# UVA CS 4774: Machine Learning

# S3: Lecture 16: Generative Bayes Classifiers

Module I

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

# 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~~

# Roadmap : 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

Extra

# Review: Bayes classifiers (BC)

- Treat each feature attribute and the class label as random variables.

- Testing: Given a sample **x** with attributes ( $x_1$, $x_2$, … , $x_p$ ):
  - Goal is to predict its class *c*.
  - Specifically, we want to find the class that maximizes $p( c \mid x_1, x_2, \ldots , x_p )$.

- Training: can we estimate $p(C_i \mid \mathbf{x}) = p( C_i \mid x_1, x_2, \ldots , x_p )$ directly from data?

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_p)$$

MAP Rule

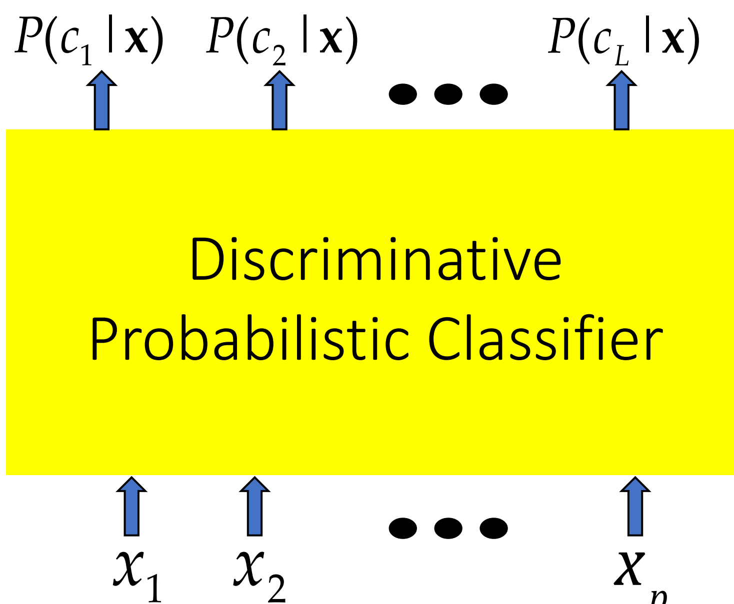
# Two kinds of Bayes classifiers via MAP classification rule

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

# Review: Discriminative BC

– (1) Discriminative model

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




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

Softmax

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

multi-class

logistic regression

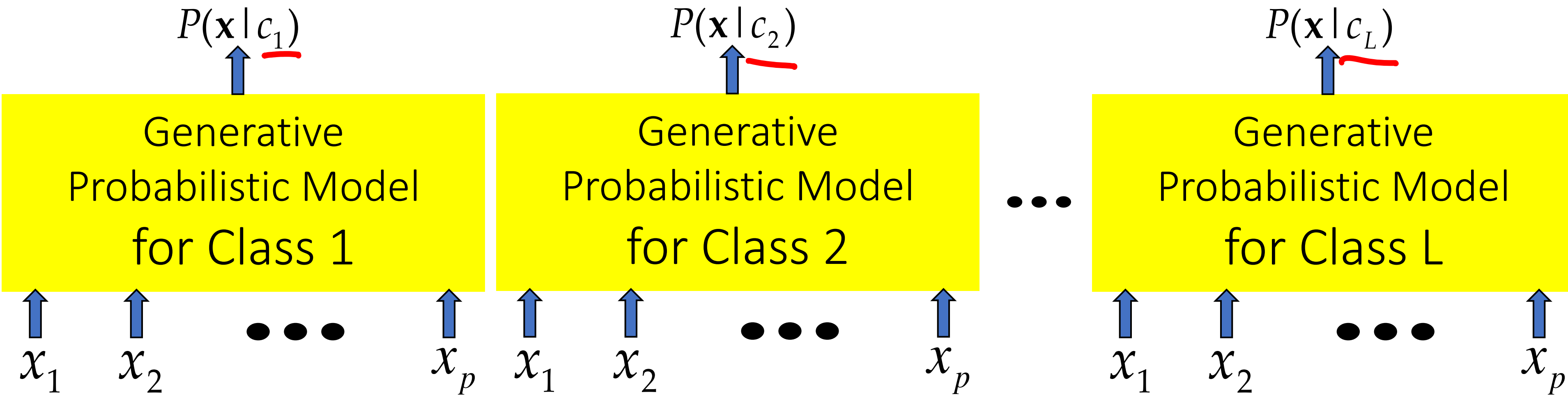


Adapt from Prof. Ke Chen NB slides

# (2) Generative BC

$\Rightarrow p(c_i | \vec{x})$

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




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



Adapt from Prof. Ke Chen NB slides

# 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

$$p(c_i|\vec{x}) = \frac{p(\vec{x}|c_i)\,p(c_i)}{p(\vec{x})}$$

$$\Downarrow$$

$$c^* = \underset{c_i \sim \{c_1, c_2, \cdots, c_L\}}{\operatorname{argmax}} p(c_i|\vec{x})$$

