NAÏVE BAYES MODELS

SECTION 5.7 OF ALPAYDIN
SECTION 3.5.1, 3.5.2 OF MURPHY

Outline

- We have learned how to model multivariate data
- But we have only considered continuous-valued data
- A lot of real data are discrete (e.g., text)
- Bag-of-words representation
- Naïve Bayes classifier
  - Employ Bayes Theorem
  - Assume feature independence
  - Multiple parameter estimation techniques

Motivating Example: Spam Filter

Bag of Words: intuitions

- Disregard grammar and even word order
- Only keep multiplicity or simply presence/absence
- 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

Bag of Words: Binary Representation

- Describe a document as a d-dimensional **binary** vector x, indicating the presence/absence of a word in a vocabulary V.
- Consider the following tiny vocabulary ($d = 5$):

$$V = \{\text{football, defence, strategy, goal, office}\}$$

- Then, a sentence "Adam from UIC Registrar's Office scored two goals in a community football game." is represented as

$$\mathbf{x} = (1, 0, 0, 1, 1),$$

since it contains only the words "football", "office", and "goal"

- We do not care about the order of the words
- We do not care about the words that are not in the vocabulary

Bag of Words: Multinomial Representation

Breaking News! Zoloft causes serious injuries, compensation available
1 message

ZoloftInjuryLawyers <mail@ragho-qws.evc.gh> Mon, Dec 17, 2012 at 9:31 PM
To: bziebart@ragho-qws.evc.gh

Settlements for Zoloft users

ZOLOFT® LAWSUIT CLAIM CENTER

Have you or a loved one used Zoloft®?
You may be entitled to Financial Compensation!

There is limited time to file your claim!
DON'T DELAY!

CLICK HERE
For More Information

You have received this notice as a potential client of ConsumerInjuryAdvocates. If you do not wish to receive these emails please contact us at ConsumerInjuryAdvocates, 8690 Aero Drive, Suite 115-34, San Diego, CA 92123 or click on the link below to unsubscribe:
Unsubscribe
This email is a commercial solicitation.

aardvark	0	$X_{aardvark} = 0$
about	0	$X_{about} = 0$
all	0	$X_{all} = 0$
an	1	$X_{an} = 1$
as	1	$X_{as} = 1$
at	2	$X_{at} = 2$
…		
claim	1	
…		
lawsuit	1	
…		
Zoloft	3	$X_{Zoloft} =$ '>2'

- Can also use multinomial variables, e.g., four levels: {0, 1, 2, '>2'}.
- Let $X_{'at'}$ be a random variable of the frequency of 'at' in a document. Then $X_{'at'} \sim$ Multinoulli (Θ)

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

A Probabilistic Classifier

Supervised Learning:

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

  d: size of the dictionary

  - Example, classify documents as being a spam (C_1) or not spam (C_2)

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

aardvark	0
about	0
all	0
an	1
as	1
at	2
…	
claim	1
…	
lawsuit	1
…	
Zoloft	3

Bayes Theorem

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

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

Building and Using Probabilistic Classifiers

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?

Naïve Bayes: Conditional Independence Assumptions

- Assume features are independent given class:

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

$$P(X_{lawsuit} = 2, X_{Zoloft} = 1 \mid spam)$$
$$= P(X_{lawsuit} = 2 \mid spam) * P(X_{Zoloft} = 1 \mid spam)$$

- How many parameters now? $(|X| - 1) * \#class$

- We introduced Naïve Bayes in Gaussian multivariate models (Chapter 5)
  - We now consider discrete variables $\mathbf{x}_{1:d}$, instead of continuous

Naïve Bayes: Independence Assumptions

Joint probability distribution:

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

<u>Estimation technique</u> (from Chapter 4)

Maximum likelihood:

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

Estimating: $\theta = \{P(Y), P(X_j | Y)\}$

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

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

(Naive Bayes: x_j are conditionally independent)

