

*CSE-590/634*  
*Data Mining Concepts and Techniques*  
*Spring 2009*

# *Bayesian Classification*

***Presented by:***

Muhammad A. Islam,	106506983
Moieed Ahmed,	106867769

***Guided by:*** Prof. Anita Wasilewska

# Bibliography

- DATA MINING Concepts and Techniques, Jiawei Han, Micheline Kamber  
Morgan Kaufman Publishers, 2<sup>nd</sup> Edition.
  - Chapter 6, Classification and Prediction, Section 6.4.
  
- Computer Science, Carnegie Mellon University
  - <http://www.cs.cmu.edu/~awm/tutorials>
  - <http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/mlbook/ch6.pdf>
  - <http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html>
  
- Wikipedia:
  - [http://en.wikipedia.org/wiki/Bayesian\\_probability](http://en.wikipedia.org/wiki/Bayesian_probability)
  - [http://en.wikipedia.org/wiki/Naive\\_Bayes\\_classifier](http://en.wikipedia.org/wiki/Naive_Bayes_classifier)

# Outline

- Introduction to Bayesian Classification
  - Probability
  - Bayes Theorem
  - Naïve Bayes Classifier
  - Classification Example
- Text Classification – an Application
- Paper: “Text Mining: Finding Nuggets in Mountains of Textual Data”

# *Introduction to Bayesian Classification*

*By*

*Muhammad A. Islam*

*106506983*

# Bayesian Classification

## ► What is it ?

- Statistical method for classification.
- Supervised Learning Method.
- Assumes an underlying probabilistic model, the Bayes theorem.
- Can solve diagnostic and predictive problems.
- Can solve problems involving both categorical and continuous valued attributes.
- Named after *Thomas Bayes*, who proposed the *Bayes Theorem*.

# Basic Probability Concepts

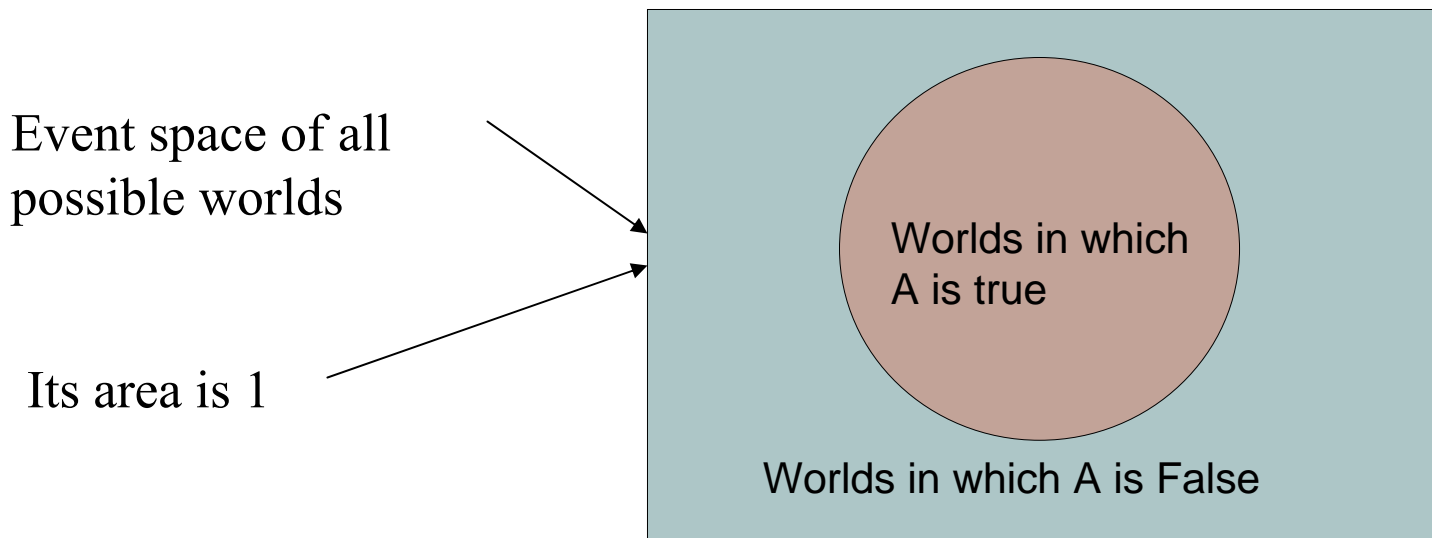
- *Sample space*  $S$  is a set of all possible outcomes
  - $S = \{1,2,3,4,5,6\}$  for a dice roll
  - $S = \{H,T\}$  for a coin toss.
- An *Event*  $A$  is any subset of the Sample Space
  - Seeing a 1 on the dice roll
  - Getting head on a coin toss