# One Example:

*Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the **set** {**r**,**r**,**r**,**b**}. What is the probability of drawing 2 red balls in the first 2 tries?*

$$\boldsymbol{P(B_1 = r, B_2 = r)} \ =$$

$$P(B_1 = r | B_2 = r)$$



Adapt from Prof. Nando de Freitas's review slides

# One Example: Joint

*Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the **set** {r,r,r,b}. What is the probability of drawing 2 red balls in the first 2 tries?*

$$\boldsymbol{P(B_1 = r, B_2 = r)} = P(B_1 = r)\, P(B_2 = r \mid B_1 = r)$$

$$= \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}$$



Adapt from Prof. Nando de Freitas's review slides

# Review : Bayes' Rule – for Generative Bayes Classifiers

$$\boxed{P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}}$$

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

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

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

# Review : Bayes' Rule
# – for Generative Bayes Classifiers

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

Prior

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

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

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

# Summary of Generative BC:

– Apply Bayes rule to get posterior probabilities

$$P(C=c_i \mid \mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x} \mid C=c_i)P(C=c_i)}{P(\mathbf{X}=\mathbf{x})}$$
$$\propto P(\mathbf{X}=\mathbf{x} \mid C=c_i)P(C=c_i)$$
$$\text{for } i=1,2,\cdots,L$$

– Then apply the MAP rule



Adapt from Prof. Ke Chen NB slides

# Summary of Generative BC:

– Apply Bayes rule to get posterior probabilities

$$\underset{c_i}{\operatorname{argmax}} \{ P(C = c_i \mid \mathbf{X} = \mathbf{x}) \} = \frac{P(\mathbf{X} = \mathbf{x} \mid C = c_i) P(C = c_i)}{\cancel{P(\mathbf{X} = \mathbf{x})}}$$

$$\propto \underline{P(\mathbf{X} = \mathbf{x} \mid C = c_i) P(C = c_i)}$$

$$\text{for } i = 1, 2, \cdots, L$$

$c_i \in \{c_1, c_2, \cdots, c_L\}$

– Then apply the MAP rule

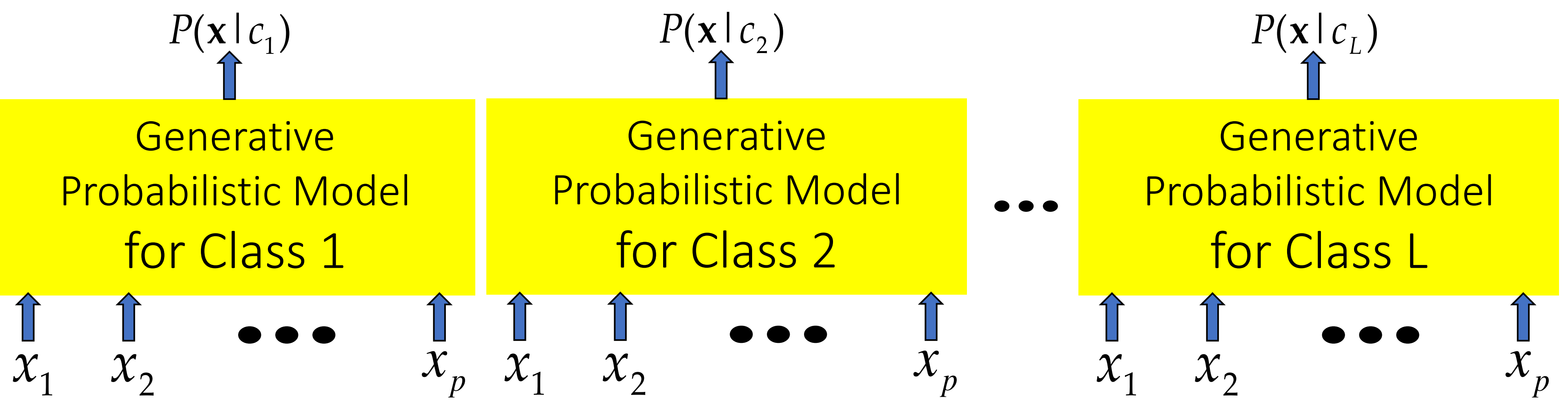
$$\begin{cases} p(\vec{x} \mid c_i) \\ p(c_i) \end{cases}$$



