UVA CS 4501: Machine Learning

Lecture 18: Generative Bayes Classifiers

Dr. Yanjun Qi

University of Virginia

Department of
Computer Science

Where are we ? → Major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory

Where are we ? → 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

2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

Today : Generative Bayes Classifiers

- ✓ Bayes Classifier
  - ▪ MAP classification rule
  - ▪ Generative Bayes Classifier
- ✓ Naïve Bayes Classifier

Review: Notations

- Inputs
  - X, X_j (jth element of vector X) : random variables written in capital letter
  - p #input features, n #observations
  - $\mathbf{X}$: matrix written in bold capital
  - Vectors are assumed to be column vectors

- Outputs
  - quantitative Y
  - qualitative C (for categorical)

Review: Bayes classifiers

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

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

MAP Rule

Review: Bayes Classifiers – MAP Rule

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

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

MAP Rule

MAP = Maximum Aposteriori Probability

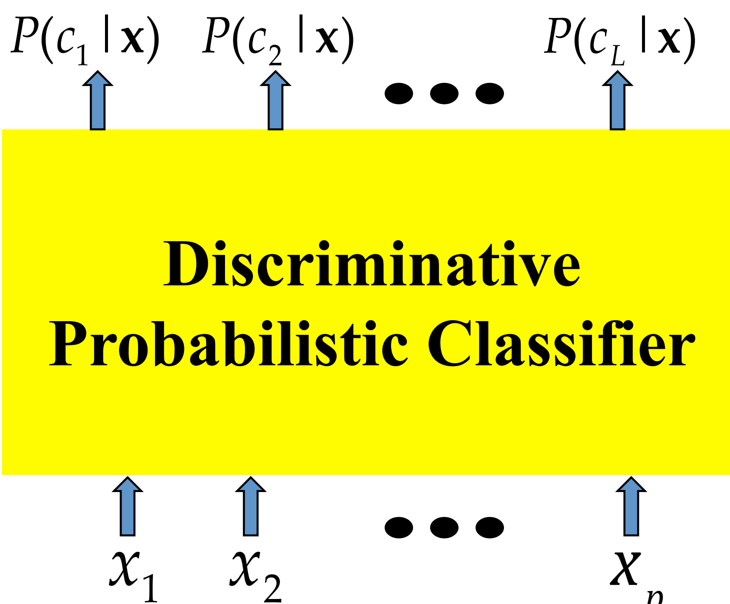
Please read the L13-Logistic for details

Adapt From Carols' prob tutorial

Review: Establishing a probabilistic model for classification

– **(1) Discriminative model**

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



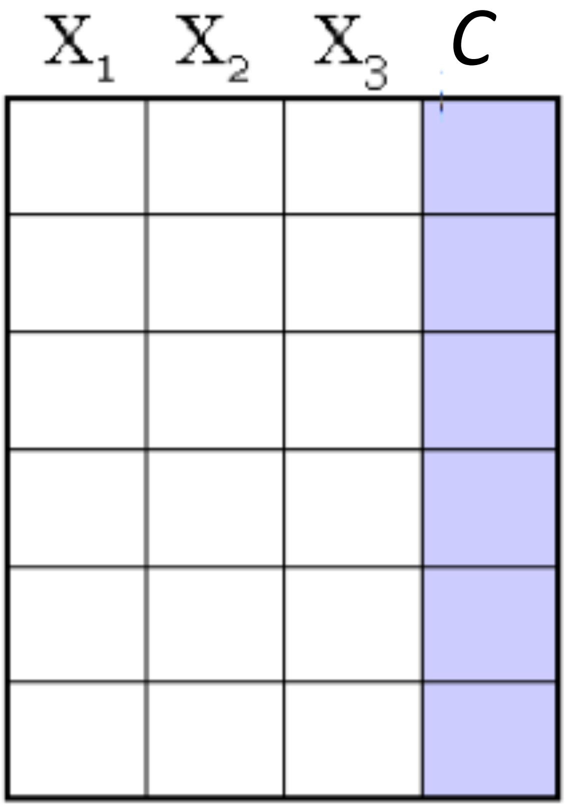
logistic regression

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

Adapt from Prof. Ke Chen NB slides

Bayes classifiers
➔ MAP classification rule

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



A Dataset for classification

$$f : \boxed{X} \longrightarrow \boxed{C}$$

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

$$\arg\max_{c \in C} P(c|X) = \arg\max_{c \in C} P(X,c) = \arg\max_{c \in C} P(X|c)P(c)$$

this lecture!