# Random Variables

- A is a **random variable** if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
  - A = The US president in 2016 will be male
  - A = You see a head on a coin toss
  - A = The weather will be sunny tomorrow
- Boolean random variables
  - A is either true or false.
- Discrete random variables
  - Weather is one of {sunny, rain, cloudy, snow}
- Continuous random variables
  - Temp=21.6.

# Probability

- We write  $P(A)$  as “the fraction of possible worlds in which  $A$  is true”



$P(A) = \text{Area of reddish oval}$

# The axioms of Probability

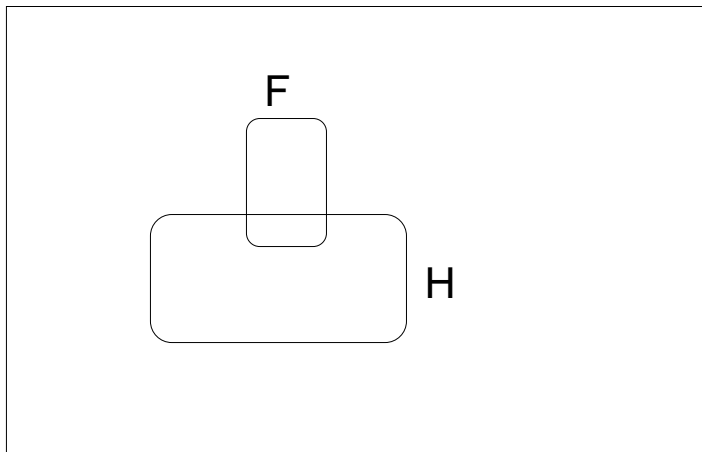
- ▶  $0 \leq P(A) \leq 1$
- ▶  $P(\text{True}) = 1$
- ▶  $P(\text{False}) = 0$
- ▶  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

From these we can prove:

- ▶  $P(\text{not } A) = P(\sim A) = 1 - P(A)$
- ▶  $P(A) = P(A \wedge B) + P(A \wedge \sim B)$

# Conditional Probability

- $P(A|B)$  = Fraction of worlds in which B is true that also have A true



H = “Have a headache”

F = “Coming down with Flu”

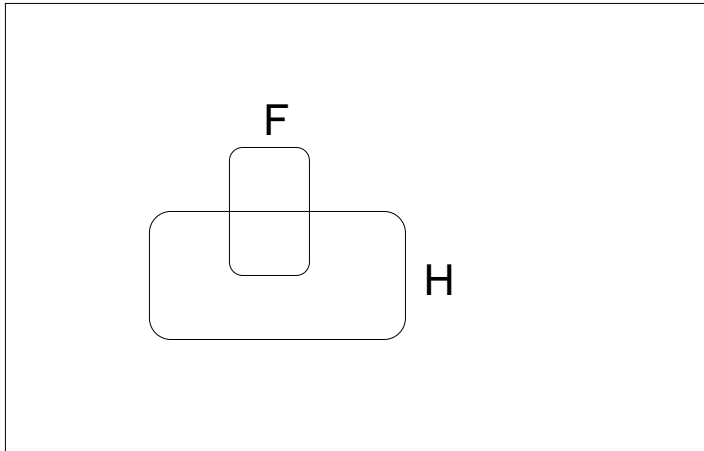
$$P(H) = 1/10$$

$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

“Headaches are rare and flu is rarer, but if you’re coming down with ‘flu there’s a 50-50 chance you’ll have a headache.”

