



# Chapter 4 Machine Learning

COMP 6721 Introduction of AI

*Russell & Norvig – Section 18.1 & 18.2*

# Supervised Learning Algorithms

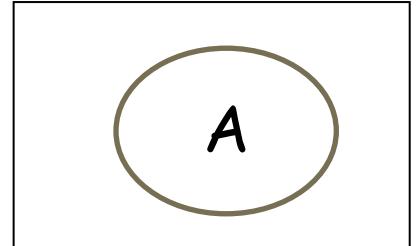
- ▶ ***Linear Regression***
- ▶ ***Logistic Regression***
- ▶ ***Decision Tree***
- ▶ ***Random Forest***
- ▶ **Naïve Bayes Classifier**

# Why Naïve Bayes ?

- How do we represent and reason about non-factual knowledge?
  - It *might* rain tonight
  - If you have red spots on your face, you *might* have the measles
  - This e-mail is *most likely* spam
  - I can't read this character, but it *looks* like a "B"
  - These 2 pictures are *very likely* of the same person
  - ...

# Remember

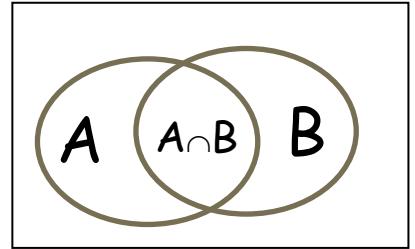
- P is a probability function:
  - $0 \leq P(A) \leq 1$
  - $P(A) = 0 \Rightarrow$  the event A will never take place
  - $P(A) = 1 \Rightarrow$  the event A must take place
  - $\sum_i P(A_i) = 1 \Rightarrow$  one of the events  $A_i$  will take place
  - $P(A) + P(\sim A) = 1$



# Remember

## ► Joint probability

- intersection  $A_1 \cap \dots \cap A_n$  is an event that takes place if **all** the events  $A_1, \dots, A_n$  take place
- denoted  $P(A \cap B)$  or  $P(A, B)$



## ► Sum Rule

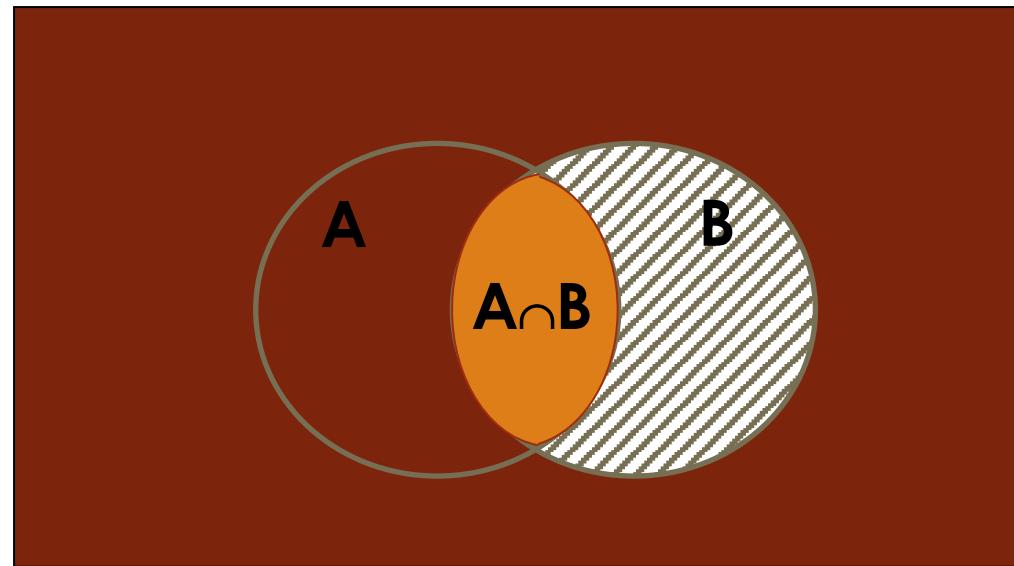
- union  $A_1 \cup \dots \cup A_n$  is an event that takes place if **at least** one of the events  $A_1, \dots, A_n$  takes place
- denoted  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

# Conditional Probability

- ▶ Prior (or unconditional) probability
  - ▶ Probability of an event before any evidence is obtained
  - ▶  $P(A) = 0.1$      $P(\text{rain today}) = 0.1$
  - ▶ i.e. Your belief about A given that you have no evidence
- ▶ Posterior (or conditional) probability
  - ▶ Probability of an event given that you know that B is true (B = some evidence)
  - ▶  $P(A | B) = 0.8$      $P(\text{rain today} | \text{cloudy}) = 0.8$
  - ▶ i.e. Your belief about A given that you know B