  - The discriminant is linear

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

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

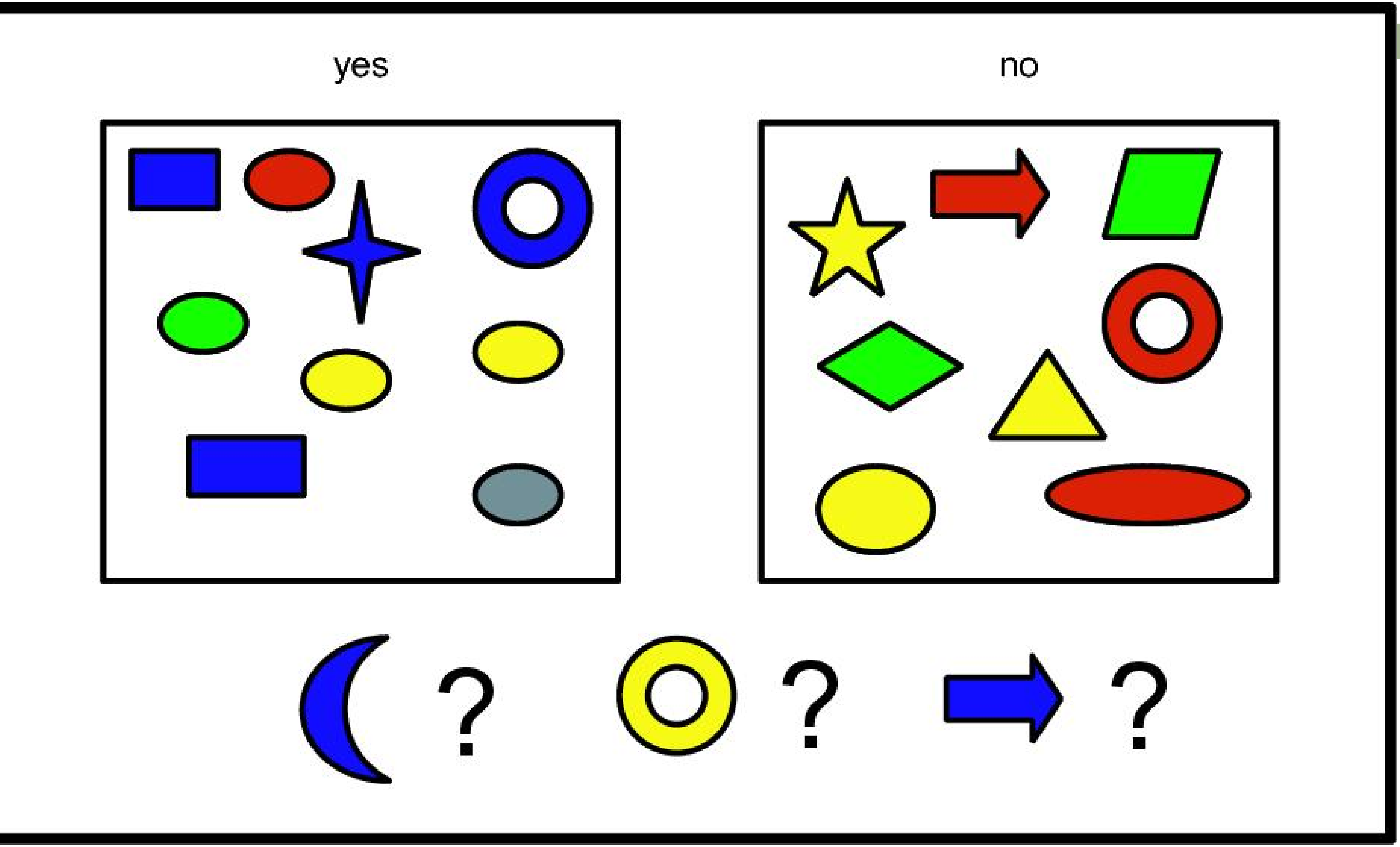
What does it mean if we drop j in p_{ijk}?