# Conditional Probability



H = “Have a headache”

F = “Coming down with Flu”

$$P(H) = 1/10$$

$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

$P(H|F)$  = Fraction of flu-inflicted worlds in which you have a headache

$$= \frac{\text{\#worlds with flu and headache}}{\text{\#worlds with flu}}$$

$$= \frac{\text{Area of “H and F” region}}{\text{Area of “F” region}}$$

$$= \frac{P(H \wedge F)}{P(F)}$$

# Conditional Probability

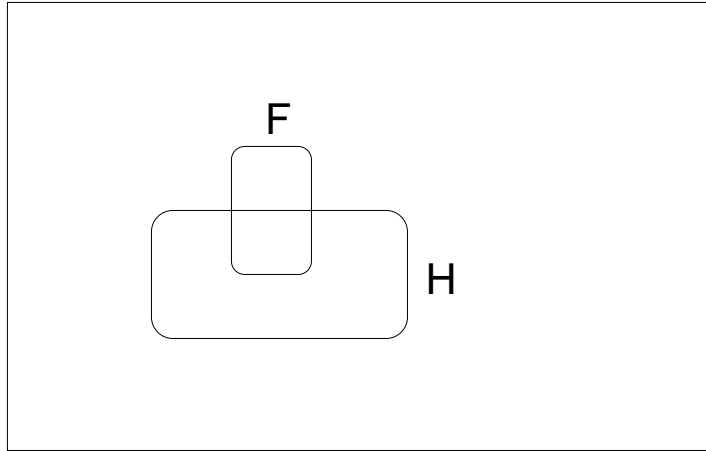
$$P(A/B) = \frac{P(A \wedge B)}{P(B)}$$

The chain rule:  $P(A \wedge B) = P(A/B) P(B)$

In general, for n random variables, the chain rule;-

$$\begin{aligned} P(A_n, A_{n-1}, \dots, A_1) &= P(A_n | A_{n-1}, \dots, A_1) \cdot P(A_{n-1}, \dots, A_1) \\ &= P(A_n | A_{n-1}, \dots, A_1) \cdot P(A_{n-1} | A_{n-2}, \dots, A_1) \cdot P(A_{n-2}, \dots, A_1) \\ &= \dots \\ &= \prod_{i=1}^n P(A_i | A_{i-1}, \dots, A_1) \end{aligned}$$

# Probabilistic Inference



H = “Have a headache”

F = “Coming down with Flu”

$$P(H) = 1/10$$

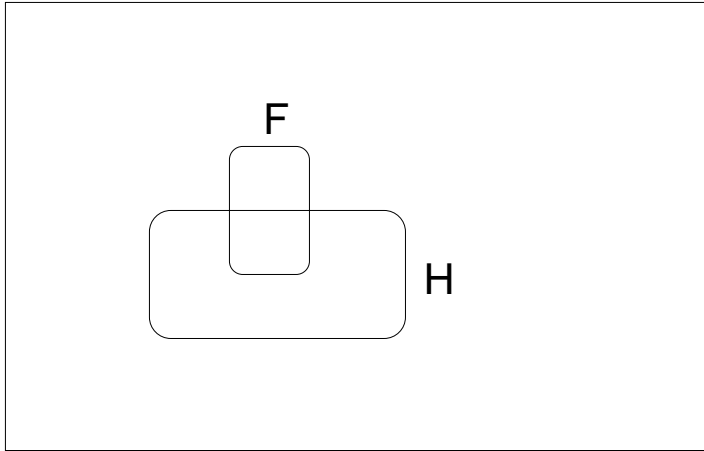
$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

One day you wake up with a headache. You think: “Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu”

Is this reasoning good?

