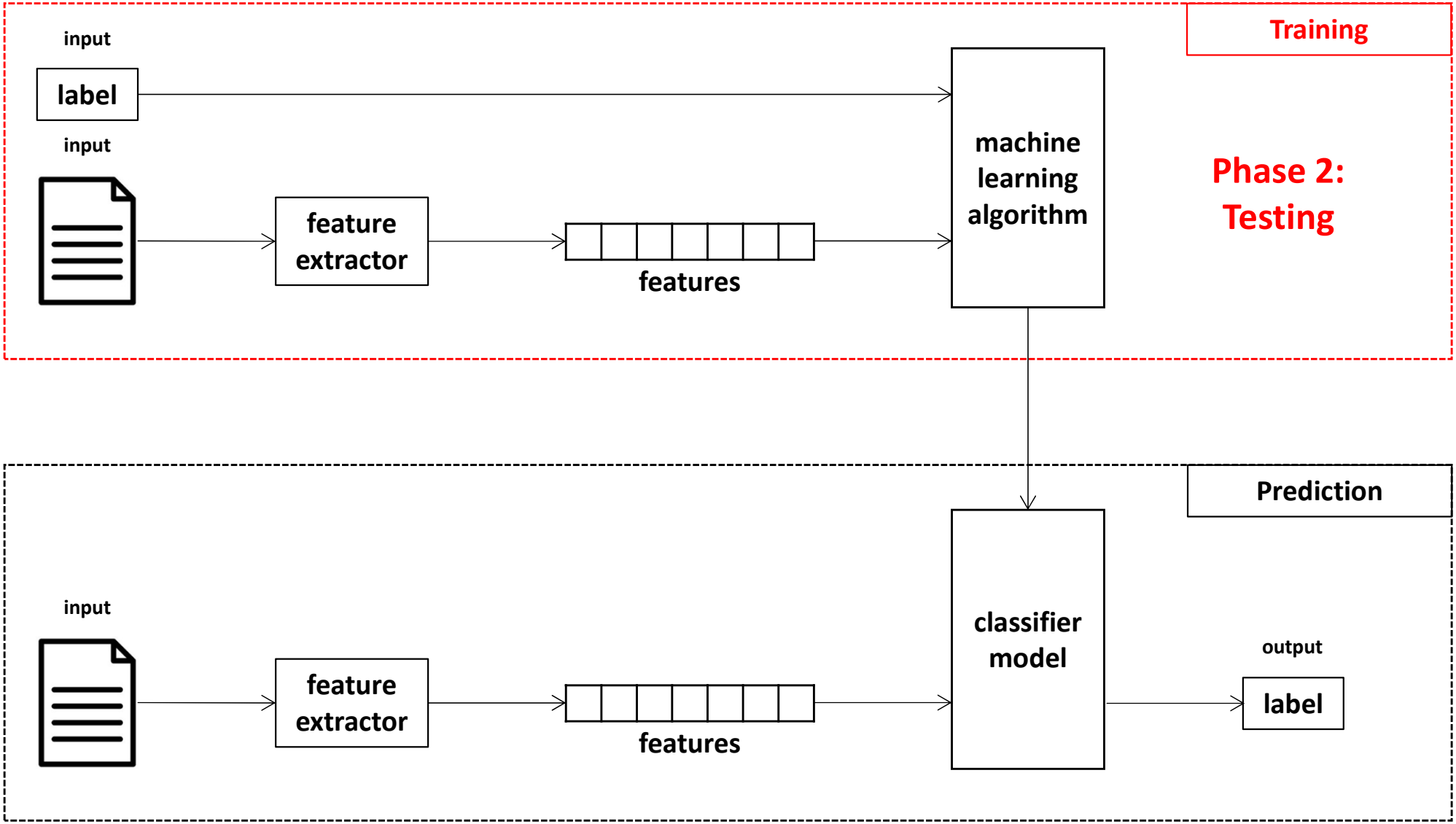
$$P(x_i = cheap \mid y = SPAM) = \frac{count(x_i = cheap, y = SPAM)}{\sum_{x \in V} count(x, y = SPAM)} = \frac{2}{7}$$

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N-2}, \mathbf{x}_{N-1}, \mathbf{x}_N$ - 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

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

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

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

Learned h()

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

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

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

category/class = $h(\mathbf{x_8})$

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Learned h()

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

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

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

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 | 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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_{N-2}}, \mathbf{x_{N-1}}, \mathbf{x_N}$ - 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()?

