

CS60050

Machine Learning

Introduction to Bayesian Classifiers

Somak Aditya

Sudeshna Sarkar

Department of CSE, IIT Kharagpur

Sep 8, 2023

Outline of lectures

Outline of Maximum likelihood estimation

2 examples of Bayesian classifiers:

- Naïve Bayes
- Logistic regression

Recap: Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$



- Why is this at all helpful?
 - Let's us build one conditional from its reverse
 - Often one conditional is tricky but the other one is simple
 - Foundation of many practical systems (e.g. ASR, MT)
- In the running for most important ML equation!

Returning to thumbtack

example...

- $P(\text{Heads}) = \theta$, $P(\text{Tails}) = 1 - \theta$



- Flips are *i.i.d.*: $D = \{x_i \mid i=1 \dots n\}$, $P(D \mid \theta) = \prod_i P(x_i \mid \theta)$
 - Independent events
 - Identically distributed according to Bernoulli distribution
- Sequence D of α_H Heads and α_T Tails
$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

Called the “likelihood” of the data under the model

Maximum Likelihood Estimation

- **Data:** Observed set D of α_H Heads and α_T Tails
- **Hypothesis:** Bernoulli distribution
- **Learning:** finding θ is an optimization problem
 - What's the objective function?

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

- **MLE:** Choose θ to maximize probability of D

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(\mathcal{D} \mid \theta) \\ &= \arg \max_{\theta} \ln P(\mathcal{D} \mid \theta)\end{aligned}$$

Your first parameter learning algorithm

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \ln P(\mathcal{D} \mid \theta) \\ &= \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero, and solve!

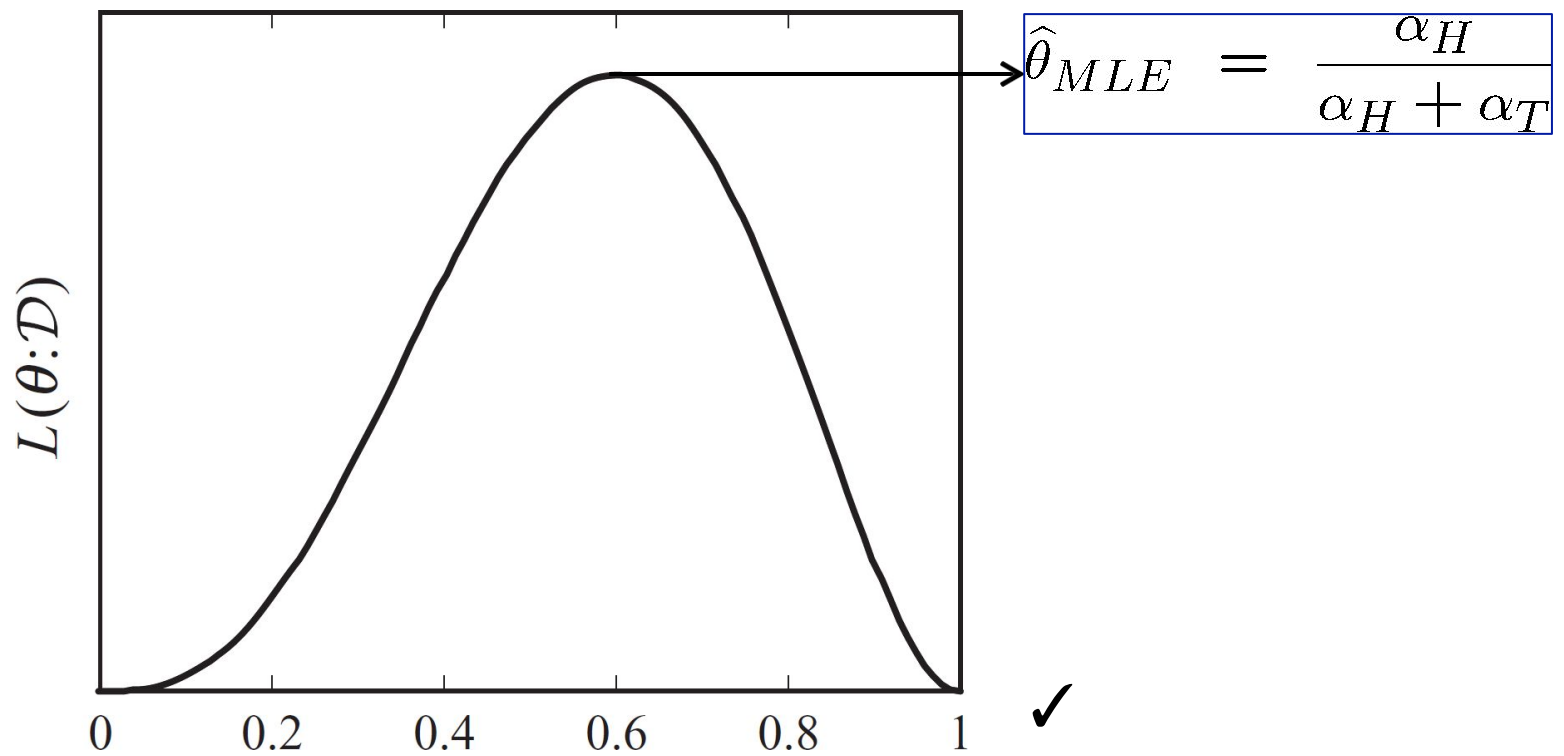
$$\begin{aligned}\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) &= \frac{d}{d\theta} [\ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}] \\ &= \frac{d}{d\theta} [\alpha_H \ln \theta + \alpha_T \ln(1 - \theta)] \\ &= \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1 - \theta) \\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1 - \theta} = 0\end{aligned}$$

$$\boxed{\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}}$$

Data

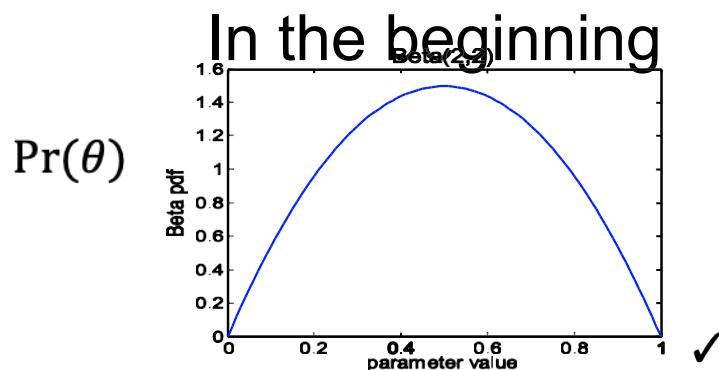


$$L(\theta; D) = \ln P(D|\theta)$$



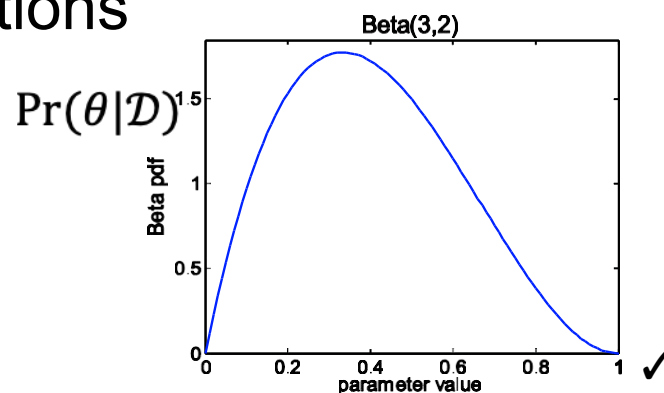
What if I have prior beliefs?

- Billionaire says: Wait, I know that the thumbtack is “close” to 50-50. What can you do for me now?
- **You say: I can learn it the Bayesian way...**
- Rather than estimating a single θ , we obtain a distribution over possible values of θ



After observations

Observe flips
e.g.: {tails, tails}



Bayesian Learning

- Use Bayes' rule!