# Conditional Probability

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



# Chain Rule

- With 2 events, the probability that A and B occur is:

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ so } P(A,B) = P(A|B) \times P(B)$$

- With 3 events, the probability that A, B and C occur is:

- The probability that A occurs
- Times, the probability that B occurs, assuming that A occurred
- Times, the probability that C occurs, assuming that A and B have occurred

- With n events, we can generalize to the Chain rule:

$$P(A_1, A_2, A_3, A_4, \dots, A_n)$$

$$= P(\cap A_i)$$

$$= P(A_1) \times P(A_2 | A_1) \times P(A_3 | A_1, A_2) \times \dots \times P(A_n | A_1, A_2, A_3, \dots, A_{n-1})$$

# Example 1

- Joint probability distribution:

**evidence**

		Toothache	~Toothache
		0.04	0.06
hypothesis	Cavity	0.04	0.06
	~Cavity	0.01	0.89

$$P(H | E) = \frac{P(H \cap E)}{P(E)}$$

$$P(\text{cavity} | \text{toothache}) = \frac{P(\text{cavity} \cap \text{toothache})}{P(\text{toothache})} = \frac{0.04}{0.04 + 0.01} = 0.8$$

# Getting the Probabilities

- in most applications, you just count from a set of observations

$$P(A) = \frac{\text{count\_of\_}A}{\text{count\_of\_all\_events}}$$

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{\text{count\_of\_}A\text{\_and\_}B\text{\_together}}{\text{count\_of\_all\_}B}$$

# Combining Evidence

- ▶ Assume now 2 pieces of evidence:
- ▶ Suppose, we know that
  - ▶  $P(\text{Cavity} \mid \text{Toothache}) = 0.12$
  - ▶  $P(\text{Cavity} \mid \text{Young}) = 0.18$
- ▶ A patient complains about Toothache and is Young...
  - ▶ what is  $P(\text{Cavity} \mid \text{Toothache} \cap \text{Young})$  ?

# Combining Evidence

		Toothache		~Toothache	
		Young	~ Young	Young	~ Young
Cavity	0.108	0.012	0.072	0.008	
~Cavity	0.016	0.064	0.144	0.576	

$P(\text{Toothache} \cap \text{Cavity} \cap \text{Young})$

- ▶ But how do we get the data ?
- ▶ In reality, we may have dozens, hundreds of variables
- ▶ We cannot have a table with the probability of all possible combinations of variables
  - ▶ Ex. with 16 binary variables, we would need  $2^{16}$  entries

# Independent Events

- In real life:
  - some variables are independent...
    - ex: living in Montreal & tossing a coin
    - $P(\text{Montreal, head}) = P(\text{Montreal}) * P(\text{head})$
    - probability of 2 heads in a row:
      - $P(\text{head, head}) = 1/2 * 1/2 = 1/4$
  - some variables are not independent...
    - ex: living in Montreal & wearing boots
    - $P(\text{Montreal, boots}) \neq P(\text{Montreal}) * P(\text{boots})$

# Independent Events

- ▶ Two events A and B are independent:
  - ▶ if the occurrence of one of them does not influence the occurrence of the other
    - ▶ i.e. A is independent of B if  $P(A) = P(A | B)$
  - ▶ If A and B are independent, then:
    - ▶  $P(A,B) = P(B) \times P(A | B)$  (by chain rule)  
 $= P(A) \times P(B)$  (by independence)
- ▶ To make things work in real applications, we often assume that events are independent
  - ▶  $P(A,B) = P(A) \times P(B)$

# Conditional Independent Events

- ▶ Two events A and B are conditionally independent given C:
  - ▶ Given that C is true, then any evidence about B cannot change our belief about A
  - ▶  $P(A, B \mid C) = P(A \mid C) \times P(B \mid C)$ .

# Bayes' Theorem

► given:

$$P(A|B) = \frac{P(A,B)}{P(B)} \text{ so } P(A,B) = P(A|B) \times P(B)$$

$$P(B|A) = \frac{P(A,B)}{P(A)} \text{ so } P(A,B) = P(B|A) \times P(A)$$

► then:

$$P(A|B) \times P(B) = P(B|A) \times P(A)$$

► and:

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

# Bayes' Theorem

- ▶ We typically want to know:  $P(\text{Hypothesis} \mid \text{Evidence})$ 
  - ▶  $P(\text{Disease} \mid \text{Symptoms}) \dots P(\text{meningitis} \mid \text{red spots})$
  - ▶  $P(\text{Cause} \mid \text{Side Effect}) \dots P(\text{misaligned brakes} \mid \text{squeaky wheels})$
- ▶ But  $P(\text{Hypothesis} \mid \text{Evidence})$  is hard to gather
  - ▶ ex: *out of all people who have red spots... how many have meningitis?*
- ▶ However  $P(\text{Evidence} \mid \text{Hypothesis})$  is easier to gather
  - ▶ ex: *out of all people who have the meningitis ... how many have red spots?*
- ▶ So

$$P(\text{Hypothesis} \mid \text{Evidence}) = \frac{P(\text{Evidence} \mid \text{Hypothesis}) \times P(\text{Hypothesis})}{P(\text{Evidence})}$$

## Example 2

- ▶ Assume we only have 1 hypothesis
- ▶ Assume:
  - ▶  $P(\text{spots}=\text{yes} \mid \text{meningitis}=\text{yes}) = 0.4$
  - ▶  $P(\text{meningitis}=\text{yes}) = 0.00003$
  - ▶  $P(\text{spots}=\text{yes}) = 0.05$

$$P(\text{meningitis} = \text{yes} \mid \text{spots} = \text{yes})$$

$$= \frac{P(\text{spots} = \text{yes} \mid \text{meningitis} = \text{yes}) \times P(\text{meningitis} = \text{yes})}{P(\text{spots} = \text{yes})}$$

$$= \frac{0.4 \times 0.00003}{0.05} = 0.00024$$

- ▶ If you have spots... you are more likely to have meningitis than if we don't know about you having spots

# Example 3

- ▶ Predict the weather tomorrow based on tonight's sunset...
- ▶ Assume we have 3 hypothesis...
  - ▶  $H_1$ : weather will be *nice*     $P(H_1) = 0.2$
  - ▶  $H_2$ : weather will be *bad*     $P(H_2) = 0.5$
  - ▶  $H_3$ : weather will be *mixed*     $P(H_3) = 0.3$
- ▶ And 1 piece of evidence with 3 possible values
  - ▶  $E_1$ : today, there's a *beautiful sunset*
  - ▶  $E_2$ : today, there's a *average sunset*
  - ▶  $E_3$ : today, there's *no sunset*

$P(E_2 | H_1)$

$P(E_x   H_i)$	$E_1$	$E_2$	$E_3$
$H_1$	0.7	0.2	0.1
$H_2$	0.3	0.3	0.4
$H_3$	0.4	0.4	0.2

## Example 3

- ▶ Observation: average sunset ( $E_2$ )
- ▶ Question: how will be the weather tomorrow?
  - ▶  $P(H_i | E_2)$  ?
  - ▶ predict the weather that maximizes the probability
  - ▶ select  $H_i$  such that  $P(H_i | E_2)$  is the greatest

$$P(H_i | E_2) = \frac{P(H_i) \times P(E_2 | H_i)}{P(E_2)}$$

$$\begin{aligned} P(E_2) &= P(H_1) \times P(E_2 | H_1) + P(H_2) \times P(E_2 | H_2) + P(H_3) \times P(E_2 | H_3) \\ &= .2 \times .2 + .5 \times .3 + .3 \times .4 = .04 + .15 + .12 = 0.31 \end{aligned}$$

$P(E_x   H_i)$	$E_1$	$E_2$	$E_3$
$H_1$	0.7	0.2	0.1
$H_2$	0.3	0.3	0.4
$H_3$	0.4	0.4	0.2

# Example 3

$$P(H_1 | E_2) = \frac{P(H_1) \times P(E_2 | H_1)}{P(E_2)} = \frac{.2 \times .2}{.31} = .129$$

$$P(H_2 | E_2) = \frac{P(H_2) \times P(E_2 | H_2)}{P(E_2)} = \frac{.5 \times .3}{.31} = .484$$

$$P(H_3 | E_2) = \frac{P(H_3) \times P(E_2 | H_3)}{P(E_2)} = \frac{.3 \times .4}{.31} = .387$$

$\Rightarrow H_2$  is the most likely hypothesis, given the evidence

$P(H_2 | E_2)$  is the highest

Tomorrow the weather will be bad

$$H_{NB} = \operatorname{argmax}_{H_i} \frac{P(H_i) \times P(E | H_i)}{P(E)}$$

$P(E_x   H_i)$	$E_1$	$E_2$	$E_3$
$H_1$	0.7	0.2	0.1
$H_2$	0.3	0.3	0.4
$H_3$	0.4	0.4	0.2

$H_1$ : weather will be nice  
 $H_2$ : weather will be bad

$H_3$ : weather will be mixed

$E_1$ : today, there's a *beautiful* sunset

$E_2$ : today, there's a *average* sunset

$E_3$ : today, there's *no* sunset

# Bayes' Reasoning

- ▶ Out of n hypothesis...
  - ▶ we want to find the most probable  $H_i$  given the evidence E
- ▶ So we choose the  $H_i$  with the largest  $P(H_i | E)$

$$H_{NB} = \operatorname{argmax}_{H_i} P(H_i | E) = \operatorname{argmax}_{H_i} \frac{P(H_i) \times P(E | H_i)}{P(E)}$$