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

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

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

For document x_8 $y = SPAM$ 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

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

For document 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	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

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

category/class = $h(\mathbf{x_9})$

Learned h()

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

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

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

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(**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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

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(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 V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

$$\boldsymbol{P}(y = HAM \mid \boldsymbol{x_9}) \propto \boldsymbol{P}(y = HAM) * \prod_{i=1}^{5} \boldsymbol{P}(x_i \mid y = HAM) =$$

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

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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

$$\boldsymbol{P}(y = HAM \mid \boldsymbol{x_9}) \propto \boldsymbol{P}(y = HAM) * \prod_{i=1}^{5} \boldsymbol{P}(x_i \mid y = HAM) =$$

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

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_{N-2}}, \mathbf{x_{N-1}}, \mathbf{x_N}$ - 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()?

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x_9}) \propto \boldsymbol{P}(y = SPAM) * \prod_{i=1}^{5} \boldsymbol{P}(x_i \mid y = SPAM) =$$

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

Learned h()

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

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

$$\boldsymbol{P}(y = SPAM \mid \boldsymbol{x_9}) \propto \boldsymbol{P}(y = SPAM) * \prod_{i=1}^{5} \boldsymbol{P}(x_i \mid y = SPAM) =$$

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

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

$$P(y = HAM \mid x_9) \approx 0.000056$$

$$P(y = SPAM \mid x_9) = 0$$

For document 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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

For document x_9 $y = HAM$ maximizes P(). Class = HAM.

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	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

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

category/class = $h(\mathbf{x_{10}})$

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

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(**replica**)

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \ldots, \mathbf{x_{N-2}}, \mathbf{x_{N-1}}, \mathbf{x_N}$ - 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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

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)

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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

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

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

For document x_{10} $y = SPAM$ maximizes P(). Class = $SPAM$.

Learned h()

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Spam Detection: Testing Classifier

Test set

Vocabulary V

	I	rolex	own	replica	watch	buy	cheap	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing

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

For document x_{10} $y = SPAM$ maximizes P(). Class = $SPAM$.

Incorrect classification! Misclassification.

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 Evaluation: Confusion Matrix

		Predicted class: Positive	Predicted class: 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	Predicted class: 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	
x_8	0	1	0	1	0	1	1	y_8=SPAM

buy cheap rolex replica rolex

x_9	1	1	1	0	1	0	1	y_9=HAM

I own cheap rolex watch

x_{10}	0	0	0	1	0	0	0	y_{10}=HAM

replica

Testing results

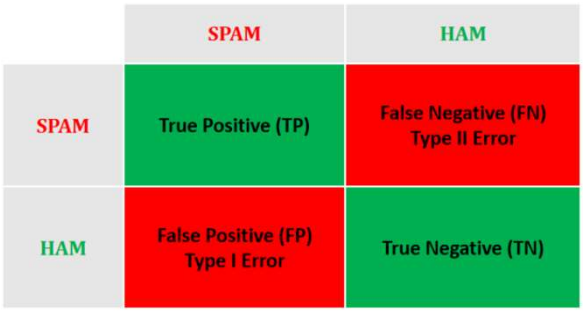
y_8=SPAM	correct
y_9=HAM	correct
y_{10}=SPAM	incorrect

