

# UVA CS 6316: Machine Learning

## Lecture 17: Generative Bayes Classifiers

Dr. Yanjun Qi  
University of Virginia  
Department of Computer Science

# Course Content Plan →

Six major sections of this course

~~Regression (supervised)~~

Y is a continuous

Classification (supervised)

Y is a discrete

Unsupervised models

NO Y

Learning theory

About  $f()$

Graphical models

About interactions among  $X_1, \dots, X_p$

Reinforcement Learning

Learn program to Interact with its environment

# Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

## 1. Discriminative

directly estimate a decision rule/boundary

e.g., support vector machine, decision tree, logistic regression,

e.g. neural networks (NN), deep NN



## 2. Generative:

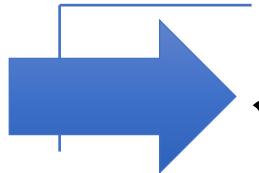
build a generative statistical model

e.g., Bayesian networks, Naïve Bayes classifier

## ~~3. Instance based classifiers~~

- Use observation directly (no models)
- e.g. K nearest neighbors

# Today : Generative Bayes Classifiers

- 
- ✓ Bayes Classifier (BC)
    - Generative Bayes Classifier
  - ✓ Naïve Bayes Classifier
  - ✓ Gaussian Bayes Classifiers
    - Gaussian distribution
    - Naïve Gaussian BC
    - Not-naïve Gaussian BC → LDA, QDA

# Review: Bayes classifiers (BC)

- Treat each feature attribute and the class label as random variables.
- **Testing:** Given a sample  $\mathbf{x}$  with attributes ( $x_1, x_2, \dots, x_p$ ):
  - Goal is to predict its class  $c$ .
  - Specifically, we want to **find the class that maximizes**  $p(c | x_1, x_2, \dots, x_p)$ .
- **Training:** can we estimate  $p(C_i | \mathbf{x}) = p(C_i | x_1, x_2, \dots, x_p)$  directly from data?

$$c_{MAP} = \operatorname{argmax}_{c_j \in C} P(c_j | x_1, x_2, \dots, x_p)$$

MAP Rule

# Two kinds of Bayes classifiers via MAP classification rule

- Establishing a probabilistic model for classification
  - **(1) Discriminative**
  - **(2) Generative**

# Review: Discriminative BC

## - (1) Discriminative model

Softmax

$$P(\hat{y}_i | \vec{x}) = \frac{e^{z_i}}{\sum_{j=1}^L e^{z_j}}$$

$$\arg \max_{c \in C} P(c | \mathbf{X}), \quad C = \{c_1, \dots, c_L\}$$

$$P(c_1 | \mathbf{x}) \quad P(c_2 | \mathbf{x}) \quad \dots \quad P(c_L | \mathbf{x})$$

Discriminative  
Probabilistic Classifier

$$x_1 \quad x_2 \quad \dots \quad x_p$$

$$\mathbf{x} = (x_1, x_2, \dots, x_p)$$

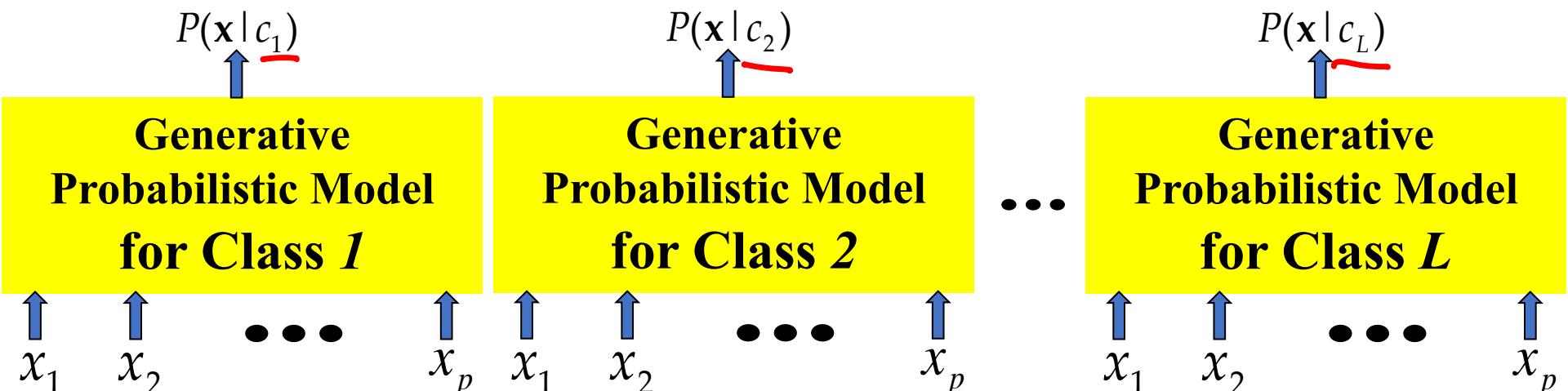
multi-class

Logistic  
regression

## (2) Generative BC

$$P(\mathbf{X}|\mathcal{C}), \quad \mathcal{C} = c_1, \dots, c_L, \mathbf{X} = (X_1, \dots, X_p)$$

$P(c_i|\vec{x})$



$$\mathbf{x} = (x_1, x_2, \dots, x_p)$$

Review Probability: If hard to directly estimate from data, most likely we can estimate

- 1. Joint probability
  - Use Chain Rule
- 2. Marginal probability
  - Use the total law of probability
- 3. Conditional probability
  - Use the Bayes Rule

$$p(c_i | \vec{x}) = \frac{p(\vec{x} | c_i) p(c_i)}{p(\vec{x})}$$
$$c^* = \operatorname{argmax}_{\{c_i\}} p(c_i | \vec{x})$$
$$\{c_1, c_2, \dots, c_L\}$$

# Review : Bayes' Rule – for Generative Bayes Classifiers

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

$P(C_1|x), P(C_2|x), \dots, P(C_L|x)$

$P(C_1), P(C_2), \dots, P(C_L)$

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

# Review : Bayes' Rule – for Generative Bayes Classifiers

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

Prior

$$P(C_1|x), P(C_2|x), \dots, P(C_L|x)$$

$$P(C_1), P(C_2), \dots, P(C_L)$$

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

# Summary of Generative BC:

- Apply Bayes rule to get posterior probabilities

$$\underset{c_i}{\operatorname{argmax}} \left\{ P(C = c_i | X = x) \right\} = \frac{P(X = x | C = c_i) P(C = c_i)}{P(X = x)}$$
$$\propto P(X = x | C = c_i) P(C = c_i)$$

for  $i = 1, 2, \dots, L$

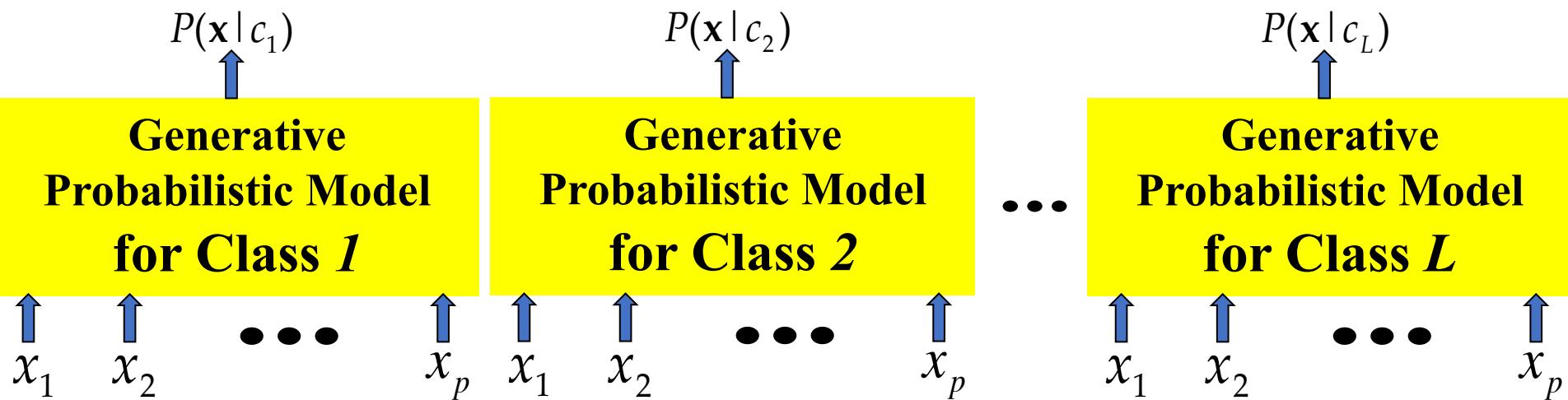
- Then apply the MAP rule

$$\left\{ \begin{array}{l} p(\vec{x} | c_i) \\ p(c_i) \end{array} \right. \quad c_i \in \{c_1, c_2, \dots, c_L\}$$