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

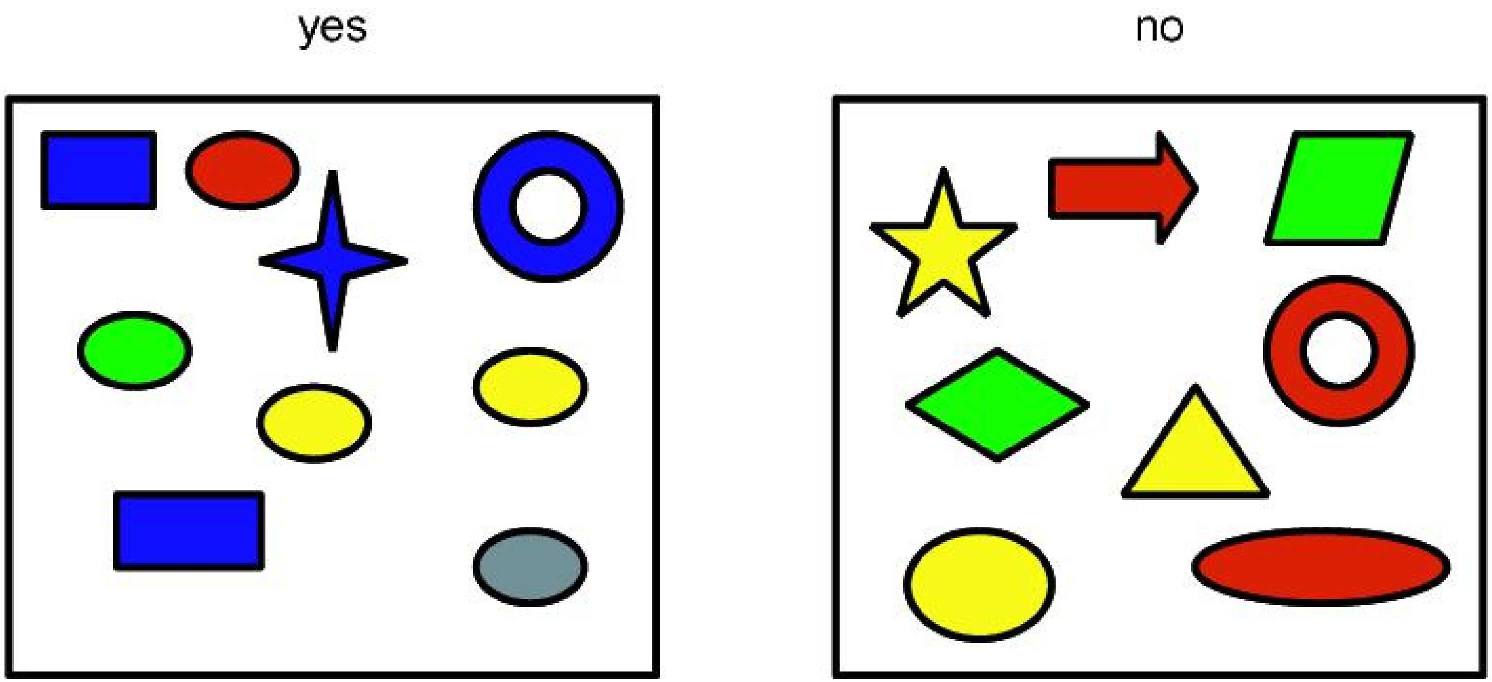
$$g_i(\mathbf{x}) = \sum_j \sum_k z_{jk} \log p_{ijk} + \log P(C_i)$$

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

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*, ...)

