

# **Introduction to Machine Learning**

CS307 --- Fall 2022

Bayesian Learning

Reading:

Sections 13.1-13.6, 20.1-20.2, R&N

Sections 6.1-6.3, 6.7, 6.9, Mitchell

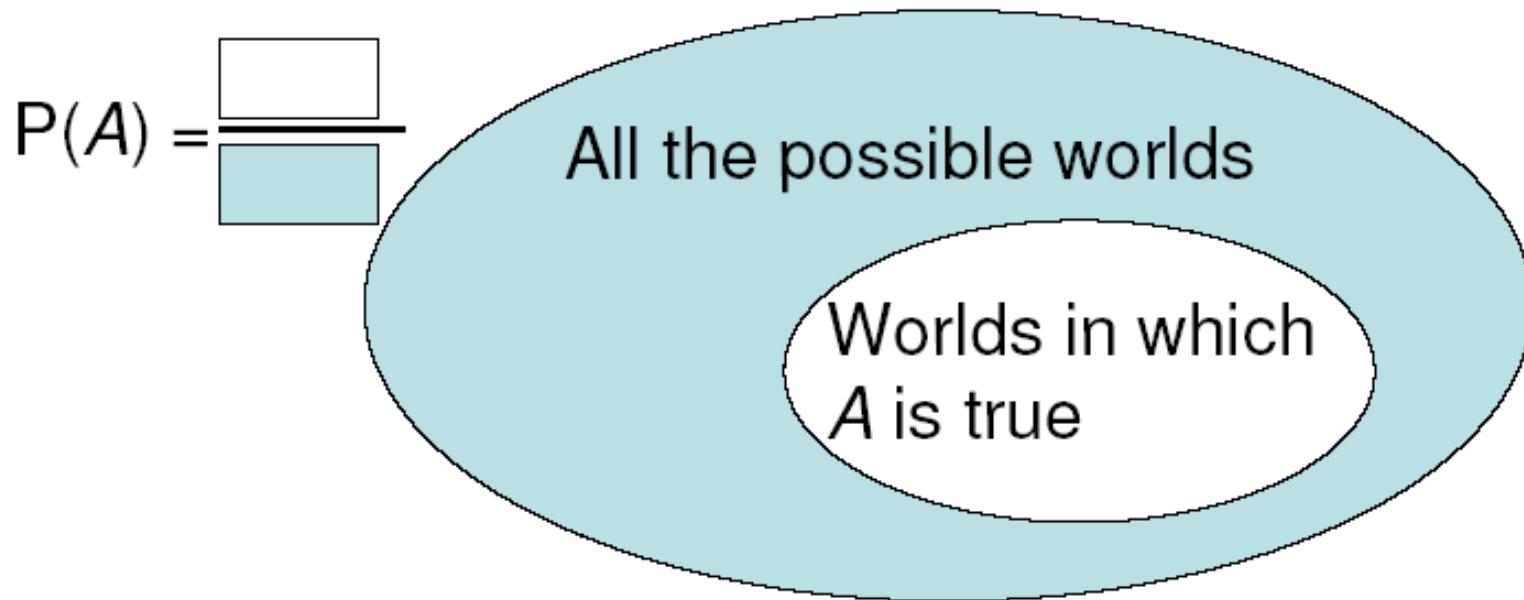
# Uncertainty

- Most real-world problems deal with uncertain information
  - Diagnosis: Likely disease given observed symptoms
  - Equipment repair: Likely component failure given sensor reading
  - Help desk: Likely operation based on past operations
  - Cannot be represented by deterministic rules  
Headache => Fever
- Correct framework for representing uncertainty:  
Probability

# Probability

- $P(A)$  = Probability of event  $A$  = fraction of all possible worlds in which  $A$  is true.

$$0 \leq P(A) \leq 1$$



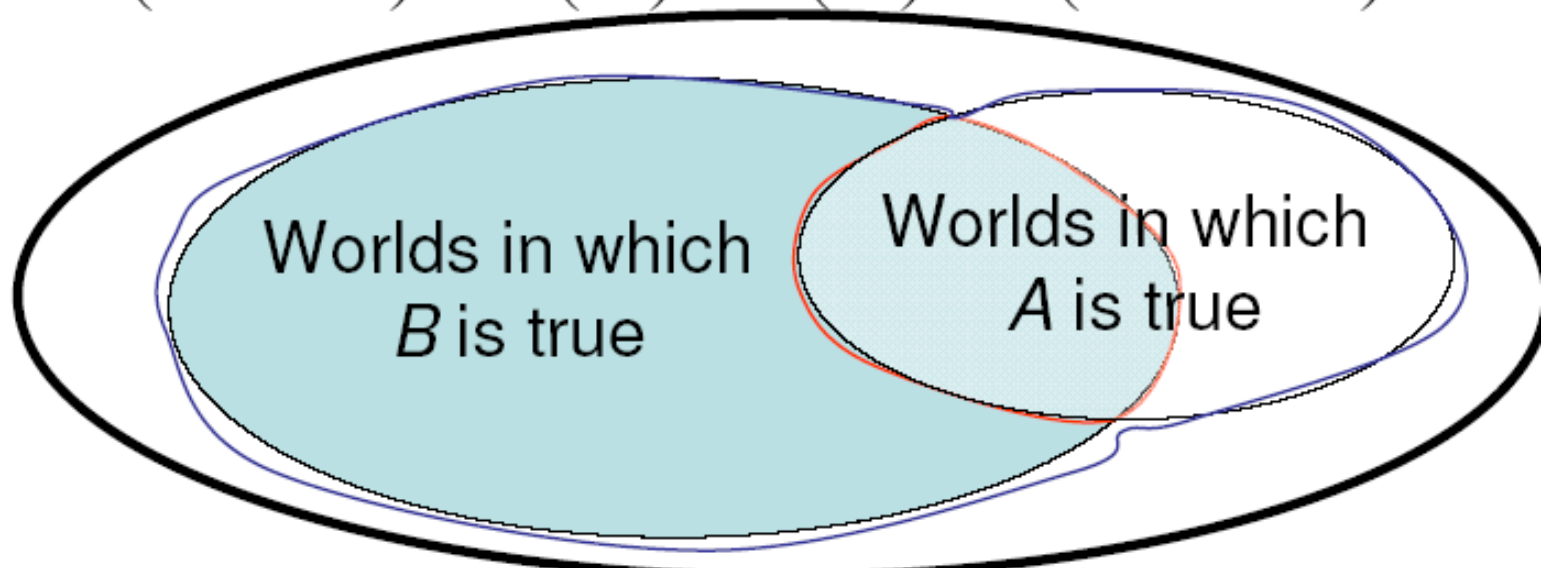
# Probability

$$0 \leq P(A) \leq 1$$

$$P(\textit{True}) = 1$$

$$P(\textit{False}) = 0$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



# Probability

- Immediately derived properties

$$P(\neg A) = 1 - P(A)$$

$$P(A) = P(A, B) + P(A, \neg B)$$

More general version of the **Theorem of Total Probability**:

IF we know that exactly one of  $B_1$ ,  $B_2$  ...,  $B_n$  are true

(i.e.  $P(B_1 \text{ or } B_2 \text{ or } \dots B_n) = 1$ , and for all  $i, j$  unequal,