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



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

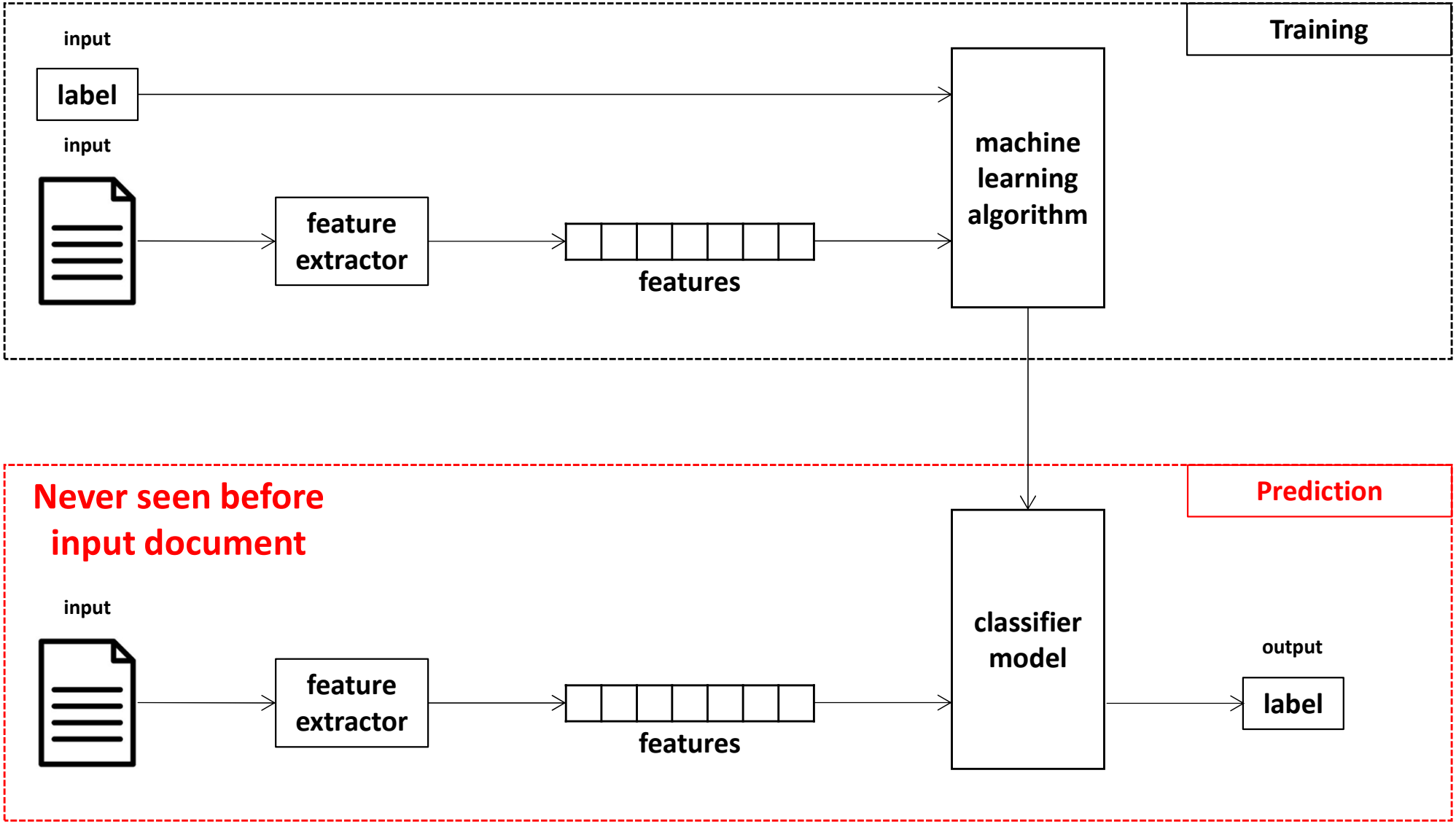
Classifier Evaluation: Confusion Matrix

		Predicted class: SPAM	Predicted class: 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
$x_?$	0	1	0	0	0	1	0

buy rolex

Label needs to be decided

$y_?$=????

Prediction

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

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

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

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

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

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

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

For document $x_?$ $y = SPAM$ 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 → Finding probabilities for y_{MAP}

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

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

Naive Bayes Classifier

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

Finding model / hypothesis h → Finding probabilities for y_{MAP}

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

- **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