VIRGINIA TECH

Classification: Naïve Bayes

Lifu Huang

Computer Science, Virginia Tech

January 27, 2021

Slides adapted from Luke Zettlemoyer, David Sontag, Bert Huang

Review – Decision Trees

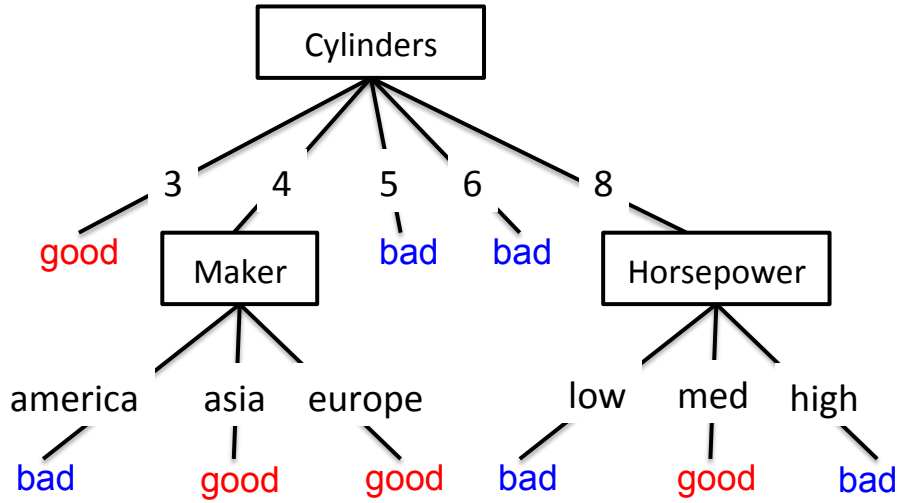
- Start from empty decision tree
- Split on next best attribute (feature)
  - Use, for example, information gain to select attribute:

$$\arg\max_i IG(X_i) = \arg\max_i H(Y) - H(Y \mid X_i)$$

- Recurse

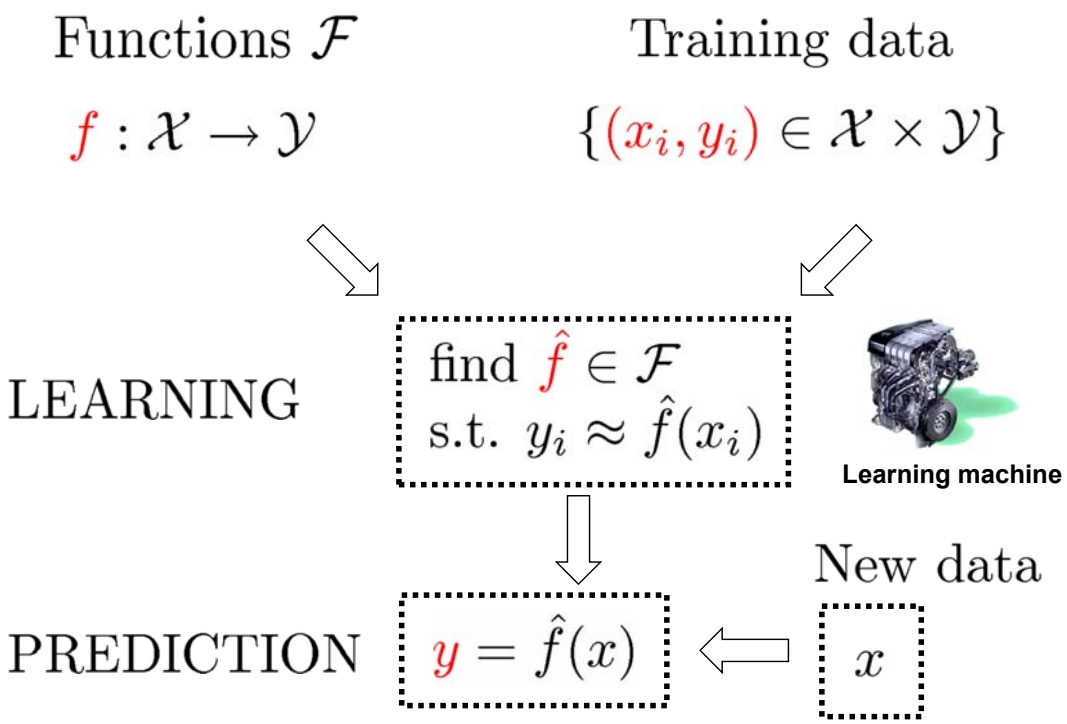
Discriminative model:
learn the boundary between classes.

mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europe
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europe
bad	5	medium	medium	medium	medium	75to78	europe



Supervised Learning: find f

- Given: Training set $\{(x_i, y_i) | i = 1 \dots n\}$
- Find: A good approximation to $f: X \to Y$
- Examples: what are X and Y?
  - Spam Detection: Map email to {Spam, Ham}
  - Digit Recognition: Map pixels to {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}



Example: Spam Filter

- Input: email
- Output: Spam/Not-spam
- Setup:
  - Get a large collection of example emails, each labeled as “spam” or “not-spam”
  - Note: someone has to manually label all the data to make sure the labels are (mostly) correct!!
  - Want to learn to predict labels of new, future emails
- Features: The attributes used to make the spam/not-spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: SenderInContacts
  - … …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

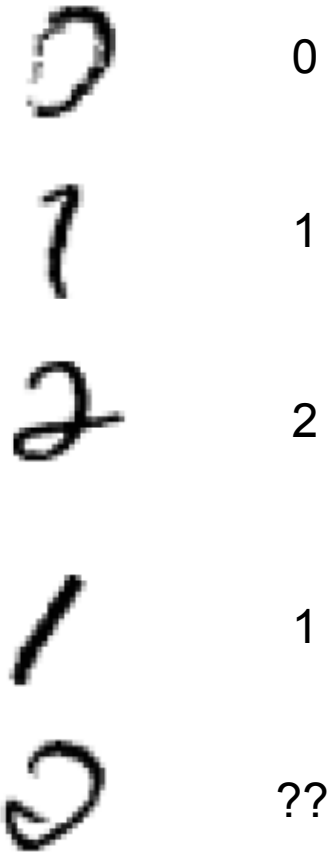
TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

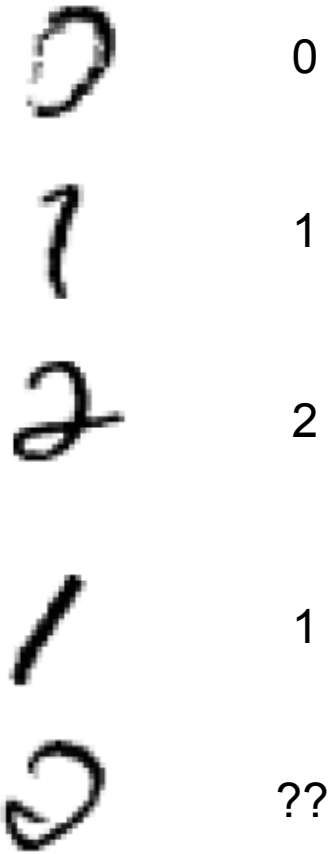
Example: Digit Recognition

- Input: images / pixel grids
- Output: a digit 0-9
- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to manually label all the data to make sure the labels are (mostly) correct!!
  - Want to learn to predict labels of new, future digit images
- Features: The attributes used to make the digit decision
  - Pixels: FREE!
  - Shape Patterns: NumComponents, NumLoops
  - … …



Example: Digit Recognition

- Input: images / pixel grids
- Output: a digit 0-9
- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to manually label all the data to make sure the labels are (mostly) correct!!
  - Want to learn to predict labels of new, future digit images
- Features: The attributes used to make the digit decision
  - Pixels: FREE!
  - Shape Patterns: NumComponents, NumLoops
  - … …



$$max_{y_i \in Y} P(y_i|X)$$

Probability Basis

- Probability $P(A)$
  - the likelihood of event A happens. $0 \leq P(A) \leq 1$
- Conditional Probability $P(A|B)$
  - the likelihood of event A happens given that B has already happened
- Joint Probability $P(A, B)$
  - the likelihood of both event A and B happen
  - $P(A, B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$
- Bayes Rule

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Prior Probability:
- from past experience

Posterior Probability:
- after considering the relevant evidence or background

Conditional Independence

- Conditional Independence: X is conditionally independent of Y given Z, if the probability distribution for X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = i | Y = j, Z = k) = P(X = i | Z = k)$$

