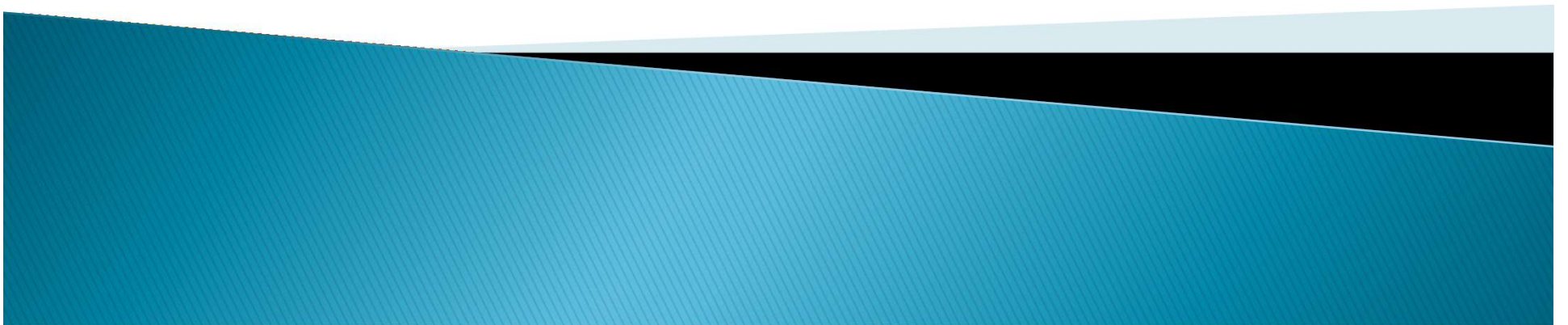


Bayesian Networks

Nazerfard, Ehsan
nazerfard@aut.ac.ir



Joint Probability Distribution

- The Wet Grass example:
 - Suppose we have 4 binary random variables:
Cloudy, Sprinkler, Rain and WetGrass
- The probabilities of all assignments is called the joint probability distribution:

C	S	R	W	Prob.
F	F	F	F	1.00
F	F	F	T	0.00
F	F	T	F	0.01
...

Joint Probability Distribution (cont.)

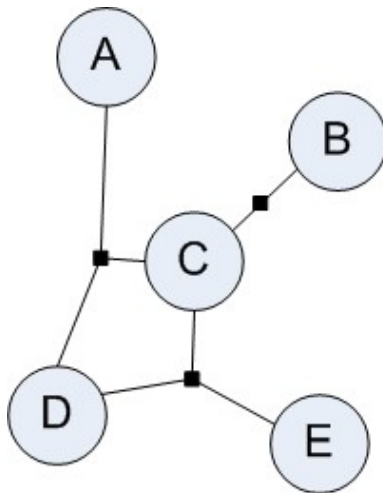
- ❑ What good is a joint probability?
- ❑ The probabilities of all assignments is called the joint probability distribution:

C	S	R	W	Prob.
F	F	F	F	1.00
F	F	F	T	0.00
F	F	T	F	0.01
...

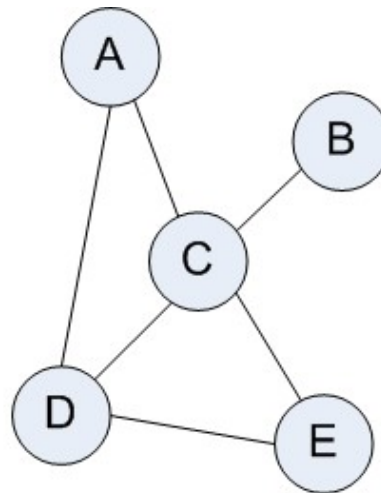
Any more compact representation ?