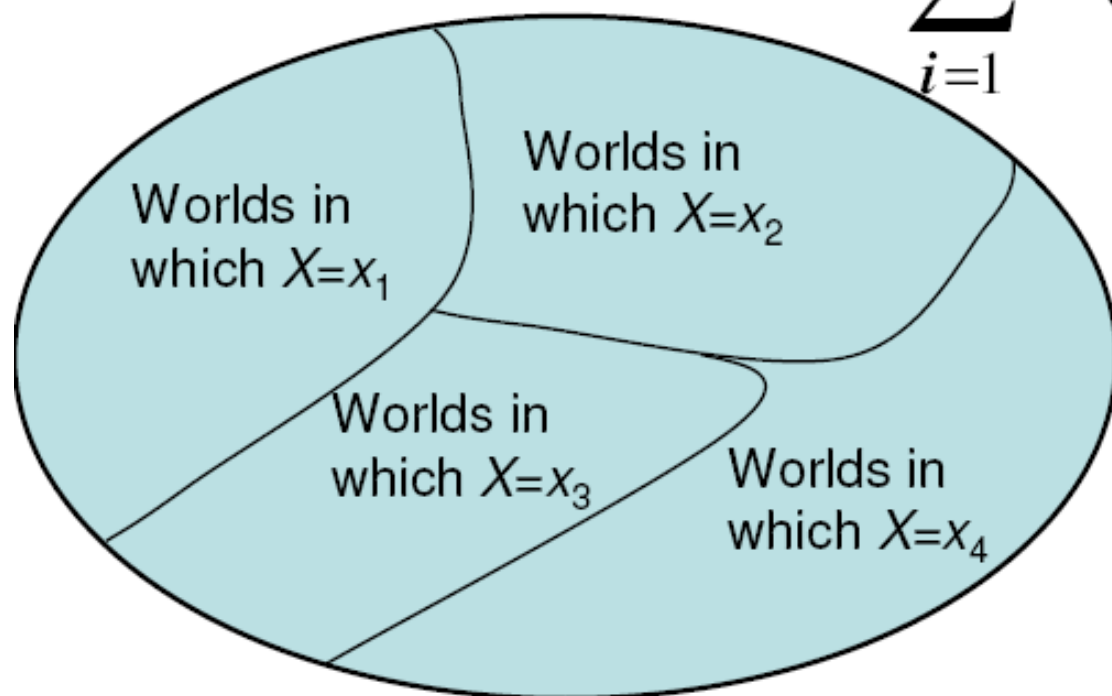
$P(B_i \text{ and } B_j) = 0$ ) THEN we know:

$$P(A) = P(A, B_1) + P(A, B_2) + \dots P(A, B_n)$$

# Probability

- A random variable is a variable  $X$  that can take values  $x_1, \dots, x_n$  with a probability  $P(X = x_i)$  attached to each  $i = 1, \dots, n$

$$\sum_{i=1}^n P(X = x_i) = 1$$



# Example

My mood can take one of two values: Happy, Sad.

The weather can take one of three values: Rainy, Sunny  
Cloudy.

Given

$$P(\text{Mood=Happy} \wedge \text{Weather=Rainy}) = 0.2$$

$$P(\text{Mood=Happy} \wedge \text{Weather=Sunny}) = 0.1$$

$$P(\text{Mood=Happy} \wedge \text{Weather=Cloudy}) = 0.4$$

Can I compute  $P(\text{Mood=Happy})$ ?

Can I compute  $P(\text{Mood=Sad})$ ?

Can I compute  $P(\text{Weather=Rainy})$  ?

# What's so great about the axioms of probability?

The axioms of probability mean you may not represent the following knowledge:

$$P(A) = 0.4$$

$$P(B) = 0.3$$

$$P(A \wedge B) = 0.0$$

$$P(A \vee B) = 0.8$$

Would you ever want to do that?

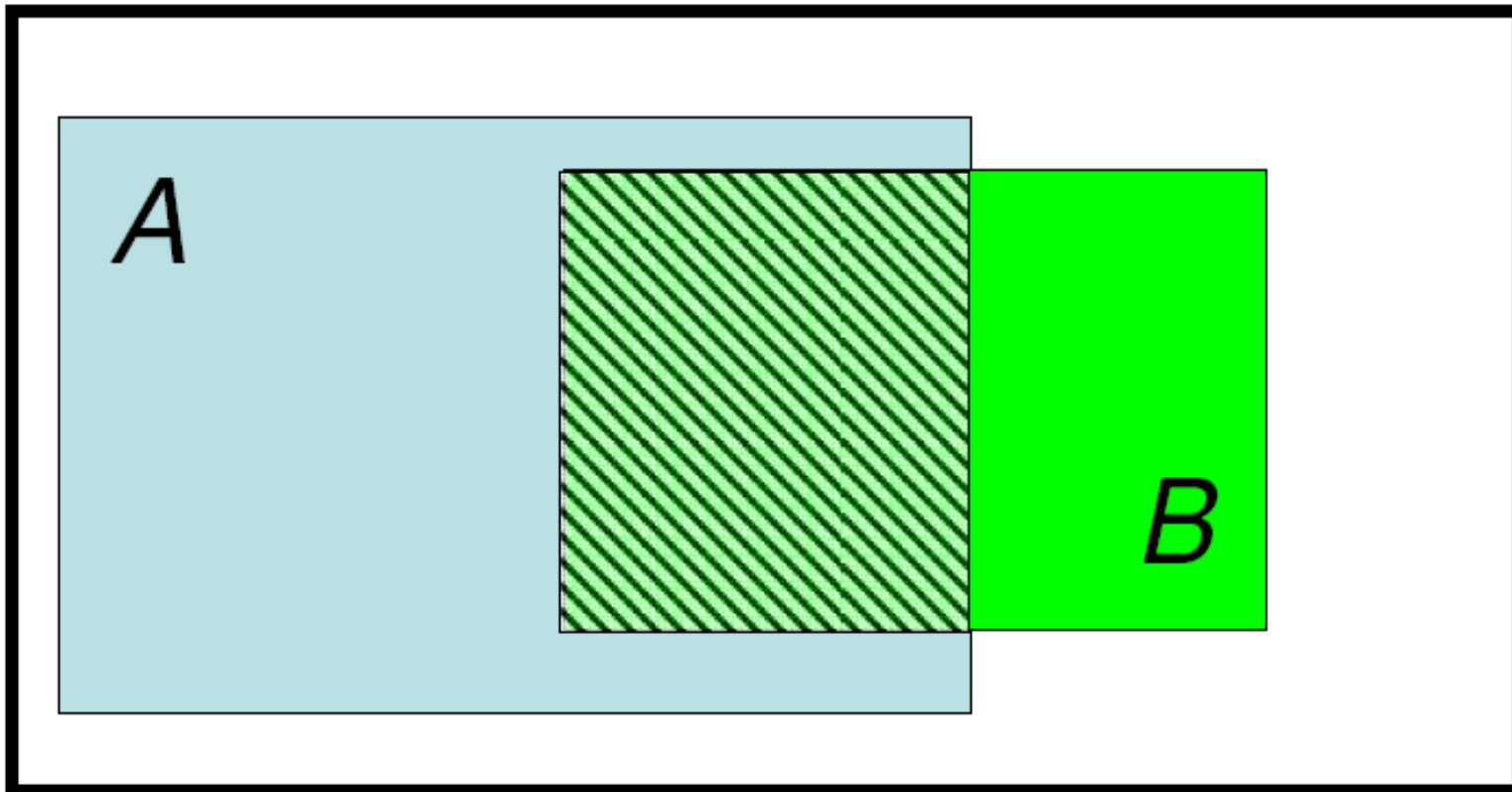
Difficult philosophical question.

Pragmatic answer: if you disobey the laws of probability in your knowledge representation, you can make suboptimal decisions.