# Probabilistic Inference



H = “Have a headache”

F = “Coming down with Flu”

$$P(H) = 1/10$$

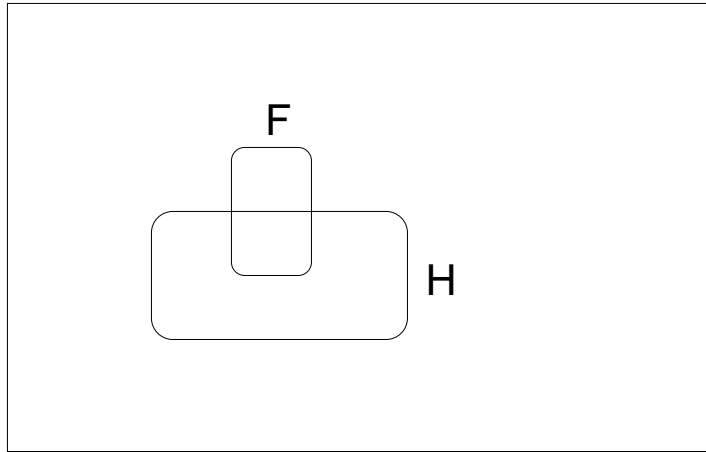
$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

$$P(F \text{ and } H) = \dots$$

$$P(F|H) = \dots$$

# Probabilistic Inference



H = “Have a headache”

F = “Coming down with Flu”

$$P(H) = 1/10$$

$$P(F) = 1/40$$

$$P(H|F) = 1/2$$

$$P(F \text{ and } H) = P(H|F) P(F) = 1/2 * 1/40 = 1/80$$

$$P(F|H) = \frac{P(F \text{ and } H)}{P(H)} = \frac{1/80}{1/10} = 1/8$$

# The Bayes Theorem

►  $P(A|B) = P(A \wedge B)/P(B)$

Chain Rule:

$$P(A \wedge B) = P(A|B).P(B)$$

Also,  $P(A \wedge B) = P(B|A).P(A)$

Therefore,

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

# Independence

- Suppose there are two events:
  - M: Manuela teaches the class (otherwise it's Andrew)
  - S: It is sunny

“The sunshine levels do not depend on and do not influence who is teaching.”

- This can be specified very simply:

$$P(S | M) = P(S)$$

“Two events A and B are statistically independent if the probability of A is the same value when B occurs, when B does not occur or when nothing is known about the occurrence of B”

# Conditional Independence

Suppose we have these three events:

- ▶ M : Lecture taught by Manuela
- ▶ L : Lecturer arrives late
- ▶ R : Lecture concerns robots

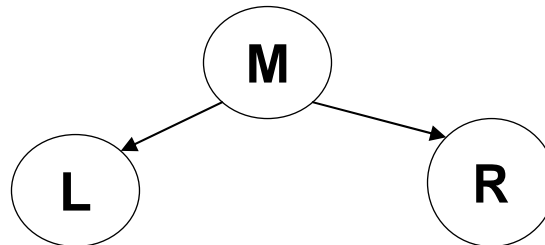
Once you know who the lecturer is, then whether they arrive late doesn't affect whether the lecture concerns robots.

$$P(R \mid M, L) = P(R \mid M)$$

We express this in the following way:

“R and L are conditionally independent given M”

..which is also noted by the following diagram.



Given knowledge of M, knowing anything else in the diagram won't help us with L, etc.