# Establishing a probabilistic model for classification through generative probabilistic models

[MAP rule]

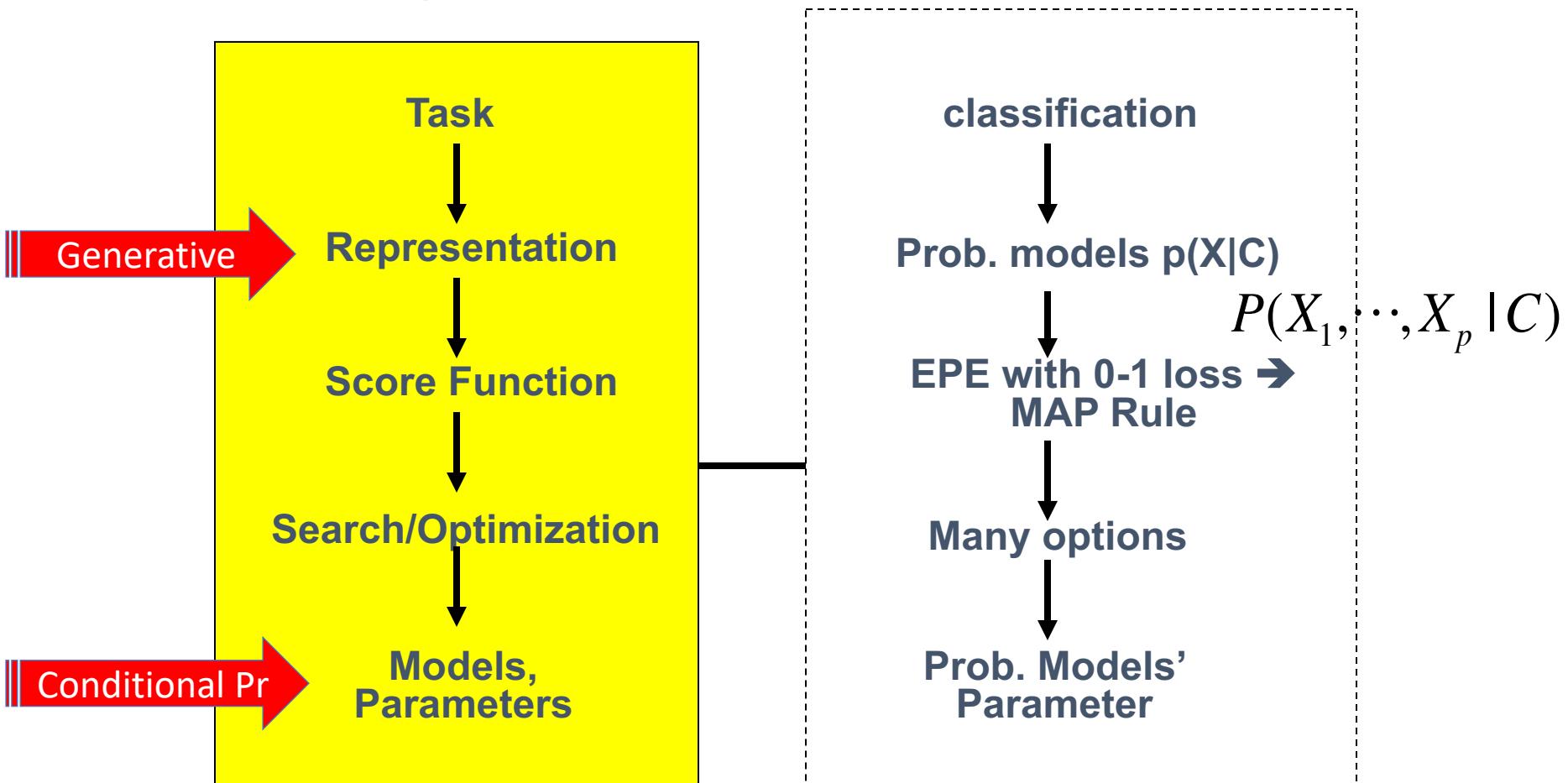
$$\operatorname{argmax}_{C_i} P(C_i | X) = \operatorname{argmax}_{C_i} P(X, C_i) = \operatorname{argmax}_{C_i} P(X | C_i)P(C_i)$$



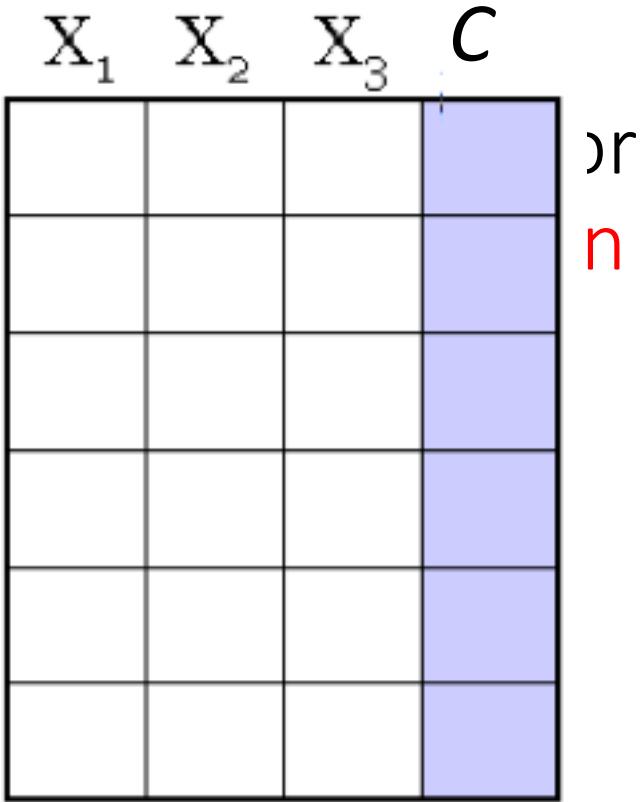
$$\mathbf{x} = (x_1, x_2, \dots, x_p)$$

Adapt from Prof. Ke Chen NB slides

# Generative Bayes Classifier



$$\operatorname{argmax}_k P(C_k | X) = \operatorname{argmax}_k P(X, C_k) = \operatorname{argmax}_k P(X | C_k)P(C_k)$$



or  
n

$$f : X \longrightarrow C$$

Output as Discrete  
Class Label  
 $C_1, C_2, \dots, C_L$

## Discriminative

$$\arg \max_{c \in \mathcal{C}} P(c | \mathbf{X}) \quad \mathcal{C} = \{c_1, \dots, c_L\}$$

# Generative

$$\operatorname{argmax}_{c \in \mathcal{C}} P(c | X) = \operatorname{argmax}_{c \in \mathcal{C}} P(X, c) = \operatorname{argmax}_{c \in \mathcal{C}} P(X | c)P(c)$$