# Conditional Probability

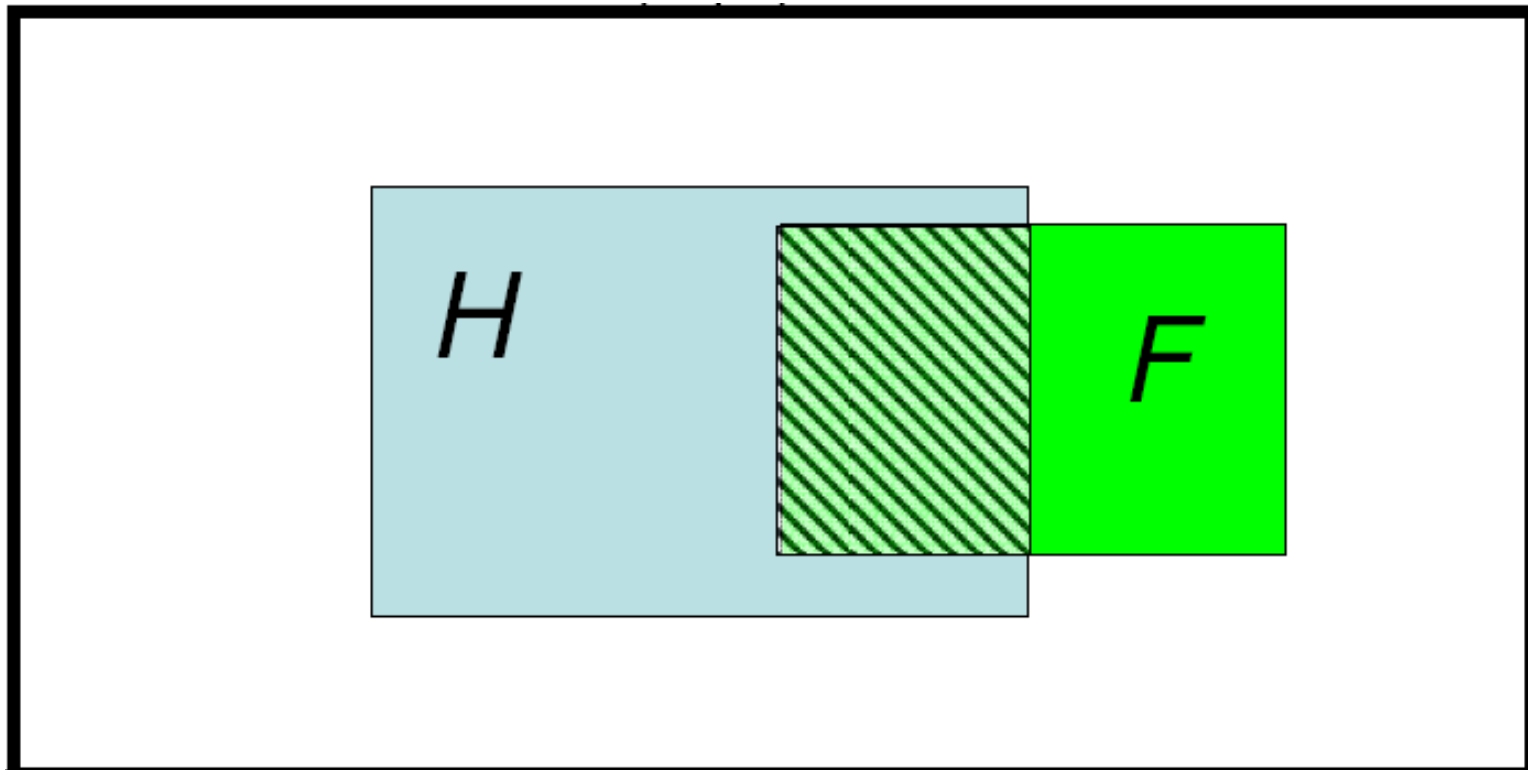
- $P(A|B)$  = Fraction of *those worlds in which B is true* for which A is also true.



# Conditional Probability Example

- $H$  = Headache  $P(H) = 1/2$
- $F$  = Flu  $P(F) = 1/8$

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



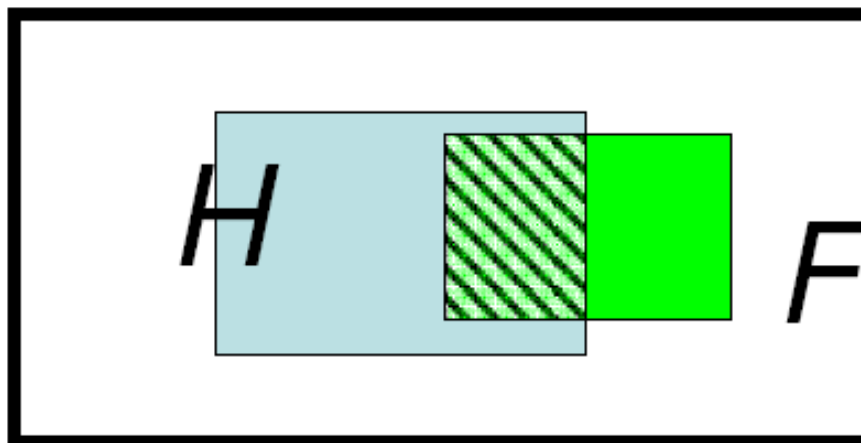
# Conditional Probability Example

- $H$  = Headache  $P(H) = 1/2$
- $F$  = Flu  $P(F) = 1/8$

$$P(H|F) = \frac{\text{Area of “}H \text{ and } F\text{” region}}{\text{Area of } F \text{ region}}$$

$$P(H|F) = P(H, F) / P(F)$$

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



# Conditional Probability

- Definition:

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

- Chain rule:

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

Can you prove that  $P(A, B) \leq P(A)$  for any events  $A$  and  $B$ ?

# Conditional Probability

- Other useful relations:

$$\mathbf{P}(A \mid B) + \mathbf{P}(\neg A \mid B) = 1$$

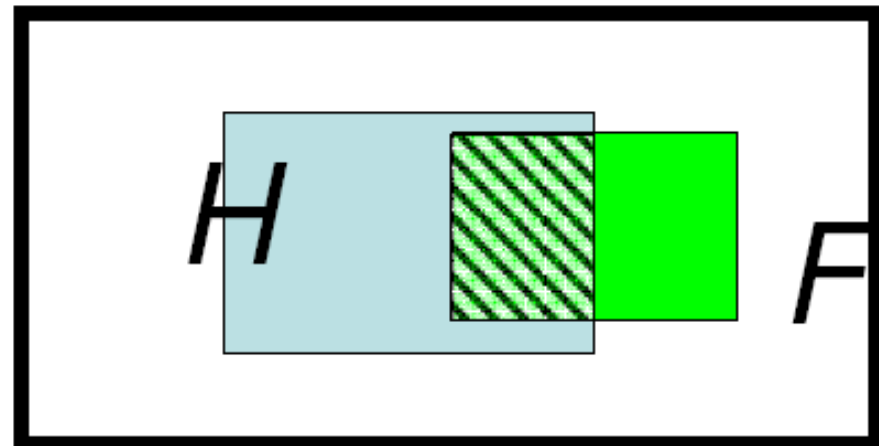
$$\sum_i \mathbf{P}(X = x_i \mid B) = 1$$

# Probabilistic Inference

- What is the probability that  $F$  is true given  $H$  is true?

Given

- $P(H) = 1/2$
- $P(F) = 1/8$
- $P(H|F) = 0.5$



# Probabilistic Inference

- Correct reasoning:
- We know  $P(H)$ ,  $P(F)$ ,  $P(H|F)$  and the two chain rules:

$$P(\mathbf{H}, \mathbf{F}) = P(\mathbf{H} \mid \mathbf{F}) P(\mathbf{F})$$

$$P(\mathbf{F} \mid \mathbf{H}) = \frac{P(\mathbf{H}, \mathbf{F})}{P(\mathbf{H})}$$

- Substituting the values:

$$P(\mathbf{H}, \mathbf{F}) = 0.5 \times 1/8 = 1/16$$

$$P(\mathbf{F} \mid \mathbf{H}) = \frac{1/16}{1/2} = \boxed{1/8}$$

## Bayes Rule

$$P(\mathbf{B} \mid \mathbf{A}) = \frac{P(\mathbf{A} \mid \mathbf{B}) P(\mathbf{B})}{P(\mathbf{A})}$$



# Bayes Rule

We want: *Posterior* probability that  $B$  occurs given that  $A$  occurs

We know: *Prior* probability that  $B$  occurs in the absence of any other information

$$P(\mathbf{B} \mid \mathbf{A}) = \frac{P(\mathbf{A} \mid \mathbf{B}) P(\mathbf{B})}{P(\mathbf{A})}$$