# Conditional Independence

$$\blacktriangleright P(A,B|C) = P(A \wedge B \wedge C)/P(C)$$

$$= P(A|B,C).P(B \wedge C)/P(C)$$

[applying chain rule,  $P(A \wedge B \wedge C) = P(A|B,C). P(B \wedge C)$ ]

$$= P(A|B,C).P(B|C)$$

$$= P(A|C).P(B|C), \quad [ \text{If } A \text{ and } B \text{ are conditionally independent given } C, \text{ then } P(A|B,C) = P(A|C) ]$$

For  $n$  random variables, it can be extended as:-

$$P(A_1, A_2 \dots A_n | C) = P(A_1 | C).P(A_2 | C) \dots P(A_n | C)$$

*If  $A_1, A_2 \dots A_n$  are conditionally independent given  $C$ .*

# Data Table

rec	Age	Income	Student	Credit_rating	Buys_computer
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31...40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31...40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>40	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31...40	Medium	No	Excellent	Yes
r13	31...40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

# An Example



I am 35 years  
old

I earn \$40,000

My credit  
rating is fair

Will he buy a  
computer?



- $X$  : 35 years old customer with an income of \$40,000 and fair credit rating.
- $H$  : Hypothesis that the customer will buy a computer.

# The Bayes Theorem

$$P(H|X) = \frac{P(X|H) P(H)}{P(X)}$$

- $P(H|X)$  : Probability that the customer will buy a computer given that we know his age, credit rating and income. (Posterior Probability of H)
- $P(H)$  : Probability that the customer will buy a computer regardless of age, credit rating, income (Prior Probability of H)
- $P(X|H)$  : Probability that the customer is 35 yrs old, have fair credit rating and earns \$40,000, given that he has bought our computer (Posterior Probability of X)
- $P(X)$  : Probability that a person from our set of customers is 35 yrs old, have fair credit rating and earns \$40,000. (Prior Probability of X)

# Bayesian Classifier

- ▶  $D$  : Set of tuples
  - Each Tuple is an ‘ $n$ ’ dimensional attribute vector
  - $X : (x_1, x_2, x_3, \dots, x_n)$
  - where  $x_i$  is the value of attribute  $A_i$
- ▶ Let there are ‘ $m$ ’ Classes :  $C_1, C_2, C_3, \dots, C_m$
- ▶ Bayesian classifier predicts  $X$  belongs to Class  $C_i$  iff
  - $P(C_i|X) > P(C_j|X)$  for  $1 \leq j \leq m, j \neq i$
- ▶ Maximum Posteriori Hypothesis
  - $$P(C_i|X) = \frac{P(X|C_i) P(C_i)}{P(X)}$$
  - Maximize  $P(X|C_i) P(C_i)$  as  $P(X)$  is constant

# Naïve Bayes Classifier

- With many attributes, it is computationally expensive to evaluate  $P(X|C_i)$
- Naïve Assumption of “class conditional independence”

$$P(X | C_i) = P(x_1, x_2, \dots, x_n | C_i)$$

$$= P(x_1 | C_i) * P(x_2 | C_i) * \dots * P(x_n | C_i)$$

$$= \prod_{k=1}^n P(x_k | C_i)$$

# Naïve Bayes Classifier (Contd..)

To compute,  $P(x_k|C_i)$

- $A_k$  is categorical:

$$P(x_k|C_i) = \frac{\text{the number of tuples of class } C_i \text{ in } D \text{ having the value } x_k \text{ for } A_k}{\text{the number of tuples of class } C_i \text{ in } D.}$$

- $A_k$  is continuous:

A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean  $\mu$  and standard deviation  $\sigma$

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}).$$

# Naïve Bayesian Classification

- Based on the [Bayesian theorem](#)
- Particularly suited when the dimensionality of the input is high
- In spite of the over-simplified assumption, it often performs better in many complex real-world situations
- Advantage: Requires a small amount of training data to estimate the parameters