## this lecture!

# An Example

- Example: Play Tennis

*PlayTennis: training examples*

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny ✓	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain ✓✓	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$\Sigma_1$	$\Sigma_2$	$\Sigma_3$	$\Sigma_4$	C
S	H	H	W	$C_1 = \text{Yes}$
O	M	N	S	$C_2 = \text{No}$
R	C			
3	3	2	2	$\Rightarrow 72 \text{ parameters}$

$\Rightarrow P(\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4 | C), P(C)$  training

$$\textcircled{1} \quad \begin{cases} P(C = \text{Yes}) = \frac{N(\text{Yes})}{N(\text{train})} = \frac{9}{14} \\ P(C = \text{No}) = 1 - P(C = \text{Yes}) = \frac{5}{14} \end{cases}$$

$$\textcircled{2} \quad \begin{aligned} P(\Sigma_1=S, \Sigma_2=H, \Sigma_3=H, \Sigma_4=W | C=\text{No}) &= \frac{N(SHHWN\text{No})}{N(\text{No})} = \frac{1}{5} \\ P(\Sigma_1=S, \Sigma_2=H, \Sigma_3=H, \Sigma_4=W | C=\text{Yes}) &= \frac{0}{9} = 0 \end{aligned}$$

X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	C

# Learning Phase:

- Example: Play Tennis

$\vec{x}_1 \downarrow$   $\vec{x}_2 \uparrow$   $\vec{x}_3 \uparrow$   $\vec{x}_4 \uparrow$   
*PlayTennis: training examples*

C

$$k_2 = 3$$

$\vec{x}_2$ :

{Hot,  
Mild,  
Cool}

$\vec{x}_3 = \{High\}$   
(Normal)

$$k_3 = 2$$

$\vec{x}_4 = (W, S)$   
 $k_4 =$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

C: {Yes,  
No}

$\vec{x}_1$ : {Sunny,  
Overc,  
Rain}  
 $k_1 = 3$

- Learning: maximum likelihood estimates
  - simply use the frequencies in the data

e.g.  $p(\text{Overcast, hot, high, weak} \mid \text{Yes}) = \frac{1}{9}$

*PlayTennis*: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Directly  
estimate from  
data

e.g.  
 $p(\text{overcast, hot, high, weak} \mid \text{No}) = \frac{0}{9}$

Check L13-MLE  
Lecture for  
Why

$$\frac{P(X_1, X_2, X_3, X_4 \mid \text{Yes})}{P(X_1, X_2, X_3, X_4 \mid \text{No})} \rightarrow 36 \times 2$$

3

## *PlayTennis*: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny ✓	Hot	High	Weak	No
D2	Sunny ✓	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain /	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny ✓	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$P(C \geq c_i)$$

{Yes}  
{No}

$$P(C = \text{Yes}) = \frac{9}{14}$$

$$P(C = N_0)$$

$$= \frac{5}{14}$$

$$3 \times 3 \times 2 \times 2 = 36$$

# Generative Bayes Classifier:

e.g English Dictionary  $\mathcal{D}_1, \dots, \mathcal{D}_P$

- Learning Phase

$$P(C_1), P(C_2), \dots, P(C_L)$$

$$\varphi(c_1)$$

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$\Rightarrow C(\text{up}, \text{down})$$

$$\varphi(c_2)$$

$$P(\text{Play}=\text{No}) = 5/14$$

$$P \sim 30k \rightarrow 2^{P \times L}$$

$$P(X_1, X_2, \dots, X_p | C_1), P(X_1, X_2, \dots, \underline{X_p} | C_2)$$

a look up table of cond. prob

Outlook (3 values)	Temperature (3 values)	Humidity (2 values)	Wind (2 values)	Play=Yes	Play>No
sunny	hot	high	weak	0/9	1/5
sunny	hot	high	strong	.../9	.../5
sunny	hot	normal	weak	.../9	.../5
sunny	hot	normal	strong	.../9	.../5
....	....	....	....	....	....
....	....	....	....	....	....
....	....	....	....	....	....
....	....	....	....	....	....

$$3^3 \cdot 3^2 \cdot 2^2 \cdot 2 \cdot 2 \cdot 2 = 72 \text{ parameters}$$

# Generative Bayes Classifier:

- **Testing Phase**

- – Given an unknown instance  $\mathbf{X}'_{ts} = (a'_1, \dots, a'_p)$ 
  - Look up tables to assign the label  $c^*$  to  $\mathbf{x}_{ts}$  if

Last  
Page

$$\hat{P}(a'_1, \dots, a'_p | c^*) \hat{P}(c^*) > \hat{P}(a'_1, \dots, a'_p | c) \hat{P}(c),$$
$$c \neq c^*, c = c_1, \dots, c_L$$

- Given a new instance,

$\mathbf{x}' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

$$\left\{ \begin{array}{l} \hat{P}(x' | \text{Yes}) \hat{P}(c=\text{Yes}) \\ \hat{P}(x' | \text{No}) \hat{P}(c=\text{No}) \end{array} \right\} \Rightarrow \arg \max_c \hat{P}(x' | c) \hat{P}(c) \Rightarrow \text{predicted } c^*$$

$$P(C=\text{Yes} \mid X_1, X_2, X_3, X_4)$$

$$P(C=\text{No} \mid X_1, X_2, X_3, X_4)$$

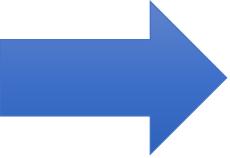
$$\rightarrow P(C=\text{Yes}) = 9/14$$

$$P(C=\text{No}) = 5/14$$

$$\rightarrow P(X_1, X_2, X_3, X_4 \mid C_i) \cdot \underbrace{72}_{\text{from train}} \xrightarrow{\text{parameter}}$$

$$\underset{i=1,2}{\operatorname{argmax}} \quad P(\bar{X}_{ts} \mid C_i) P(C_i) \quad \text{Generative BC}$$

# Today Recap: Generative Bayes Classifiers

- 
- ✓ Bayes Classifier
    - Generative Bayes Classifier
  - ✓ Naïve Bayes Classifier
  - ✓ Gaussian Bayes Classifiers
    - Gaussian distribution
    - Naïve Gaussian BC
    - Not-naïve Gaussian BC → LDA, QDA

# Naïve Bayes Classifier

- Bayes classification

$$\operatorname{argmax}_{c_j \in C} P(x_1, x_2, \dots, x_p | c_j) P(c_j)$$

$2 \times 2 \times \dots \times 2 = 2^P \times L$

Difficulty: learning the joint probability

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!  
*(when given  $c_i$ )*

# Naïve Bayes Classifier

- Bayes classification

$$\underset{c_j \in C}{\operatorname{argmax}} P(x_1, x_2, \dots, x_p | c_j) P(c_j)$$

$$= p(x_1 | c_j) p(x_2 | c_j) \cdot \dots \cdot p(x_p | c_j)$$

Difficulty: learning the joint probability

- Naïve Bayes classification

– Assumption that all input attributes are conditionally independent!  
given C variable

# Naïve Bayes Classifier

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!