We know: *Likelihood* that  $A$  occurs given that  $B$  occurs

# Bayes Rule

- What if we do not know  $P(A)$ ???
- Use the relation:

$$P(A) = P(A | B) P(B) + P(A | \neg B) P(\neg B)$$

- More general Bayes rule:

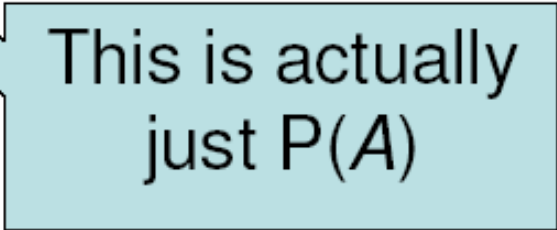
$$P(B | A) = \frac{P(A | B) P(B)}{P(A | B) P(B) + P(A | \neg B) P(\neg B)}$$

# Bayes Rule

- Same rule for a non-binary random variable, except we need to sum over all the possible events

$$P(X = x_i | A) = \frac{P(A | X = x_i) P(X = x_i)}{P(A)}$$

$$P(X = x_i | A) = \frac{P(A | X = x_i) P(X = x_i)}{\sum_k P(A | X = x_k) P(X = x_k)}$$



This is actually  
just  $P(A)$

# Generalizing Bayes Rule

If we know that exactly one of  $A_1, A_2, \dots, A_n$  are true, then:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_n) P(A_n)$$

and in general

$$P(B|X) = P(B|A_1, X)P(A_1|X) + \dots + P(B|A_n, X) P(A_n|X)$$

so

$$P(A_k | B, X) = \frac{P(A_k|X) P(B | A_k, X)}{\sum_i P(A_i|X) P(B | A_i, X)}$$

# Medical Diagnosis

A doctor knows that meningitis causes a stiff neck 50% of the time.

The doctor knows that if a person is randomly selected from the US population, there's a  $1/50,000$  chance the person will have meningitis.

The doctor knows that if a person is randomly selected from the US population, there's a 5% chance the person will have a stiff neck.

You walk into the doctor complaining of the **symptom** of a stiff neck. What's the probability that the **underlying cause** is meningitis?

# Joint Distribution

## Joint Distribution Table

- Given a set of variables A,B,C,....
- Generate a table with all the possible combinations of assignments to the variables in the rows
- For each row, list the corresponding joint probability
- For  $M$  binary variables  $\rightarrow$  size  $2^M$

<b>A</b>	<b>B</b>	<b>C</b>	<b>Prob</b>
<b>0</b>	<b>0</b>	<b>0</b>	<b>0.30</b>
<b>0</b>	<b>0</b>	<b>1</b>	<b>0.05</b>
<b>0</b>	<b>1</b>	<b>0</b>	<b>0.10</b>
<b>0</b>	<b>1</b>	<b>1</b>	<b>0.05</b>
<b>1</b>	<b>0</b>	<b>0</b>	<b>0.05</b>
<b>1</b>	<b>0</b>	<b>1</b>	<b>0.10</b>
<b>1</b>	<b>1</b>	<b>0</b>	<b>0.25</b>
<b>1</b>	<b>1</b>	<b>1</b>	<b>0.10</b>

# Using the Joint Distribution

Compute the probability  
of event  $E$ :

$$P(E) = \sum_{\substack{\text{all rows} \\ \text{containing } E}} P(\text{row})$$

$$P(A, B) = 0.25 + 0.10 = 0.35$$

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

# Inference Using the Joint Distribution

Given that event  $E_1$  occurs,  
what is the probability that  $E_2$   
occurs:

$$P(E_2 | E_1) = \frac{P(E_2, E_1)}{P(E_1)}$$

<b>A</b>	<b>B</b>	<b>C</b>	<b>Prob</b>
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



# Inference Using the Joint Distribution

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

$$= \frac{0.10}{0.05+0.05+0.10+0.10} = \frac{0.10}{0.30}$$

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

# Inference

- General view: I have some evidence (Headache) how likely is a particular conclusion (Fever)

# Generating the Joint Distribution

- Three possible ways of generating the joint distribution:
  1. Human experts
  2. Using known conditional probabilities (e.g.,  
if we know  $P(C|A,B)$ ,  $P(B|A)$ , and  $P(A)$ , we  
know  $P(A,B,C) = P(C|A,B)P(B|A)P(A) \dots$ )
  3. *Learning from data*

# Learning the Joint Distribution

Suppose that we have recorded a lot of training data:

(0,1,1)  
(1,0,1)  
(1,1,0)  
(0,0,0)  
(1,1,0).....

The entry for  $P(A,B,\sim C)$  in the table is:

A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

# of data entries with  $A=1, B=1, C=0$   
Total number of data entries

# Learning the Joint Distribution

Suppose that we have recorded a lot of training data:

(0,1,1)  
(1,0,1)  
(1,1,0)  
(0,0,0)  
(1,1,0).....

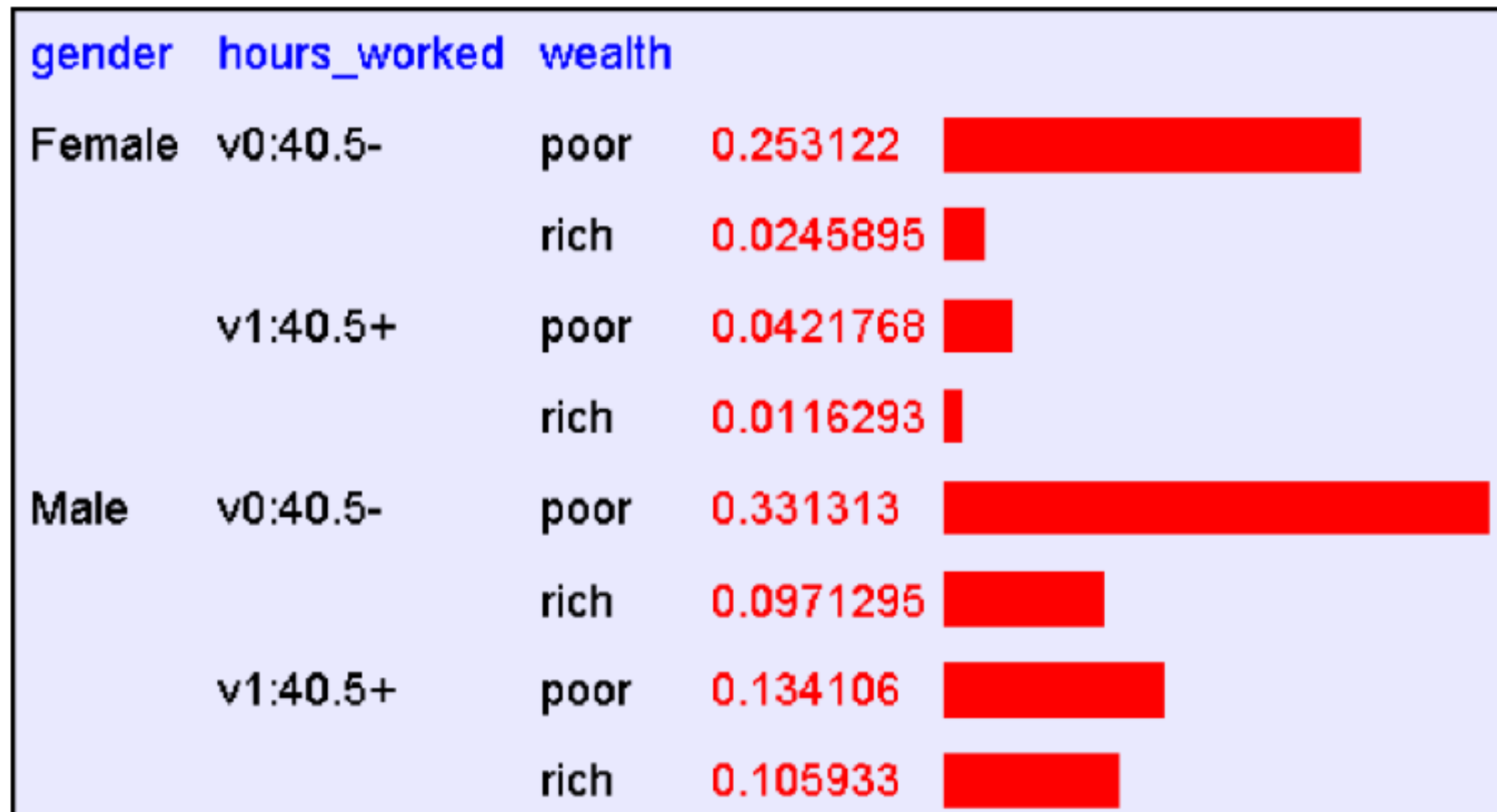
More generally, the entry for  $P(E)$  in the table is:

$$\frac{\text{\# of data entries with } E}{\text{Total number of data entries}}$$