# Data Table

rec	Age	Income	Student	Credit_rating	Buys_computer
r1	<=30	High	No	Fair	No
r2	<=30	High	No	Excellent	No
r3	31...40	High	No	Fair	Yes
r4	>40	Medium	No	Fair	Yes
r5	>40	Low	Yes	Fair	Yes
r6	>40	Low	Yes	Excellent	No
r7	31...40	Low	Yes	Excellent	Yes
r8	<=30	Medium	No	Fair	No
r9	<=30	Low	Yes	Fair	Yes
r10	>40	Medium	Yes	Fair	Yes
r11	<=30	Medium	Yes	Excellent	Yes
r12	31...40	Medium	No	Excellent	Yes
r13	31...40	High	Yes	Fair	Yes
r14	>40	Medium	No	Excellent	No

# Example

▶  $X = ( \text{Age} = \leq 30, \text{Income} = \text{medium}, \text{Student} = \text{yes}, \text{Credit\_rating} = \text{fair} )$

▶  $P(C1) = P(\text{Buys\_computer} = \text{yes}) = 9/14 = 0.643$

▶  $P(C2) = P(\text{Buys\_computer} = \text{no}) = 5/14 = 0.357$

▶  $P(\text{Age} = \leq 30 \mid \text{Buys\_computer} = \text{yes}) = \frac{\text{number of tuples with Buys\_computer} = \text{yes and Age} \leq 30}{\text{number of tuples with Buys\_computer} = \text{yes}}$

▶  $P(\text{Age} = \leq 30 \mid \text{Buys\_computer} = \text{yes}) = 2/9 = 0.222$

Similarly,

▶  $P(\text{Age} = \leq 30 \mid \text{Buys\_computer} = \text{no}) = 3/5 = 0.600$

▶  $P(\text{Income} = \text{medium} \mid \text{Buys\_computer} = \text{yes}) = 4/9 = 0.444$

▶  $P(\text{Income} = \text{medium} \mid \text{Buys\_computer} = \text{no}) = 2/5 = 0.400$

▶  $P(\text{Student} = \text{yes} \mid \text{Buys\_computer} = \text{yes}) = 6/9 = 0.667$

▶  $P(\text{Student} = \text{yes} \mid \text{Buys\_computer} = \text{no}) = 1/5 = 0.200$

▶  $P(\text{Credit\_rating} = \text{fair} \mid \text{Buys\_computer} = \text{yes}) = 6/9 = 0.667$

▶  $P(\text{Credit\_rating} = \text{fair} \mid \text{Buys\_computer} = \text{no}) = 2/5 = 0.400$

# Example (contd..)

- ▶  $P(X \mid \text{Buys a computer} = \text{yes})$   
 $= P(\text{Age} = \leq 30 \mid \text{buys\_computer} = \text{yes}) * P(\text{Income} = \text{medium} \mid \text{buys\_computer} = \text{yes}) * P(\text{Student} = \text{yes} \mid \text{buys\_computer} = \text{yes}) * P(\text{Credit rating} = \text{fair} \mid \text{buys\_computer} = \text{yes})$   
 $= 0.222 * 0.444 * 0.667 * 0.667 = 0.044$
- ▶  $P(X \mid \text{Buys a computer} = \text{No})$   
 $= 0.600 * 0.400 * 0.200 * 0.400 = 0.019$
- ▶ Find class  $C_i$  that Maximizes  $P(X|C_i) * P(C_i)$   
 $\rightarrow P(X \mid \text{Buys a computer} = \text{yes}) * P(\text{Buys\_computer} = \text{yes}) = 0.028$   
 $\rightarrow P(X \mid \text{Buys a computer} = \text{No}) * P(\text{Buys\_computer} = \text{no}) = 0.007$
- ▶ Prediction : Buys a computer for Tuple X

# *Text Classification – An application of Naïve Bayes Classifier*

*By*

*Moieed Ahmed*

*106867769*

# Why text classification?

- Learning which articles are of interest
- Classify web pages by topic
- Internet filters
- Recommenders
- Information extraction