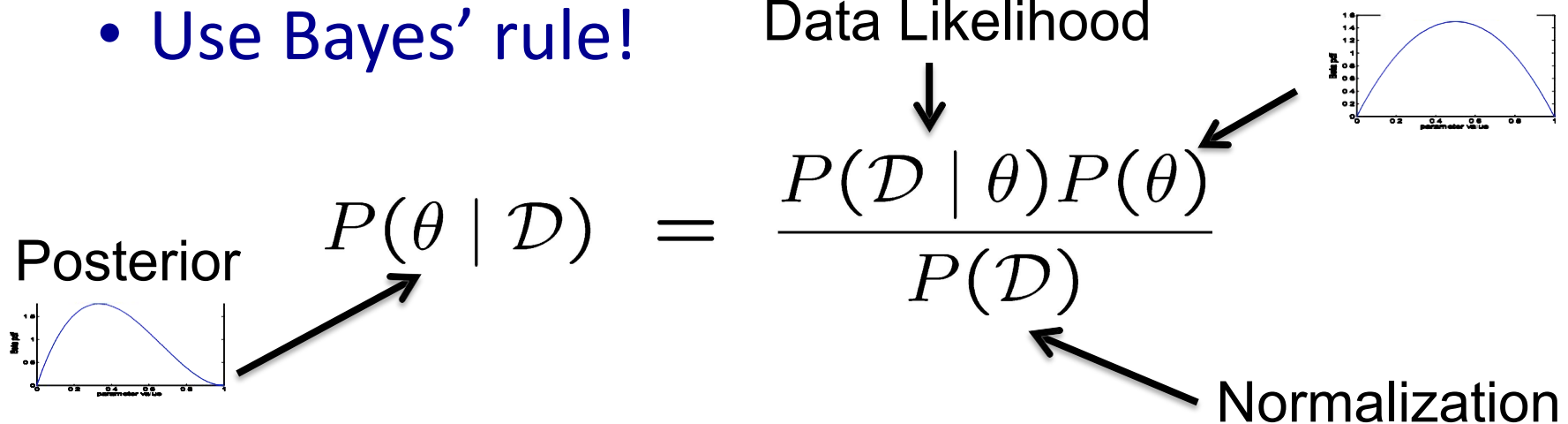
Diagram illustrating Bayes' rule for parameter estimation:

Data Likelihood (arrow pointing to $P(\mathcal{D} | \theta)$)

Prior (arrow pointing to $P(\theta)$)

Posterior (arrow pointing to $P(\theta | \mathcal{D})$)

Normalization (arrow pointing to $P(\mathcal{D})$)

$$P(\theta | \mathcal{D}) = \frac{P(\mathcal{D} | \theta) P(\theta)}{P(\mathcal{D})}$$


- Or equivalently: $P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta) P(\theta)$
- For *uniform* priors, this reduces to maximum likelihood estimation!

$$P(\theta) \propto 1 \quad P(\theta | \mathcal{D}) \propto P(\mathcal{D} | \theta)$$

Bayesian Learning for Thumbtacks

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

Likelihood: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$

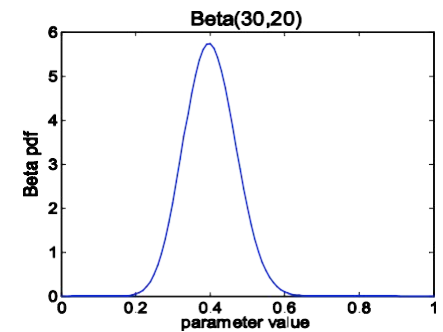
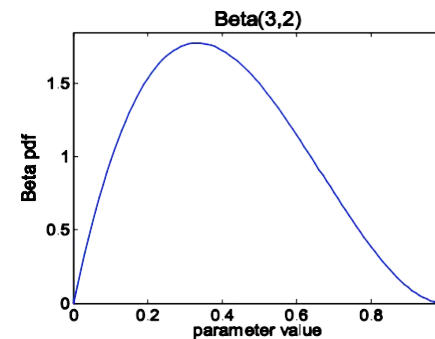
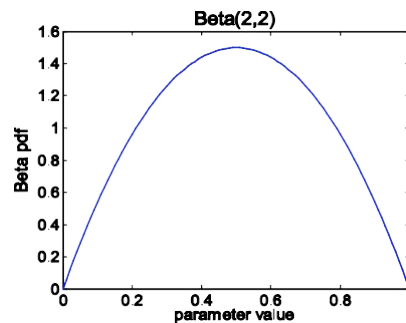
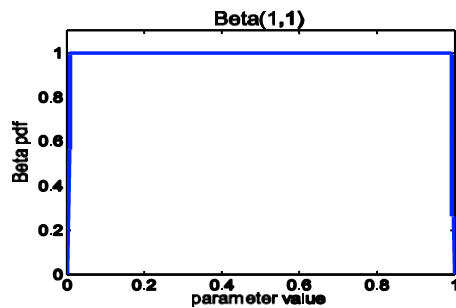
- What should the prior be?
 - Represent expert knowledge
 - Simple posterior form
- For binary variables, commonly used prior is the

Beta distribution:

$$P(\theta) = \frac{\theta^{\beta_H-1} (1 - \theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$

Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T)$$



- Since the Beta distribution is *conjugate* to the Bernoulli distribution, the posterior distribution has a particularly simple form:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

$$\propto \theta^{\alpha_H}(1-\theta)^{\alpha_T} \theta^{\beta_H-1}(1-\theta)^{\beta_T-1}$$

$$= \theta^{\alpha_H+\beta_H-1} (1-\theta)^{\alpha_T+\beta_T-1}$$

$$= \text{Beta}(\alpha_H+\beta_H, \alpha_T+\beta_T)$$

Using Bayesian inference for prediction

- We now have a **distribution** over parameters
- For any specific f , a function of interest, compute the expected value of f :

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid \mathcal{D}) d\theta$$

- Integral is often hard to compute
 - *As more data is observed, posterior is more concentrated*
- **MAP (Maximum a posteriori approximation)**: use most likely parameter to approximate the expectation

$$\hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D})$$

$$E[f(\theta)] \approx f(\hat{\theta})$$

Bayesian Classification

- Problem statement:

- Given features

$$X_1, X_2, \dots, X_n$$

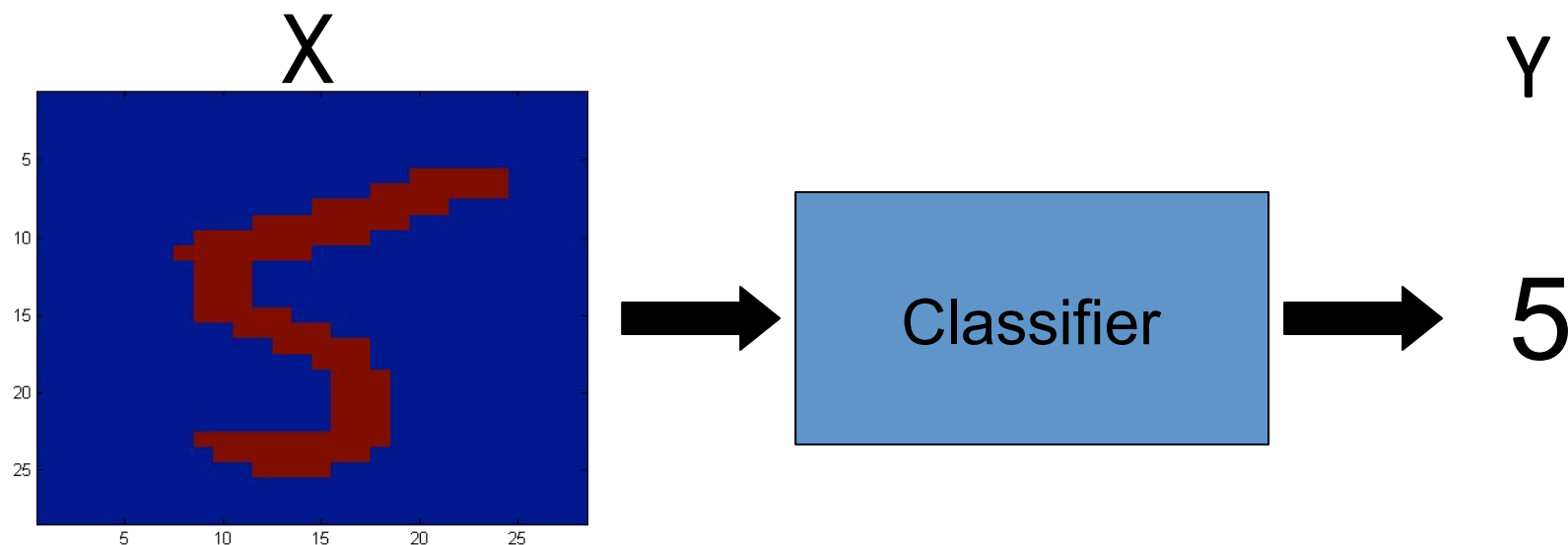
- Predict a label Y

[Next several slides adapted from:

Vibhav Gogate, Jonathan Huang, Luke Zettlemoyer, Carlos Guestrin, and Dan Weld]

Example Application

- **Digit Recognition**



- $X_1, \dots, X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{0,1,2,3,4,5,6,7,8,9\}$

The Bayes Classifier

- If we had the joint distribution on $\mathbf{X}_1, \dots, \mathbf{X}_n$ and \mathbf{Y} , could predict using:

$$\arg \max_Y P(Y | X_1, \dots, X_n)$$

– (for example: what is the probability that the image represents a 5 given its pixels?)

- So ... How do we compute that?