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

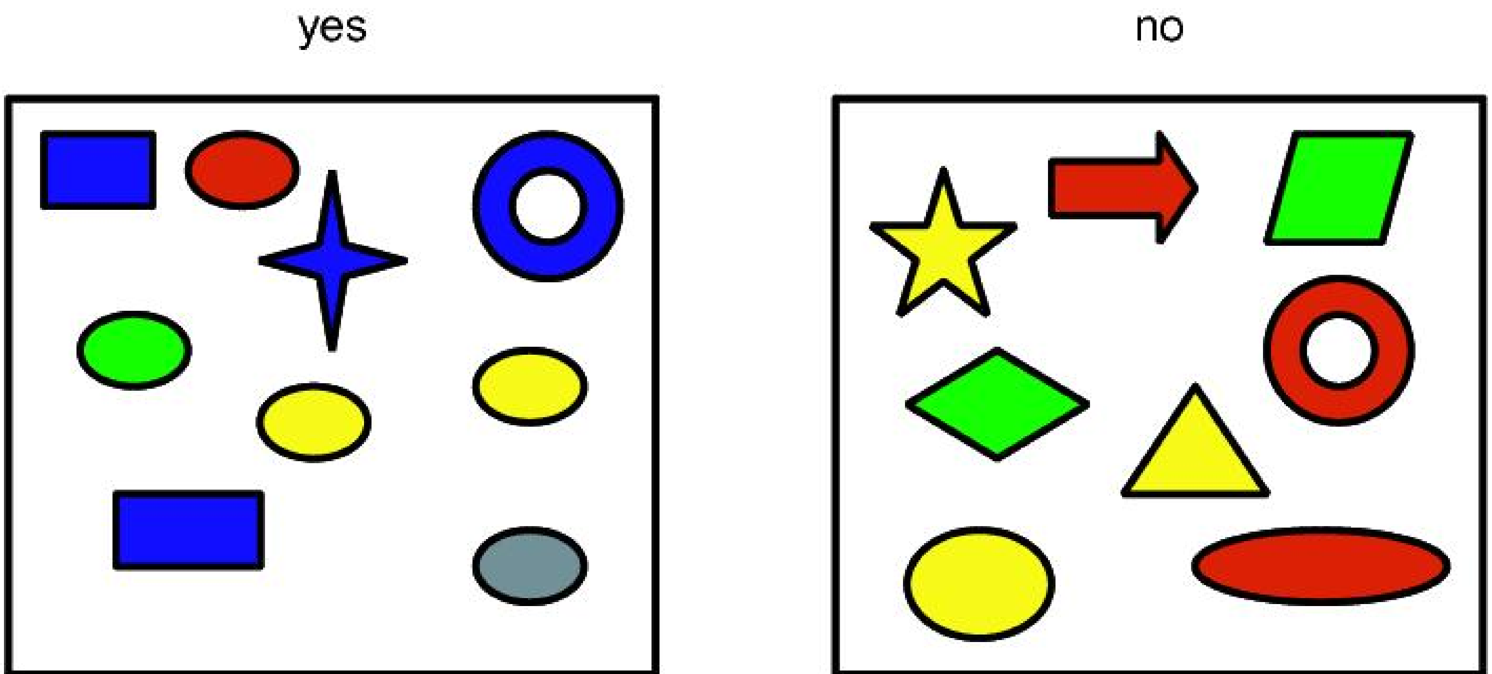
$P(no)=8/17$, $p(blue \mid no) = 0$, $p(ellipse \mid no) = 3/8$, $p(green \mid no)=2/8$

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

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

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

Prediction



What is the prediction for red star? 

$$P(yes \mid x_1 \cdots x_3) \propto P(yes)\, P(x_1 \mid yes) \cdot P(x_2 \mid yes)\, P(x_3 \mid yes)$$

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

$$P(no \mid x_1 \cdots x_3) \propto \frac{8}{17} \cdot 1 \cdot \frac{5}{8} \cdot \frac{6}{8} = 0.221 \ \checkmark$$