Adapt from Prof. Ke Chen NB slides

# Establishing a probabilistic model for classification through generative probabilistic models

$$\underset{C_i}{\operatorname{argmax}} P(C_i \mid X) = \underset{C_i}{\operatorname{argmax}} P(X, C_i) = \underset{C_i}{\operatorname{argmax}} P(X \mid C_i) P(C_i)$$




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



Adapt from Prof. Ke Chen NB slides

# Generative Bayes Classifier

Generative → Representation

Conditional Pr → Models, Parameters

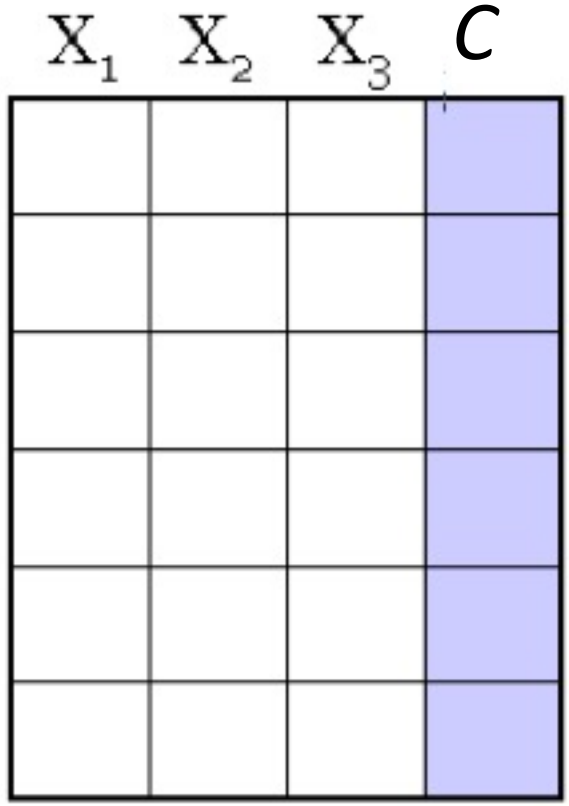
- Data → Task → Representation → Score Function → Search/Optimization → Models, Parameters

this lecture!

- Tabular → classification → Prob. models p(X|C) → EPE with 0-1 loss ➔ MAP Rule → MLE based (mostly) → Prob. Models' Parameter

$$P(X_1, \cdots, X_p \mid C)$$

$$\underset{k}{\operatorname{argmax}} P(C\_k \mid X) = \underset{k}{\operatorname{argmax}} P(X, C) = \underset{k}{\operatorname{argmax}} P(X \mid C) P(C)$$




# A Dataset for classification

$$f : X \longrightarrow C$$

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

Discriminative

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

Generative

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

this lecture!

# An Example

- Example: Play Tennis

| $X_1$ | $X_2$ | $X_3$ | C |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

*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 |

$x_1$ $x_2$ $x_3$ $x_4$ $C$

C

# Learning Phase:

- Example: Play Tennis

$X_1$ $X_2$ $X_3$ $X_4$ C

$k_2=3$

$X_2$: {Hot, Mild, Cool}

$X_3$ = {High, Normal}

$k_3=2$

$X_4=(W,S)$

$K_4=$

*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 |

C: {Yes, No}

$L=2$

$X_1$: {Sunny, overc, rain}

$k_1=3$

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $C$ |
|---|---|---|---|---|
| S<br>O<br>R | H<br>M<br>C | H<br>N | W<br>S | $C_1$=Yes<br>$C_2$=No |
| 3 × | 3 × | 2 × | 2 × | 2 ⇒ 72 parameters |

$\Rightarrow p(X_1 X_2 X_3 X_4 | C)$, $p(C)$ training

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

② $p(X_1=S, X_2=H, X_3=H, X_4=W | C=\text{No}) = \dfrac{N(SHHW\text{No})}{N(\text{No})} = \dfrac{1}{5}$

$p(X_1=S, X_2=H, X_3=H, X_4=W | C=\text{Yes}) = \dfrac{0}{9} = 0$

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

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

Directly estimate from data via counting

*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 |

e.g.