[assumption]

$$P(X_1, X_2, \dots, X_p | C) = P(X_1 | C)P(X_2 | C) \cdots P(X_p | C)$$

$$\begin{aligned} P(C_j | X_1, \dots, X_p) &\propto P(X_1, X_2, \dots, X_p | C_j) P(C_j) \\ &= \underbrace{P(X_1 | C_j)}_{\text{1}} \underbrace{P(X_2 | C_j)}_{\text{2}} \cdots \underbrace{P(X_p | C_j)}_{\text{3}} P(C_j) \end{aligned}$$

$3 \times L$	$P(X_1   C)$
$X_1 = \text{Sum}$	$C = \text{Yes}$
$X_1 = 0$	$C = \text{Yes}$
$X_1 = R$	$C = \text{Yes}$
$X_1 = S$	$C = \text{No}$
0	$C = \text{No}$
R	$C = \text{No}$

$$X_1 : 3 \times L$$

$$X_2 : 3 \times L$$

$$X_3 : 2 \times L$$

$$X_4 : 2 \times L$$

$L=2$	$P(C=\text{Ye})$	$P(\text{EN})$
	$9/14$	$5/14$

No Naive

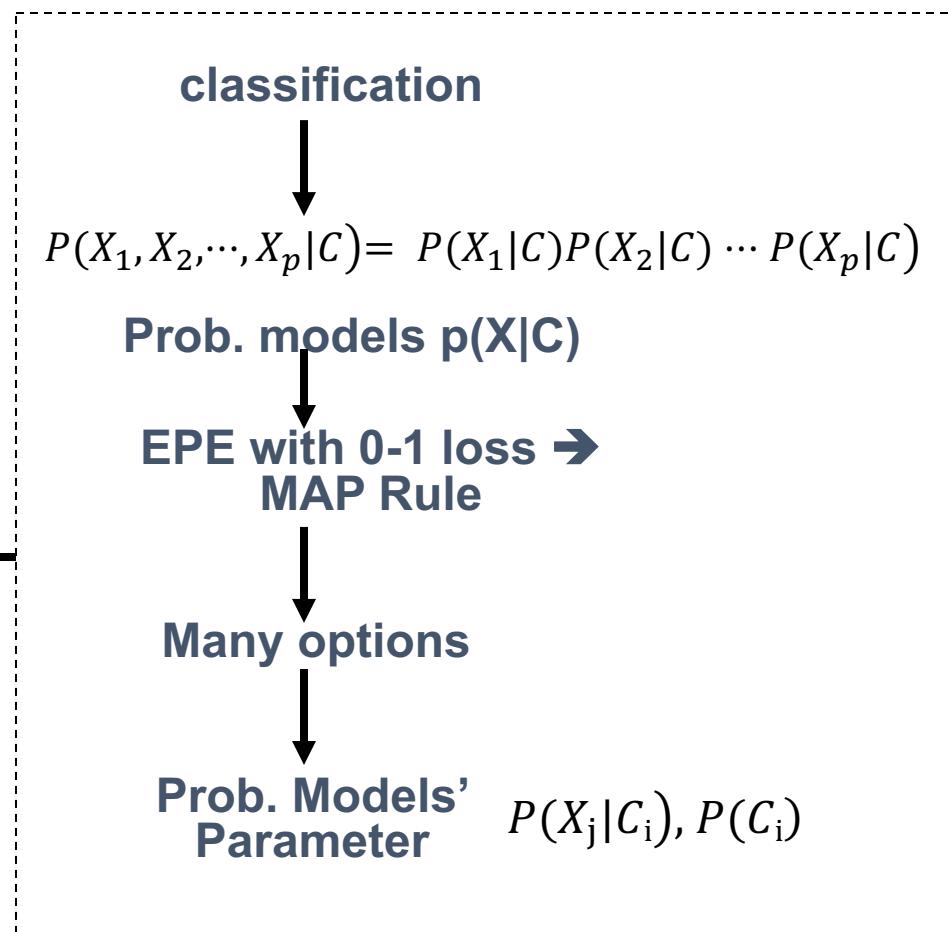
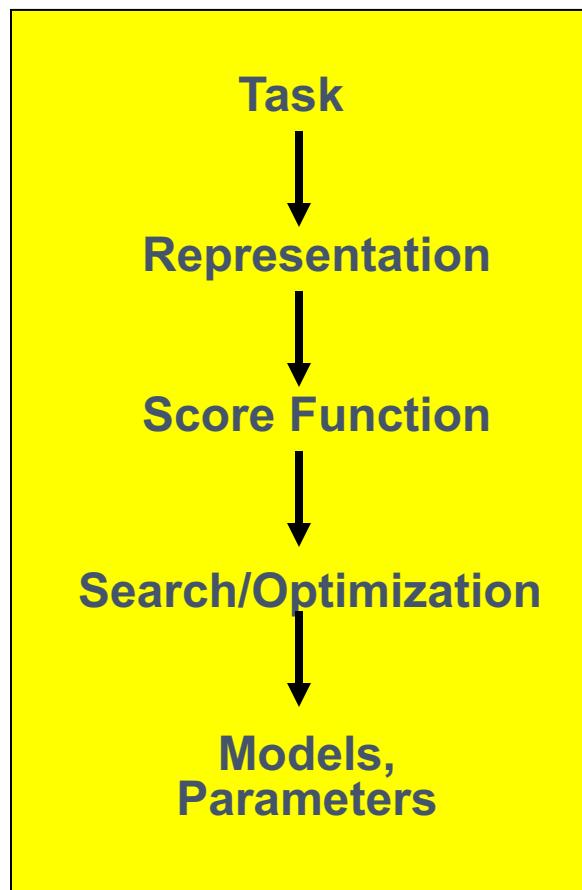
$$72 + 2 = 74 \text{ para}$$

$$2^P \cdot L$$

$$(3 + 3 + 2 + 2) \times L = 20$$

$$\Rightarrow X_1 + X_2 + X_P \Rightarrow 2^P \times L \quad P \approx 30k$$

# Naive Bayes Classifier



Estimate  $P(X_j = x_{jk} | C = c_i)$  with examples in training;

$$P(X_2|C_1), P(X_2|C_2)$$

Outlook	Play=Yes	Play=No
Sunny		
Overcast		
Rain		

Temperature	Play=Yes	Play=No
Hot		
Mild		
Cool		

Humidity	Play=Yes	Play=No
High		
Normal		

$$P(X_4|C_1), P(X_4|C_2)$$

Wind	Play=Yes	Play=No
Strong		
Weak		

$$P(\text{Play}=Yes) = ??$$

$$P(\text{Play}=No) = ??$$

$$P(C_1), P(C_2), \dots, P(C_L)$$

# Naïve Bayes Classifier

- Naïve Bayes classification
  - Assumption that **all input attributes are conditionally independent!**

$$P(X_1, X_2, \dots, X_p | C) = P(X_1 | C)P(X_2 | C)\cdots P(X_p | C)$$

- MAP classification rule: for a sample  $\mathbf{x} = (x_1, x_2, \dots, x_p)$

$$[P(x_1 | c^*) \cdots P(x_p | c^*)]P(c^*) > [P(x_1 | c) \cdots P(x_p | c)]P(c),$$

$$c \neq c^*, c = c_1, \dots, c_L$$

# Naïve Bayes Classifier

- Naïve Bayes classification

- Assumption that **all input attributes are conditionally independent!**

$$P(X_1, X_2, \dots, X_p | C) = P(X_1 | C)P(X_2 | C) \cdots P(X_p | C)$$

- MAP classification rule: for a sample  $\mathbf{x} = (x_1, x_2, \dots, x_p)$

$$\underbrace{[P(x_1 | c^*) \cdots P(x_p | c^*)]P(c^*)}_{c \neq c^*, c = c_1, \dots, c_L} > [P(x_1 | c) \cdots P(x_p | c)]P(c),$$

$$\Rightarrow \arg \max_{i=1, \dots, L} P(c_i) P(x_1 | c_i) P(x_2 | c_i) \cdots P(x_p | c_i)$$

# Naïve Bayes Classifier

- Naïve Bayes classification

- Assumption that **all input attributes are conditionally independent!**

$$P(X_1, X_2, \dots, X_p | C) = P(X_1 | C)P(X_2 | C)\cdots P(X_p | C)$$

- MAP classification rule: for a sample  $\mathbf{x} = (x_1, x_2, \dots, x_p)$

$$[P(x_1 | c^*) \cdots P(x_p | c^*)]P(c^*) > [P(x_1 | c) \cdots P(x_p | c)]P(c),$$

$$c \neq c^*, c = c_1, \dots, c_L$$

$\underbrace{P(x_i | c_i)}_{\substack{\{i=1,2,\dots,p\} \\ \{c=c_1, \dots, c_L\}}}$

Bernoulli (P)  
Binomial (K,P)  
Multinomial  
Gaussian

# Naïve Bayes Classifier (for discrete input attributes) – training/ Learning phase

- Learning Phase: Given a training set  $S$ ,

For each target value of  $c_i$  ( $c_i = c_1, \dots, c_L$ )