- Equivalent to

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

Probability for Classification

$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

- Can we directly estimate the data distribution $P(Y|X) \propto P(X,Y) = P(X|Y)P(Y)$?
- Naïve Bayes assumption
  - Features are independent given class

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y) = P(X_1|Y)P(X_2|Y)$$

  - More generally

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

mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europe
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europe
bad	5	medium	medium	medium	medium	75to78	europe

The Naïve Bayes Classifier

- Given:
  - Prior $P(Y)$
  - n conditionally independent features $\boldsymbol{X}$ given the class Y
  - For each X_i, we have likelihood $P(X_i|Y)$
- Decision Rule:

$$P(Y|X) \propto P(X,Y) = P(X|Y)P(Y)$$

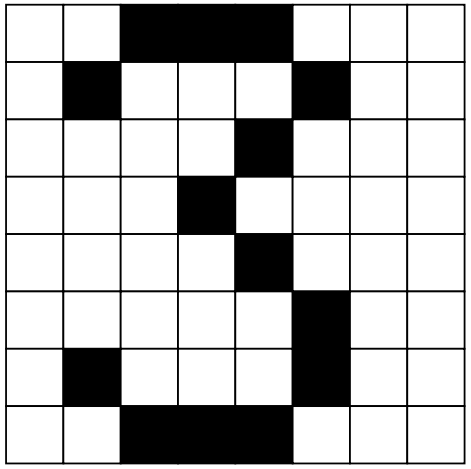
$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

$$y^* = h_{NB}(\mathbf{x}) = \arg\max_y P(y)P(x_1, \ldots, x_n \mid y)$$
$$= \arg\max_y P(y)\prod_i P(x_i|y)$$

mpg	cylinders	displacement	horsepower	weight	acceleration	modelyear	maker
good	4	low	low	low	high	75to78	asia
bad	6	medium	medium	medium	medium	70to74	america
bad	4	medium	medium	medium	low	75to78	europe
bad	8	high	high	high	low	70to74	america
bad	6	medium	medium	medium	medium	70to74	america
bad	4	low	medium	low	medium	70to74	asia
bad	4	low	medium	low	low	70to74	asia
bad	8	high	high	high	low	75to78	america
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
bad	8	high	high	high	low	70to74	america
good	8	high	medium	high	high	79to83	america
bad	8	high	high	high	low	75to78	america
good	4	low	low	low	low	79to83	america
bad	6	medium	medium	medium	high	75to78	america
good	4	medium	low	low	low	79to83	america
good	4	low	low	medium	high	79to83	america
bad	8	high	high	high	low	70to74	america
good	4	low	medium	low	medium	75to78	europe
bad	5	medium	medium	medium	medium	75to78	europe

Example: A Digit Recognizer

- Input: pixel grids



- Output: a digit 0-9

Naïve Bayes for Digits

- Simple Version:
  - One feature F_{ij} for each grid position $< i, j >$
  - Possible feature values are On/Off or 1/0, indicating whether intensity is more or less than 0.5 in underlying image
  - Each input maps to a feature vector, e.g.,

$$\rightarrow \langle F_{0,0} = 0 \;\; F_{0,1} = 0 \;\; F_{0,2} = 1 \;\; F_{0,3} = 1 \;\; F_{0,4} = 0 \;\; \ldots F_{15,15} = 0 \rangle$$

- Naïve Bayes Model

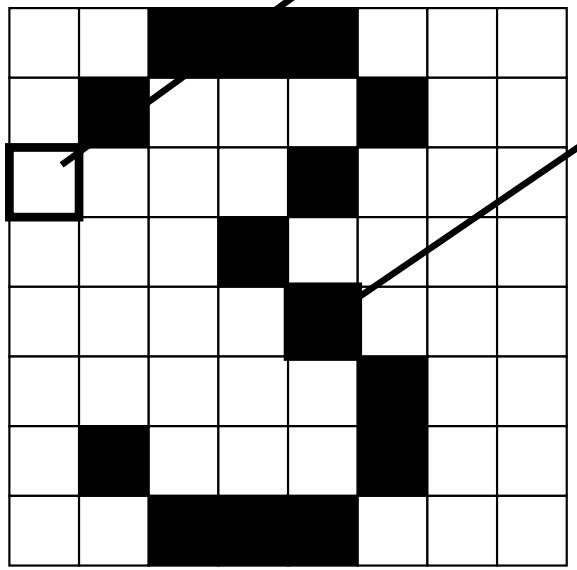
$$P(Y|F_{0,0} \ldots F_{15,15}) \propto P(Y) \prod_{i,j} P(F_{i,j}|Y)$$

- Are the features independent given class?
- What do we need to learn?