(Probabilistic) Graphical Models

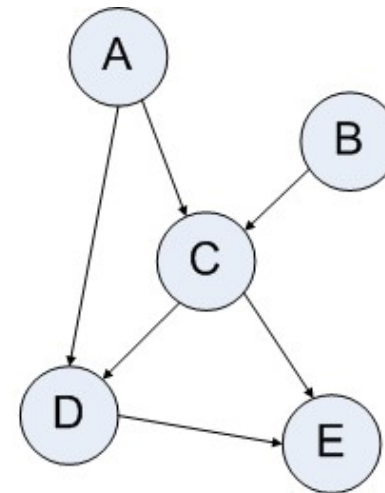
- Three main types of graphical models [3].
- They all provide a compact way of representing the joint probability distributions.
 - Nodes: random variables
 - Edges: statistical dependencies



Factor Graph

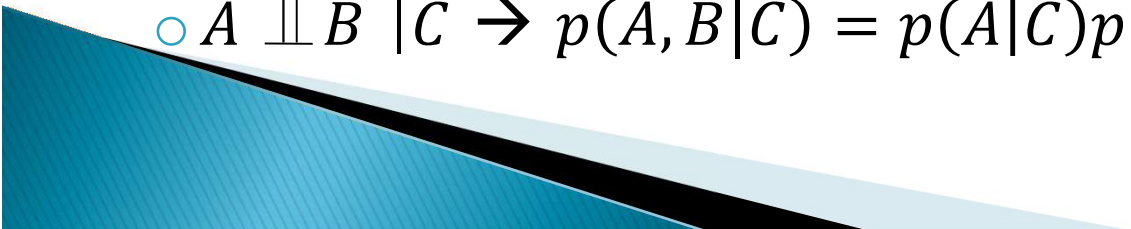


Undirected Graph



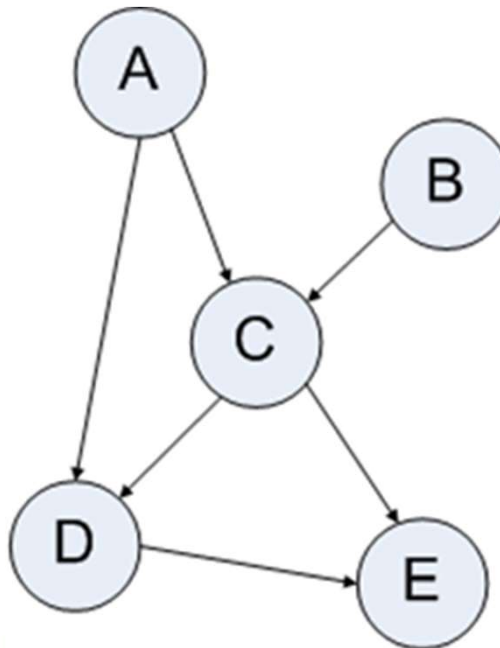
Directed Graph

Why Graphical Models?

- Why we need graphical models?
 - Intuitive way of representation of the relations between variables
 - Abstract out the conditional independence relations between variables
 - Conditional independence
 - “Is A dependent on B , given C ?”
 - $A \perp\!\!\!\perp B \mid C \rightarrow p(A|B, C) = p(A|C)$
 - $A \perp\!\!\!\perp B \mid C \rightarrow p(A, B|C) = p(A|C)p(B|C)$
- 

Directed Graphical Models

- Directed Graphical Models
 - Nodes: random variables
 - Edges: direct influence (causality relation)
 - The graph has to be a DAG (Directed Acyclic Graph).

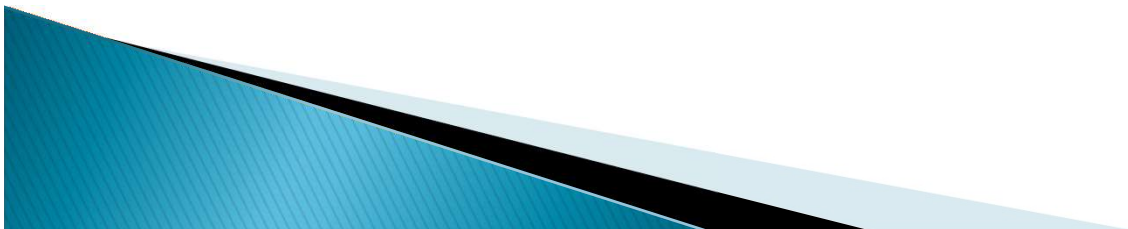


Bayesian/Belief Networks

- Given $X = (X_1, X_2, \dots, X_n)$, a Bayesian network is an annotated DAG that represents a unique joint probability distribution over X :

$$p(X_1, X_2, \dots, X_n) = \prod_i p(X_i | pa(X_i))$$

- Each node is annotated with a conditional probability table that represents $p(X_i | pa(X_i))$.