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

Today : Generative Bayes Classifiers

- ✓ Bayes Classifier
  - MAP classification rule
  - Generative Bayes Classifier
- ✓ Naïve Bayes Classifier

(2) Generative

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

$P(\mathbf{x}\mid c_1)$	$P(\mathbf{x}\mid c_2)$		$P(\mathbf{x}\mid c_L)$
Generative Probabilistic Model for Class *1*	**Generative Probabilistic Model for Class *2***	**•••**	**Generative Probabilistic Model for Class *L***
x_1 x_2 ••• x_p	x_1 x_2 ••• x_p		x_1 x_2 ••• x_p

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

Adapt from Prof. Ke Chen NB slides

Generative BC

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

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

Generative Probabilistic Model for Class *1*   **Generative Probabilistic Model for Class *2***   •••   **Generative Probabilistic Model for Class *L***

x_1 x_2 ••• x_p x_1 x_2 ••• x_p x_1 x_2 ••• x_p

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

Adapt from Prof. Ke Chen NB slides

Generative BC

$p(x|c_1), p(x|c_2), \ldots, p(x|c_L)$

$P(\mathbf{X}|C)$,

$C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_p)$ $\Rightarrow p(c|x)$

MAP rule

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

Generative Probabilistic Model for Class *1* … **Generative Probabilistic Model for Class *2*** … **Generative Probabilistic Model for Class *L***

$x_1 \quad x_2 \quad \bullet\bullet\bullet \quad x_p$

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

Adapt from Prof. Ke Chen NB slides

Review : Bayes' Rule – for Generative Bayes Classifiers

$$P(C,X) = P(C|X)P(X) = P(X|C)P(C)$$

$$\Downarrow$$

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

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

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

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

Posterior

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

Prior

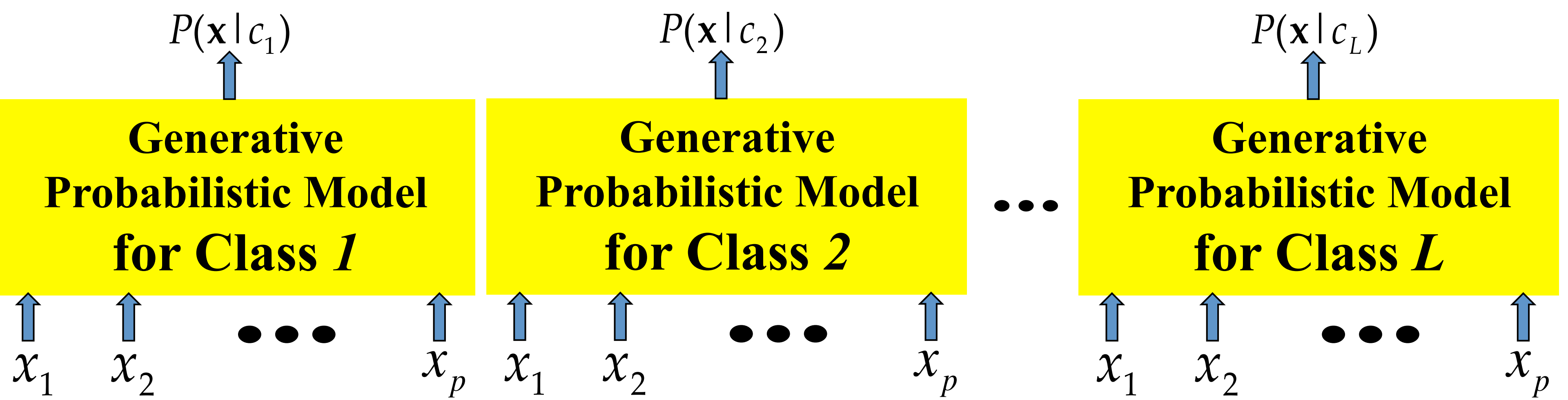
$P(C_1), P(C_2), \ldots, P(C_L)$

$$P(C / \mathbf{X}) = \frac{P(\mathbf{X} / C)P(C)}{P(\mathbf{X})}$$

Establishing a probabilistic model for classification through generative modeling