The Bayes Classifier

- Use Bayes Rule!

$$P(Y|X_1, \dots, X_n) = \frac{\overset{\text{Likelihood}}{P(X_1, \dots, X_n|Y)} \overset{\text{Prior}}{P(Y)}}{\underset{\text{Normalization Constant}}{P(X_1, \dots, X_n)}}$$

- Why did this help? Well, we think that we might be able to specify how features are “generated” by the class label

The Bayes Classifier

- Let's expand this for our digit recognition task:

$$P(Y = 5|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n|Y = 5)P(Y = 5)}{P(X_1, \dots, X_n|Y = 5)P(Y = 5) + P(X_1, \dots, X_n|Y = 6)P(Y = 6)}$$
$$P(Y = 6|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n|Y = 6)P(Y = 6)}{P(X_1, \dots, X_n|Y = 5)P(Y = 5) + P(X_1, \dots, X_n|Y = 6)P(Y = 6)}$$

- To classify, we'll simply compute these probabilities, one per class, and predict based on which one is largest

Model

Parameters

- How many parameters are required to specify the likelihood, $P(X_1, \dots, X_n|Y)$?
 - (Supposing that each image is 30x30 pixels)
- The problem with explicitly modeling $P(X_1, \dots, X_n|Y)$ is that there are usually way too many parameters:
 - We'll run out of space
 - We'll run out of time
 - And we'll need tons of training data (which is usually not available)

Naïve

Bayes

- Naïve Bayes assumption:

- Features are independent given class:

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

- More generally:

$$P(X_1 \dots X_n|Y) = \prod_i P(X_i|Y)$$

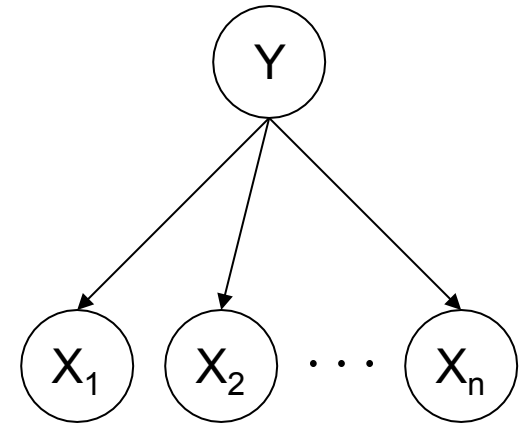
- How many parameters now?

- Suppose \mathbf{X} is composed of n binary features

The Naïve Bayes Classifier

- Given:

- Prior $P(Y)$
- n conditionally independent features X_1, \dots, X_n , given the class Y
- For each feature i , we specify $P(X_i|Y)$



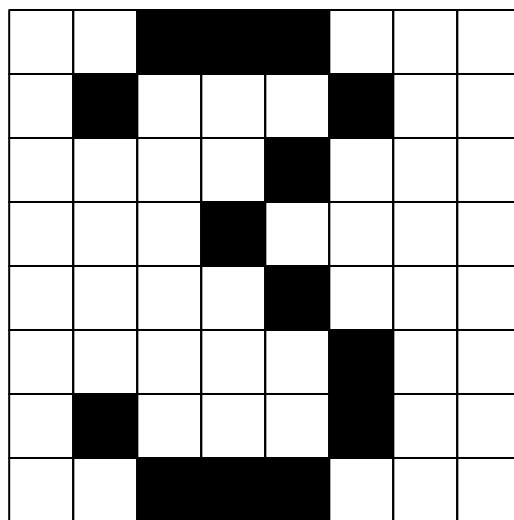
- Classification decision rule:

$$\begin{aligned} y^* = h_{NB}(\mathbf{x}) &= \arg \max_y P(y) P(x_1, \dots, x_n | y) \\ &= \arg \max_y P(y) \prod_i P(x_i | y) \end{aligned}$$

If certain assumption holds, NB is optimal classifier!
(they typically don't)

A Digit Recognizer

- Input: pixel grids



- Output: a digit 0-9

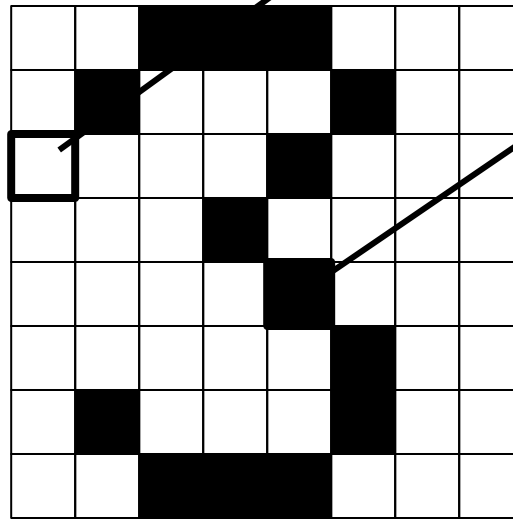
Are the naïve Bayes assumptions realistic here?



What has to be learned?

$P(Y)$

1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



$P(F_{3,1} = on|Y)$

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

$P(F_{5,5} = on|Y)$

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

MLE for the parameters of NB

- Given dataset
 - $\text{Count}(A=a, B=b)$ □ number of examples where $A=a$ and $B=b$
- MLE for discrete NB, simply:
 - Prior:

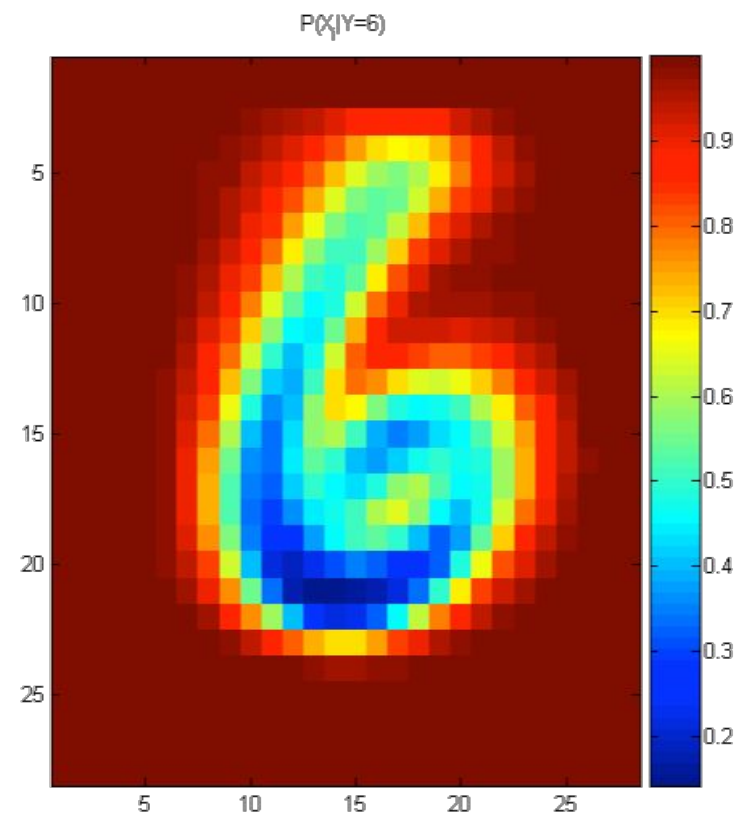
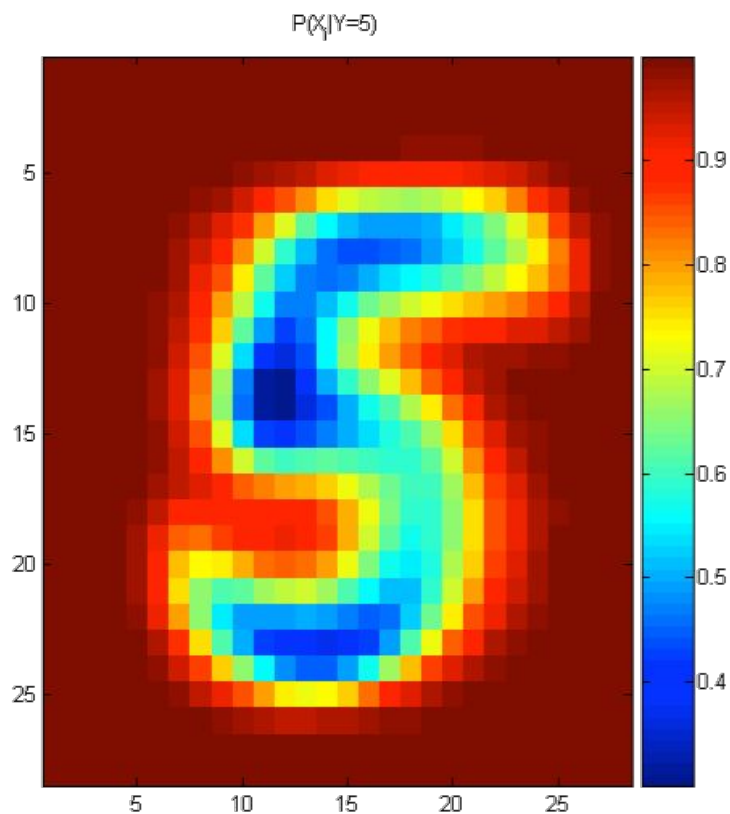
$$P(Y = y) = \frac{\text{Count}(Y = y)}{\sum_{y'} \text{Count}(Y = y')}$$

