CS 412 Introduction to Machine Learning

Naïve Bayes

Instructor: Wei Tang

Department of Computer Science
University of Illinois at Chicago
Chicago IL 60607

https://tangw.people.uic.edu
tangw@uic.edu

Slides credit: Xinhua Zhang

Outline

- KNN only works for continuous-valued features
- A lot of real data are discrete (e.g., text)
- Bag-of-words representation
- Naïve Bayes classifier
  - Employ Bayes Theorem
  - Assume feature independence

Motivating Example: Spam Filter

DONATION OF €1.5 MILLION EURO!!! External Spam ×

Frances & Patrick Connolly <silvio.confalone@com... Sep 2, 2021, 5:07 PM

to

This message seems dangerous

Similar messages were used to steal people's personal information. Avoid clicking links, downloading attachments, or replying with personal information.

Looks safe

Dear Beneficiary

Our Names are Frances and Patrick Connolly from County Armagh in Northern Ireland. We Just Won €115 Million Euro from the EuroMillions lottery jackpot Lottery draw. We are therefore giving out a Grant donation of €1.5 Million Euro each to (15) Lucky international recipients worldwide to show God our appreciation. You received this message because you have been listed as one of the (15) lucky Individual Selected. Send your name, Address, Phone, Country, for claims Now.

Bag of Words

Word order does not matter

- Sounds silly, but often works well!

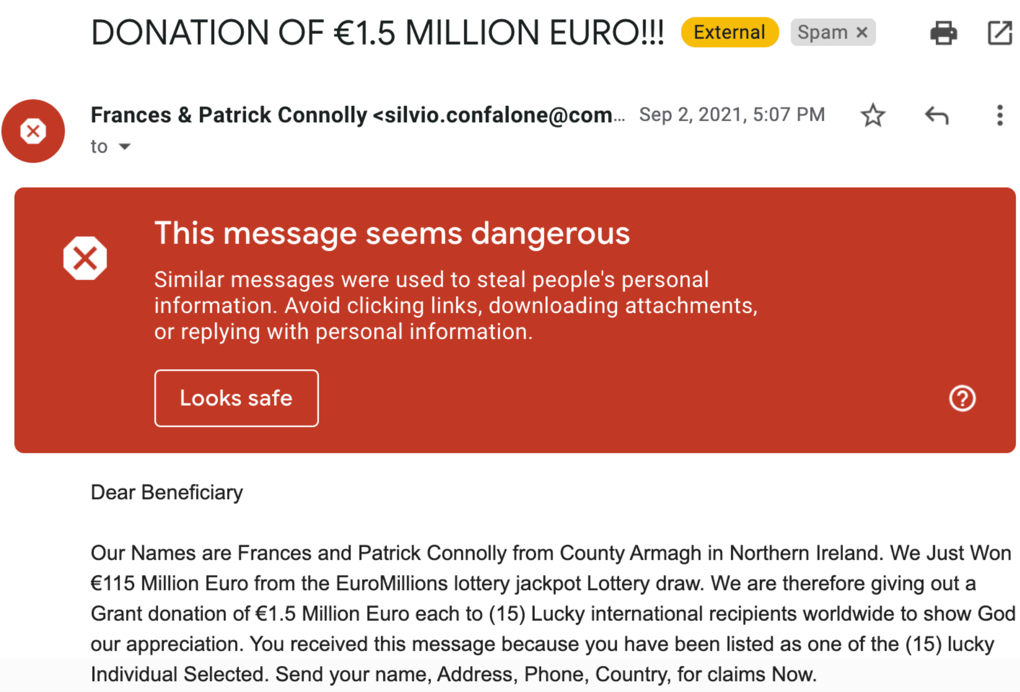
When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

in is lecture lecture next over person remember room sitting the the the to to up wake when you

Let $X_{'at'}$ be a random variable of the frequency of 'at' in a document.
Then $X_{'at'} \sim$ Multinoulli (Θ)

e.g., $P(X_{'at'}=0)=0.1$, $P(X_{'at'}=1)=0.2$, $P(X_{'at'}=2)=0.1$, $P(X_{'at'}>2)=0.6$

Motivating Example: Spam Filter



abandon	0	$X_{abandon} = 0$
about	0	$X_{about} = 0$
all	0	$X_{all} = 0$
an	1	
as	1	
at	2	
…		
donation	1	
…		
lottery	1	
…		
million	3	$X_{million} = $ '>2'

Dictionary can be just the collection of words that appear in the dataset.