- ▶ But...  $P(E)$ 
  - ▶ is the same for all possible  $H_i$  (and is hard to gather anyways)
  - ▶ so we can drop it
- ▶ So Bayesian reasoning:

$$\begin{aligned} P(E_2) &= P(H_1) \times P(E_2 | H_1) + P(H_2) \times P(E_2 | H_2) + P(H_3) \times P(E_2 | H_3) \\ &=.2 \times .2 + .5 \times .3 + .3 \times .4 = .04 + .15 + .12 = 0.31 \end{aligned}$$

$$H_{NB} = \operatorname{argmax}_{H_i} \frac{P(H_i) \times P(E | H_i)}{P(E)} = \operatorname{argmax}_{H_i} P(H_i) \times P(E | H_i)$$

# Representing the Evidence

- ▶ The evidence is typically represented by many attributes/features
  - ▶ beautiful sunset? clouds? temperature? summer?, ...
- ▶ so often represented as a feature/attribute vector

evidence						hypothesis
	sunset a <sub>1</sub>	clouds a <sub>2</sub>	temp a <sub>3.</sub>	summer a <sub>4</sub>		weather tomorrow
e1	beautiful	no	high	yes		Nice

- ▶ e1 = <a<sub>1</sub>, ..., a<sub>n</sub>>
- ▶ e1 = <sunset:beautiful, clouds:no, temp:high, summer:yes>

# Combining Evidence

toothache	young	cavity
yes	yes	?

$$P(\text{Cavity} = \text{yes} | \text{Toothache} = \text{yes} \cap \text{Young} = \text{yes}) = ?$$

with Bayes Rule :

$$= \frac{P(\text{Toothache} = \text{yes} \cap \text{Young} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes} \cap \text{Young} = \text{yes})}$$

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

with independence assumption :  $P(A, B) = P(A) \times P(B)$

$$= \frac{P(\text{Toothache} = \text{yes} \cap \text{Young} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

with conditional independence assumption :  $P(A, B | C) = P(A | C) \times P(B | C)$

$$= \frac{P(\text{Toothache} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Young} = \text{yes} | \text{Cavity} = \text{yes}) \times P(\text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

Now we have decomposed the joint probability distribution into much smaller pieces...

# Combining Evidence

toothache	young	cavity
yes	yes	yes? or no?

But since we only care about ranking the hypothesis...

$$P(\text{Cavity} = \text{yes} \mid \text{Toothache} = \text{yes} \cap \text{Young} = \text{yes})$$

&gt;

$$P(\text{Cavity} = \text{no} \mid \text{Toothache} = \text{yes} \cap \text{Young} = \text{yes})$$

$$\frac{P(\text{Cavity} = \text{yes}) \times P(\text{Toothache} = \text{yes} \mid \text{Cavity} = \text{yes}) \times P(\text{Young} = \text{yes} \mid \text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

&gt;

$$\frac{P(\text{Cavity} = \text{no}) \times P(\text{Toothache} = \text{yes} \mid \text{Cavity} = \text{no}) \times P(\text{Young} = \text{yes} \mid \text{Cavity} = \text{no})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

$$\frac{P(\text{Cavity} = \text{yes}) \times P(\text{Toothache} = \text{yes} \mid \text{Cavity} = \text{yes}) \times P(\text{Young} = \text{yes} \mid \text{Cavity} = \text{yes})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

&gt;

$$\frac{P(\text{Cavity} = \text{no}) \times P(\text{Toothache} = \text{yes} \mid \text{Cavity} = \text{no}) \times P(\text{Young} = \text{yes} \mid \text{Cavity} = \text{no})}{P(\text{Toothache} = \text{yes}) \times P(\text{Young} = \text{yes})}$$

$$H_{NB} = \operatorname{argmax}_{H_i} \frac{P(H_i) \times P(E \mid H_i)}{P(E)} = \operatorname{argmax}_{H_i} P(H_i) \times P(E \mid H_i) = \operatorname{argmax}_{H_i} P(H_i) \times P(a_1, a_2, a_3, \dots, a_n \mid H_i) = \operatorname{argmax}_{H_i} P(H_i) \times \prod_{j=1}^n P(a_j \mid H_i)$$