A	B	C	Prob
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

# Real-Life Joint Distribution

- UCI Census Database



# Real-Life Joint Distribution

gender	hours_worked	wealth		
Female	v0:40.5-	poor	0.253122	
		rich	0.0245895	
	v1:40.5+	poor	0.0421768	
		rich	0.0116293	
Male	v0:40.5-	poor	0.331313	
		rich	0.0971295	
	v1:40.5+	poor	0.134106	
		rich	0.105933	

$$P(\text{Male}|\text{Poor}) = 0.4654/0.7604 = 0.612$$

## So Far ...

- Basic probability concepts
- Bayes rule
- What are joint distributions
- Inference using joint distributions
- Learning joint distributions from data
- Problem: If we have  $M$  variables, we need  $2^M$  entries in the joint distribution table → An independence assumption leads to an efficient way to learn and to do inference



# Independence

- $A$  and  $B$  are independent iff:

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

- In words: Knowing  $B$  does not affect how likely we think that  $A$  is true

# Key Properties

- Symmetry:

$$P(\mathbf{A} \mid \mathbf{B}) = P(\mathbf{A}) \Leftrightarrow P(\mathbf{B} \mid \mathbf{A}) = P(\mathbf{B})$$

- Joint distribution:

$$P(\mathbf{A}, \mathbf{B}) = P(\mathbf{A}) P(\mathbf{B})$$

- Independence of complements:

$$P(\neg \mathbf{A} \mid \mathbf{B}) = P(\neg \mathbf{A}) \qquad P(\mathbf{A} \mid \neg \mathbf{B}) = P(\mathbf{A})$$

# Independence

- Suppose that  $A$ ,  $B$ ,  $C$  are *independent*
- Then any value of the joint distribution can be computed easily:

$$P(\mathbf{A}, \mathbf{B}, \mathbf{C}) = P(\mathbf{A}) P(\mathbf{B}) P(\mathbf{C})$$

$$P(\mathbf{A}, \neg \mathbf{B}, \mathbf{C}) = P(\mathbf{A}) P(\neg \mathbf{B}) P(\mathbf{C})$$

- In fact, we need only  $M$  numbers instead of  $2^M$  for binary variables!!

# Independence: General Case

- If  $X_1, \dots, X_M$  are independent variables:

$$P(X_1 = \mathbf{x}_1, X_2 = \mathbf{x}_2, \dots, X_M = \mathbf{x}_M) = \\ P(X_1 = \mathbf{x}_1) P(X_2 = \mathbf{x}_2) \dots P(X_M = \mathbf{x}_M)$$

- Under the independence assumption, we can compute any value of the joint distribution
- We can answer any inference query.
- How do we learn the distributions?

# Learning with the Independence Assumption

$$P(X_i = x) = \frac{\text{Number of observations with } X_i = x}{\text{Total Number of observations}}$$

- Learning the distributions from data is simple and efficient
- In practice, the independence assumption may not be met but it is often a very useful approximation

# Conditional Independence

- $A$  is conditionally independent of  $B$  given  $C$  iff:

$$P(A \mid B, C) = P(A \mid C)$$

- In words: When  $C$  is present, knowing  $B$  does not affect how likely we think that  $A$  is true

# Key Properties

- Symmetry:

$$P(A \mid B, C) = P(A \mid C) \iff P(B \mid A, C) = P(B \mid C)$$

- Joint distribution:

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

- Independence of complements:

$$P(\sim A \mid B, C) = P(\sim A \mid C) \qquad P(A \mid \sim B, C) = P(A \mid C)$$

# Conditional Independence

- Suppose that  $A$ ,  $B$ ,  $C$  are **conditionally independent** *given*  $D$

- Then:

$$P(A, B, C \mid D) = P(A \mid D)P(B \mid D)P(C \mid D)$$

$$P(A, \sim B, C \mid D) = P(A \mid D)P(\sim B \mid D)P(C \mid D)$$



# Conditional Independence: General Case

- If  $X_1, \dots, X_M$  are conditionally independent given  $Y$ :

$$P(X_1 = x_1, X_2 = x_2, \dots, X_M = x_M \mid Y = y) =$$

$$P(X_1 = x_1 \mid Y = y)P(X_2 = x_2 \mid Y = y) \dots P(X_M = x_M \mid Y = y)$$

- We can learn these distributions from training data

## So Far ...

- Basic probability concepts
  - Bayes rule
  - What are joint distributions
  - Inference using joint distributions
  - Learning joint distributions from data
  - Independence assumption
  - Conditional independence assumption
- 
- Problem: We now have the joint distribution.  
How can we use it to make decision → Bayes Classifier

# Problem Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye, Height, Country) collected over population

Joint Distribution Table:

(B,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(D,T,P)
(B,T,G)	(D,T,P)
(B,S,G)	(D,S,P)
(B,S,G)	(B,S,P)
(D,S,G)	(D,S,P)

# Learn Joint Probabilities

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye, Height, Country) collected over population

(B,T,G) (B,T,P)

$$P(B,S,G) = 2/16$$

(D,T,G) (B,T,P)

$$P(B,T,G) = 2/16$$

(D,T,G) (B,T,P)

$$P(D,S,G) = 1/16$$

(D,T,G) (D,T,P)

$$P(D,T,G) = 3/16$$

(B,T,G) (D,T,P)

$$P(B,S,P) = 1/16$$

(B,S,G) (D,S,P)

$$P(B,T,P) = 3/16$$

(B,S,G) (B,S,P)

$$P(D,S,P) = 2/16$$

(D,S,G) (D,S,P)

$$P(D,T,P) = 2/16$$

# Compute other Joint or Conditional Distributions

$$P(B, S, G) = 2/16$$

$$P(B, T, G) = 2/16$$

$$P(D, S, G) = 1/16$$

$$P(D, T, G) = 3/16$$

$$P(B, S, P) = 1/16$$

$$P(B, T, P) = 3/16$$

$$P(D, S, P) = 2/16$$

$$P(D, T, P) = 2/16$$

$$P(\text{Hair} = B, \text{Height} = S | \text{Country} = G) =$$

$$\frac{P(\text{Hair} = B, \text{Height} = S, \text{Country} = G)}{P(\text{Country} = G)} =$$

$$P(\text{Country} = G)$$

$$\frac{2/16}{1/2} = 4/16$$

# Bayes Classifier Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye, Height, Country) collected over population

(B,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(B,T,P)
(D,T,G)	(D,T,P)
(B,T,G)	(D,T,P)
(B,S,G)	(D,S,P)
(B,S,G)	(B,S,P)
(D,S,G)	(D,S,P)

If I observe a new individual tall with blond hair, what is the most likely country of origin?