[MAP rule]

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



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

Adapt from Prof. Ke Chen NB slides

Summary: Generative classification with the MAP rule

- MAP classification rule
  - **MAP**: **M**aximum **A** **P**osterior
  - Assign x to c^* if

$$P(C=c^* \mid \mathbf{X}=\mathbf{x}) > P(C=c \mid \mathbf{X}=\mathbf{x}) \quad c \neq c^*,\ c=c_1,\cdots,c_L$$

Summary: Generative classification with the MAP rule

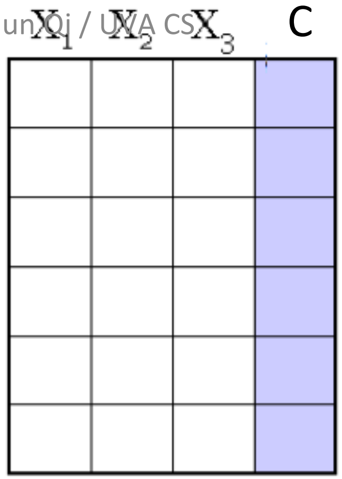
$$P(C=c^* \mid \mathbf{X}=\mathbf{x}) > P(C=c \mid \mathbf{X}=\mathbf{x}) \quad c \neq c^*, \; c = c_1, \cdots, c_L$$

- Generative classification with the MAP rule
  - Apply Bayes rule to convert them into 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

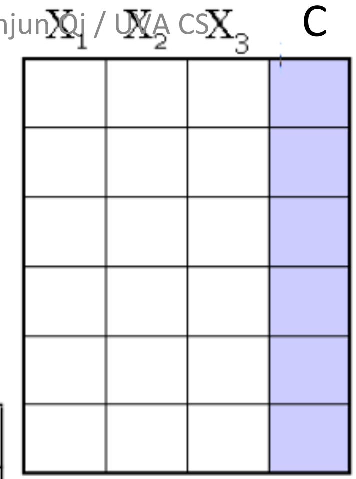


An Example

- 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



An Example

- Example: Play Tennis

X_1 X_2 X_3 X_4 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

$k_2=3$

X_2: {Hot, Mild, Cool}

X_3={High, Normal}

$k_3=2$

X_4=(W, S)

k_4=

C: {Yes, No}

L=2

X_1: {sunny, overc, rain}

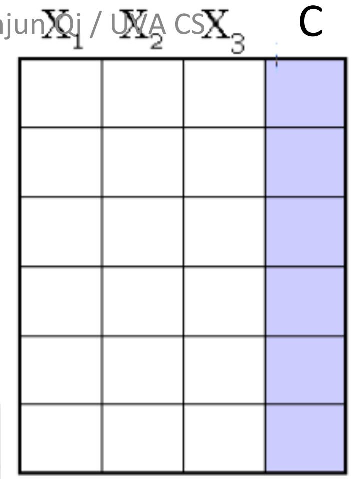
$k_1=3$

$P(X_1, X_2, X_3, X_4 \mid Yes)$

$P(X_1, X_2, X_3, X_4 \mid No)$

36×2

Example



- Example: Play Tennis

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

$P(C = C_i)$

$\{Yes, No\}$

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

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

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

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

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

$$\rightarrow \quad P(C_1=\text{'Yes'}) = 9/14$$

$$P(C_2=\text{No}) = 5/14$$

$$\rightarrow \quad P(X_1, X_2, X_3, X_4 \mid C_i)$$

$$3 \times 3 \times 2 \times 2 \times 2 \Rightarrow 72$$

parameter from train

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

Generative BC

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

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

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

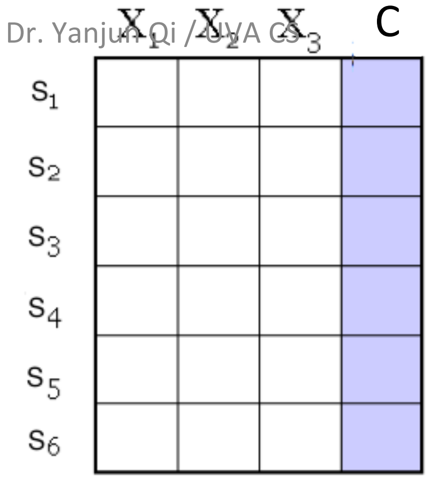
e.g.