# Example 4

evidence

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

# Example 4

evidence

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No



Frequency Table		Play	
		Yes	No
Outlook	Sunny	2	3
	Overcast	4	0
	Rainy	3	2

Frequency Table		Play	
		Yes	No
Humidity	High	3	4
	Normal	6	1

Frequency Table		Play	
		Yes	No
Wind	Strong	3	3
	Weak	6	2

## Example 4

- Goal: Given a new instance  $X = \langle a_1, \dots, a_n \rangle$ , classify as Yes/No

$$H_{NB} = \operatorname{argmax}_{H_i} \frac{P(H_i) \times P(E|H_i)}{P(E)} = \operatorname{argmax}_{H_i} P(H_i) \times P(E|H_i) = \operatorname{argmax}_{H_i} P(H_i) \times P(\langle a_1, a_2, a_3, \dots, a_n \rangle | H_i) = \operatorname{argmax}_{H_i} P(H_i) \times \prod_{j=1}^n P(a_j | H_i)$$

- Naïve Bayes: Assumes that the attributes/features are conditionally independent

## Example 4

► Goal: Given a new instance  $X = \langle a_1, \dots, a_n \rangle$ , classify as Yes/No

$$H_{NB} = \underset{H_i}{\operatorname{argmax}} P(H_i) \times \prod_{j=1}^n P(a_j | H_i)$$

1. 1st estimate the probabilities from the training examples:
  - a) For each hypothesis  $H_i$  **estimate  $P(H_i)$**
  - b) For each attribute value  $a_j$  of each instance (evidence)  
**estimate  $P(a_j | H_i)$**

# Example 4

## 1. TRAIN:

-- compute the probabilities from the training set

$$P(\text{Play Tennis} = \text{yes}) = 9/14 = 0.64$$

$$P(\text{Play Tennis} = \text{no}) = 5/14 = 0.36$$

}

prior probabilities  $P(H_i)$

$$P(\text{Out} = \text{sunny} | \text{Play Tennis} = \text{yes}) = 2/9 = 0.22$$

$$P(\text{Out} = \text{sunny} | \text{Play Tennis} = \text{no}) = 3/5 = 0.60$$

$$P(\text{Out} = \text{rain} | \text{Play Tennis} = \text{yes}) = 3/9 = 0.33$$

$$P(\text{Out} = \text{rain} | \text{Play Tennis} = \text{no}) = 2/5 = 0.4$$

...

$$P(\text{Wind} = \text{strong} | \text{Play Tennis} = \text{yes}) = 3/9 = 0.33$$

$$P(\text{Wind} = \text{strong} | \text{Play Tennis} = \text{no}) = 3/5 = 0.60$$

conditional probabilities  
 $P(a_j | H_i)$

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

# Example 4

Likelihood Table		Play		
		Yes	No	
Outlook	Sunny	2/9	3/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	3/9	2/5	5/14
		9/14	5/14	

$$P(\text{Sunny} \mid \text{Yes}) = \dots$$

$$P(\text{Overcast} \mid \text{Yes}) = \dots$$

$$P(\text{Rainy} \mid \text{Yes}) = \dots$$

Likelihood Table		Play		
		Yes	No	
Wind	Strong	3/9	3/5	6/14
	Weak	6/9	2/5	8/14
		9/14	5/14	

$$P(\text{Strong} \mid \text{Yes}) = \dots$$

$$P(\text{Weak} \mid \text{Yes}) = \dots$$

Likelihood Table		Play		
		Yes	No	
Humidity	High	3/9	4/5	7/14
	Normal	6/9	1/5	7/14
		9/14	5/14	

$$P(\text{High} \mid \text{Yes}) = \dots$$

$$P(\text{Normal} \mid \text{Yes}) = \dots$$

# Example 4

## 2. TEST:

classify the new case:  $X = (\text{Outlook: Sunny}, \text{Temp: Cool}, \text{Hum: High}, \text{Wind: Strong})$

$$\begin{aligned}
 H_{NB} &= \underset{H_i \in \{\text{yes, no}\}}{\operatorname{argmax}} P(H_i) \times P(X | H_i) \\
 &= \underset{H_i \in \{\text{yes, no}\}}{\operatorname{argmax}} P(H_i) \times \prod_j P(a_j | H_i) \\
 &= \underset{H_i \in \{\text{yes, no}\}}{\operatorname{argmax}} P(H_i) \times P(\text{Outlook} = \text{sunny} | H_i) \times P(\text{Temp} = \text{cool} | H_i) \\
 &\quad \times P(\text{Humidity} = \text{high} | H_i) \times P(\text{Wind} = \text{strong} | H_i)
 \end{aligned}$$