- Observation distribution:

$$P(X_i = x | Y = y) = \frac{\text{Count}(X_i = x, Y = y)}{\sum_{x'} \text{Count}(X_i = x', Y = y)}$$

MLE for the parameters of NB

- Training amounts to, for each of the classes, averaging all of the examples together:



Using the Naïve Bayes Classifier

- Now, we have

$$P(Y = y_k \mid X = \mathbf{x}_i) = \frac{P(Y = y_k) \prod_{j=1}^d P(X_j = x_{i,j} \mid Y = y_k)}{P(X = \mathbf{x}_i)}$$

This is constant for a given instance,
and so irrelevant to our prediction

- In practice, we use log-probabilities to prevent underflow

- To classify a new point \mathbf{x} ,

$$\begin{aligned} h(\mathbf{x}) &= \arg \max_{y_k} P(Y = y_k) \prod_{j=1}^d P(X_j = \underbrace{x_j}_{j^{\text{th}} \text{ attribute value of } \mathbf{x}} \mid Y = y_k) \\ &= \arg \max_{y_k} \log P(Y = y_k) + \sum_{j=1}^d \log P(X_j = x_j \mid Y = y_k) \end{aligned}$$

The Naïve Bayes Classifier Algorithm

- For each class label y_k
 - Estimate $P(Y = y_k)$ from the data
 - For each value $x_{i,j}$ of each attribute X_i
 - Estimate $P(X_i = x_{i,j} | Y = y_k)$

- Classify a new point via:

$$h(\mathbf{x}) = \arg \max_{y_k} \log P(Y = y_k) + \sum_{j=1}^d \log P(X_j = x_j | Y = y_k)$$

- In practice, the independence assumption doesn't often hold true, but Naïve Bayes performs very well despite this

Naïve Bayes: Subtlety #1

Often the X_i are not really conditionally independent

- We use Naïve Bayes in many cases anyway, and it often works pretty well
 - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])
- What is effect on estimated $P(Y|X)$?
 - Extreme case: what if we add two copies: $X_i = X_k$

Naïve Bayes: Subtlety 2 Zero Counting

- Notice that some probabilities estimated by counting might be zero
 - Possible overfitting!
- Fix by using Laplace Smoothing
 - Adds 1 to each count

$$P(X_j = v \mid Y = y_k) = \frac{c_v + 1}{\sum_{v' \in \text{values}(X_j)} c_{v'} + |\text{values}(X_j)|}$$

where

- c_v is the count of training instances with a value of v for attribute j and class label y_k
- $|\text{values}(X_j)|$ is the number of values X_j can take on

MAP estimation for NB

- Given dataset

- $\text{Count}(A=a, B=b)$ □ number of examples where $A=a$ and $B=b$

- MAP estimation for discrete NB, simply:

- Prior:

$$P(Y = y) = \frac{\text{Count}(Y = y)}{\sum_{y'} \text{Count}(Y = y')}$$

- Observation distribution:

$$P(X_i = x | Y = y) = \frac{\text{Count}(X_i = x, Y = y) + \mathbf{a}}{\sum_{x'} \text{Count}(X_i = x', Y = y) + |\mathbf{X}_i| * \mathbf{a}}$$

- Called “smoothing”. Corresponds to Dirichlet prior!

Training Naïve Bayes (Example 2)

Estimate $P(X_j | Y)$ and $P(Y)$ directly from the training data by counting!

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Wind</u>	<u>Water</u>	<u>Forecast</u>	<u>Play?</u>
sunny	warm	normal	strong	warm	same	yes
sunny	warm	high	strong	warm	same	yes
rainy	cold	high	strong	warm	change	no
sunny	warm	high	strong	cool	change	yes

$P(\text{play}) = ?$

$P(\text{Sky} = \text{sunny} \mid \text{play}) = ?$

$P(\text{Humid} = \text{high} \mid \text{play}) = ?$

...

$P(\neg \text{play}) = ?$

$P(\text{Sky} = \text{sunny} \mid \neg \text{play}) = ?$

$P(\text{Humid} = \text{high} \mid \neg \text{play}) = ?$

...

Training Naïve Bayes with Laplace Smoothing

Estimate $P(X_j | Y)$ and $P(Y)$ directly from the training data by counting with Laplace smoothing:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Wind</u>	<u>Water</u>	<u>Forecast</u>	<u>Play?</u>
sunny						yes
sunny						yes
rainy	cold	high	strong	warm	change	no
sunny						yes

$$P(\text{play}) = 3/4 \quad P(\neg \text{play}) = 1/4$$

$$P(\text{Sky} = \text{sunny} | \text{play}) = 4/5$$

$$P(\text{Humid} = \text{high} | \text{play}) = ?$$

...

$$P(\text{Sky} = \text{sunny} | \neg \text{play}) = ?$$

$$P(\text{Humid} = \text{high} | \neg \text{play}) = ?$$

...

Training Naïve Bayes with Laplace Smoothing

Estimate $P(X_j | Y)$ and $P(Y)$ directly from the training data by counting with Laplace smoothing:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Wind</u>	<u>Water</u>	<u>Forecast</u>	<u>Play?</u>
sunny	warm	normal	strong	warm	same	yes
sunny	warm	high	strong	warm	same	yes
rainy						no
sunny	warm	high	strong	cool	change	yes

$$P(\text{play}) = \frac{3}{4}$$

$$P(\neg \text{play}) = \frac{1}{4}$$

$$P(\text{Sky} = \text{sunny} | \text{play}) = \frac{4}{5}$$

$$P(\text{Sky} = \text{sunny} | \neg \text{play}) = \frac{1}{3}$$

$$P(\text{Humid} = \text{high} | \text{play}) = ?$$

$$P(\text{Humid} = \text{high} | \neg \text{play}) = ?$$

...

...

Training Naïve Bayes with Laplace Smoothing

Estimate $P(X_j | Y)$ and $P(Y)$ directly from the training data by counting with Laplace smoothing:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Wind</u>	<u>Water</u>	<u>Forecast</u>	<u>Play?</u>
		normal				yes
		high				yes
rainy	cold	high	strong	warm	change	no
		high				yes

$$P(\text{play}) = \frac{3}{4}$$

$$P(\neg \text{play}) = \frac{1}{4}$$

$$P(\text{Sky} = \text{sunny} | \text{play}) = \frac{4}{5}$$

$$P(\text{Humid} = \text{high} | \text{play}) = \frac{3}{5}$$

...

$$P(\text{Sky} = \text{sunny} | \neg \text{play}) = \frac{1}{3}$$

$$P(\text{Humid} = \text{high} | \neg \text{play}) = ?$$

...

Training Naïve Bayes with Laplace Smoothing

Estimate $P(X_j | Y)$ and $P(Y)$ directly from the training data by counting with Laplace smoothing:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Wind</u>	<u>Water</u>	<u>Forecast</u>	<u>Play?</u>
sunny	warm	normal	strong	warm	same	yes
sunny	warm	high	strong	warm	same	yes
		<u>high</u>				
sunny	warm	high	strong	cool	change	yes