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

Check L12-MLE Lecture for Why

Generative Bayes Classifier:

- Learning Phase



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

$P(\text{Play}=Yes) = 9/14 \quad P(\text{Play}=No) = 5/14$

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

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:

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

- Test Phase
  - Given an unknown instance $\mathbf{X}_{ts}' = (a_1', \cdots, a_p')$
  - Look up tables to assign the label c^* to $\mathbf{X}_{ts}$ if

$$\hat{P}(a_1',\cdots a_p'|c^*)\hat{P}(c^*)>\hat{P}(a_1',\cdots a_p'|c)\hat{P}(c),$$
$$c\neq c^*,\ c=c_1,\cdots,c_L$$

Last Page

  - Given a new instance,

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

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

Today : Generative Bayes Classifiers

- ✓ Bayes Classifier
  - ▪ MAP classification rule
  - ▪ Generative Bayes Classifier
- ✓ Naïve Bayes Classifier

Naïve Bayes Classifier

- Bayes classification

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

Difficulty: learning the joint probability

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

Naïve Bayes Classifier

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

- Bayes classification

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

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!

$$
\begin{aligned}
P(X_1, X_2, \cdots, X_p \mid C) &= P(X_1 \mid X_2, \cdots, X_p, C) P(X_2, \cdots, X_p \mid C) \\
&= P(X_1 \mid C) P(X_2, \cdots, X_p \mid C) \\
&= P(X_1 \mid C) P(X_2 \mid C) \cdots P(X_p \mid C)
\end{aligned}
$$

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!

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

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

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

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

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

Naïve Bayes Classifier

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

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

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

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

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

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

$P(X_j | C_i)$

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

Naïve Bayes Classifier (for discrete input attributes) - training

- Naïve Bayes Algorithm (for discrete input attributes)
  - 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

- Naïve Bayes Algorithm (for discrete input attributes)
  - 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**;

$P(c_1)$ $P(c_2)$ .. $P(c_L)$

L parameters

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

Naïve Bayes Classifier (for discrete input attributes) - training

- Naïve Bayes Algorithm (for discrete input attributes)
  - 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**;

L

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**;

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

$K_1, K_2, \dots, K_p$

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

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

Naïve Bayes (for discrete input attributes) - testing

- Naïve Bayes Algorithm (for discrete input attributes)
  - Test Phase: Given an unknown instance $\mathbf{X}' = (a'_1, \cdots, a'_p)$

    Look up tables to assign the label c^* to $\mathbf{X}'$ if

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

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

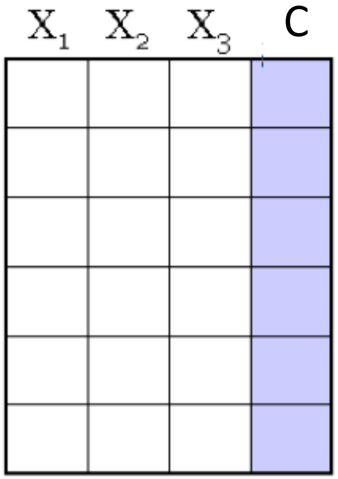
$P(X' | C_i) P(C_i)$

$= P(a'_1 | C_i) P(a'_2 | C_i) \ldots P(a'_p | C_i) P(C_i)$

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

Today : Generative Bayes Classifiers

- ✓ Bayes Classifier
  - MAP classification rule
  - Generative Bayes Classifier
- ✓ Naïve Bayes Classifier
  - ✓ NBC for discrete input variables



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

X_1	X_2	X_3	C

An Example

- Example: Play Tennis

X_1 X_2 X_3 X_4 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

$k_2=3$

X_2: {Hot, Mild, Cool}

X_3 = {High, Normal}

$k_3=2$

X_4 = (W, S)