$\hat{P}(C = c_i) \leftarrow$  estimate  $P(C = c_i)$  with examples in  $S$ ;

# Naïve Bayes Classifier (for discrete input attributes) – training/ Learning phase

- Learning Phase: Given a training set  $S$ ,

For each target value of  $c_i$  ( $c_i = c_1, \dots, c_L$ )

$\hat{P}(C = c_i) \leftarrow$  estimate  $P(C = c_i)$  with examples in  $S$ ;

For every attribute value  $x_{jk}$  of each attribute  $X_j$  ( $j = 1, \dots, p$ ;  $k = 1, \dots, K_j$ )

$\hat{P}(X_j = x_{jk} | C = c_i) \leftarrow$  estimate  $P(X_j = x_{jk} | C = c_i)$  with examples in  $S$ ;

Output: conditional probability tables; for  $X_j: K_j \times L$  elements

# Naïve Bayes Classifier (for discrete input attributes) - training

- Learning Phase: Given a training set  $S$ ,

For each target value of  $c_i$  ( $c_i = c_1, \dots, c_L$ )

—  $\hat{P}(C = c_i) \leftarrow$  estimate  $P(C = c_i)$  with examples in  $S$ ;

— For every attribute value  $x_{jk}$  of each attribute  $X_j$  ( $j = 1, \dots, p$ ;  $k = 1, \dots, K_j$ )

—  $\hat{P}(X_j = x_{jk} | C = c_i) \leftarrow$  estimate  $P(X_j = x_{jk} | C = c_i)$  with examples in  $S$ ;

$\checkmark$   $K_1, K_2, \dots, K_p$   
 $\{X_1, X_2, \dots, X_p\}$

$K_1 \times L +$   
 $K_2 \times L +$   
 $K_p \times L$

$\checkmark$

# Naïve Bayes (for discrete input attributes) - testing

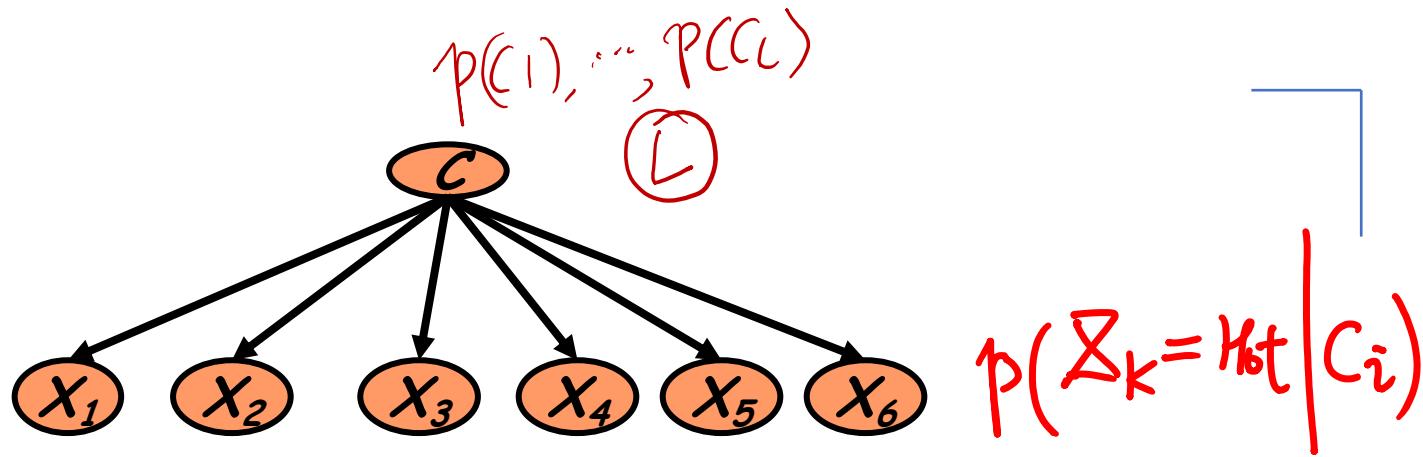
- **Test Phase:** Given an unknown instance

Look up tables to assign the label  $c^*$  to  $\mathbf{x}'$  if  $\mathbf{X}' = (a'_1, \dots, a'_p)$

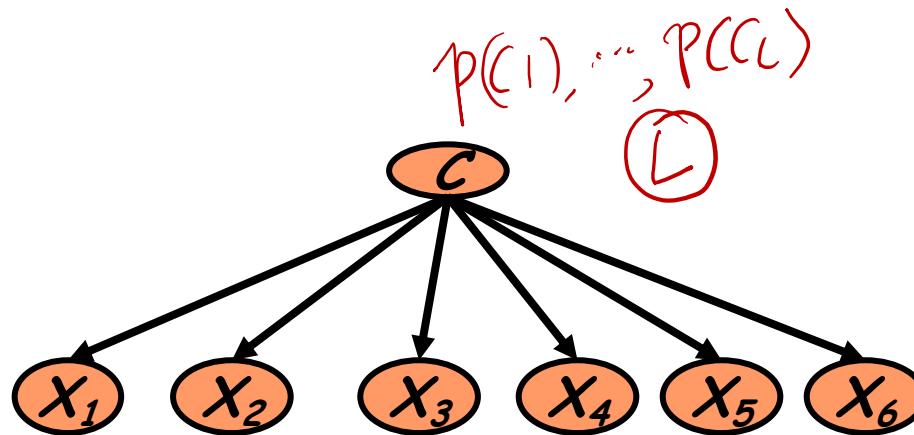
$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_p | c^*)] \hat{P}(c^*) > [\underbrace{\hat{P}(a'_1 | c) \cdots \hat{P}(a'_p | c)}_{c \neq c^*, c = c_1, \dots, c_L} ] \hat{P}(c),$$

$$\begin{aligned} & P(\mathbf{x}' | c_i) P(c_i) \\ &= P(a'_1 | c_i) P(a'_2 | c_i) \cdots P(a'_p | c_i) P(c_i) \\ & \quad i=1, 2, \dots, L \end{aligned}$$

# Learning (training) the NBC Model



# Learning (training) the NBC Model

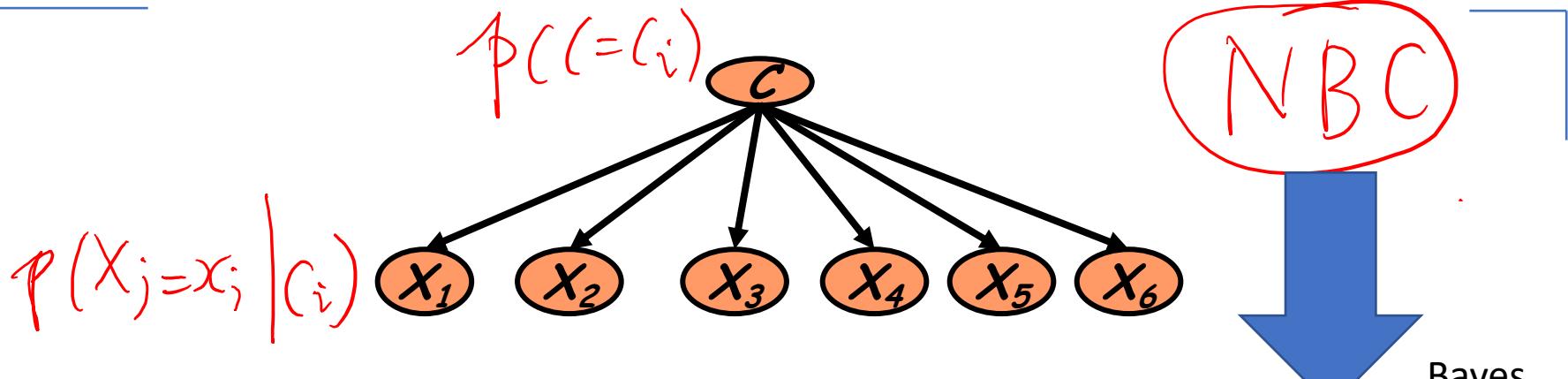


- maximum likelihood estimates:
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