1)  $P(\text{PlayTennis} = \text{yes})$

$$\begin{aligned}
 &\times P(\text{Outlook} = \text{sunny} | \text{PlayTennis} = \text{yes}) \times P(\text{Temp} = \text{cool} | \text{PlayTennis} = \text{yes}) \times P(\text{Hum} = \text{high} | \text{PlayTennis} = \text{yes}) \times P(\text{Wind} = \text{strong} | \text{PlayTennis} = \text{yes}) \\
 &= 0.0053
 \end{aligned}$$

2)  $P(\text{PlayTennis} = \text{no})$

$$\begin{aligned}
 &\times P(\text{Outlook} = \text{sunny} | \text{PlayTennis} = \text{no}) \times P(\text{Temp} = \text{cool} | \text{PlayTennis} = \text{no}) \times P(\text{Hum} = \text{high} | \text{PlayTennis} = \text{no}) \times P(\text{Wind} = \text{strong} | \text{PlayTennis} = \text{no}) \\
 &= 0.0206
 \end{aligned}$$

$\Rightarrow \text{answer : PlayTennis}(X) = \text{no}$

# Application of Bayesian Reasoning

- ▶ Categorization:  $P(\text{Category} \mid \text{Features of Object})$ 
  - ▶ Diagnostic systems:  $P(\text{Disease} \mid \text{Symptoms})$
  - ▶ Text classification:  $P(\text{sports\_news} \mid \text{text})$
  - ▶ Character recognition:  $P(\text{character} \mid \text{bitmap})$
  - ▶ Speech recognition:  $P(\text{words} \mid \text{acoustic signal})$
  - ▶ Image processing:  $P(\text{face\_person} \mid \text{image features})$
  - ▶ Spam filter:  $P(\text{spam\_message} \mid \text{words in e-mail})$
  - ▶ ...

# Naïve Bayes Classifier

- ▶ A simple probabilistic classifier based on Bayes' theorem
  - ▶ with strong (naive) independence assumption
  - ▶ i.e. the features/attributes are conditionally independent
- ▶ The assumption of conditional independence, often does not hold...
- ▶ But Naïve Bayes works very well in many applications anyways!
  - ▶ ex: Medical Diagnosis
  - ▶ ex: Text Categorization (spam filtering)

# Application 1 – Spam Filtering

- ▶ Task: classify e-mails (documents) into a pre-defined class
  - ▶ ex: spam / ham
  - ▶ ex: sports, recreation, politics, war, economy,...
- ▶ Given
  - ▶ N sets of training texts (1 set for each class)
  - ▶ Each set is already tagged by the class name



Strictly speaking, what we will see is called a multinomial Naïve Bayes classifier, because we will count the number of words, as opposed to just using binary values for the presence/absence of words... but it's a cool application!

# Email Representation

- ▶ each e-mail is represented by a vector of feature/value:
  - ▶ feature = actual words in the e-mail
  - ▶ value = number of times that word appears in the e-mail
- ▶ each e-mail in the training set is tagged with the correct category.

data instance	features / evidence / $X$						$f(X)$
	offer	money	viagra	laptop	exam	study	
email 1	3	2	5	1	0	1	SPAM
email 2	1	1	0	5	4	3	HAM
email 3	0	3	2	1	0	1	SPAM
...							

- ▶ task: correctly tag a new e-mail

	offer	money	viagra	laptop	exam	study	category
new email	2	1	0	1	1	2	?

# Naïve Bayes Algorithm

```

// 1. training

for all classes  $c_i$  // ex. ham or spam

    for all words  $w_j$  in the vocabulary
        compute  $P(w_j | c_i) = \frac{\text{count}(w_j, c_i)}{\sum_j \text{count}(w_j, c_i)}$ 

for all classes  $c_i$ 
    compute  $P(c_i) = \frac{\text{count}(\text{documents in } c_i)}{\text{count}(\text{all documents})}$ 

// 2. testing a new document D

for all classes  $c_i$  // ex. ham or spam

    score( $c_i$ ) =  $P(c_i)$ 

    for all words  $w_j$  in the D
        score( $c_i$ ) = score( $c_i$ )  $\times P(w_j | c_i)$ 

choose  $c^*$  = with the greatest score( $c_i$ )

```

	$w_1$	$w_2$	$w_3$	$w_4$	$w_5$	$w_6$
$c1 : SPAM$	$p(w_1 c_1)$	$p(w_2 c_1)$	$p(w_3 c_1)$	$p(w_4 c_1)$	$p(w_5 c_1)$	$p(w_6 c_1)$
$c2 : HAM$	$p(w_1 c_2)$	$p(w_2 c_2)$	$p(w_3 c_2)$	$p(w_4 c_2)$	$p(w_5 c_2)$	$p(w_6 c_2)$

# Example

- ▶ Dataset
  - ▶ c1: SPAM
    - doc1: "cheap viagra for sale"
    - doc2: "click here for the best viagra"
    - doc3: "book your trip"
  - ▶ c2: HAM
    - doc4: "cheap book sale, not viagra"
    - doc5: "here is the book for you"
- ▶ Question:
  - ▶ doc6: "the cheap book"
  - ▶ should it be classified as HAM or SPAM?