Building Bayesian Networks

□ Structure Learning

- To build a Bayes net, we start with the random variables in the problem definition as our nodes. After that, a directed edge is added from variable i to variable j if we perceive variable i to be a direct cause of variable j .

Building Bayesian Networks (cont.)

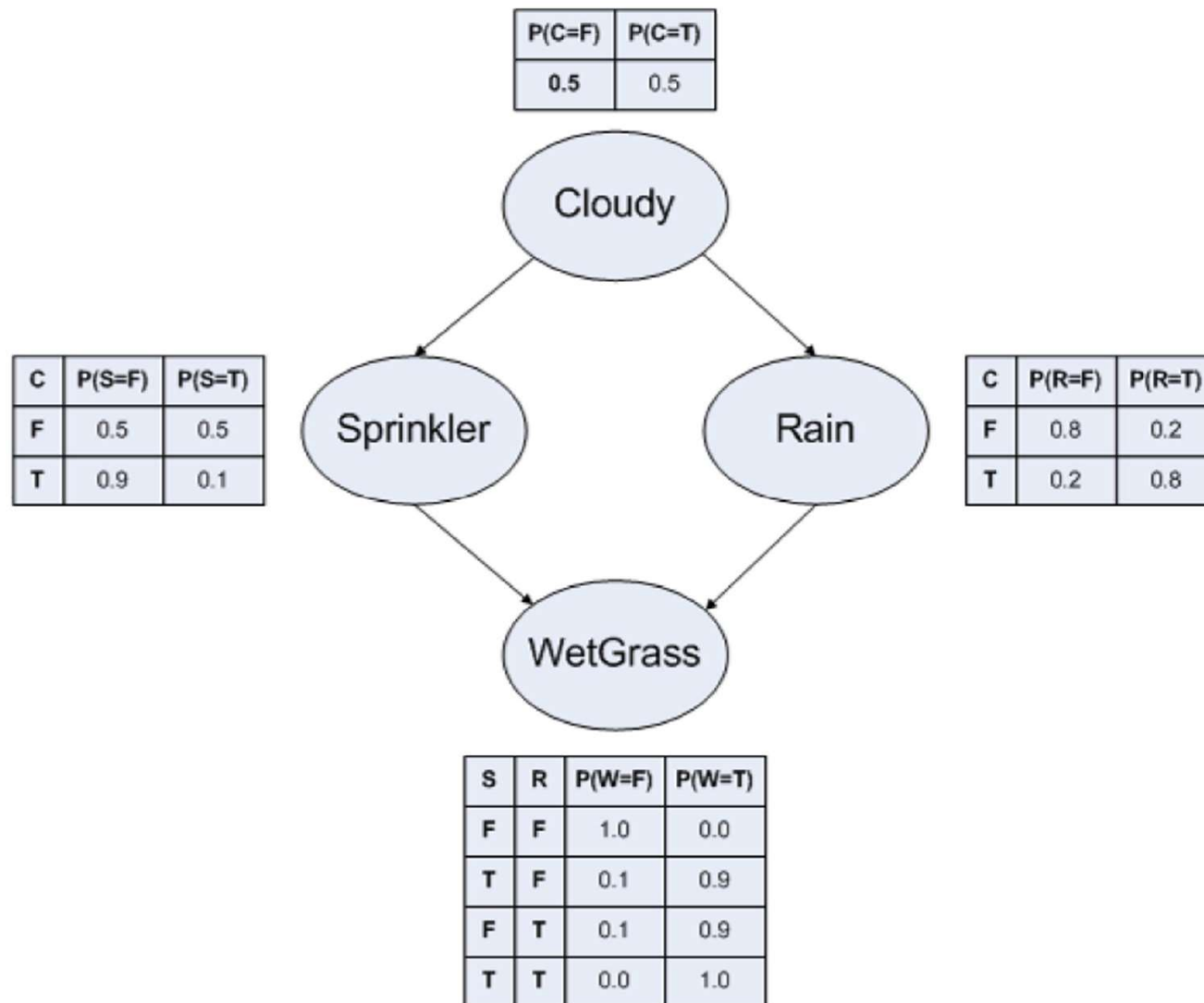
□ Structure Learning

- To build a Bayes net, we start with the random variables in the problem definition as our nodes. After that, a directed edge is added from variable i to variable j if we perceive variable i to be a direct cause of variable j .

□ Parameter Learning

- After building the Bayes net structure, we associate with each variable a table called the conditional probability (CPT). A CPT of a variable is a conditional probability distribution of this variable given its parents.