Probabilistic Spam Filter

$$\hat{P}(\mathrm{spam} \mid \quad)$$

abandon	0
about	0
all	0
an	1
as	1
at	2
...	
donation	1
...	
lottery	1
…	
million	3

A Probabilistic Classifier

Supervised Learning:

Predict (binary) class Y given feature values $\mathbf{x}_{1:d}$

d: size of the dictionary

$$P(\text{spam} \mid \quad)$$

abandon	0
an	1
as	1
at	2
...	
donation	1
...	
lottery	1
…	
million	3

A Probabilistic Classifier

Supervised Learning:

Predict (binary) class Y given feature values $\mathbf{x}_{1:d}$

Training: Estimate the value of $P(\mathbf{x}_{1:d} | Y)$ and $P(Y)$

Testing:

1. Compute $P(Y | \mathbf{x}_{1:d})$ for all $\mathbf{x}_{1:d}$ by using the Bayes theorem on $P(\mathbf{x}_{1:d} | Y)$ and $P(Y)$

2. Predict $y = \text{argmax}_y P(y | \mathbf{x}_{1:d})$

Big problem: Too many parameters to estimate

If $|X| = 10$ (possible values) and $d = 7$,
how many parameters do we need to estimate?

Bayes Theorem

$$P(Y \mid X_{1:d}) = \frac{P(X_{1:d} \mid Y)\, P(Y)}{P(X_{1:d})}$$

$$= \frac{P(X_{1:d} \mid Y)\, P(Y)}{\sum_{Y'} P(X_{1:d} \mid Y')\, P(Y')}$$

Naïve Bayes: Independence Assumptions

Assume features are independent given class:

$$P(\mathbf{x}_{1:d} \mid y) = \prod_{j=1:d} P(x_j \mid y)$$

$$P(X_{\text{lottery}} = 1, X_{\text{million}} = 2 \mid \text{spam})$$
$$= P(X_{\text{lottery}} = 1 \mid \text{spam}) * P(X_{\text{million}} = 2 \mid \text{spam})$$

equivalently:

$$\forall\, j, x, y: \quad P(x_j \mid y, x_{-j}) = P(x_j \mid y)$$
$$\forall\, j: \quad X_j \perp X_{-j} \mid Y$$

How many parameters now? $\quad d\,(|X| - 1)$

Naïve Bayes: Independence Assumptions

Joint probability distribution:

$$P(\mathbf{x}_{1:d}, y) = P(y) \prod_{j=1:d} P(x_j \mid y)$$

<u>Learning</u>

Maximum likelihood:

$$\mathrm{argmax}_\theta \, P(X,Y \mid \theta)$$

Estimating: $\Theta = \{P(Y), P(X_j \mid Y)\}$

Discrete Features for bag of words

- Binary features:
  - Dictionary has d words $x_1, \ldots, x_d$
  - Only model whether a word appeared in a doc: $x_i \in \{0,1\}$
  - $x_2=1$ if the 2nd word in the dictionary appeared in a doc

$$p_{ij} \equiv p(x_j = 1 \mid C_i) \qquad p(x \mid C_i) = \prod_{j=1}^{d} p_{ij}^{x_j} (1 - p_{ij})^{(1-x_j)}$$

(Naive Bayes: x_j are conditionally independent)

  - The prediction function is linear

$$g_i(\mathbf{x}) = \log p(\mathbf{x} \mid C_i) + \log P(C_i)$$
$$= \sum_j \left[x_j \log p_{ij} + (1 - x_j) \log (1 - p_{ij}) \right] + \log P(C_i)$$

Estimated parameters $\hat{p}_{ij} = \dfrac{\sum_t x_j^t r_i^t}{\sum_t r_i^t}$

What about $P(C_i)$?

