

## Document Classification

- ◆ Given:
  - Corpus of pre-classified documents
- ◆ Task:
  - Classify unseen documents

1

## Spam Filtering

- ◆ Given:
  - Spam messages
  - Non-spam messages
- ◆ Task:
  - Predict whether new message is spam

2

## Representation

- ◆ How to represent message?
- ◆ Unordered list of words
  - High dimensionality
  - Useless words (“and”, “the”)
- ◆ Dictionary
  - Small set of useful words (“mortgage”)
  - Ignore all other words

3

## Representation

- ◆ Dictionary:  $n$ -tuple of words
- ◆ Document:  $n$ -tuple of booleans:
$$\vec{x} = \langle x_1, x_2, \dots, x_n \rangle$$
$$x_i = \begin{cases} 1 & \text{if word } i \text{ appears in message} \\ 0 & \text{otherwise} \end{cases}$$

4

## Example

- ◆ Spam Messages:
  - “Get a *mortgage*, *buy* a house!”
  - “*Buy Viagra* from the *Viagra* store”
- ◆ Non-spam Messages:
  - “*Machine learning* homework...”
  - “Your Linux *machine* is fixed”

5

## ◆ Dictionary:

$\langle \text{buy, mortgage, viagra, machine, learning} \rangle$

## ◆ Documents:

$$\vec{d}_1 = \langle 1, 1, 0, 0, 0 \rangle \quad \vec{d}_3 = \langle 0, 0, 0, 1, 1 \rangle$$

$$\vec{d}_2 = \langle 1, 0, 1, 0, 0 \rangle \quad \vec{d}_4 = \langle 0, 0, 0, 1, 0 \rangle$$

## ◆ Corpus:

$$X = \{ \vec{d}_1, \vec{d}_2, \vec{d}_3, \vec{d}_4 \}$$

6

## Learning

- ◆ Learner  $L$  builds a classifier  $h$  given corpus  $X$  and classifications  $C$

$$h = L(X, C)$$

$$c_j = \begin{cases} 1 & \text{if document } \vec{x}_j \text{ is spam} \\ 0 & \text{otherwise} \end{cases}$$

7

## Classification

- ◆ Classifier  $h$  predicts class of new message  $\vec{x}$

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

$$y = \begin{cases} 1 & \text{if } \vec{x} \text{ is predicted spam} \\ 0 & \text{otherwise} \end{cases}$$

8

## Example

- ◆ **FIND-WORD**: learner that finds word  $w_s$  that only appears in spam messages

- ◆  $h = \text{FIND-WORD}(X, C)$

$$h(\bar{x}) = \begin{cases} 1 & \text{if } \bar{x} \text{ contains } w_s \\ 0 & \text{otherwise} \end{cases}$$

- ◆ In our corpus,  $w_s$  could be “buy”

9

## Where To Go From Here?

- ◆ Use word frequency

- ◆ How to build dictionary?

- How to know “mortgage” and “buy” are good spam words?

- ◆ What happens when spam words change?

- How to predict what words are useful tomorrow?

10