A Bayes Net for the Wet Grass Example



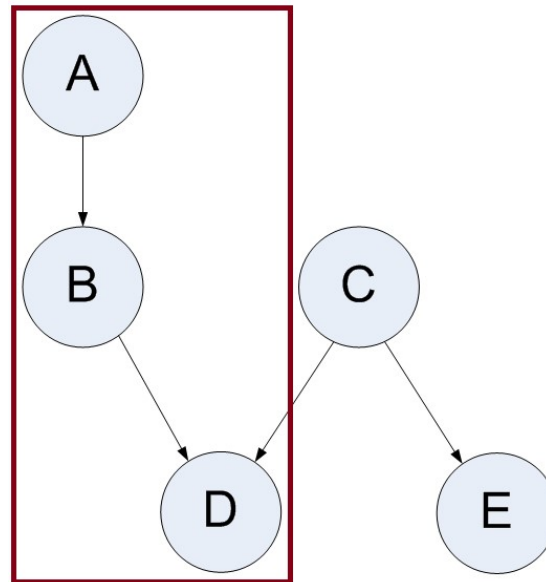
Conditional Independence (CI)

- “Is A dependent on B , given C ?”
- Examples
 - child’s genes $\perp\!\!\!\perp$ grandparents’ genes \mid parents’ genes
 - amount of speeding fine $\perp\!\!\!\perp$ type of car \mid speed
 - lung cancer $\perp\!\!\!\perp$ yellow teeth \mid smoker
 - ...

CI Test – Serial (Head–Tail) Connection

□ Example

- Child's genes and his grandparents' genes
- $A \perp\!\!\!\perp D \mid B$
- Variable B d-separates A and D
- $p(A, B, D) = p(A)p(B|A)p(D|B)$



CI Test – Converging (Head–Head) Connection

- Example: Rolling two dices...