$p(\text{overcast}, \text{hot}, \text{high}, \text{weak} \mid \text{No})$

$= \frac{0}{9}$

Check MLE Lecture for Why

C

$P(X_1, X_2, X_3, X_4 \mid Yes)$
$p(X_1, X_2, X_3, X_4 \mid No)$ → $36 \times 2$

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 |

1

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

$P(C = C_i)$

$\{Yes, No\}$

$P(C = Yes) = 9/14$

$P(C = No) = 5/14$

# Generative Bayes Classifier:

- Learning Phase → a look up table of cond. prob

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

P(Play=Yes) = 9/14        P(Play=No) = 5/14

$P(X_1,X_2,\ldots,X_p|C_1), P(X_1,X_2,\ldots,X_p|C_2)$

| 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*2*2 [conjunctions of attributes] * 2 [two classes]=72 parameters

# Generative Bayes Classifier:

e.g English Dictionary $X_1, \dots, X_p$

$p \sim 30k$ } $2^p \times L$

$\Rightarrow C(\underbrace{up, down}_{L})$

- Learning Phase

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

$p(C_1)$ $p(C_2)$

P(Play=Yes) = 9/14 P(Play=No) = 5/14

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

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*2*2 [conjunctions of attributes] * 2 [two classes]=72 parameters

# Generative Bayes Classifier:

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

Last Page: the learned model

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

  - – Given a new instance,

x'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

$$\left\{ \begin{array}{l} p(x' \mid Yes)\, p(C=Yes) \\ p(x' \mid No)\, p(C=No) \end{array} \right\} \Rightarrow \arg\max_{C} \Rightarrow \text{predicted } C^*$$

# Thank You

Thank you

# UVA CS 4774: Machine Learning

# S3: Lecture 16: Generative Bayes Classifiers

Module II

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

# 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

# Generative Bayes Classifier

- Bayes classification

$$\underset{c_j \in C}{\operatorname{argmax}} \underbrace{P(x_1, x_2, \ldots, x_p}_{2\times2\times\cdots\times2 = 2^p \times L} | \underbrace{c_j}_{L})P(c_j)$$

Difficulty: learning the joint probability

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

(when given $C_i$)

# Naïve Bayes Classifier

$= p(x_1|c_j)\,p(x_2|c_j)\ldots p(x_p|c_j)$

- Bayes classification

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

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]

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

$$P(C_j | X_1, \cdots X_p) \propto P(X_1, X_2, \cdots X_p | C_j) P(C_j)$$
$$= \underline{P(X_1|C_j)} \, \underline{P(X_2|C_j)} \cdots \underline{P(X_p|C_j)} \, \underline{P(C_j)}$$

# Learning Phase:

- Example: Play Tennis

$X_1$ $X_2$ $X_3$ $X_4$ $C$

$k_2=3$

$X_2$: {Hot, Mild, Cool}

$X_3$ = {High, Normal}

$k_3=2$

$X_4=(W,S)$

$K_4=$

*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 |

C: {Yes, No}

$L=2$

$X_1$: {Sunny, overc, rain}

$k_1=3$

| $3 \times L$ | | $P(X_1 \mid C)$ |
|---|---|---|
| $X_1 =$ Sunny | C=Yes | $\frac{N(S,Yes)}{N(Yes)} = \frac{2}{9}$ |
| $X_1 = O$ | C=Yes | |
| $X_1 = R$ | C=Yes | |
| $X_1 = S$ | C=No | $\frac{N(S,N)}{N(No)} = \frac{3}{5}$ |
| O | C=No | |
| R | C=No | |

$X_1 : 3 \times L$

$X_2 : 3 \times L$

$X_3 : 2 \times L$

$X_4 : 2 \times L$

$L=2$

| $P(C=Ye)$ | $P(C=N)$ |
|---|---|
| $9/14$ | $5/14$ |

No Naive

$72+2=74$ para

$2^p \times L$

$(3+3+2+2) \times \underbrace{L}_{=2} = 20$

$+2$

$\Rightarrow X_1 + X_2 + \dots X_p \Rightarrow \underbrace{2^p \times L}$ $p \sim 30k$

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(Play=Yes) = ??          P(Play=No) = ??

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

# Naïve Bayes Classifier

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

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

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

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

$$c \neq c^*, \; c = c_1, \cdots, c_L$$

# Naïve Bayes Classifier

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

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

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

$$\underline{[P(x_1 \mid c^*)\cdots P(x_p \mid c^*)]P(c^*)} > [P(x_1 \mid c)\cdots P(x_p \mid c)]P(c),$$

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

$$\Rightarrow \underset{i=1,\cdots,L}{\operatorname{argmax}} P(C_i)P(x_1 \mid C_i)P(x_2 \mid C_2)\cdots P(x_p \mid C_i)$$

# Naïve Bayes Classifier

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

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

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

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

$$c \neq c^*, \; c = c_1, \cdots, c_L$$

$\{j = 1, 2, \cdots, p\}$ $\{i = 1, 2, \cdots, L\}$

$\underline{P(X_j | C_i)}$

Bernolli $(p)$
Bionomial $(K, p)$
Multinomial
Gaussian

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

- Learning Phase: Given a training set S,

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

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

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

– Learning Phase: Given a training set S,

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

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

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

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

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 $c_i$ $(c_i = c_1, \cdots, c_L)$