$$P(\text{play}) = \frac{3}{4}$$

$$P(\neg \text{play}) = \frac{1}{4}$$

$$P(\text{Sky} = \text{sunny} | \text{play}) = \frac{4}{5}$$

$$P(\text{Sky} = \text{sunny} | \neg \text{play}) = \frac{1}{3}$$

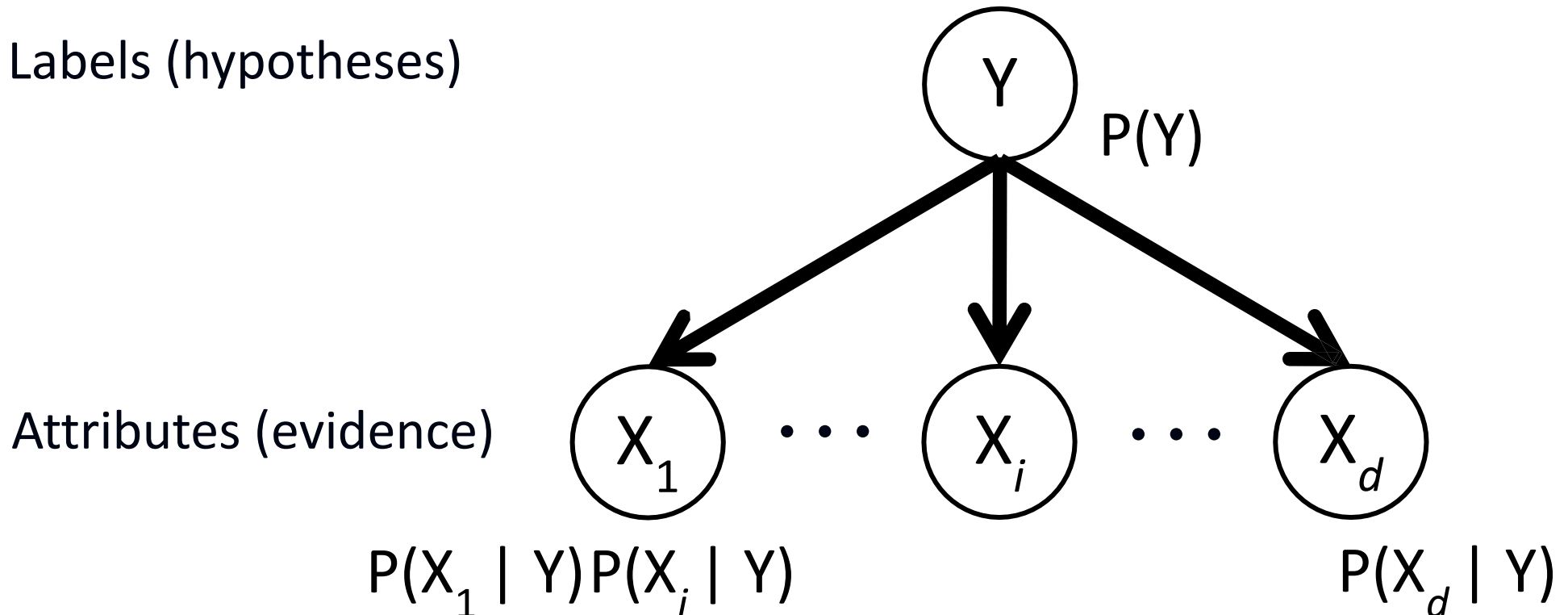
$$P(\text{Humid} = \text{high} | \neg \text{play}) = \frac{2}{3}$$

...

...

Extra Slides with Full Example

The Naïve Bayes Graphical Model

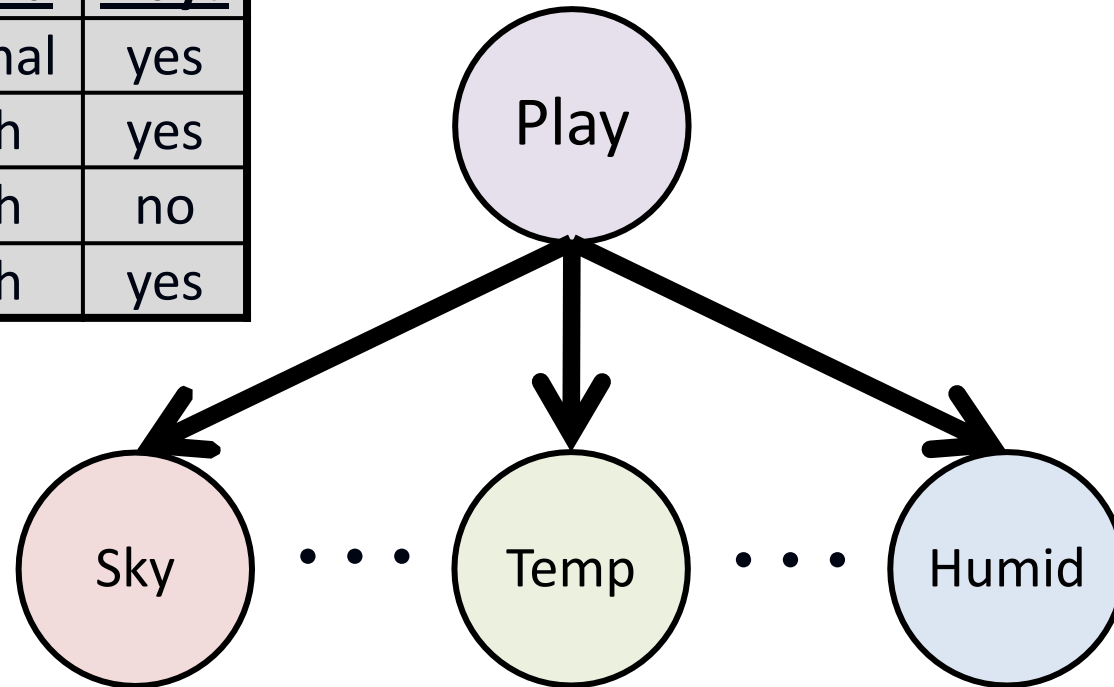


- Nodes denote random variables
- Edges denote dependency
- Each node has an associated conditional probability table (CPT), conditioned upon its parents

Example NB Graphical Model

Data:

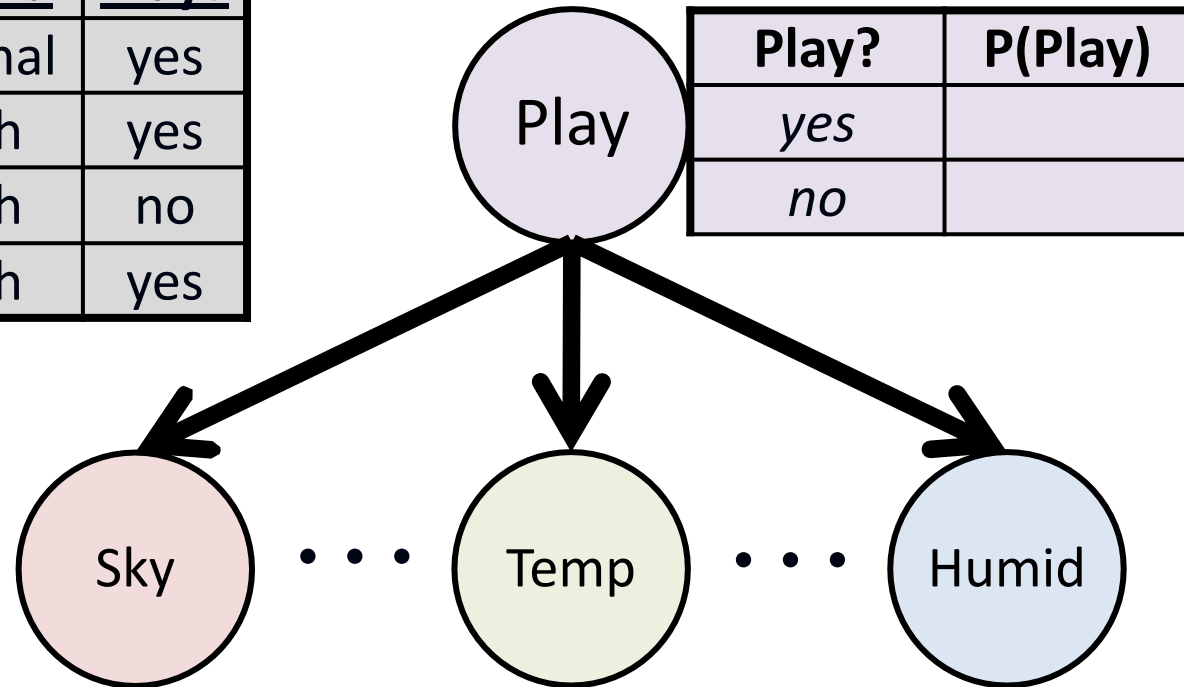
<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



Example NB Graphical Model

Data:

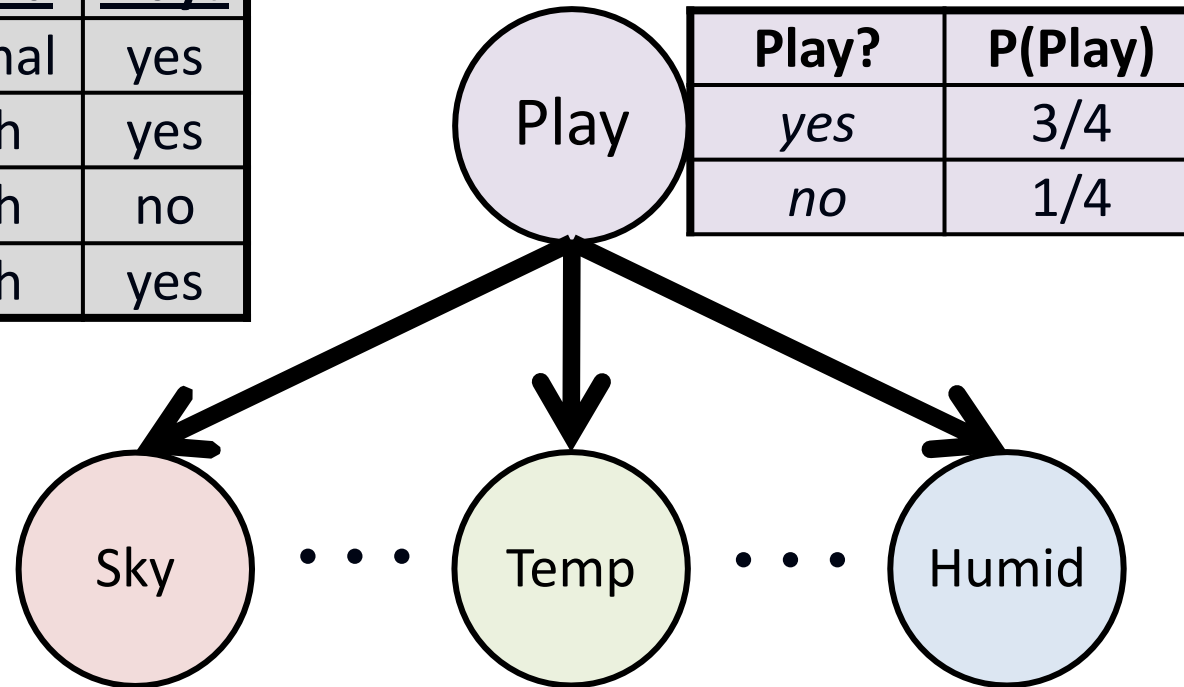
<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