# Bayes Classifiers

- We want to find the value of  $Y$  that is the most probable, given the observations  $X_1, \dots, X_n$
- Find  $y$  such that this is maximum:

$$P(Y = y \mid X_1 = x_1, \dots, X_n = x_n)$$

The maximum is called the *Maximum A Posteriori (MAP)* estimator

# Bayes Classifiers

- We want to find the value of  $Y$  that is the most probable, given the observations  $X_1, \dots, X_n$
- Find  $y$  such that this is maximum:

$$\begin{aligned} P(Y = y \mid X_1 = x_1, \dots, X_n = x_n) = \\ \frac{P(X_1 = x_1, \dots, X_n = x_n \mid Y = y) P(Y = y)}{P(X_1 = x_1, \dots, X_n = x_n)} \end{aligned}$$



# Bayes Classifier

- We want to find the value of  $Y$  that is the most probable, given the observations  $X_1, \dots, X_n$
- Find  $y$  such that this is maximum:

$$P(X_1 = \mathbf{x}_1, \dots, X_n = \mathbf{x}_n | Y = y) P(Y = y)$$

# Bayes Classifier

- Learning:
  - Collect all the observations  $(x_1, \dots, x_n)$  for each class  $y$  and estimate:

$$P(X_1 = x_1, \dots, X_n = x_n | Y = y) = \frac{\text{\# observations with } (X_1 = x_1, \dots, X_n = x_n) \text{ in class } y}{\text{Total Number of observations in class } y}$$
$$P(Y = y) = \frac{\text{\# observations in class } y}{\text{Total Number of observations}}$$

- Classification:
  - Given a new input  $(x_1, \dots, x_n)$ , compute the best class:  
$$y^{best} = \arg \max_y P(X_1 = x_1, \dots, X_n = x_n | Y = y) P(Y = y)$$

# Classifier Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye, Height, Country) collected over population

(B,T,G) (B,T,P)

(D,T,G) (B,T,P)

(D,T,G) (B,T,P)

(D,T,G) (D,T,P)

(B,T,G) (D,T,P)

(B,S,G) (D,S,P)

(B,S,G) (B,S,P)

(D,S,G) (D,S,P)

$$P(B,T|G)P(G) = 2/8 \times 1/2 = 2/16$$

$$P(B,T|P)P(P) = 3/8 \times 1/2 = 3/16$$

Conclusion: Country = P

If I observe a new individual tall with blond hair, what is the most likely country of origin?

# Classifier Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye, Height, Country) collected over population

If I observe a new individual tall with blond hair, what is the most likely country of origin?

(B,T,G)	(B,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(B,T,P)
(D,T,G)	(D,T,G)	(D,T,P)
(B,T,G)	(B,T,G)	(D,T,P)
(B,S,G)	(B,S,G)	(D,S,P)
(B,S,G)	(B,S,G)	(B,S,P)
(D,S,G)	(D,S,G)	(D,S,P)

$$P(B,T|G)P(G) = 2/8 \times 2/3 = 4/24$$

$$P(B,T|P)P(P) = 3/8 \times 1/3 = 3/24$$

Conclusion: Country = G

# Naïve Bayes Assumption

To make the problem tractable, we often need to make the following **conditional independence** assumption:

$$\begin{aligned} P(x_1, x_2, \dots, x_n \mid y) &= P(x_1 \mid y)P(x_2 \mid y) \dots P(x_n \mid y) \\ &= \prod_i P(x_i \mid y) \end{aligned}$$

which allows us to define the **Naïve Bayes Classifier**:

$$y_{NB} = \arg \max_{y \in C} P(y) \prod_i P(x_i \mid y)$$

# Naïve Bayes Classifier

- Learning:
  - Collect all the observations  $(x_1, \dots, x_n)$  for each class  $y$  and estimate:

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

$$\frac{\text{Number of observations with } X_i = x_i \text{ in class } y}{\text{Total Number of observations in class } y}$$

$$P(Y = y) = \frac{\text{Number of observations in class } y}{\text{Total Number of observations}}$$

- Classification:

$$y^{best} =$$

$$\arg \max_y P(X_1 = x_1 | Y = y) \dots P(X_n = x_n | Y = y) P(Y = y)$$

# Naïve Bayes Classifier

- Learning:
  - Collect all the observations  $(x_1, \dots, x_n)$  for each class  $y$  and estimate:

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

$$\frac{\text{Number of observations with } X_i = x_i \text{ in class } y}{\text{Total Number of observations in class } y}$$

$P(Y = y)$  Note that we need only  $k \times n$  numbers ( $P(X_i = x_i | Y = y)$ ) to implement this classifier, instead of  $k^n$  if we were to use the full model, without independence assumptions.

- Classification

$$y^{best} =$$

$$\arg \max_y P(X_1 = x_1 | Y = y) \dots P(X_n = x_n | Y = y) P(Y = y)$$

# Naïve Bayes Implementation

- Small (but important) implementation detail: If  $n$  is large, we may be taking the product of a large number of small floating-point values. Underflow avoided by taking log.
- Take the max of:

$$\log P(X_1 = x_1 | Y = y) + \dots \\ + \log P(X_n = x_n | Y = y) + \log P(Y = y)$$

- Instead of:

$$P(X_1 = x_1 | Y = y) \dots P(X_n = x_n | Y = y) P(Y = y)$$



# Same Example, the Naïve Bayes Way

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {Gromland, Polvia}
- Training data: Values of (Eye, Height, Country) collected over population

(B,T,G)	(B,T,G)	(B,T,P)	$P(B,T G)P(G) \approx P(B G)P(T G)P(G)$
(D,T,G)	(D,T,G)	(B,T,P)	$8/16 \times 10/16 \times 2/3 \approx 160/768 = 40/192$
(D,T,G)	(D,T,G)	(B,T,P)	
(D,T,G)	(D,T,G)	(D,T,P)	
(B,T,G)	(B,T,G)	(D,T,P)	$P(B,T P)P(P) \approx 4/8 \times 5/8 \times 1/3 = 20/192$
(B,S,G)	(B,S,G)	(D,S,P)	
(B,S,G)	(B,S,G)	(B,S,P)	
(D,S,G)	(D,S,G)	(D,S,P)	Conclusion: Country = G

# Same Example, the Naïve Bayes Way

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {Gromland, Polvi
- Training data: Values of (Eye, population

The values are of course different, but the conclusion remains the same  
 0.17 vs. 0.2 for Country = G  
 0.125 vs. 0.1 for Country = P

(B,T,G)	(B,T,G)
(D,T,G)	(D,T,G)
(D,T,G)	(D,T,G)
(D,T,G)	(D,T,G)
(B,T,G)	(B,T,G)
(B,S,G)	(B,S,G)
(B,S,G)	(B,S,G)
(D,S,G)	(D,S,G)

The variables are not independent so it is only an approximation.

$P(B,T|G)P(G) \approx 16/768 = 40/192$   
 $P(B,T|P)P(P) \approx 4/8 \times 5/8 \times 1/3 = 20/192$   
 Conclusion: **Country = G**

# Naïve Bayes: Subtleties

Conditional independence assumption is often violated

$$P(x_1, x_2, \dots, x_n \mid y) = \prod_i P(x_i \mid y)$$

... but it works surprisingly well anyway.

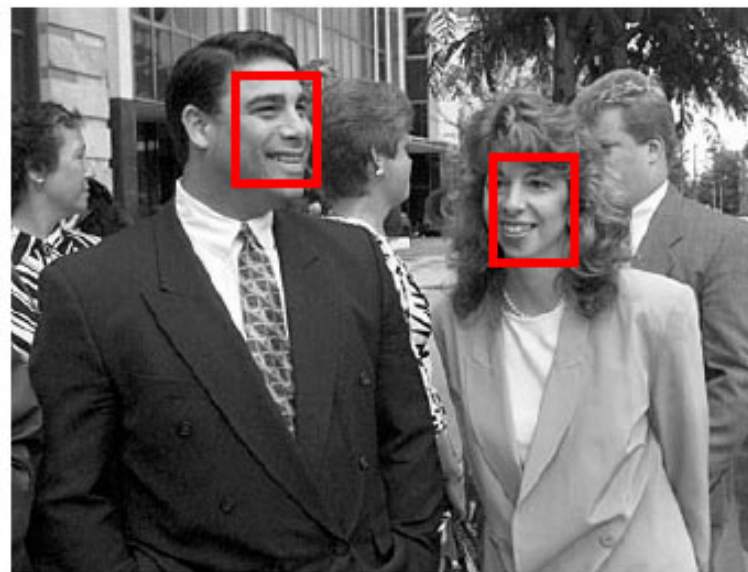
A plausible reason is that to make correct predictions,