# EXAMPLES OF TEXT CLASSIFICATION

- CLASSES=BINARY
  - “spam” / “not spam”
- CLASSES =TOPICS
  - “finance” / “sports” / “politics”
- CLASSES =OPINION
  - “like” / “hate” / “neutral”
- CLASSES =TOPICS
  - “AI” / “Theory” / “Graphics”
- CLASSES =AUTHOR
  - “Shakespeare” / “Marlowe” / “Ben Jonson”

# EXAMPLES OF TEXT CLASSIFICATION

- Classify news stories as world, US, business, SciTech, Sports ,Health etc
- Classify email as spam / not spam
- Classify business names by industry
- Classify email to tech stuff as Mac, windows etc
- Classify pdf files as research , other
- Classify movie reviews as favourable, unfavourable, neutral
- Classify documents as WrittenByReagan, GhostWritten
- Classify technical papers as Interesting, Uninteresting
- Classify Jokes as Funny, NotFunny
- Classify web sites of companies by Standard Industrial Classification (SIC)

# Naïve Bayes Approach

- Remove stop words and markings
- Build the Vocabulary as the list of all distinct words that appear in all the documents of the training set.
- The words in the vocabulary become the attributes, *assuming that classification is independent of the positions of the words*
- Each document in the training set becomes a record with frequencies for each word in the Vocabulary.
- Train the classifier based on the training data set, by computing the prior probabilities for each class and attributes.
- Evaluate the results on Test data

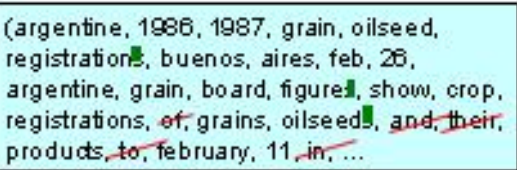
# Representing Text: a list of words

$f(\text{[raw text]}) = y$



ARGENTINA: 1986, 1987, grain, oilseed, registration, buenos, aires, feb, 26, argentine, grain, board, figure, show, crop, registrations, of, grains, oilseed, and, their, products, to, february, 11, in, ...

$f(\text{[refined list of words]}) = y$



(argentine, 1986, 1987, grain, oilseed, registration, buenos, aires, feb, 26, argentine, grain, board, figure, show, crop, registrations, of, grains, oilseed, and, their, products, to, february, 11, in, ...)

- Common Refinements: Remove Stop Words, Symbols

# Text Classification with Naïve Bayes

- ▶ Represent document  $X$  as a list of words  $W_1, W_2$  etc
- ▶ For each  $y$ , build a probabilistic model  $\Pr(x|y)$  of documents in class  $y$ .
  - $\Pr(X = \{\text{argentine, grain..}\} \mid Y = \text{wheat})$
  - $\Pr(X = \{\text{Stocks, rose.....}\} \mid Y = \text{Nonwheat})$
- ▶ To classify find the  $y$  which is most likely to generate  $x$ -i.e, which gives  $x$  the best score according to  $\Pr(x|y)$ 
  - $f(x) = \operatorname{argmax}_y \Pr(x \mid y) * \Pr(y)$

# Text Classification Algorithm: Learning

- From training documents, extract *Vocabulary*
- Calculate required  $P(c_j)$  and  $P(x_k / c_j)$  terms
  - For each  $c_j$  in  $C$  do
    - $docs_j \leftarrow$  subset of documents for which the target class is  $c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- for each word  $x_k$  in *Vocabulary*
  - $n_k \leftarrow$  total number of occurrences of  $x_k$  where target class is  $c_j$
  - $n \leftarrow$  total number of words in all training examples whose target value is  $c_j$

$$P(x_k | c_j) \leftarrow \frac{n_k + 1}{n + |Vocabulary|}$$

# Text Classification Algorithm: Classifying

- ▶ How to estimate  $P(x_i | c_j)$
- ▶ Simplest useful process
  - Pick a word 1 according to  $P(W|Y)$
  - Repeat for word 2,3..
  - each word is generated independently of the others

$$\Pr(w_1, \dots, w_n | Y = y) = \prod_{i=1}^n \underbrace{\Pr(w_i | Y = y)}$$

$$\Pr(W = w | Y = y) = \frac{\text{count}(W = w \text{ and } Y = y)}{\text{count}(Y = y)} + \frac{1}{|\text{Vocabulary}|}$$