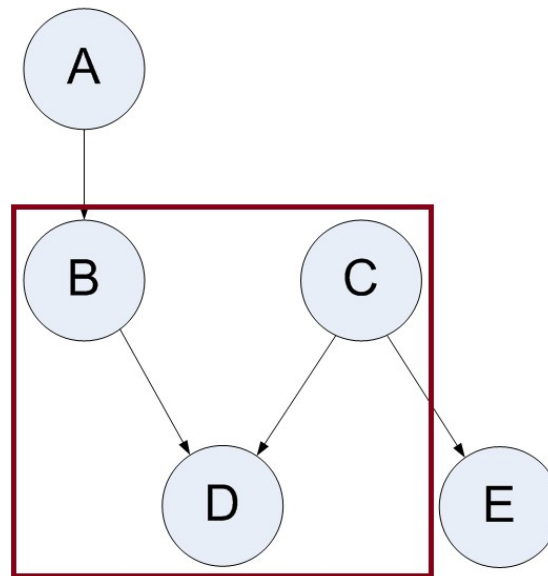
- $B \perp\!\!\!\perp C \mid \phi$

- $B \not\perp\!\!\!\perp C \mid D$



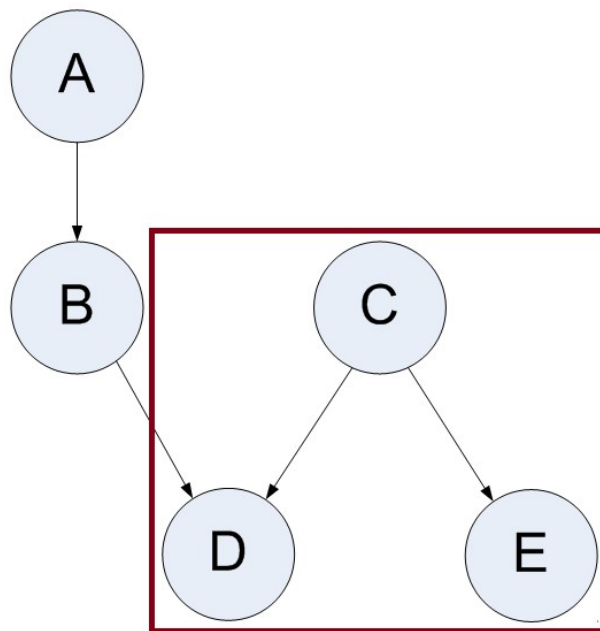
- V-Structure

- $p(B, C, D) = p(B)p(C)p(D|B, C)$



CI Test – Diverging (Tail–Tail) Connection

- Example: Dice example ...
 - C : random numbers are in $[1,6]$
 - $D \perp\!\!\!\perp E \mid C$
- $p(C, D, E) = p(c)p(D|C)p(E|C)$

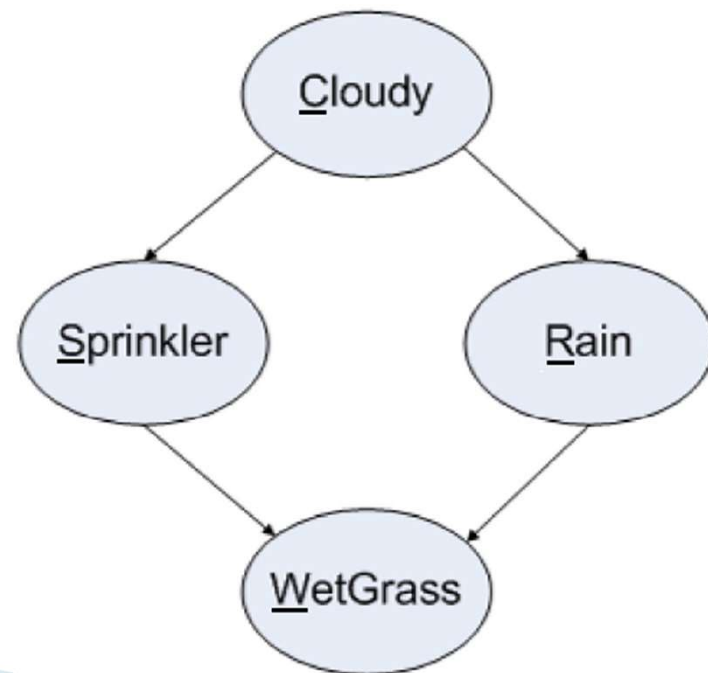


The Joint Prob. for the Wet Grass Example

□ General Chain rule:

$$p(C, S, R, W) = p(C)p(S|C)p(R|C, S)p(W|C, S, R)$$

- How to simplify the above joint probability, considering the conditional independence relations?

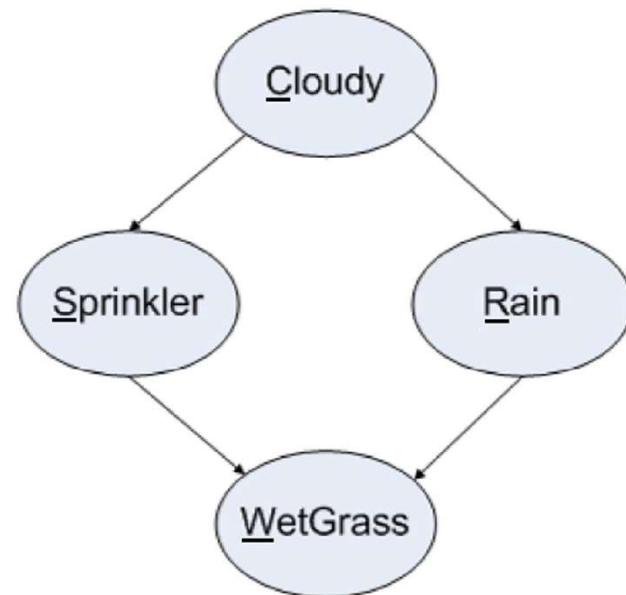


The Joint Prob. for the Wet Grass Example

□ Considering the CI relations:

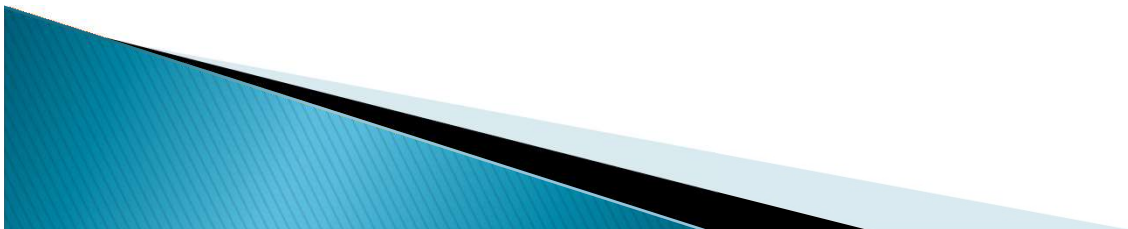
$$p(C, S, R, W) = p(C)P(S|C)p(R|C)p(W|S, R)$$

- So, the CI relations allow the joint probability to be presented in a more compact form. The more compact form/factorized form has fewer parameters, which makes the learning easier.



Inference in Bayesian Networks

- As a general statement, inference in Bayesian networks boils down to marginalizing the joint probability distribution (JPD).
- Given the JPD, we can answer all possible inference queries by marginalizing the irrelevant variables.



Inference Example

- An inference example in the WetGrass graph:
 - We have observed that the grass is wet. What would be the most likely cause for this observation?

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

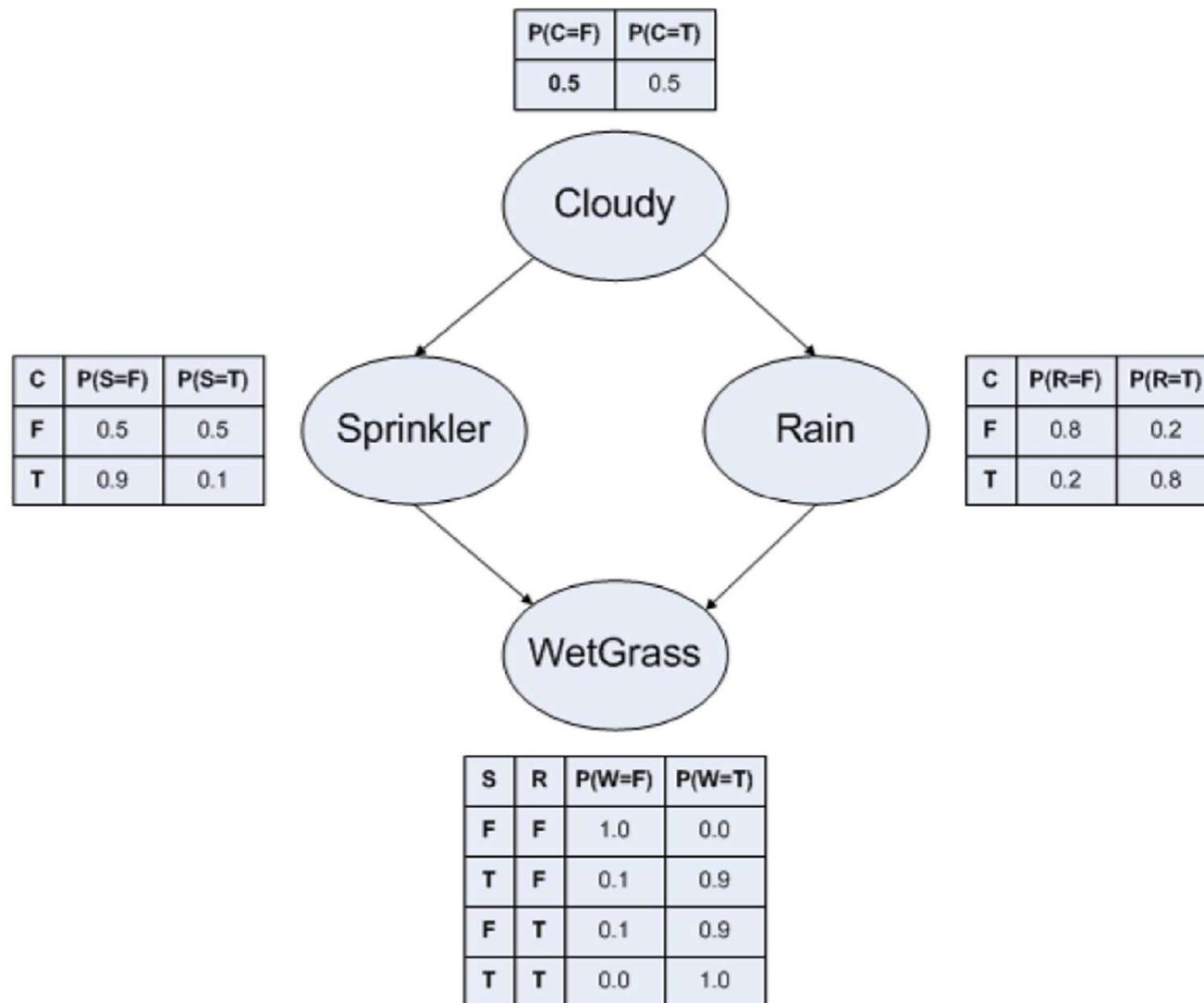
Diagram illustrating the components of Bayes' theorem:

- likelihood: $p(B|A)$
- prior: $p(A)$
- posterior: $p(A|B)$
- marginalization: $p(B)$



Thomas Bayes
1702 - 1761

Inference Example (cont.)




Inference Example (cont.)

- An inference example in the WetGrass graph:
 - We have observed that the grass is wet. What would be the most likely cause for this observation?

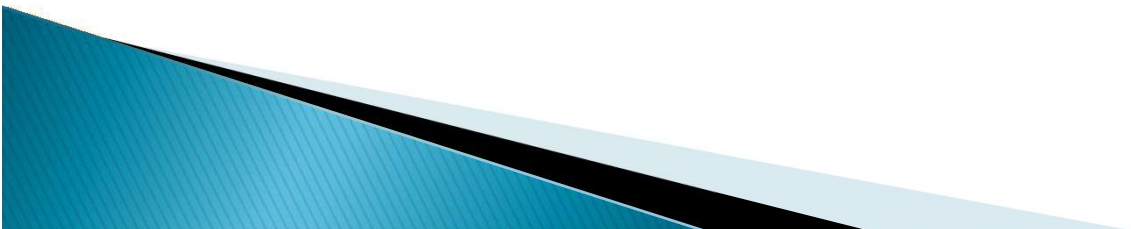
$$p(S = 1|W = 1) = \frac{p(S = 1, W = 1)}{p(W = 1)} = \frac{0.28}{0.65}$$

$$p(R = 1|W = 1) = \frac{p(R = 1, W = 1)}{p(W = 1)} = \frac{0.46}{0.65}$$

where the denominator, i.e., $p(W = 1) = \sum_{c,s,r} p(W = 1|S, R, C)p(S, R, C)$ is the normalization factor.