# Example

Assume

vocabulary = {best, book, cheap, sale, trip, clearance}

If not in vocabulary, ignore word

1. Training:

- ▶  $P(\text{best} | \text{SPAM}) = 1/7 \quad P(\text{best} | \text{HAM}) = 0/5$
  - ▶  $P(\text{book} | \text{SPAM}) = 1/7 \quad P(\text{book} | \text{HAM}) = 2/5$
  - ▶  $P(\text{cheap} | \text{SPAM}) = 1/7 \quad P(\text{cheap} | \text{HAM}) = 1/5$
  - ▶  $P(\text{sale} | \text{SPAM}) = 1/7 \quad P(\text{sale} | \text{HAM}) = 1/5$
  - ▶  $P(\text{trip} | \text{SPAM}) = 1/7 \quad P(\text{trip} | \text{HAM}) = 0/5$
  - ▶  $P(\text{clearance} | \text{SPAM}) = 2/7 \quad P(\text{clearance} | \text{HAM}) = 1/5$
- 
- ▶  $P(\text{SPAM}) = 3/5 \quad P(\text{HAM}) = 2/5$

2. Testing: "the cheap book"

- ▶  $\text{Score}(\text{HAM}) = P(\text{HAM}) \times P(\text{cheap} | \text{HAM}) \times P(\text{book} | \text{HAM})$
- ▶  $\text{Score}(\text{SPAM}) = P(\text{SPAM}) \times P(\text{cheap} | \text{SPAM}) \times P(\text{book} | \text{SPAM})$

# Smooth Probabilities

- ▶ normally:  $P(w_i | c_j) = \frac{\text{(frequency of } w_i \text{ in } c_j)}{\text{total number of words in } c_j}$
- ▶ what if we have a  $P(w_i | c_j) = 0 \dots ?$ 
  - ▶ ex. the word "fake" never appeared in the class SPAM?
  - ▶ then  $P(\text{"fake"} | \text{SPAM}) = 0$
- ▶ so if a text contains the word "fake", the class SPAM is completely ruled out !
- ▶ to solve this: we assume that every word always appears at least once (or a smaller value)
  - ▶ ex: add-1 smoothing:

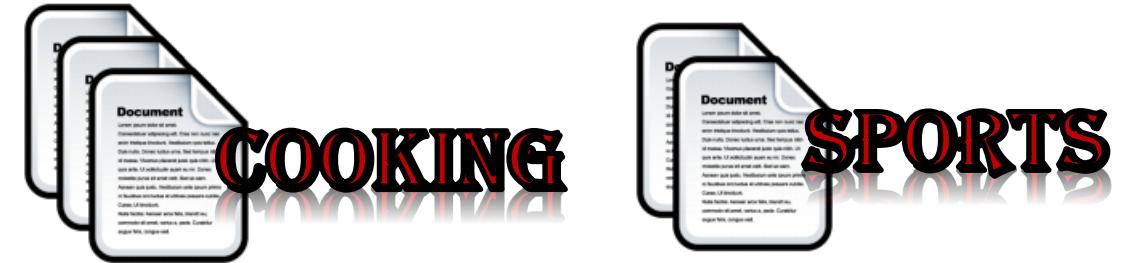
$$P(w_i | c_j) = \frac{\text{(frequency of } w_i \text{ in } c_j) + 1}{\text{total number of words in } c_j + \text{size of vocabulary}}$$

# Using logs

- ▶ if we really do the product of probabilities...
  - ▶  $\operatorname{argmax}_{c_j} P(c_j) \prod P(w_i | c_j)$
  - ▶ we soon have numerical underflow...
  - ▶ ex:  $0.01 \times 0.02 \times 0.05 \times \dots$
- ▶ so instead, we add the log of the probs
  - ▶  $\operatorname{argmax}_{c_j} \log(P(c_j)) + \sum \log(P(w_i | c_j))$
  - ▶ ex:  $\log(0.01) + \log(0.02) + \log(0.05) + \dots$

# Example

► Dataset



c1: COOKING

doc<sub>1</sub>: ... stove... kitchen... the... heat  
 doc<sub>2</sub>: ... kitchen... pasta... stove...  
 ...  
 doc<sub>100000</sub>: ... stove...heat... ball...

c2: SPORTS

doc<sub>1</sub>: ... ball... heat...  
 doc<sub>2</sub>: ... the... referee... player...  
 ...  
 doc<sub>75000</sub>: goal... injury ...