Example NB Graphical Model

Data:

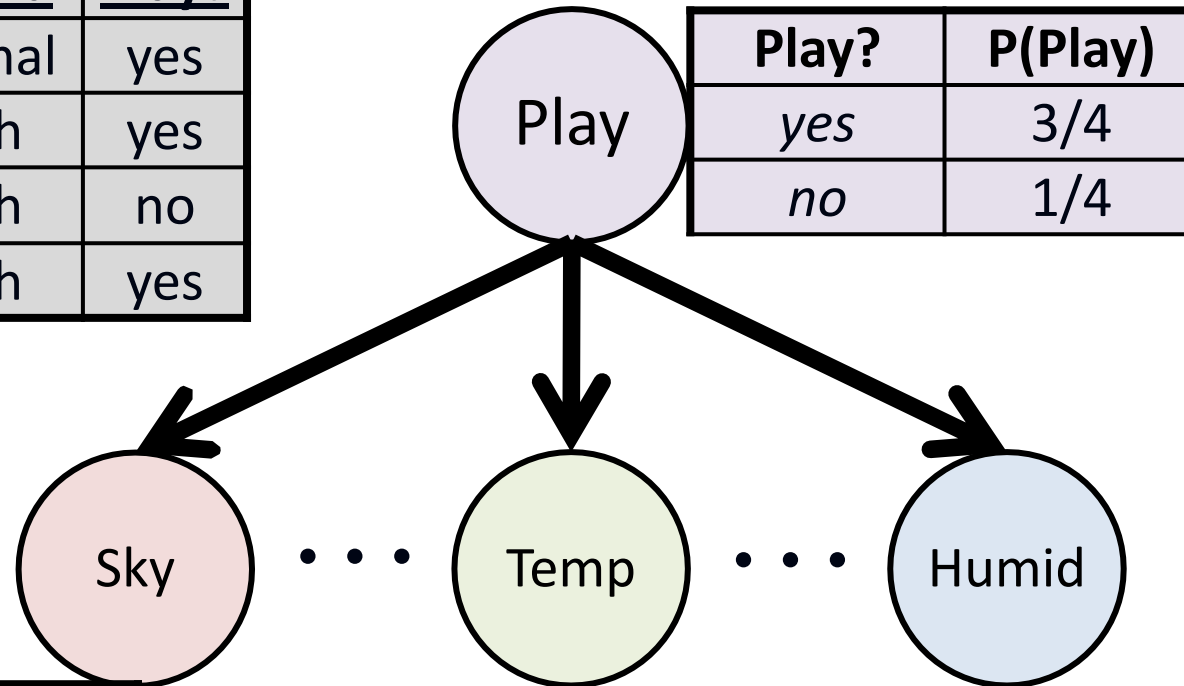
<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



Example NB Graphical Model

Data:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



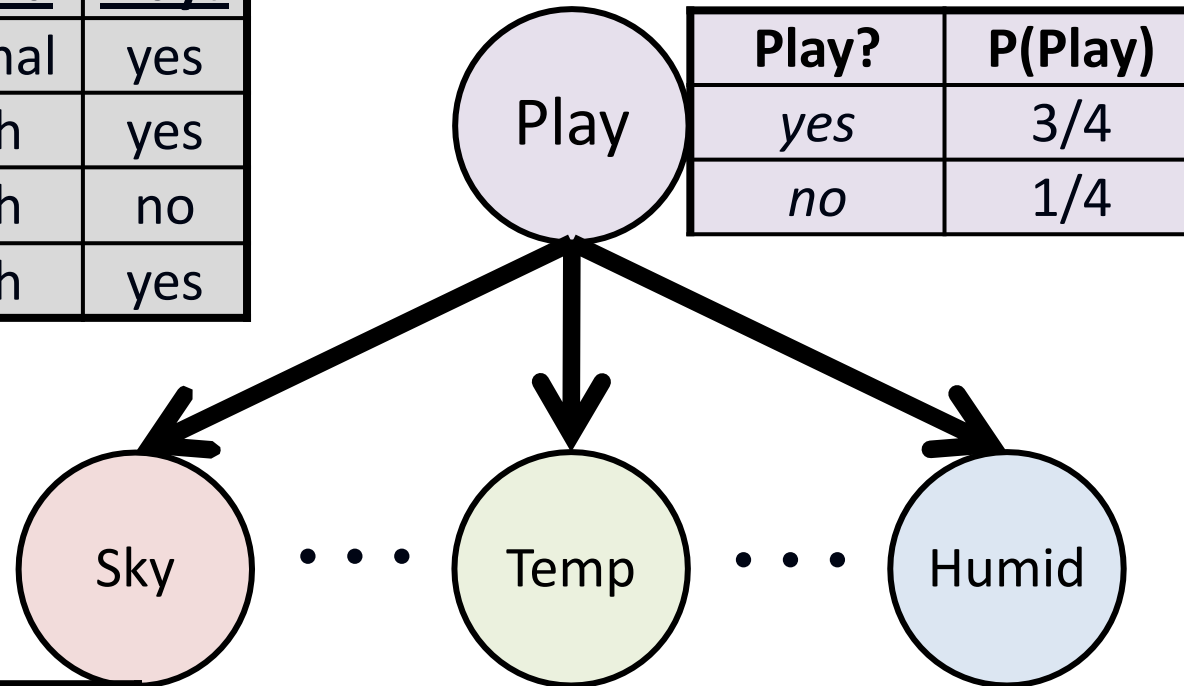
Play?	P(Play)
<i>yes</i>	3/4
<i>no</i>	1/4

Sky	Play?	P(Sky Play)
<i>sunny</i>	<i>yes</i>	
<i>rainy</i>	<i>yes</i>	
<i>sunny</i>	<i>no</i>	
<i>rainy</i>	<i>no</i>	

Example NB Graphical Model

Data:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



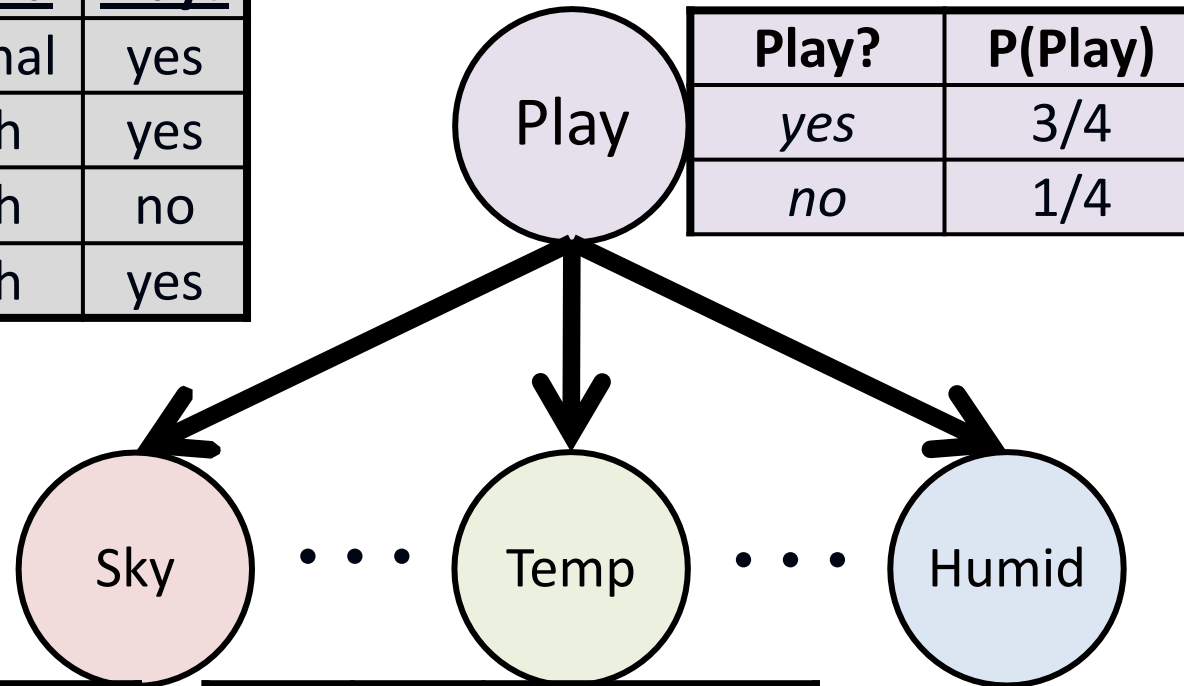
Play?	P(Play)
<i>yes</i>	3/4
<i>no</i>	1/4