- Don't need the probabilities to be estimated correctly
- Only need the posterior of the correct class to be largest among the class posteriors

# Naïve Bayes at Work: Face Detection



Input Image



Find the faces (quickly)

Approach:

- Model the likelihood of an image window assuming face/non-face
- Use independence assumption along the way to make computations tractable

# Naïve Bayes at Work

Move a window over an input image

At every position of the window:



1. Compute the values  $x_1, \dots, x_n$  of a bunch of features  $X_1, \dots, X_n$  from the image content within the window

2. Retrieve the probabilities:

$$P(X_i = x_i \mid \text{Face}), P(X_i = x_i \mid \neg \text{Face}) \quad i = 1, \dots, n$$

from tables learned off-line

3. Assuming independence, compute:

$$(1) P(\text{Face}) P(X_1 = x_1 \mid \text{Face}) \dots P(X_n = x_n \mid \text{Face})$$

$$(2) P(\neg \text{Face}) P(X_1 = x_1 \mid \neg \text{Face}) \dots P(X_n = x_n \mid \neg \text{Face})$$

4. Classify the window as a face if  $(1) > (2)$

# Learning

- Collect the values of the features for training data in tables that approximate the probabilities



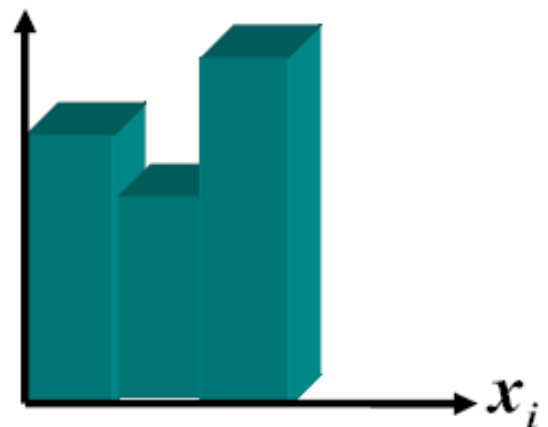
Face examples:  
50-2,000 original images

$$P(X_i = x_i \mid \text{Face})$$



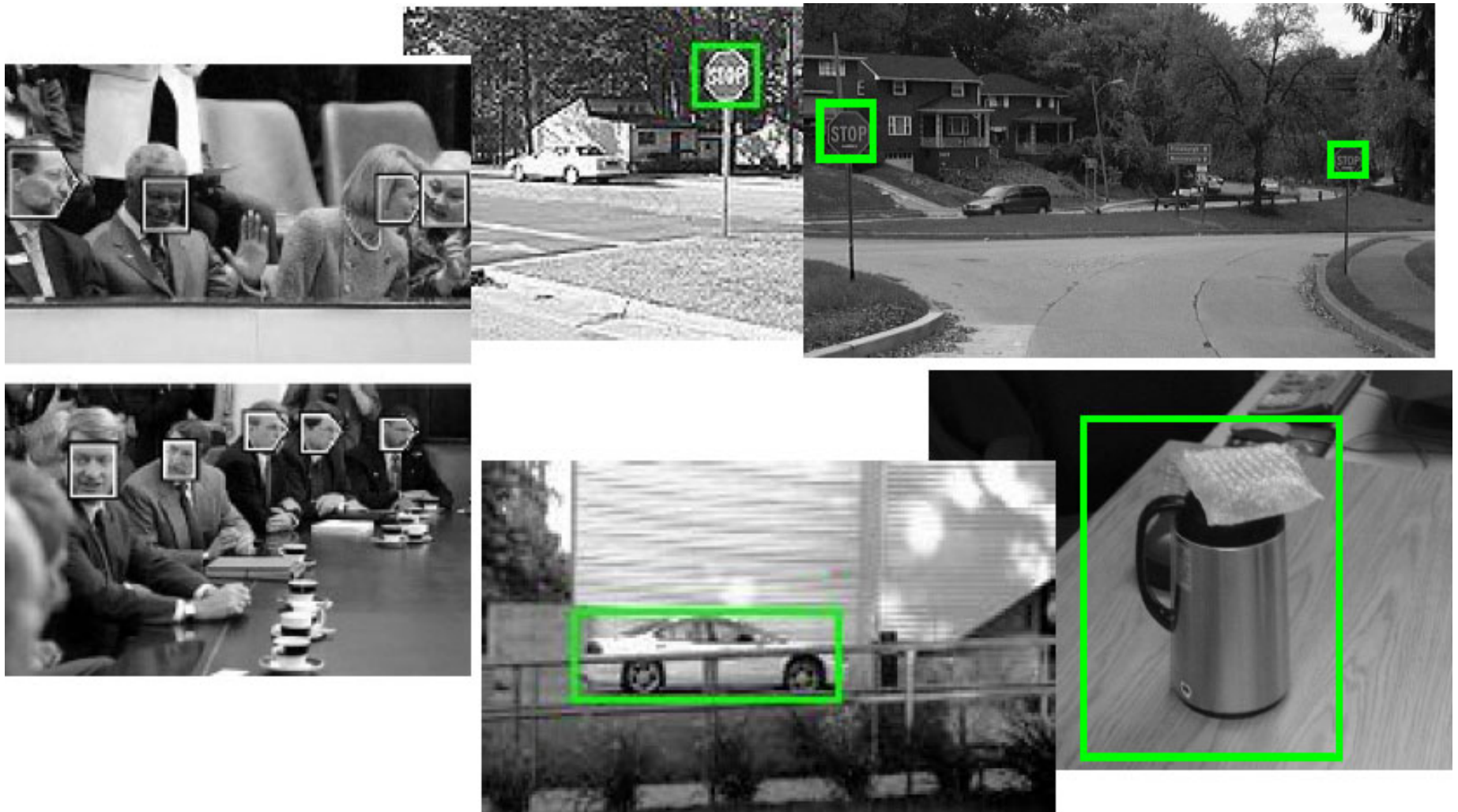
Non-face examples:  
~10,000,000 examples

$$P(X_i = x_i \mid \neg \text{Face})$$





- Yes it works. And in real-time. And with other objects than faces also....



# Generative vs. Discriminative Models

Given training examples  $(x_1, y_1), \dots, (x_n, y_n)$ ,

## **Discriminative Models**

Select hypothesis space  $H$  to consider

Find  $h$  from  $H$  with lowest training error

Argument: low training error leads to low prediction error

Examples: decision trees, perceptrons, SVMs

## **Generative Models**

Select set of distributions to consider for modeling  $P(X, Y)$

Find distribution that best matches  $P(X, Y)$  on training data

Argument: If match is close enough, we can use Bayes decision rule

Examples: naïve Bayes, HMMs



# Hypothesis Selection: An Example

I have three identical boxes labeled H1, H2 and H3.

Into H1 I place 1 black bead, 3 white beads.

Into H2 I place 2 black beads, 2 white beads.

Into H3 I place 4 black beads, no white beads.

I draw a box at random. I remove a bead at random from that box, take note of its color, and put it back into the box. I repeat this process to generate a sequence of colors. Which box is most likely to have yielded this color sequence?

# Bayesian Ranting

A nice way to look at this:

- H1, H2 and H3 were my prior models of the world.
- The fact that  $P(H1) = 1/3$ ,  $P(H2) = 1/3$ ,  $P(H3) = 1/3$  was my **prior distribution**.
- The color of the bead was a piece of **evidence** about the true model of the world.
- The use of bayes' rule was a piece of probabilistic inference, giving me a **posterior distribution** on possible worlds.
- Learning is prior + evidence ---> posterior
- A piece of evidence decreases my ignorance about the world.

# Bayesian Methods for Hypothesis Selection

**Goal:** find the best hypothesis from some space  $H$  of hypotheses, **given** the observed data  $D$ .

Define best to be: most probable hypothesis in  $H$

In order to do that, we need to assume a probability distribution **over the class  $H$** .