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

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

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

$K_1, K_2, \ldots, K_p$

$\{X_1, X_2, \ldots, X_p\}$

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

# Naïve Bayes
# (for discrete input attributes) - testing

- Test Phase: Given an unknown instance

  Look up tables to assign the label c* to X' if $\mathbf{X}' = (a_1', \cdots, a_p')$

$$[\hat{P}(a_1' \mid c^*) \cdots \hat{P}(a_p' \mid c^*)]\hat{P}(c^*) > \underline{[\hat{P}(a_1' \mid c) \cdots \hat{P}(a_p' \mid c)]\hat{P}(c)},$$

$$c \neq c^*, \; c = c_1, \cdots, c_L$$

$$P(X' \mid C_i)\, P(C_i)$$

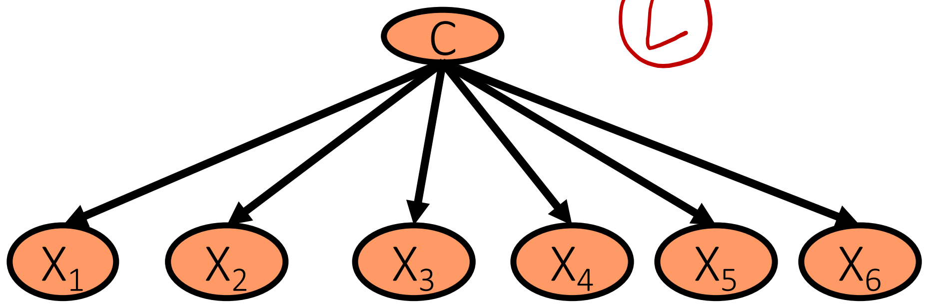
$$= P(a_1' \mid C_i)\, P(a_2' \mid C_i) \cdots P(a_p' \mid C_i)\, P(C_i)$$

$$i = 1, 2, \ldots, L$$

# Learning (training) the NBC Model

$p(C_1), \ldots, p(C_L)$

$L$

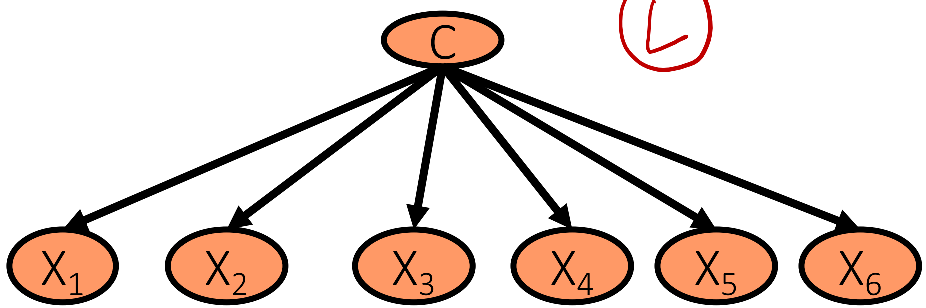



$p(X_k = \text{Hot} \mid C_i)$

# Learning (training) the NBC Model

$p(c_1), \ldots, p(c_L)$

$L$




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

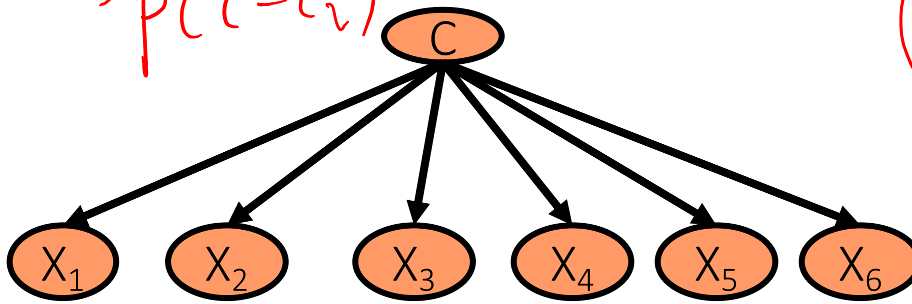
$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