Discrete Features for bag of words

- Multinomial (1-of-n_j) features: $x_j \in \{v_1, v_2, \ldots, v_{n_j}\}$
  - E.g., model the frequency of 0, 1, 2, >2

$$p_{ijk} \equiv p(z_{jk}=1 \mid C_i) = p(x_j = v_k \mid C_i)$$

$Z_{jk} = 1$ if $x_j = v_k$
0 else

if x_j are independent

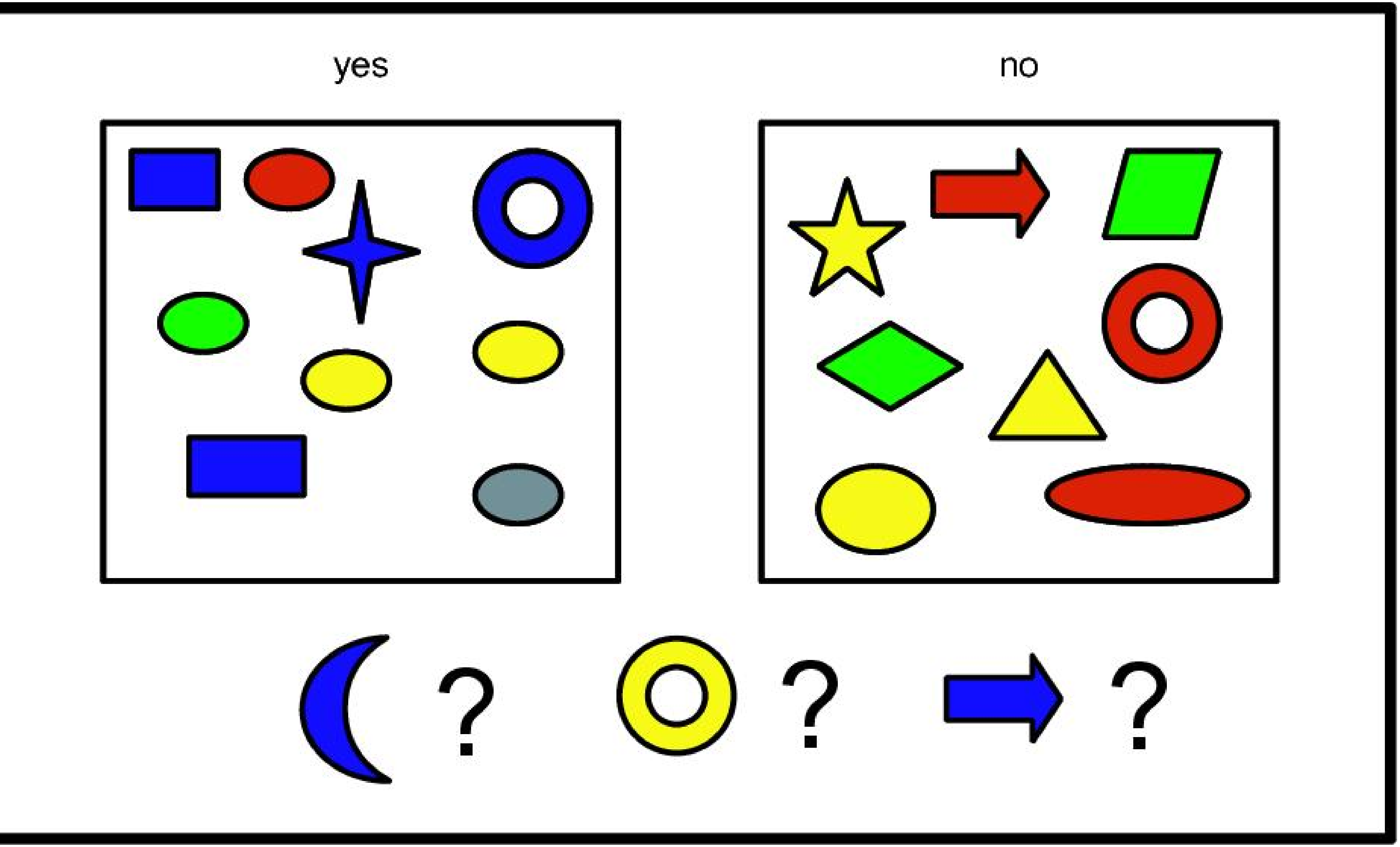
What does it mean if we drop j in p_{ijk}?