$$\hat{P}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

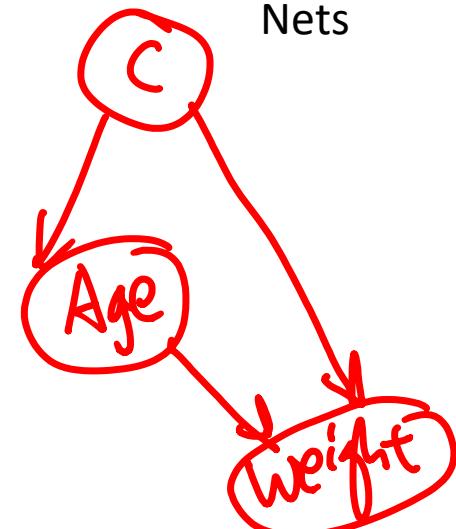
# Learning (training) the NBC Model



- maximum likelihood estimates
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

$$\hat{P}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$



## PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$P(X_1 = \text{Rain} \mid C = \text{Yes}) = \frac{3}{9}$$

$$P(X_1 = \text{Rain} \mid C = \text{No}) = \frac{2}{5}$$

Counting  
↑

## • Learning Phase

Estimate  $P(X_j = x_{jk} | C = c_i)$  with examples in training;

$$P(X_2|C_1), P(X_2|C_2)$$

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

$$P(X_4|C_1), P(X_4|C_2)$$

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

3+3+2+2 [naïve assumption] \* 2 [two classes]= 20 parameters

$$P(\text{Play}=Yes) = 9/14$$

$$P(\text{Play}=No) = 5/14$$

$$P(C_1), P(C_2), \dots, P(C_L)$$

Counting  
↑

## Learning Phase

$$P(X_1, X_2, X_3, X_4 | C_i)$$

Estimate  $P(X_j = x_{jk} | C = c_i)$  with examples in training;

$$P(X_2|C_1), P(X_2|C_2)$$

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

3+3+2+2 [naïve assumption] \* 2 [two classes] = 20 parameters

$$P(\text{Play}=Yes) = 9/14$$

$$P(\text{Play}=No) = 5/14$$

$$P(C_1), P(C_2), \dots, P(C_L)$$

$$P(C_i)$$

# Testing the NBC Model

look up

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_p | c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_p | c)] \hat{P}(c)$$

- Test Phase

- Given a new instance,

$\mathbf{x}' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

# Using the NBC Model

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_p | c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_p | c)] \hat{P}(c)$$

- Test Phase

- Given a new instance,

$\mathbf{x}' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

$$\begin{aligned} &\rightarrow P(C_1) P(\text{Sunny} | C_1) P(\text{Cool} | C_1) P(\text{High} | C_1) P(\text{Strong} | C_1) \\ &= \frac{9}{14} \times \frac{2}{9} \times \dots = \\ &\rightarrow P(C_2) P(\text{Su} | C_2) P(\text{Co} | C_2) P(\text{hi} | C_2) P(\text{St} | C_2) \\ &= \frac{5}{14} \times \frac{3}{5} \times \dots = \end{aligned}$$

# Testing the NBC Model

- **Test Phase**

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_p | c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_p | c)] \hat{P}(c)$$

- Given a new instance,  
 $x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$
- Look up in conditional-prob tables

$$P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{Yes}) = 2/9$$

$$P(\text{Temperature}=\text{Cool} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Humidity}=\text{High} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Wind}=\text{Strong} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{No}) = 3/5$$

$$P(\text{Temperature}=\text{Cool} | \text{Play}=\text{No}) = 1/5$$

$$P(\text{Humidity}=\text{High} | \text{Play}=\text{No}) = 4/5$$

$$P(\text{Wind}=\text{Strong} | \text{Play}=\text{No}) = 3/5$$

$$P(\text{Play}=\text{No}) = 5/14$$

- Test Phase

$$[\hat{P}(a'_1 | c^*) \cdots \hat{P}(a'_p | c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 | c) \cdots \hat{P}(a'_p | c)] \hat{P}(c)$$

- Given a new instance,

$\mathbf{x}' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})$

- Look up in conditional-prob tables

$$P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{Yes}) = 2/9$$

$$P(\text{Temperature}=\text{Cool} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Humidity}=\text{High} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Wind}=\text{Strong} | \text{Play}=\text{Yes}) = 3/9$$

$$P(\text{Play}=\text{Yes}) = 9/14$$

$$P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{No}) = 3/5$$

$$P(\text{Temperature}=\text{Cool} | \text{Play}=\text{No}) = 1/5$$

$$P(\text{Humidity}=\text{High} | \text{Play}=\text{No}) = 4/5$$

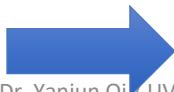
$$P(\text{Wind}=\text{Strong} | \text{Play}=\text{No}) = 3/5$$

$$P(\text{Play}=\text{No}) = 5/14$$

- MAP rule

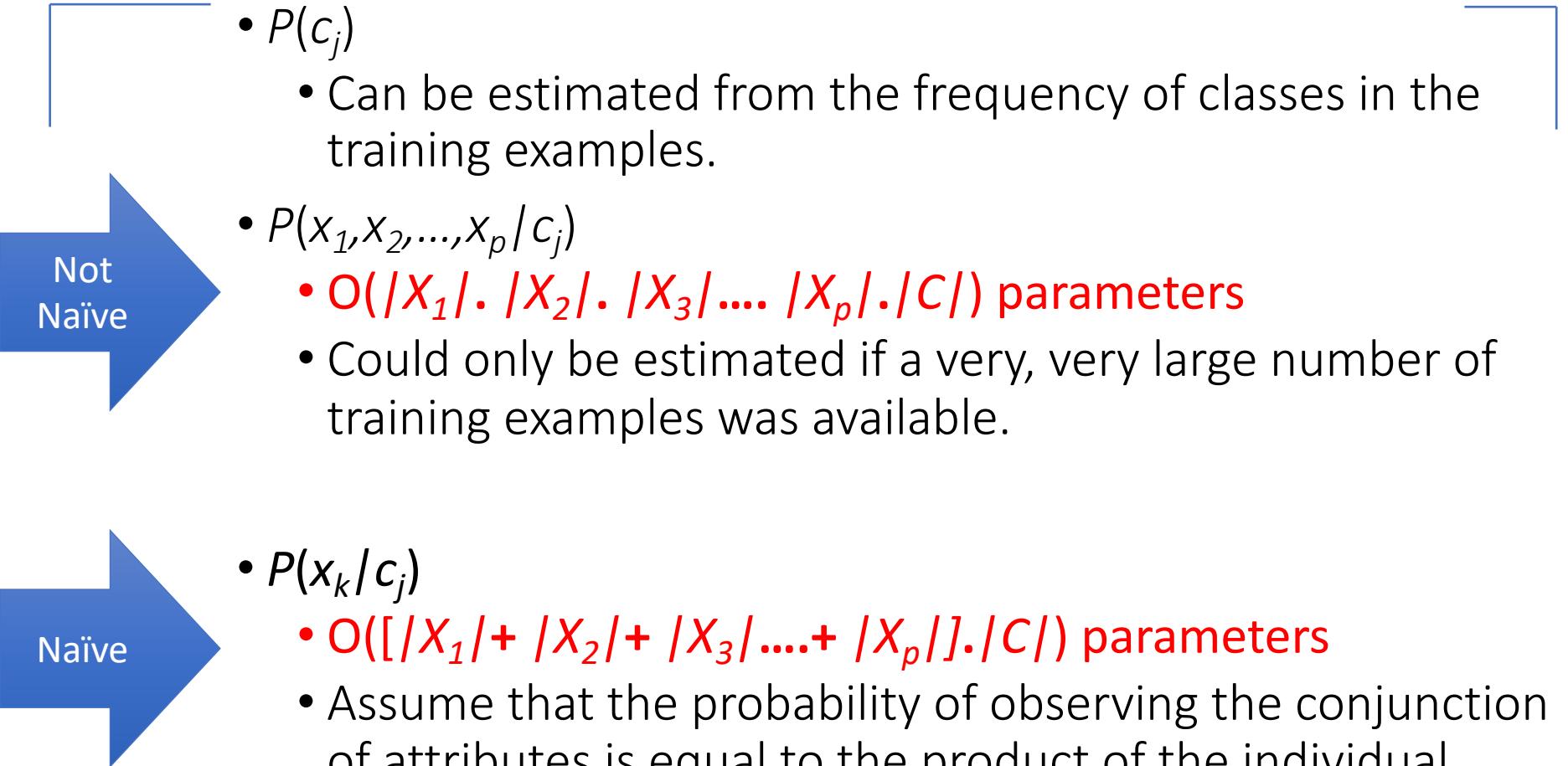
$$P(\text{Yes} | \mathbf{x}') = [P(\text{Sunny} | \text{Yes}) P(\text{Cool} | \text{Yes}) P(\text{High} | \text{Yes}) P(\text{Strong} | \text{Yes})] P(\text{Play}=\text{Yes}) = 0.0053$$