k_4 =

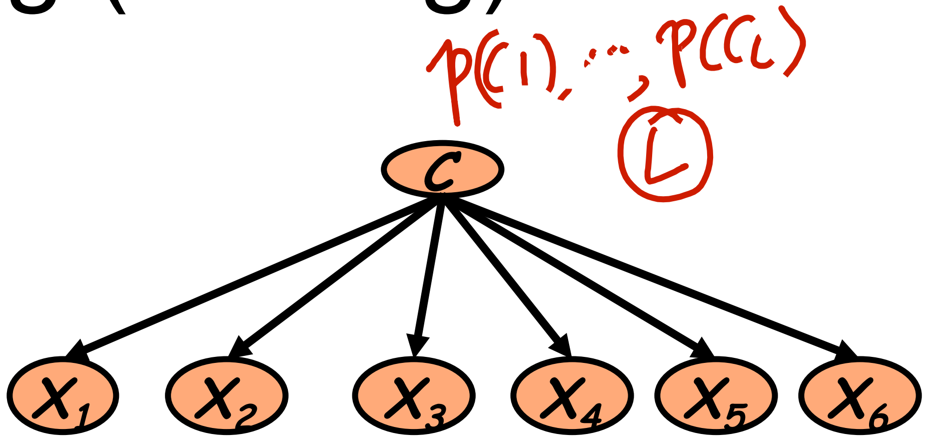
C: {Yes, No}

L=2

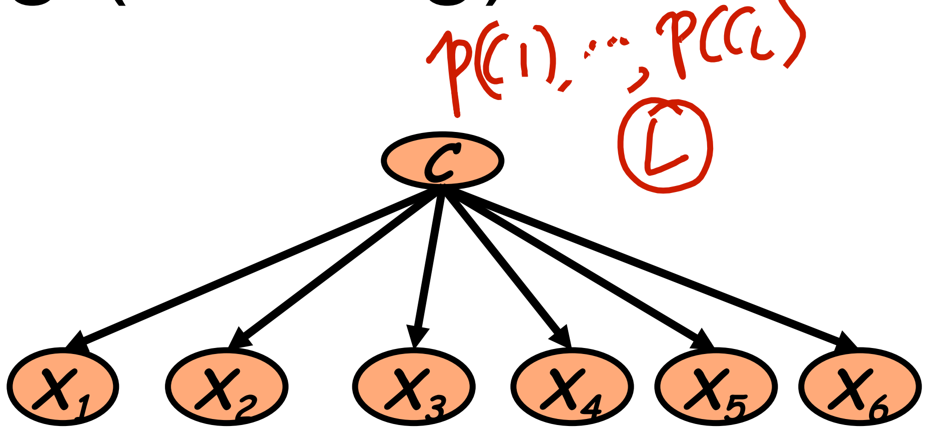
X_1: {sunny, overc, rain}

$k_1=3$

Learning (training) the NBC Model



Learning (training) the NBC Model



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

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

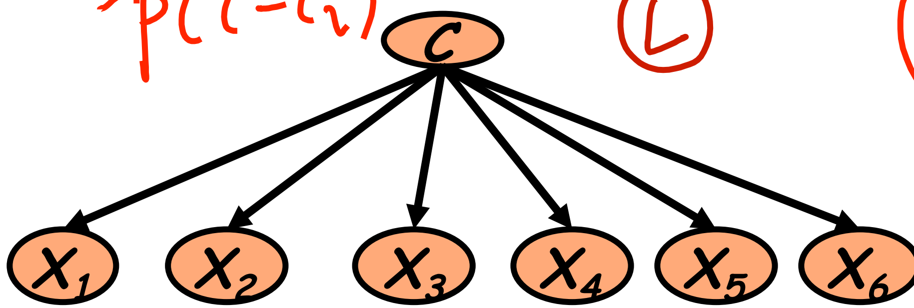
$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