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

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

    - d: number of words in the dictionary

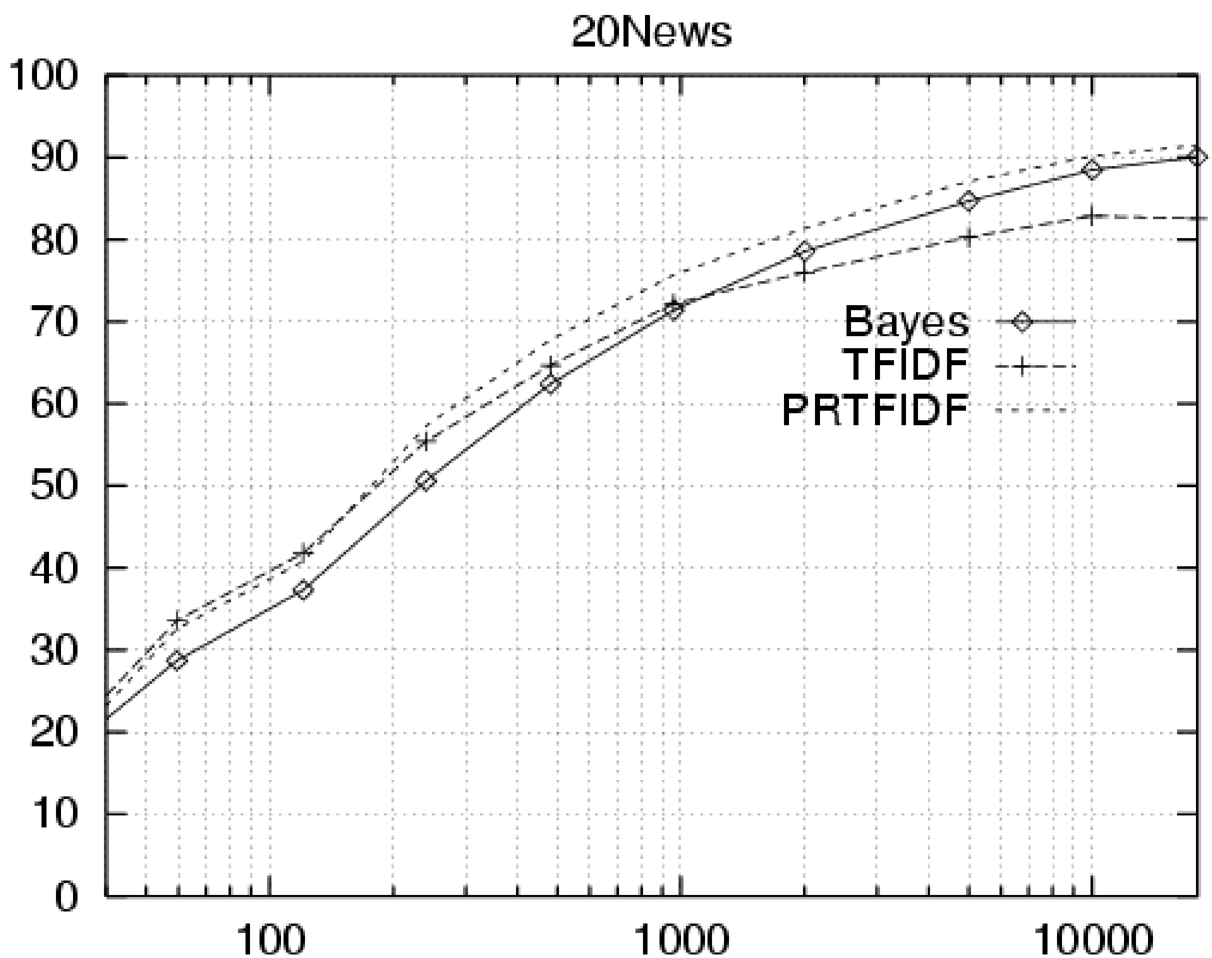
Twenty News Groups results

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)

Another representation of document (to be used in the Lab)

Recall the bag of words representation of a text document

aardvark	0	$X_{aardvark} = 0$
about	0	$X_{about} = 0$
all	0	$X_{all} = 0$
an	1	
as	1	
at	2	
…		
claim	1	
…		
lawsuit	1	
…		
Zoloft	3	X_{zoloft} = '>2'

Issue: need to manually discretize the frequency bin ('>2' or '[2, 5] and >5'?)

Naïve Bayes with word sequence

- X_j: the *j-th* word in a document and it can take value in a dictionary with words $\{w_1, \ldots, w_N\}$
  - Assume conditional independence (now word order matters)

$$P(\mathbf{x}_{1:T} \mid y) = \prod_{j=1:T} P(x_j \mid y) \quad \text{(T: length of the document)}$$

- Learning phase:
  - $P(Y = C_i)$: count how many emails are spam/not spam
  - $P(X_j = w_k \mid Y = C_i)$
    - For each {spam, not spam}, count how often the j-th word of a doc is the k-th word in the dictionary, among documents of the i-th category
- Test phase

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

- Problem: what if a test document has 101 words, but all training documents have at most 100 words?
  - $P(X_{101}|Y)$ never learned

Naïve Bayes with word sequence

- X_j: the *j-th* word in a document taking value in $\{w_1, \ldots, w_N\}$
- Solution: assume $P(X_j \mid Y)$ share the same distribution across j
  - We can drop j from subscript
  - Order of words ignored

$$P(\text{'cat is cute'} \mid \text{spam}) = P(\text{'is cute cat'} \mid \text{spam})$$

  - Still different from Bag-of-Words
- Learning phase:
  - $P(Y = C_i)$: count how many emails are spam/not spam
  - $P(X_j = w_k \mid Y = C_i)$
    - For each {spam, not spam}, count how often the k-th word in the dictionary appears among documents of the i-th category (see Lab 5)
- Test phase

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{j=1}^{\textbf{length of doc}} P(X_j \mid y)$$

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
- Posterior prob P(Y | **X**) often committed towards 0 or 1

$$P(Y|X_1, X_2, \ldots) = P(Y) \prod_j P(X_j|Y)$$

- 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