Sky	Play?	P(Sky Play)
<i>sunny</i>	<i>yes</i>	4/5
<i>rainy</i>	<i>yes</i>	1/5
<i>sunny</i>	<i>no</i>	1/3
<i>rainy</i>	<i>no</i>	2/3

Example NB Graphical Model

Data:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



Play?	P(Play)
<i>yes</i>	3/4
<i>no</i>	1/4

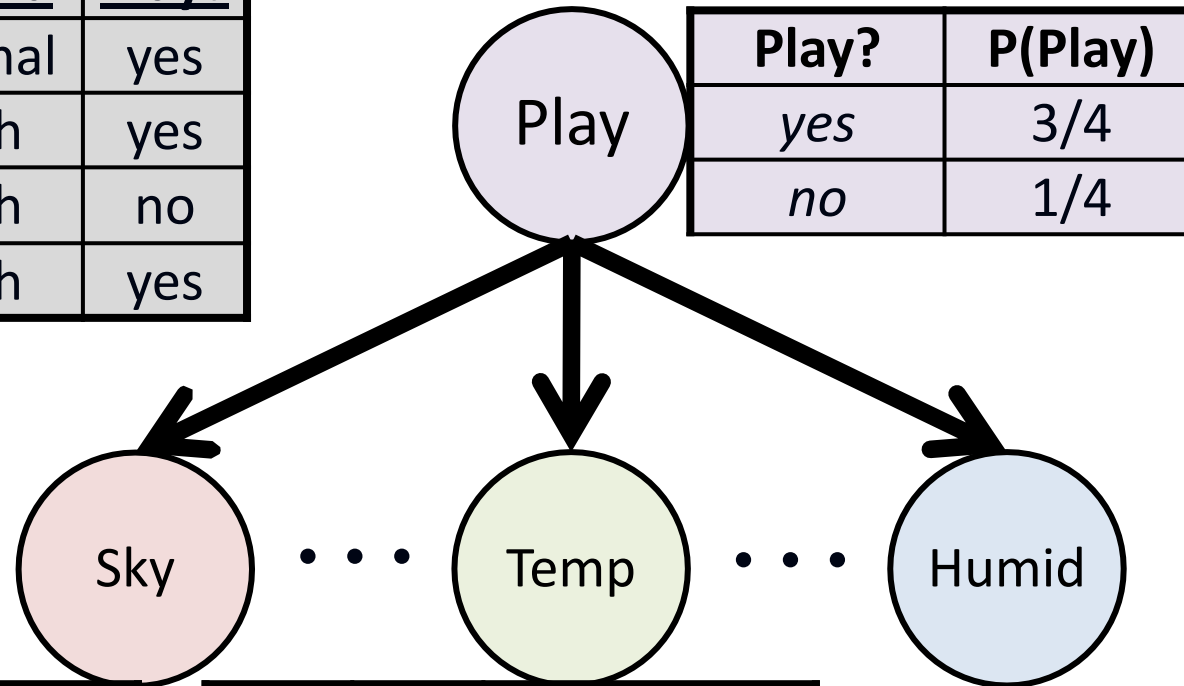
Sky	Play?	P(Sky Play)
<i>sunny</i>	<i>yes</i>	4/5
<i>rainy</i>	<i>yes</i>	1/5
<i>sunny</i>	<i>no</i>	1/3
<i>rainy</i>	<i>no</i>	2/3

Temp	Play?	P(Temp Play)
<i>warm</i>	<i>yes</i>	
<i>cold</i>	<i>yes</i>	
<i>warm</i>	<i>no</i>	
<i>cold</i>	<i>no</i>	

Example NB Graphical Model

Data:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



Play?	P(Play)
<i>yes</i>	3/4
<i>no</i>	1/4

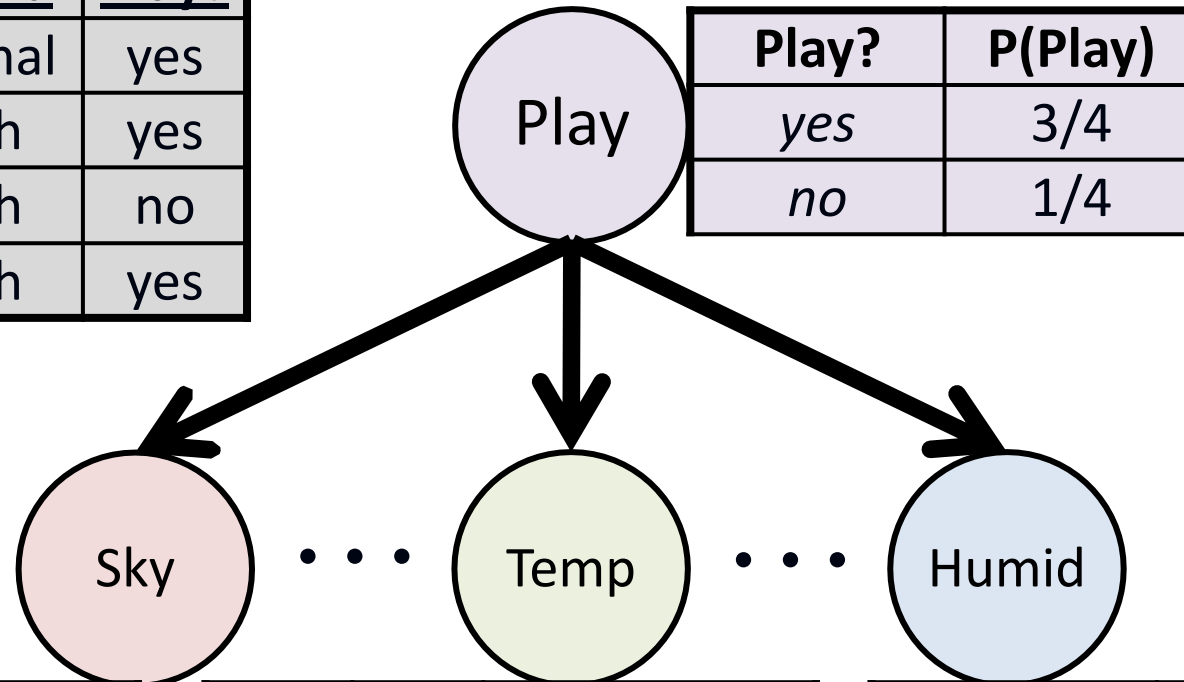
Sky	Play?	P(Sky Play)
<i>sunny</i>	<i>yes</i>	4/5
<i>rainy</i>	<i>yes</i>	1/5
<i>sunny</i>	<i>no</i>	1/3
<i>rainy</i>	<i>no</i>	2/3

Temp	Play?	P(Temp Play)
<i>warm</i>	<i>yes</i>	4/5
<i>cold</i>	<i>yes</i>	1/5
<i>warm</i>	<i>no</i>	1/3
<i>cold</i>	<i>no</i>	2/3

Example NB Graphical Model

Data:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



Play?	P(Play)
<i>yes</i>	3/4
<i>no</i>	1/4

Sky	Play?	P(Sky Play)
<i>sunny</i>	<i>yes</i>	4/5
<i>rainy</i>	<i>yes</i>	1/5
<i>sunny</i>	<i>no</i>	1/3
<i>rainy</i>	<i>no</i>	2/3