Learning (training) the NBC Model

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

$p(C=c_i)$ L

NBC



$p(X_j = x_j \mid c_i)$

$p(X_i, X_j \mid c)$
$= p(X_i \mid c)\, p(X_j \mid c)$

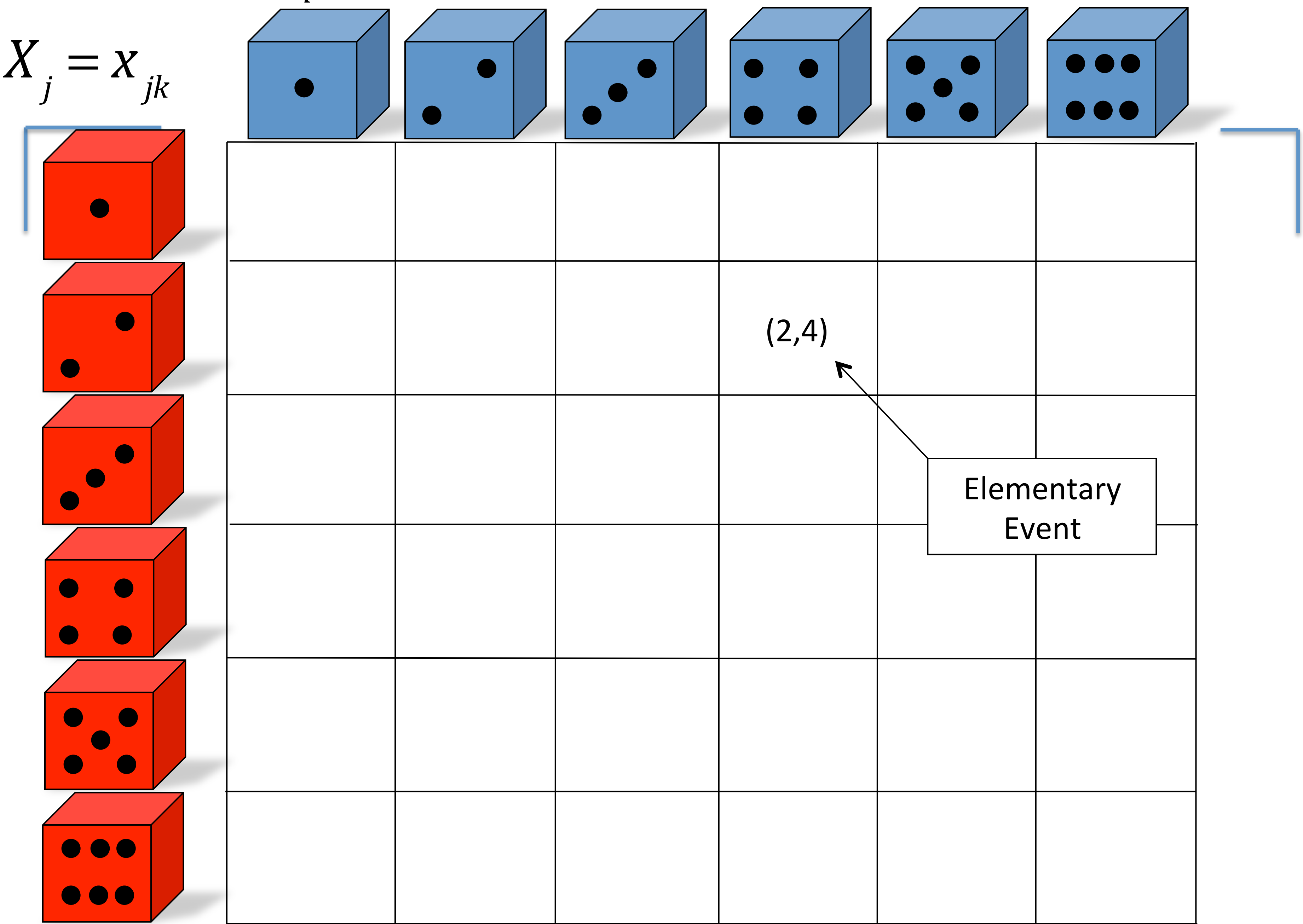
- maximum likelihood estimates
  - simply use the frequencies in the data

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

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

$C = c_i$ 2 – Dimensional

$X_j = x_{jk}$



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=N*o*
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

$P(\text{Play}=Yes) = 9/14$ $P(\text{Play}=No) = 5/14$

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

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

Counting

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

- Learning Phase

$P(X_2|C_1), P(X_2|C_2)$ 2

X_1 3

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

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

$P(\text{Play}=Yes) = 9/14 \qquad P(\text{Play}=No) = 5/14$

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

$P(C_i)$

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*)

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*)

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

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

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

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

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

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

  - 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*".