In addition, we need to know something about the relation between the data observed and the hypotheses (E.g., a bead problem.)

# Notations

$P(h)$  - the prior probability of a hypothesis  $h$

Reflects background knowledge; before data is observed.  
If no information - uniform distribution.

$P(D)$  - The probability that this sample of the Data is observed. (No knowledge of the hypothesis)

$P(D|h)$ : The probability of observing the sample  $D$ , given that the hypothesis  $h$  holds

$P(h|D)$ : The posterior probability of  $h$ . The probability  $h$  holds, given that  $D$  has been observed.

# Bayes Theorem

$$P(\mathbf{h} \mid \mathbf{D}) = P(\mathbf{D} \mid \mathbf{h}) \frac{P(\mathbf{h})}{P(\mathbf{D})}$$

$P(h|D)$  increases with  $P(h)$  and with  $P(D|h)$

$P(h|D)$  decreases with  $P(D)$

# Learning Scenario

$$\mathbf{P(h \mid D)} = \mathbf{P(D \mid h) P(h) / P(D)}$$

The learner considers a set of candidate hypotheses  $H$  (**models**), and attempts to find the most probable one  $h \in H$ , given the observed data.

Such maximally probable hypothesis is called maximum a posteriori hypothesis (MAP); Bayes theorem is used to compute it:

$$\begin{aligned} \mathbf{h_{MAP}} &= \mathbf{argmax_{h \in H} P(h \mid D)} = \mathbf{argmax_{h \in H} P(D \mid h) P(h) / P(D)} \\ &= \mathbf{argmax_{h \in H} P(D \mid h) P(h)} \end{aligned}$$

## Learning Scenario (2)

$$\mathbf{h}_{\text{MAP}} = \operatorname{argmax}_{\mathbf{h} \in H} \mathbf{P}(\mathbf{h} \mid \mathbf{D}) = \operatorname{argmax}_{\mathbf{h} \in H} \mathbf{P}(\mathbf{D} \mid \mathbf{h})\mathbf{P}(\mathbf{h})$$

We may assume that a priori, hypotheses are equally probable

$$\mathbf{P}(\mathbf{h}_i) = \mathbf{P}(\mathbf{h}_j), \forall \mathbf{h}_i, \mathbf{h}_j \in H$$

We get the **Maximum Likelihood hypothesis**:

$$\mathbf{h}_{\text{ML}} = \operatorname{argmax}_{\mathbf{h} \in H} \mathbf{P}(\mathbf{D} \mid \mathbf{h})$$

Here we just look for the hypothesis that best explains the data

## Example

$$\mathbf{h}_{\text{MAP}} = \operatorname{argmax}_{\mathbf{h} \in \mathbf{H}} \mathbf{P}(\mathbf{h} \mid \mathbf{D}) = \operatorname{argmax}_{\mathbf{h} \in \mathbf{H}} \mathbf{P}(\mathbf{D} \mid \mathbf{h})\mathbf{P}(\mathbf{h})$$

A given coin is either **fair** or has a **60%** bias in favor of Head.  
**Decide** what is the bias of the coin



# Notations

$P(h)$  - the prior probability of a hypothesis  $h$

Reflects background knowledge; before data is observed.  
If no information - uniform distribution.

$P(D)$  - The probability that this sample of the Data is observed. (No knowledge of the hypothesis)

$P(D|h)$ : The probability of observing the sample  $D$ , given that the hypothesis  $h$  holds

$P(h|D)$ : The posterior probability of  $h$ . The probability  $h$  holds, given that  $D$  has been observed.

## Example

$$\mathbf{h}_{\text{MAP}} = \operatorname{argmax}_{\mathbf{h} \in \mathbf{H}} \mathbf{P}(\mathbf{h} \mid \mathbf{D}) = \operatorname{argmax}_{\mathbf{h} \in \mathbf{H}} \mathbf{P}(\mathbf{D} \mid \mathbf{h})\mathbf{P}(\mathbf{h})$$

A given coin is either **fair** or has a **60%** bias in favor of Head.  
**Decide** what is the bias of the coin

Two hypotheses:  **$h_1$** :  $P(H)=0.5$ ;  **$h_2$** :  $P(H)=0.6$

**Prior:  $P(h)$** :  $P(h_1)=0.75$   $P(h_2)=0.25$

Now we **need Data**. 1<sup>st</sup> Experiment. coin toss is H.

**$P(D|h)$** :  $P(D|h_1)=0.5$  ;  $P(D|h_2)=0.6$

**$P(D)$** :  $P(D)=P(D|h_1)P(h_1)+P(D|h_2)P(h_2)=0.5 \bullet 0.75 + 0.6 \bullet 0.25 = 0.525$

**$P(h|D)$** :

$P(h_1|D)=P(D|h_1)P(h_1)/P(D)=0.5 \bullet 0.75 / 0.525 = 0.714$

$P(h_2|D)=P(D|h_2)P(h_2)/P(D)=0.6 \bullet 0.25 / 0.525 = 0.286$

## Example (2)

$$\mathbf{h}_{\text{MAP}} = \operatorname{argmax}_{\mathbf{h} \in \mathbf{H}} \mathbf{P}(\mathbf{h} \mid \mathbf{D}) = \operatorname{argmax}_{\mathbf{h} \in \mathbf{H}} \mathbf{P}(\mathbf{D} \mid \mathbf{h})\mathbf{P}(\mathbf{h})$$

A given coin is either **fair** or has a **60%** bias in favor of Head.

**Decide** what is the bias of the coin

Two hypotheses:  $h_1$ :  $P(H)=0.5$ ;  $h_2$ :  $P(H)=0.6$

**Prior:  $P(h)$ :**  $P(h_1)=0.75$

After 1<sup>st</sup> coin toss is H we still think that the coin is more likely to be fair

If we were to use **Maximum Likelihood** approach (i.e., assume equal priors) we would think otherwise. The data supports the biased coin better.

Try: 100 coin tosses; 70 heads.

Now you will believe that the coins is biased.

## Example (2)

$$\mathbf{h}_{\text{MAP}} = \operatorname{argmax}_{\mathbf{h} \in \mathbf{H}} \mathbf{P}(\mathbf{h} \mid \mathbf{D}) = \operatorname{argmax}_{\mathbf{h} \in \mathbf{H}} \mathbf{P}(\mathbf{D} \mid \mathbf{h})\mathbf{P}(\mathbf{h})$$

A given coin is either **fair** or has a **60%** bias in favor of Head.

**Decide** what is the bias of the coin

Two hypotheses:  **$h_1$** :  $P(H)=0.5$ ;  **$h_2$** :  $P(H)=0.6$

**Prior:  $P(h)$** :  $P(h_1)=0.75$

Case of 100 coin tosses; 70 heads.

$$\begin{aligned} P(\mathbf{D}) &= P(\mathbf{D} \mid \mathbf{h}_1)P(\mathbf{h}_1) + P(\mathbf{D} \mid \mathbf{h}_2)P(\mathbf{h}_2) \\ &= 0.5^{100} \bullet 0.75 + 0.6^{70} 0.4^{40} \bullet 0.25 = \\ &= 7.9 \bullet 10^{-31} \bullet 0.75 + 3.4 \bullet 10^{-28} \bullet 0.25 \end{aligned}$$

$$P(\mathbf{h}_1 \mid \mathbf{D}) = P(\mathbf{D} \mid \mathbf{h}_1) \frac{P(\mathbf{h}_1)}{P(\mathbf{D})} \ll P(\mathbf{D} \mid \mathbf{h}_2) \frac{P(\mathbf{h}_2)}{P(\mathbf{D})} = P(\mathbf{h}_2 \mid \mathbf{D})$$

**0.0057**

**0.9943**

# Summary

- Basic probability concepts
- Bayes rule
- What are joint distributions
- Inference using joint distributions
- Learning joint distributions from data
- Independence
- Conditional independence
- Bayes classifiers
- Naïve Bayes approach
- Selecting hypothesis using Bayesian methods