Example Distributions

$P(Y)$

1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



$P(F_{3,1} = on|Y)$

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

$P(F_{5,5} = on|Y)$

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

MLE (Maximum Likelihood Estimation) for the parameters of NB

- Given Data
  - Count (A=a, B=b): number of examples with A=a and B=b
- MLE for discrete NB, simply:
  - Prior

$$P(Y = y) = \frac{Count(Y = y)}{\sum_{y'} Count(Y = y')}$$

  - Likelihood

$$P(X_i = x|Y = y) = \frac{Count(X_i = x, Y = y)}{\sum_{x'} Count(X_i = x', Y = y)}$$

Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:

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

- Actual probabilities $P(Y|\boldsymbol{X})$ often biased towards 0 or 1
- 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

Subtleties of NB classifier 2 – Overfitting

$P(\text{features}, C = 2)$

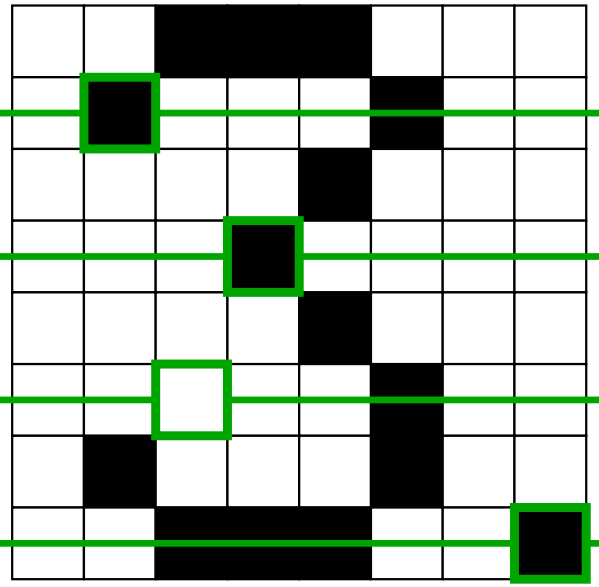
$P(C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.8$

$P(\text{on}|C = 2) = 0.1$

$P(\text{off}|C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.01$



$P(\text{features}, C = 3)$

$P(C = 3) = 0.1$

$P(\text{on}|C = 3) = 0.8$

$P(\text{on}|C = 3) = 0.9$

$P(\text{off}|C = 3) = 0.7$

$P(\text{on}|C = 3) = 0.0$

2 wins !!

Laplace Smoothing

- Laplace's estimation
  - Handle the zero probability in Naïve Bayes
  - Pretend you saw every outcome *k* extra times

$$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$

  - *k* is the strength of the prior
- Laplace for conditionals
  - Smooth each condition independently

$$P_{LAP,k}(x|y) = \frac{c(x,y) + k}{c(y) + k|X|}$$

$$P_{LAP,0}(X) = \left\langle \frac{2}{3}, \frac{1}{3} \right\rangle$$

$$P_{LAP,1}(X) = \left\langle \frac{3}{5}, \frac{2}{5} \right\rangle$$

$$P_{LAP,100}(X) = \left\langle \frac{102}{203}, \frac{101}{203} \right\rangle$$

Text Classification

- Classify emails
  - Y = {Spam, Not Spam}
- Classify news articles
  - Y = {what is the topic of the article?}
- Classify webpages
  - Y = {Student, Professor, Project}

- What about the features **X**?
  - The text!