$$P(\text{No} | \mathbf{x}') = [P(\text{Sunny} | \text{No}) P(\text{Cool} | \text{No}) P(\text{High} | \text{No}) P(\text{Strong} | \text{No})] P(\text{Play}=\text{No}) = 0.0206$$



Given the fact  $P(\text{Yes} | \mathbf{x}') < P(\text{No} | \mathbf{x}')$ , we label  $\mathbf{x}'$  to be "No".

# WHY ? Naïve Bayes Assumption

- 
- $P(c_j)$ 
    - Can be estimated from the frequency of classes in the training examples.
  - $P(x_1, x_2, \dots, x_p | c_j)$ 
    - $O(|X_1| \cdot |X_2| \cdot |X_3| \dots |X_p| \cdot |C|)$  parameters
    - Could only be estimated if a very, very large number of training examples was available.
  - $P(x_k | c_j)$ 
    - $O(|X_1| + |X_2| + |X_3| \dots + |X_p| \cdot |C|)$  parameters
    - Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities  $P(x_i | c_j)$ .

# WHY ? Naïve Bayes Assumption

- $P(c_j)$

- Can be estimated from the frequency of classes in the training examples.

- $P(x_1, x_2, \dots, x_p | c_j)$

- $O(|X_1| \cdot |X_2| \cdot |X_3| \dots |X_p| \cdot |C|)$  parameters

- Could only be estimated if a very, very large number of training examples was available.

Not  
Naïve

- $P(x_k | c_j)$

- $O(|X_1| + |X_2| + |X_3| \dots + |X_p| \cdot |C|)$  parameters

- Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities  $P(x_i | c_j)$ .

Assuming  $|C| = L$   
num of unique values

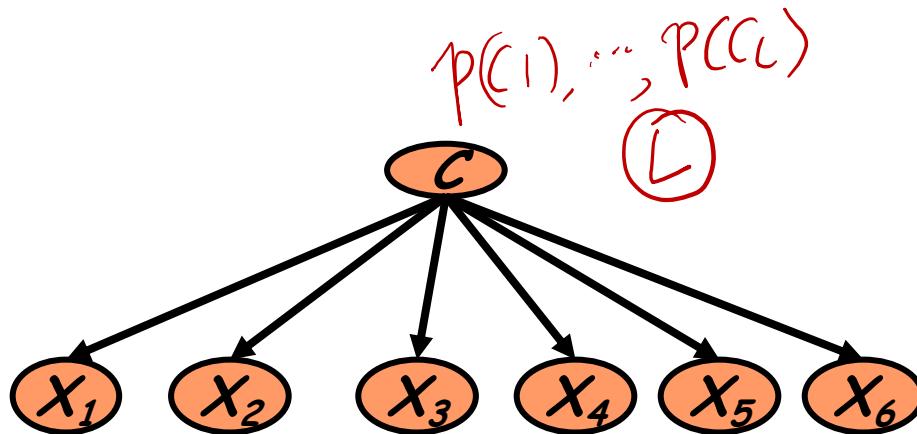
Assuming  $|X_i| = 2, i=1, 2, \dots, P$   
 $\Rightarrow 2^P \times L$

(Exp)

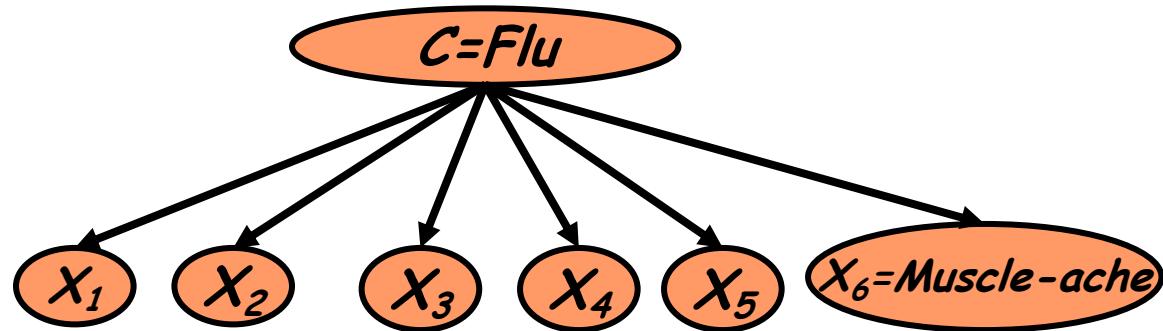
$$(2+2+2+\dots+2) \times L = 2^P \times L$$

Linear

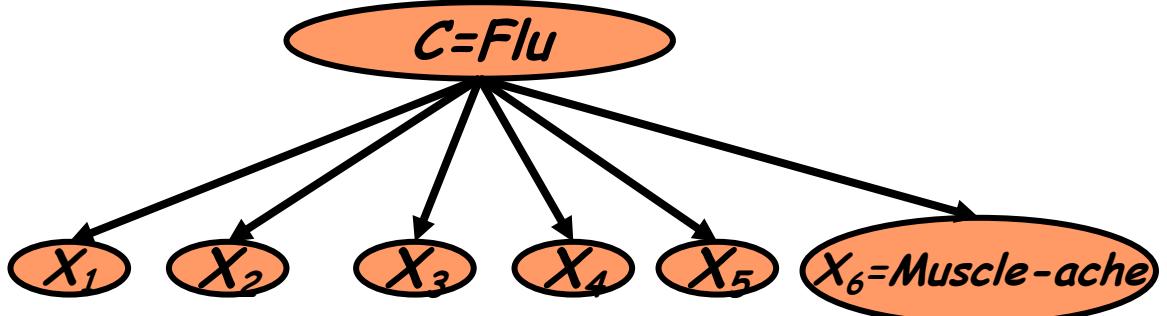
# Challenges during Learning (training) the NBC Model



For instance:



For instance:



- What if we have seen no training cases where patient had no flu and muscle aches?

$$\hat{P}(X_6 = T | C = \text{not\_flu}) = \frac{N(X_6 = T, C = \text{nf})}{N(C = \text{nf})} = 0$$

Muscle-Ache Yes/No      Flu/nf

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$?? = \arg \max_c \hat{P}(c) \prod_i \hat{P}(x_i | c)$$

$$\delta_f = p(c=f|u) p(x_1|f) p(x_2|f) p(x_3|f) p(x_4|f) p(x_5|f) p(x_6|f)$$

$$\delta_{nf} = p(c=nf) p(x_1|nf) p(x_2|nf) p(x_3|nf) p(x_4|nf) p(x_5|nf) p(x_6|nf)$$

if any term gives 0,

$$\Rightarrow \delta_{nf} = 0$$

no matter other terms' value

# Smoothing to Avoid Overfitting

Why necessary ??