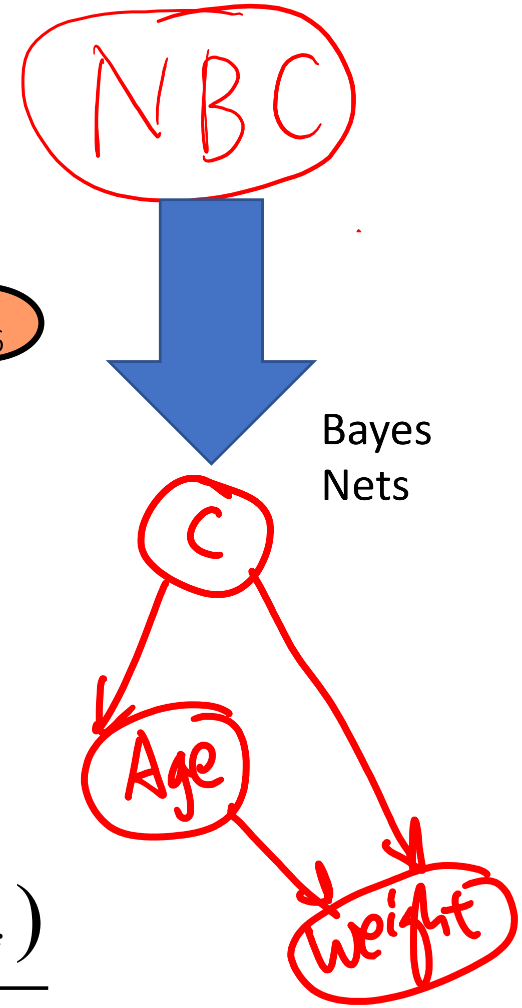
$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

# Learning (training) the NBC Model

$p(C=c_i)$

$p(X_j=x_i \mid c_i)$




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

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

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

# *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(Play=Yes) = 9/14 P(Play=No) = 5/14

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

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;

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

2

2

$X_1$

3

| 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 |

3

2

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

2

2

| 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 |

2

<< 72

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

P(Play=Yes) = 9/14    P(Play=No) = 5/14

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

$P(C_i)$

# 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,

    x'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=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,

    x'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

$\rightarrow P(C_1)P(Sunny|C_1)P(Cool|C_1)P(High|C_1)P(Strong|C_1)$

$= \frac{9}{14} \times \frac{2}{9} \cdots\cdots =$

$\rightarrow P(C_2)P(Su|C_2)P(Co|C_2)P(hi|C_2)P(St|C_2)$

$= \frac{5}{14} \times \frac{3}{5} \times \cdots\cdots =$

# Testing 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,

    x'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)
  - Look up in conditional-prob tables

| | |
|---|---|
| P(Outlook=Sunny\|Play=Yes) = 2/9 | P(Outlook=Sunny\|Play=No) = 3/5 |
| P(Temperature=Cool\|Play=Yes) = 3/9 | P(Temperature=Cool\|Play==No) = 1/5 |
| P(Huminity=High\|Play=Yes) = 3/9 | P(Huminity=High\|Play=No) = 4/5 |
| P(Wind=Strong\|Play=Yes) = 3/9 | P(Wind=Strong\|Play=No) = 3/5 |
| P(Play=Yes) = 9/14 | P(Play=No) = 5/14 |

$$[\hat{P}(a_1'|c^*)\cdots\hat{P}(a_p'|c^*)]\hat{P}(c^*) > [\hat{P}(a_1'|c)\cdots\hat{P}(a_p'|c)]\hat{P}(c)$$

- Test Phase
  - Given a new instance,

    x'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)
  - Look up in conditional-prob tables

| | |
|---|---|
| P(Outlook=Sunny\|Play=Yes) = 2/9 | P(Outlook=Sunny\|Play=No) = 3/5 |
| P(Temperature=Cool\|Play=Yes) = 3/9 | P(Temperature=Cool\|Play==No) = 1/5 |
| P(Huminity=High\|Play=Yes) = 3/9 | P(Huminity=High\|Play=No) = 4/5 |
| P(Wind=Strong\|Play=Yes) = 3/9 | P(Wind=Strong\|Play=No) = 3/5 |
| P(Play=Yes) = 9/14 | P(Play=No) = 5/14 |

  - MAP rule

    P(Yes|x'): [P(Sunny|Yes)P(Cool|Yes)P(High|Yes)P(Strong|Yes)]P(Play=Yes) = 0.0053

    P(No|x'): [P(Sunny|No) P(Cool|No)P(High|No)P(Strong|No)]P(Play=No) = 0.0206

Given the fact P(Yes|x') < P(No|x'), we label x' to be "No".

# Thank You

# UVA CS 4774: Machine Learning

# S3: Lecture 16: Generative Bayes Classifiers

Module III

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

# WHY ? Naïve Bayes Assumption

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.

Not Naïve

- $P(x_1,x_2,\ldots,x_p|c_j)$
  - $O(|X_1|.\ |X_2|.\ |X_3|\ldots.\ |X_p|.|C|)$ parameters
  - Could only be estimated if a very, very large number of training examples was available.

Naïve

- $P(x_k|c_j)$
  - $O([|X_1|+\ |X_2|+\ |X_3|\ldots.+\ |X_p|].|C|)$ parameters
  - Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(x_i|c_j)$.