Features X are entire document - X_i for i^{th} word in article

Article from rec.sport.hockey

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most
obvious candidate for pleasant surprise is Alex
Zhitnik. He came highly touted as a defensive
defenseman, but he's clearly much more than that.
Great skater and hard shot (though wish he were
more accurate). In fact, he pretty much allowed
the Kings to trade away that huge defensive
liability Paul Coffey. Kelly Hrudey is only the
biggest disappointment if you thought he was any
good to begin with. But, at best, he's only a
mediocre goaltender. A better choice would be
Tomas Sandstrom, though not through any fault of
his own, but because some thugs in Toronto decided
```

NB for Text Classification

- Bag of Words Model
  - Article at least 1000 words, $\boldsymbol{X} = \{X_1, X_2, \ldots, X_{1000}\}$
  - X_i represents the i^{th} word in document, i.e., the domain of X_i is entire vocabulary
- NB assumption helps a lot!!
  - $P(X_i = x_i | Y = y)$ is just the probability of observing word x_i in a document on topic y

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Bag of words model

- Typical additional assumption
  - Position in document doesn't matter

$$P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$$

  - "Bag of Words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

When the lecture is over, remember to say bye to your classmates

is lecture remember classmates your When to say the bye over, to

Bag of Words Approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

NB with Bag of Words for Text Classification

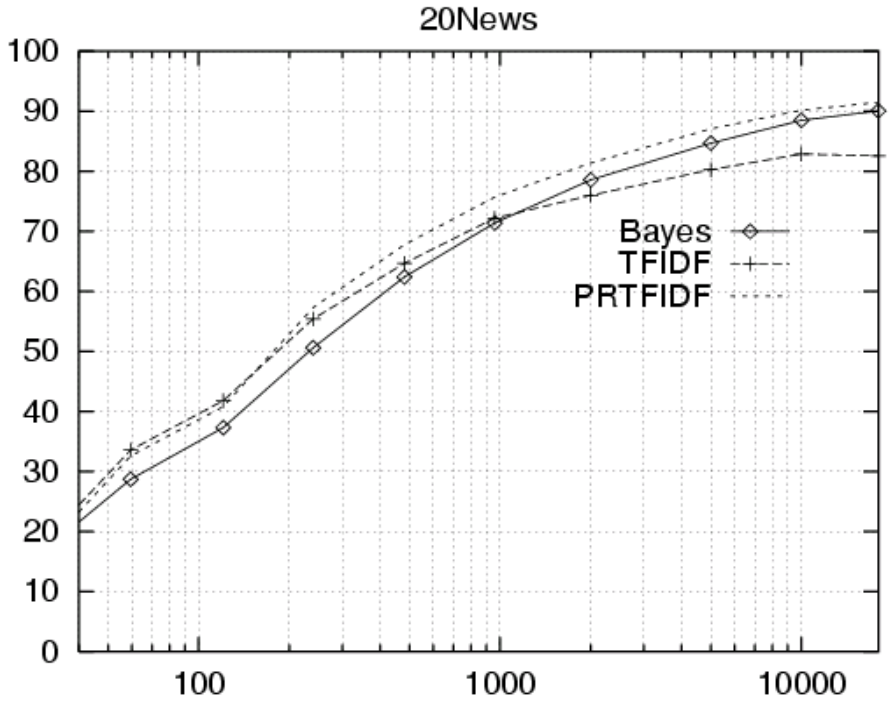
- Learning Phase
  - Prior $P(Y)$
    - Count how many documents from each category or topic
  - $P(X_i|Y)$
    - For each topic or category, count how many times your saw the word X_i in documents of this topic (Y), remember to ignore the positions across documents
- Test Phase
  - For each document
    - Use Naïve Bayes decision rule

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

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	



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

Naive Bayes: 89% classification accuracy

What if we have continuous X_i

- Split the continuous variable into several intervals
  - e.g. $X \sim [0, 1] \rightarrow [0, 0.05], (0.05, 0.2], (0.2, 1]$
- Gaussian Naïve Bayes
  - Each feature has a Gaussian distribution given a class

Gaussian Naïve Bayes (GNB):

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \quad e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

mean

standard deviation

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

- Mean:

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

jth training example

$\delta(x)$=1 if x true, else 0

- Variance:

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

What you need to know about Naïve Bayes

- Naïve Bayes Classifier
  - What's the assumption
  - Why we use it
  - How do we learn it
- Text Classification
  - Bag of words model
- Gaussian Naïve Bayes Classifier
  - Features are still conditionally independent
  - Each feature has a Gaussian distribution given class