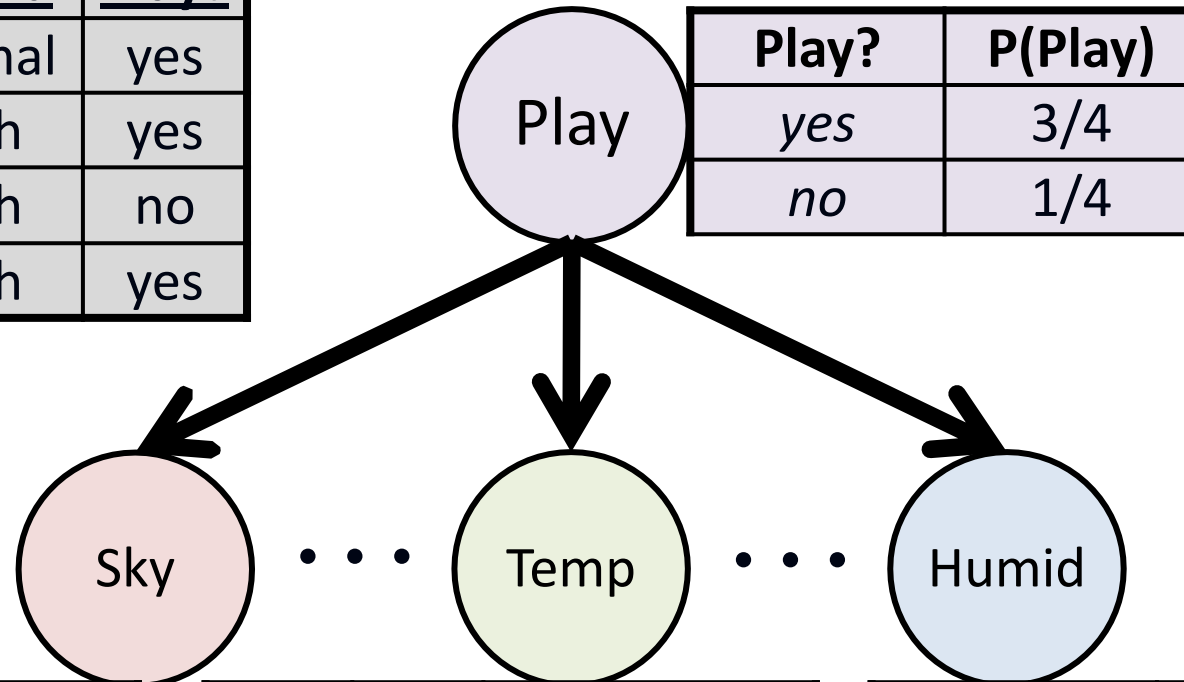
Temp	Play?	P(Temp Play)
<i>warm</i>	<i>yes</i>	4/5
<i>cold</i>	<i>yes</i>	1/5
<i>warm</i>	<i>no</i>	1/3
<i>cold</i>	<i>no</i>	2/3

Humid	Play?	P(Humid Play)
<i>high</i>	<i>yes</i>	
<i>norm</i>	<i>yes</i>	
<i>high</i>	<i>no</i>	
<i>norm</i>	<i>no</i>	

Example NB Graphical Model

Data:

<u>Sky</u>	<u>Temp</u>	<u>Humid</u>	<u>Play?</u>
sunny	warm	normal	yes
sunny	warm	high	yes
rainy	cold	high	no
sunny	warm	high	yes



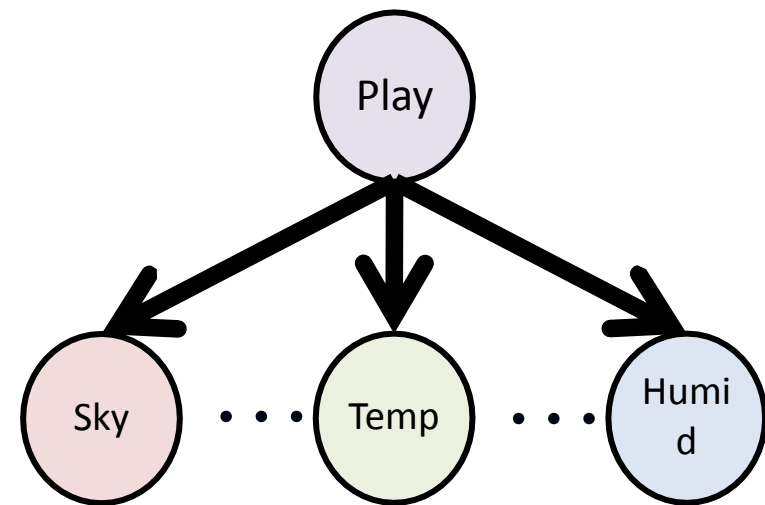
Play?	P(Play)
<i>yes</i>	3/4
<i>no</i>	1/4

Sky	Play?	P(Sky Play)
<i>sunny</i>	<i>yes</i>	4/5
<i>rainy</i>	<i>yes</i>	1/5
<i>sunny</i>	<i>no</i>	1/3
<i>rainy</i>	<i>no</i>	2/3

Temp	Play?	P(Temp Play)
<i>warm</i>	<i>yes</i>	4/5
<i>cold</i>	<i>yes</i>	1/5
<i>warm</i>	<i>no</i>	1/3
<i>cold</i>	<i>no</i>	2/3

Humid	Play?	P(Humid Play)
<i>high</i>	<i>yes</i>	3/5
<i>norm</i>	<i>yes</i>	2/5
<i>high</i>	<i>no</i>	2/3
<i>norm</i>	<i>no</i>	1/3

Example Using NB for Classification



Play?	P(Play)
<i>yes</i>	3/4
<i>no</i>	1/4

Temp	Play?	P(Temp Play)
<i>warm</i>	<i>yes</i>	4/5
<i>cold</i>	<i>yes</i>	1/5
<i>warm</i>	<i>no</i>	1/3
<i>cold</i>	<i>no</i>	2/3

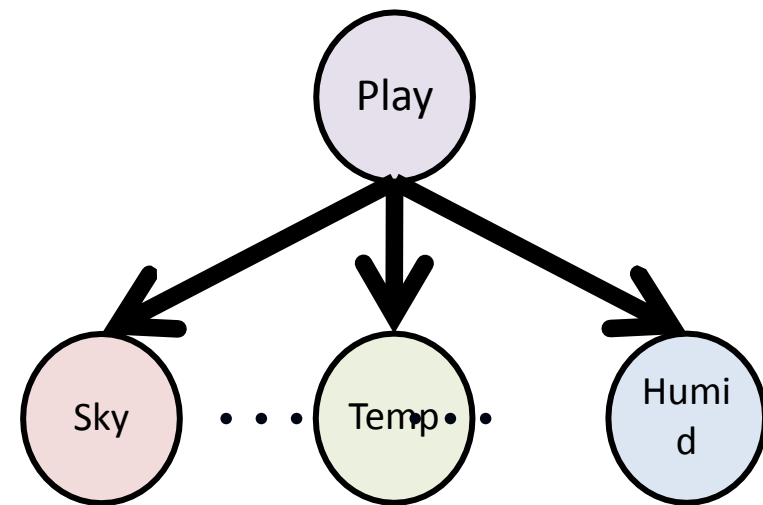
Sky	Play?	P(Sky Play)
<i>sunny</i>	<i>yes</i>	4/5
<i>rainy</i>	<i>yes</i>	1/5
<i>sunny</i>	<i>no</i>	1/3
<i>rainy</i>	<i>no</i>	2/3

Humid	Play?	P(Humid Play)
<i>high</i>	<i>yes</i>	3/5
<i>norm</i>	<i>yes</i>	2/5
<i>high</i>	<i>no</i>	2/3
<i>norm</i>	<i>no</i>	1/3

$$h(\mathbf{x}) = \arg \max_{y_k} \log P(Y = y_k) + \sum_{j=1}^d \log P(X_j = x_j \mid Y = y_k)$$

Goal: Predict label for $\mathbf{x} = (\text{rainy, warm, normal})$

Example Using NB for Classification



Predict label for:
 $\mathbf{x} = (\text{rainy}, \text{warm}, \text{normal})$

Play?	P(Play)
yes	3/4
no	1/4

Temp	Play?	P(Temp Play)
warm	yes	4/5
cold	yes	1/5
warm	no	1/3
cold	no	2/3

Sky	Play?	P(Sky Play)
sunny	yes	4/5
rainy	yes	1/5
sunny	no	1/3
rainy	no	2/3

Humid	Play?	P(Humid Play)
high	yes	3/5
norm	yes	2/5
high	no	2/3
norm	no	1/3

$$\begin{aligned}
 P(\text{play} \mid \mathbf{x}) &\propto \log P(\text{play}) + \log P(\text{rainy} \mid \text{play}) + \log P(\text{warm} \mid \text{play}) + \log P(\text{normal} \mid \text{play}) \\
 &\propto \log 3/4 + \log 1/5 + \log 4/5 + \log 2/5 = -1.319
 \end{aligned}$$

predict
PLA
Y

$$\begin{aligned}
 P(\neg \text{play} \mid \mathbf{x}) &\propto \log P(\neg \text{play}) + \log P(\text{rainy} \mid \neg \text{play}) + \log P(\text{warm} \mid \neg \text{play}) + \log P(\text{normal} \mid \neg \text{play}) \\
 &\propto \log 1/4 + \log 2/3 + \log 1/3 + \log 1/3 = -1.732
 \end{aligned}$$