Recap

| 3 | × L | $P(X_1 \mid C)$ |
|---|---|---|
| $X_1$ = Sunn | C=Yes | $\frac{N(S,Yes)}{N(Yes)} = \frac{2}{9}$ |
| $X_1$=O | C=Yes | |
| $X_1$=R | C=Yes | |
| $X_1$=S | C=No | $\frac{N(S,N)}{N(No)} = \frac{3}{5}$ |
| O | C=No | |
| R | C=No | |

$X_1$ : 3×L

$X_2$ : 3×L

$X_3$ : 2×L

$X_4$ : 2×L

| L=2 | P(C=Ye) | P(C=N) |
|---|---|---|
| | 9/14 | 5/14 |

No Naive

72+2=74 para

$2^p \times L$

$(3+3+2+2) \times \underbrace{L}_{=2} = 20$

+2

$\Rightarrow X_1+X_2+\ldots X_p \Rightarrow \underbrace{2^p \times L}$

p~30k

Recap

| $3 \times L$ | | $P(X_1 \mid C)$ |
|---|---|---|
| $X_1 =$ Sunny | $C =$ Yes | $\frac{N(S, Yes)}{N(Yes)} = \frac{2}{9}$ |
| $X_1 = O$ | $C =$ Yes | |
| $X_1 = R$ | $C =$ Yes | |
| $X_1 = S$ | $C =$ No | $\frac{N(S, N)}{N(No)} = \frac{3}{5}$ |
| O | $C =$ No | |
| R | $C =$ No | |

$X_1 : 3 \times L$

$X_2 : 3 \times L$

$X_3 : 2 \times L$

$X_4 : 2 \times L$

| $L=2$ $P(C=Ye)$ | $P(C=N)$ |
|---|---|
| $9/14$ | $5/14$ |

No Naive
$72 + 2 = 74$ param
$2^p \times L$

$$\underbrace{(3+3+2+2) \times L}_{=2} = 20 \underbrace{+2}$$

$\Rightarrow X_1 + X_2 + \dots X_p \Rightarrow \underbrace{2p \times L}$ $p \sim 30k$

# WHY ? Naïve Bayes Assumption

Assuming $\underbrace{|C|}_{\text{num of unique values}} = L$

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.

Assuming $|X_i| = 2$, $i = 1, 2, \cdots, p$

$\Rightarrow 2^p \times L$ Exp

Not Naïve

- $P(x_1, x_2, \ldots, x_p / c_j)$
  - O(/$X_1$/. /$X_2$/. /$X_3$/.... /$X_p$/./C/) parameters
  - Could only be estimated if a very, very large number of training examples was available.