■ Assume:

- |V| = 100   vocabulary = {ball, heat, kitchen, referee, stove, the, ... }
- 500,000 words in Cooking
- 300,000 words in Sports
- 100,000 docs in Cooking
- 75,000 docs in Sports

# Example

► Training – Unsmoothed / Smoothed probs:

<ul style="list-style-type: none"> <li>► <math>P(\text{ball} \mid \text{COOKING}) = \frac{10,000}{500,000}</math> ??</li> <li>► <math>P(\text{heat} \mid \text{COOKING}) = \frac{255}{500,000}</math> ??</li> <li>► <math>P(\text{kitchen} \mid \text{COOKING}) = \frac{2,600}{500,000}</math> ??</li> <li>► <math>P(\text{referee} \mid \text{COOKING}) = \frac{0}{500,000}</math> ??</li> <li>► <math>P(\text{stove} \mid \text{COOKING}) = \frac{3,600}{500,000}</math> ??</li> <li>► <math>P(\text{the} \mid \text{COOKING}) = \frac{400,000}{500,000}</math> ??</li> <li>► ...</li> <li>► <math>P(\text{COOKING}) = \frac{100,000}{175,000}</math></li> </ul>	$P(\text{ball} \mid \text{SPORTS}) = \frac{10,000}{300,000}$ ?? $P(\text{heat} \mid \text{SPORTS}) = \frac{1,8000}{300,000}$ ?? $P(\text{kitchen} \mid \text{SPORTS}) = \frac{0}{300,000}$ ?? $P(\text{referee} \mid \text{SPORTS}) = \frac{1,500}{300,000}$ ?? $P(\text{stove} \mid \text{SPORTS}) = \frac{4}{300,000}$ ?? $P(\text{the} \mid \text{SPORTS}) = \frac{19,000}{300,000}$ ?? $P(\text{SPORTS}) = \frac{75,000}{175,000}$
--	--

# Example

- ▶ Training – Unsmoothed / **Smoothed** probs:
  - ▶  $P(COOKING) = \frac{100,000}{175,000}$        $P(SPORTS) = \frac{75,000}{175,000}$
- ▶ Testing: “the referee hit the ~~blue bird~~”
  - ▶  $\text{Score}(COOKING) = \log\left(\frac{100,000}{175,000}\right) + \log(P(\text{the} | COOKING)) + \log(P(\text{referee} | COOKING)) + \log(P(\text{hit} | COOKING)) + \log(P(\text{the} | COOKING))$
  - ▶  $\text{Score}(SPORTS) = \log\left(\frac{75,000}{175,000}\right) + \log(P(\text{the} | SPORTS)) + \log(P(\text{referee} | SPORTS)) + \log(P(\text{hit} | SPORTS)) + \log(P(\text{the} | SPORTS))$

## Application 2 – Post Code Recognition

BAM BAM  
42 T-REX RD.  
PANGAEA, RB 48016

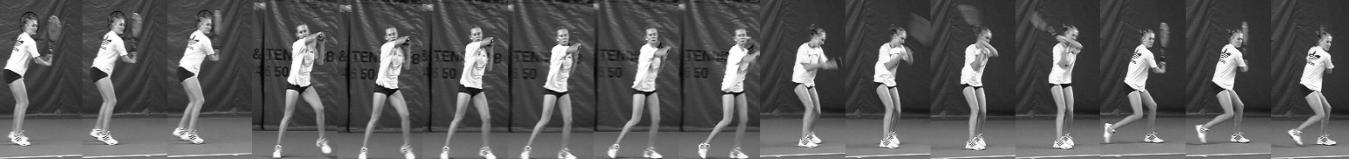
FRED FLINSTONE  
69 OLD SCHOOL AVE  
BEDROCK, OLDEN-TOWN  
77005

# Digital Recognition

- ▶ USPS data set contains handwritten digits, scanned by the U.S. Postal Service.
- ▶ 16 x 16 grayscale images
- ▶ 7291 training and 2007 test observations
- ▶ Features: each pixel is used as a feature, so there are 16 by 16, i.e., 256 features
- ▶ Task: classify new digits into one of the 10 classes

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

# Application 3 – Picture Classification



# Naïve Bayes Classification

- ▶ Makes a strong assumption of conditional independence
  - ▶ that is often incorrect
  - ▶ ex: the word *ambulance* is not conditionally independent of the word *accident* given the class SPORTS
- ▶ BUT:
  - ▶ surprisingly very effective on real-world tasks
  - ▶ basis of many spam filters
  - ▶ fast, simple
  - ▶ gives confidence in its class predictions (i.e., the scores)

# The End