$$\hat{P}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k_i}$$

# of values of feature  $X_i$

To make  
 $\sum_i P(x_i | C_j) = 1$

$$|X_i| = k_i$$

# Smoothing to Avoid Overfitting

$$\hat{P}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k_i}$$

# of values of  $X_i$

- Somewhat more subtle version

overall fraction in data  
where  $X_i = x_{i,k}$

$$\hat{P}(x_{i,k} | c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m}$$

$\rightarrow k \in \{1, 2, \dots, k_i\}$

extent of  
“smoothing” 54

# Summary:

## Generative Bayes Classifier

Task: Classify a new instance  $X$  based on a tuple of attribute values  $X = \langle X_1, X_2, \dots, X_p \rangle$  into one of the classes

$$c_{MAP} = \operatorname{argmax}_{c_j \in C} P(c_j | x_1, x_2, \dots, x_p)$$

$$= \operatorname{argmax}_{c_j \in C} \frac{P(x_1, x_2, \dots, x_p | c_j) P(c_j)}{P(x_1, x_2, \dots, x_p)}$$

$$= \operatorname{argmax}_{c_j \in C} P(x_1, x_2, \dots, x_p | c_j) P(c_j)$$

$j=1, 2, \dots, L$

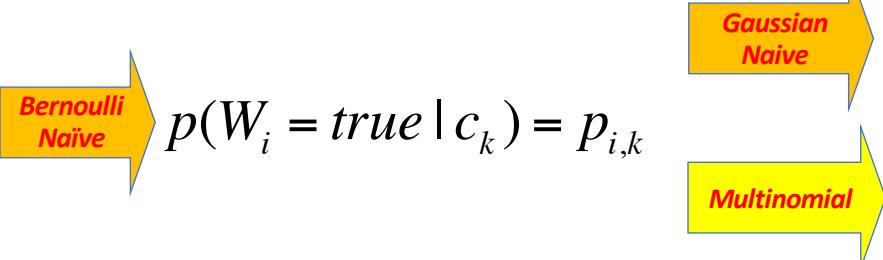
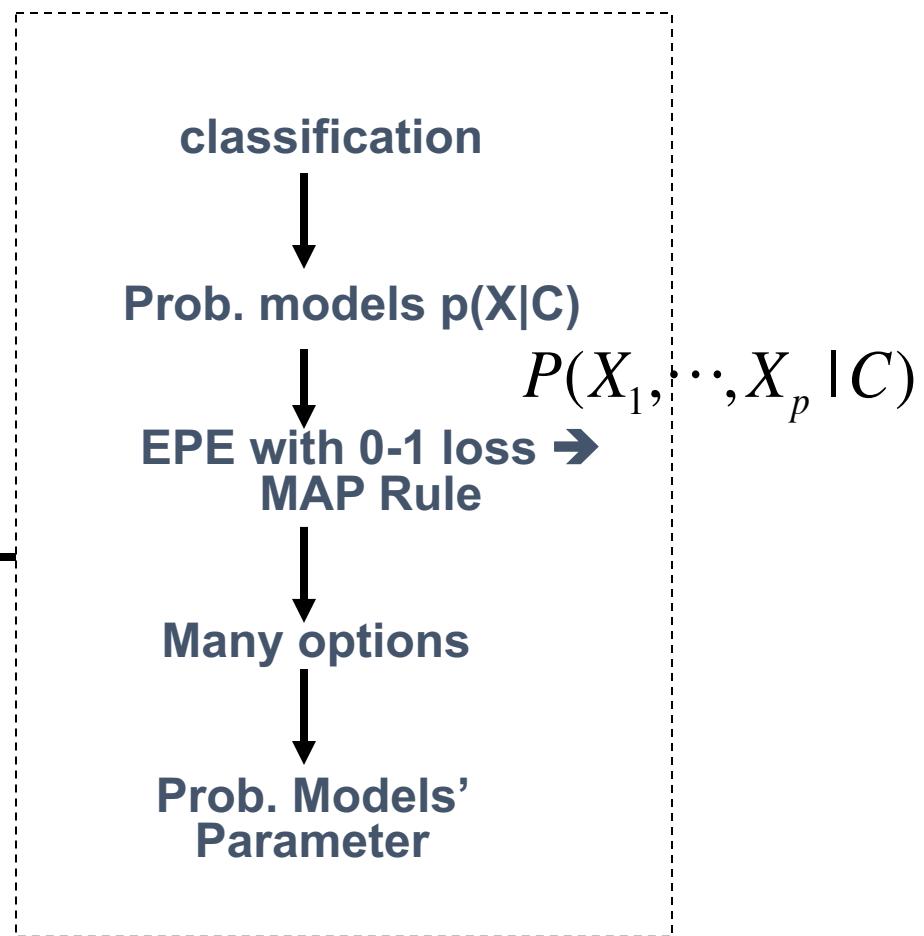
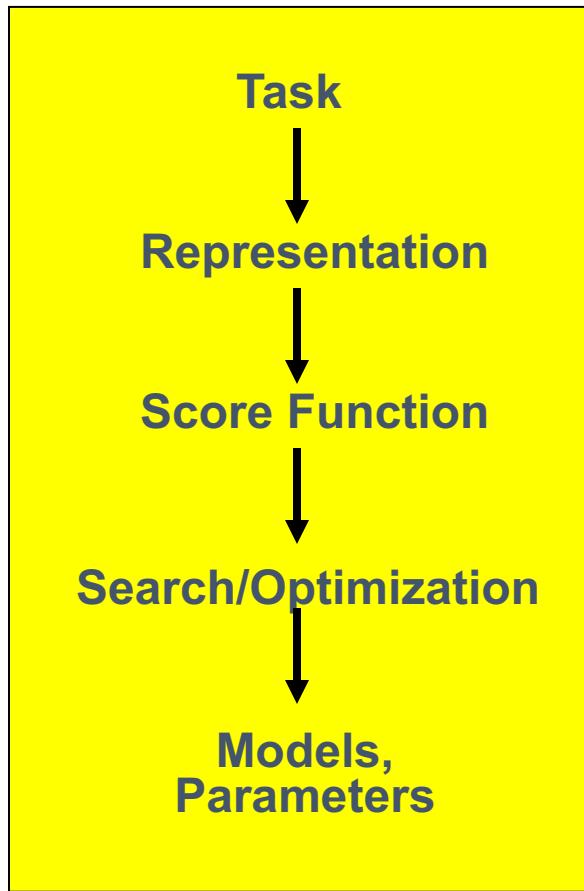
MAP = Maximum A Posteriori

## NEXT: More Generative Bayes Classifiers

- ✓ Generative Bayes Classifier
- ✓ Naïve Bayes Classifier
- ✓ Gaussian Bayes Classifiers
  - Gaussian distribution
  - Naïve Gaussian BC
  - Not-naïve Gaussian BC → LDA, QDA
- ✓ Discriminative vs. Generative

$$\underset{k}{\operatorname{argmax}} P(C_k | X) = \underset{k}{\operatorname{argmax}} P(X, C) = \underset{k}{\operatorname{argmax}} P(X | C)P(C)$$

## Generative Bayes Classifier



$$\hat{P}(X_j | C = c_k) = \frac{1}{\sqrt{2\pi}\sigma_{jk}} \exp\left(-\frac{(X_j - \mu_{jk})^2}{2\sigma_{jk}^2}\right)$$

$$P(W_1 = n_1, \dots, W_v = n_v | c_k) = \frac{N!}{n_{1k}! n_{2k}! \dots n_{vk}!} \theta_{1k}^{n_{1k}} \theta_{2k}^{n_{2k}} \dots \theta_{vk}^{n_{vk}}$$

# References

- Prof. Andrew Moore's review tutorial
- Prof. Ke Chen NB slides
- Prof. Carlos Guestrin recitation slides