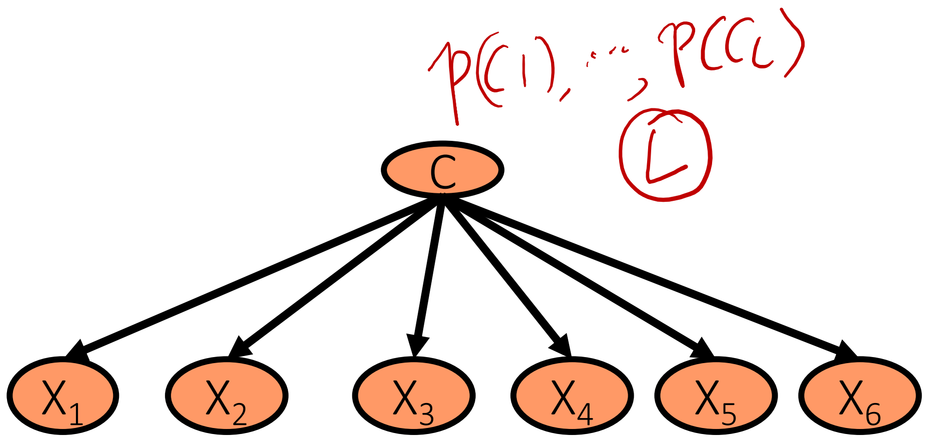
$(2+2+2+\cdots 2) \times L$

$= 2 \cdot p \times L$ Linear

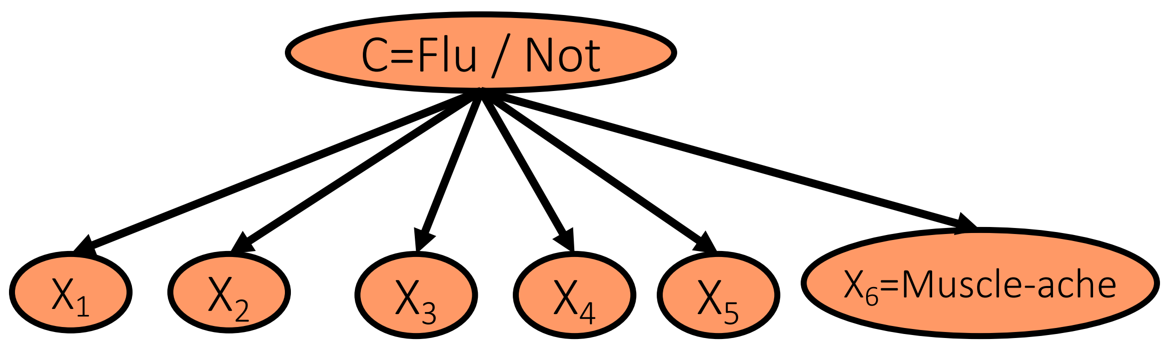
Naïve

- $P(x_k / c_j)$
  - O([/$X_1$/+ /$X_2$/+ /$X_3$/....+ /$X_p$/]./C/) parameters
  - Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(x_i | c_j)$.

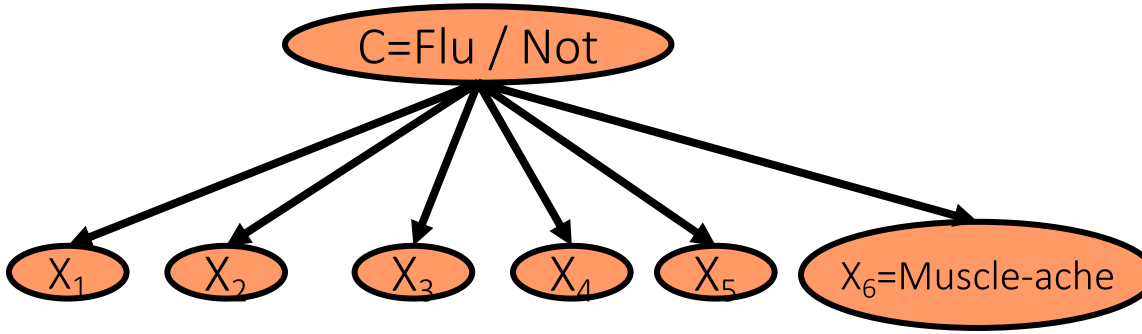
# Challenges during Learning (training) the GBC Model




For instance:



For instance:




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

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

muscle-ache=Yes/no flu/nf

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

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

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

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

if any term gives 0,

$\Rightarrow \delta_{nf} = 0$

no matter other terms' value

# Smoothing to Avoid Overfitting

Why necessary ??

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

# of values of feature $X_i$

To make
sum_i (P(xi|Cj)=1

$|X_i| = k_i$



Adapt From Manning' textCat tutorial

# Smoothing to Avoid Overfitting

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

# of values of $X_i$

- Somewhat more subtle version

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

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

$k \in \{1, 2, \ldots, k_i\}$

extent of "smoothing"

# Summary:
## Generative Bayes Classifier

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

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}}\, P(c_j \mid x_1, x_2, \ldots, x_p)$$

$$= \underset{c_j \in C}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_p \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_p)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}}\, \underbrace{P(x_1, x_2, \ldots, x_p \mid c_j)}\, \underbrace{P(c_j)}$$

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

MAP = Maximum A Posteriori



Adapt From Carols' prob tutorial

# Today Recap: Generative Bayes Classifier and Naïve BC

Task

Representation

Score Function

Search/Optimization

Models, Parameters

classification

Prob. models p(X|C)

$$P(X_1, \cdots, X_p \mid C)$$

EPE with 0-1 loss ➔ MAP Rule

Most MLE based

Prob. Models' Parameter

$$\underset{k}{\operatorname{argmax}} P(C\_k \mid X) = \underset{k}{\operatorname{argmax}} P(X, C) = \underset{k}{\operatorname{argmax}} P(X \mid C) P(C)$$

Bernoulli Naïve

$$p(W_i = true \mid c_k) = p_{i,k}$$

# Thank You

Thank you

# NEXT: More Generative Bayes Classifiers

EXTRA

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

$$\underset{k}{\operatorname{argmax}} P(C\_k \mid X) = \underset{k}{\operatorname{argmax}} P(X, C) = \underset{k}{\operatorname{argmax}} P(X \mid C) P(C)$$

# Generative Bayes Classifier

| | |
|---|---|
| Task | classification |
| Representation | Prob. models p(X\|C) |
| Score Function | EPE with 0-1 loss → MAP Rule |
| Search/Optimization | Many options |
| Models, Parameters | Prob. Models' Parameter |

$$P(X_1, \cdots, X_p \mid C)$$

**Gaussian Naive**

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

**Bernoulli Naïve**

$$p(W_i = true \mid c_k) = p_{i,k}$$

**Multinomial**

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

# References

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