$$p(\mathbf{x} \mid C_i) = \prod_{j=1}^{d} \prod_{k=1}^{n_j} p_{ijk}^{z_{jk}}$$

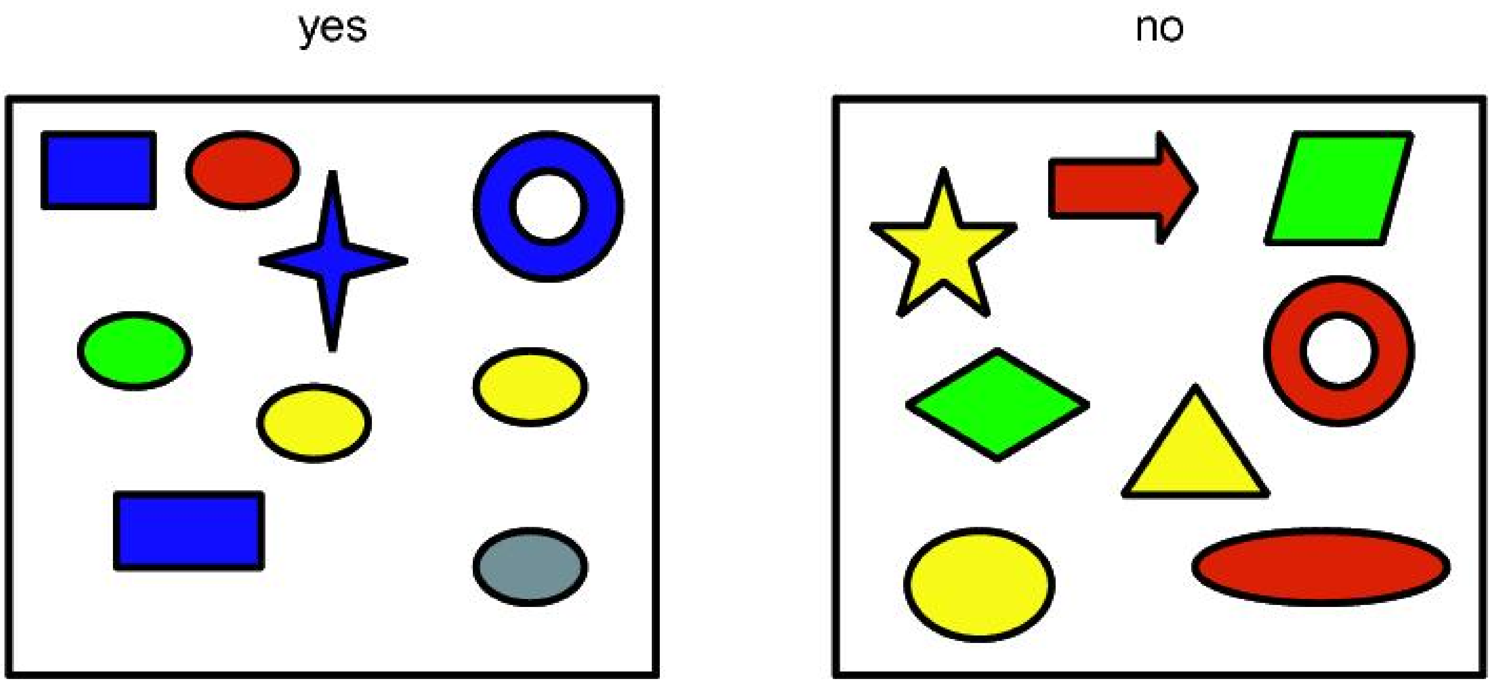
$$g_i(\mathbf{x}) = \sum_j \sum_k z_{jk} \log p_{ijk} + \log P(C_i)$$

$$\hat{p}_{ijk} = \frac{\sum_t z_{jk}^t r_i^t}{\sum_t r_i^t}$$

Estimation



Estimation



1. Choose binary-valued (for simplicity) property of objects
2. Estimate $P(X_i = yes|Class = yes)$ and $P(X_i = yes|Class=no)$