WHY ? Naïve Bayes Assumption

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

If no naïve assumption

Adapt From Manning' textCat tutorial

WHY ? Naïve Bayes Assumption

Not Naïve

Naïve

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

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

WHY ? Naïve Bayes Assumption

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

Not Naïve

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

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

$\Rightarrow 2^p * L$ Exp

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

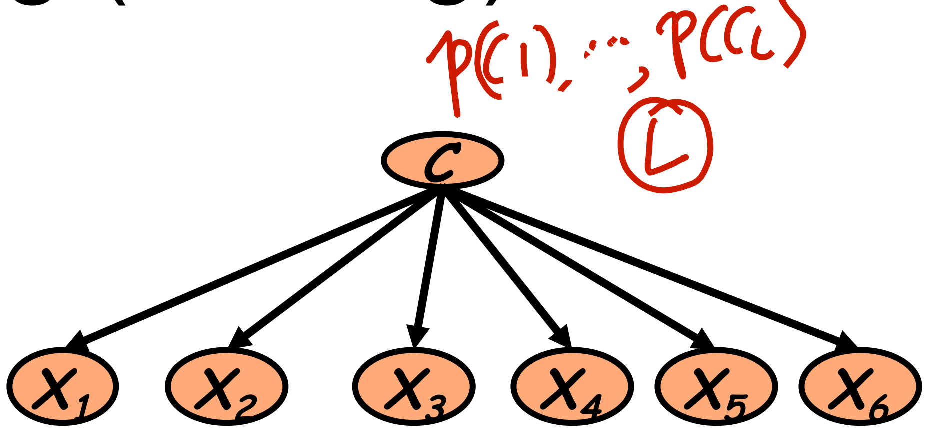
$(2+2+2+\cdots 2) * L$

$= 2 * p * L$ Linear

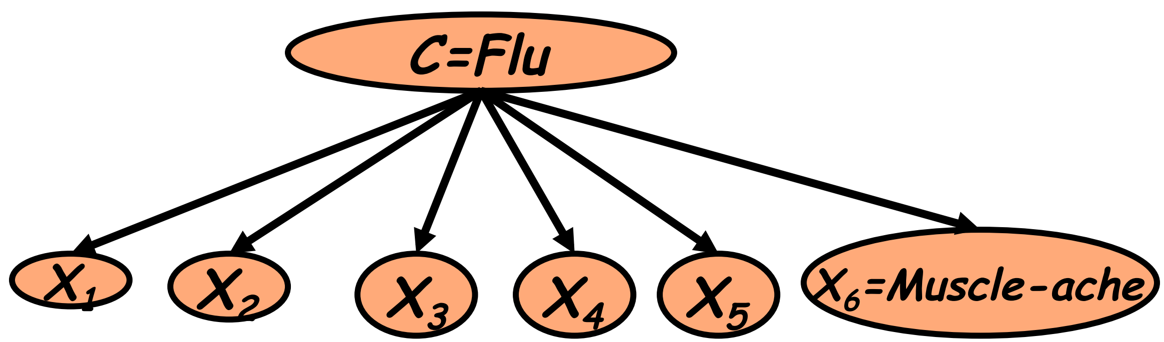
Naïve

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

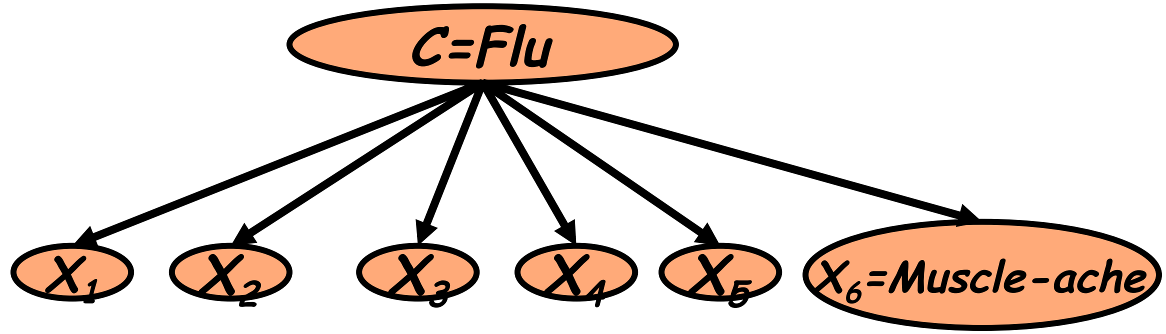
Learning (training) the NBC Model



For instance:



For instance:



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

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

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

muscle-ache=Yes/no flu/nf

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

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

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

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 $X_i = x_{i,k}$

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

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

extent of "smoothing"

Summary: Generative Bayes Classifier with the MAP rule

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

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

Generative Bayes Classifier

Task	classification
Representation	Prob. models p(X|C)
Score Function	EPE with 0-1 loss → likelihood
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