This gives score of zero if  $x$  contains a brand-new word  $w_{new}$

# Text Classification : Example

Word	Soccer	Election	Chip	Hockey	hardware	Policy	CPU	Vot	Cricket	class
Doc1	10	1	2	12	2	3	1	2	15	Sports
Doc2	1	1	15	4	12	2	13	2	5	Comp
Doc3	1	15	1	2	4	12	1	12	1	Politics
Doc4	13	1	2	14	1	2	1	2	12	Sports
Doc5	14	3	1	12	4	2	5	2	14	Sports
Doc6	2	4	14	4	13	1	14	5	3	Comp
Doc7	1	13	1	2	2	12	5	14	5	Politics

Class Attribute	Meaning
Sports	Sports related document
Comp	Computers related document
Politics	Politics related document

# Calculating Probabilities - Example

Compute the prior probabilities for each class.

- $P(\text{Comp}) \rightarrow \text{Number of Computers Related Documents} / \text{Number of Documents}$   
 **$P(\text{Comp}) = 2 / 7$**
- $P(\text{Politics}) \rightarrow \text{Number of Politics Related Documents} / \text{Number of Documents}$   
 **$P(\text{Politics}) = 2 / 7$**
- $P(\text{Sports}) \rightarrow \text{Number of Sports Related Documents} / \text{Number of Documents}$   
 **$P(\text{Sports}) = 3 / 7$**

# Calculating Probabilities - Example

Now lets calculate Probabilities for every word..

$P(\text{"soccer"} | \text{Comp})$

The word "soccer" occurs  $1+2 = 3$  times in comp docs

The total number of words in "comp" doc =  $1+1+15+4+\dots = 115$

then  **$P(\text{"soccer"} | \text{Comp}) = 3/115$**

Word "soccer" occurs  $10+13+14 = 37$  times in Sports docs.

then  **$P(\text{"soccer"} | \text{Sports}) = 37/153$**

Word "soccer" occurs  $1+1+1 = 3$  times in politics docs.

then  **$P(\text{"soccer"} | \text{politics}) = 3/104$**

# Calculating Probabilities - Example

Word	Class	P(Word Class)
Soccer	Sports	<b>37/153</b>
Soccer	Comp	3/115
Soccer	Politics	3/104
Chip	Comp	<b>29/115</b>
Chip	Sports	5/153
Chip	Politics	2/104
Policy	Comp	3/115
Policy	Sports	7/153
Policy	Politics	<b>24/104</b>

Class Attribute	P(Class)
Sports	3/7
Comp	2/7
Politics	2/7

# Text Classification Algorithm: Classifying

- Words  $\leftarrow$  all words in current document which contain tokens found in *Vocabulary*
- For each word in Word, lookup the previously computed prior probability and
- Return  $c_{NB}$ , where

$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j)$$

# TEST Document

Word	Soccer	Election	chip	Hockey	hardware	Policy	cpu	Vote	Cricket	class
Docx	11	1	2	10	2	3	1	2	15	C = sports

$C_j = \text{Sports}$

$$\Rightarrow \frac{3}{7} * 0.034 = 0.0145$$

$C_j = \text{Comp}$

$$\Rightarrow \frac{2}{7} * 0.0104 = 0.0029$$

$C_j = \text{Politics}$

$$\Rightarrow \frac{2}{7} * 0.0014 = 0.0004$$

Then the class with the highest posterior probability, is selected

$C = \text{Sports}$

THE record is correctly classified

Thank You