   e.g., X_1*: Blue,* X_2*: Ellipse,* X_3*: Green, (or further,* X_4*: Arrow, …)*

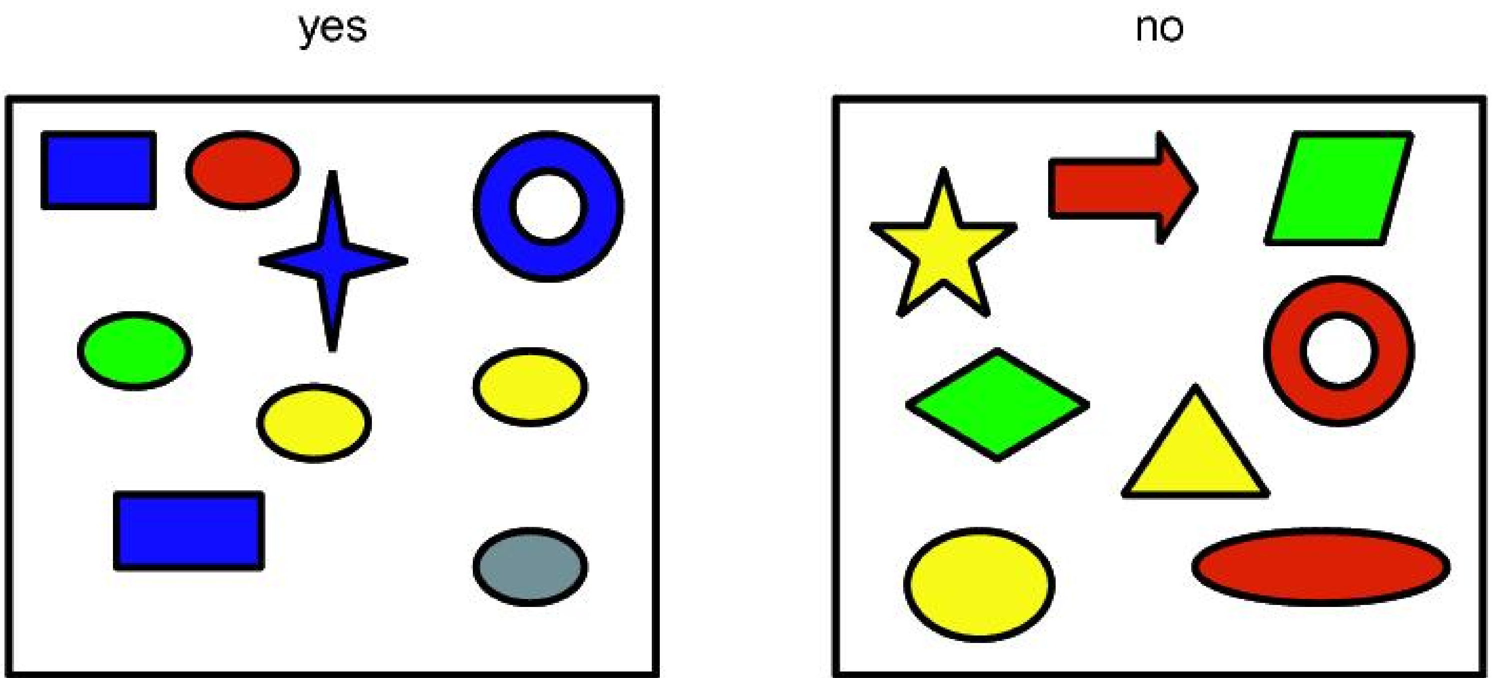
$P(yes)=9/17$, $p(blue \mid yes)=4/9$, $p(ellipse \mid yes)=6/9$, $p(green \mid yes)=1/9$

$P(no)=8/17$, $p(blue \mid no) = 0$, $p(ellipse \mid no) = 3/8$, $p(green \mid no)=2/8$

NB: $p(blue|yes)$ is a shorthand of $p(Blue = yes \mid Class = yes)$

P(yes)=9/17, p(blue | yes)=4/9, p(ellipse | yes)=6/9, p(green | yes)=1/9
P(no)=8/17, p(blue | no) = 0, p(ellipse | no) = 3/8, p(green | no)=2/8

Prediction



What is the prediction for red star? 

$P(yes|X_1 \cdots X_3) \propto P(yes) \cdot P(X_1|yes) \cdot P(X_2|yes) P(X_3|yes)$

$= \frac{9}{17} \cdot \frac{5}{9} \cdot \frac{3}{9} \cdot \frac{8}{9} = 0.087$

$P(no|X_1 \cdots X_3) \propto \frac{8}{17} \cdot 1 \cdot \frac{5}{8} \cdot \frac{6}{8} = 0.221$ ✓

Naïve Bayes with bag of word

- Learning phase:
  - Prior P(Y = C_i)
    - Count how many emails are spam/ not spam
  - P($X_j = v_k$ | Y = C_i)
    - For each {spam, not spam}, count how often the j-th word of a dictionary appears for v_k frequency in docs of the category
- Test phase:
  - For each document
    - Use naïve Bayes decision rule

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{j=1}^{size\ of\ dictionary} P(X_j|y)$$

    - d: number of words in the dictionary

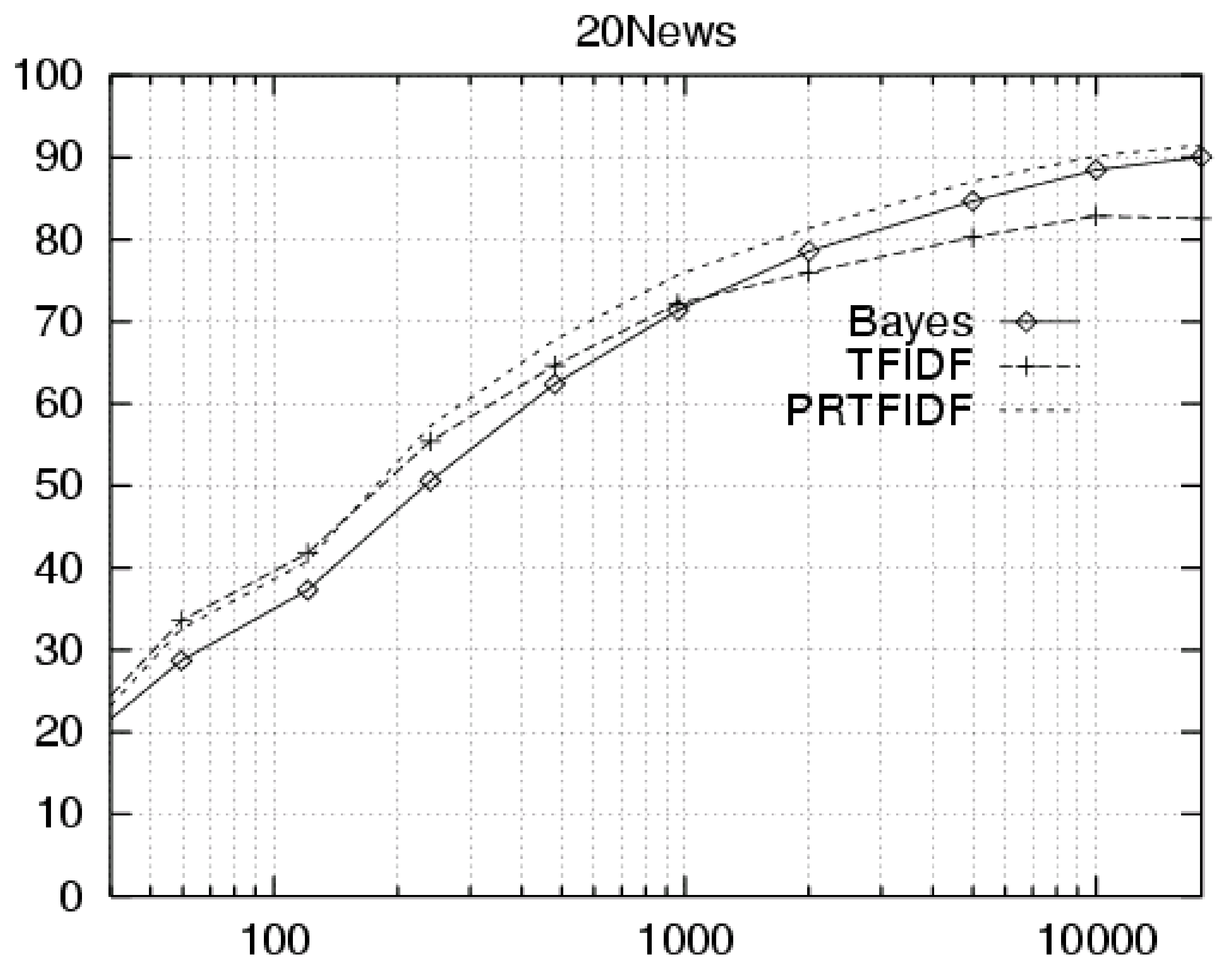
Twenty News Groups results

Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy

Learning curve for Twenty News Groups



Accuracy vs. Training set size (1/3 withheld for test)

Violating the NB assumption

- Usually, features are not conditionally independent:

$$P(X_1 \ldots X_n|Y) \neq \prod_i P(X_i|Y)$$

- Word not observed in training data

- Nonetheless, NB is the single most used classifier out there
  - NB often performs well, even when assumption is violated
  - [Domingos & Pazzani '96] discuss some conditions for good performance