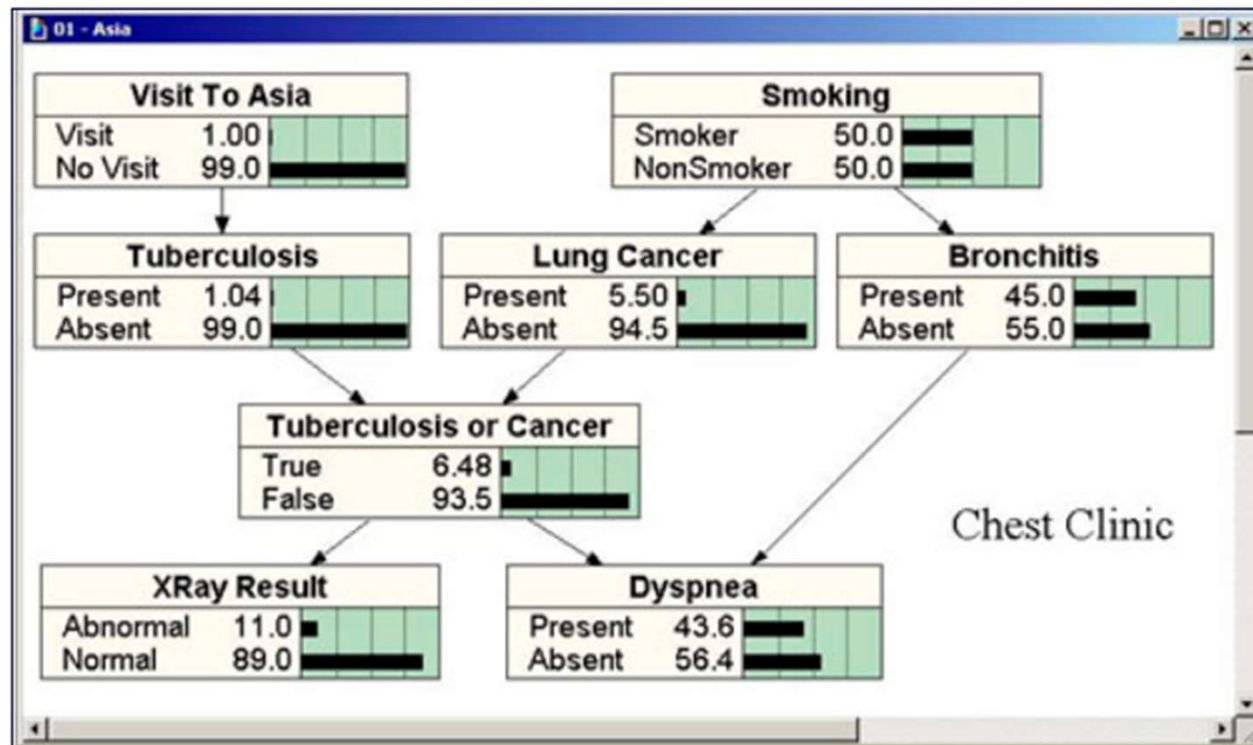


Inference Example (cont.)

$$\begin{aligned}
 p(W = 1) &= \\
 \sum_{c,s,r} p(W = 1|S, R, C) p(S, R, C) &= \\
 \sum_{c,s,r} p(W = 1|S, R, C) p(S|R, C) p(R|C)P(C) &= \\
 \sum_{c,s,r} p(W = 1|S, R) p(S|C) p(R|C)p(C) &= \\
 p(W = 1| S = 0, R = 0) p(S = 0| C = 0) p(R = 0| C = 0) p(C = 0) + \\
 p(W = 1| S = 0, R = 0) p(S = 0| C = 1) p(R = 0| C = 1) p(C = 1) + \\
 p(W = 1| S = 0, R = 1) p(S = 0| C = 0) p(R = 1| C = 0) p(C = 0) + \\
 p(W = 1| S = 0, R = 1) p(S = 0| C = 1) p(R = 1| C = 1) p(C = 1) + \\
 p(W = 1| S = 1, R = 0) p(S = 1| C = 0) p(R = 0| C = 0) p(C = 0) + \\
 p(W = 1| S = 1, R = 0) p(S = 1| C = 1) p(R = 0| C = 1) p(C = 1) + \\
 p(W = 1| S = 1, R = 1) p(S = 1| C = 0) p(R = 1| C = 0) p(C = 0) + \\
 p(W = 1| S = 1, R = 1) p(S = 1| C = 1) p(R = 1| C = 1) p(C = 1) &= 0.65
 \end{aligned}$$


Bayes Net Application

□ Daily Medical Practice



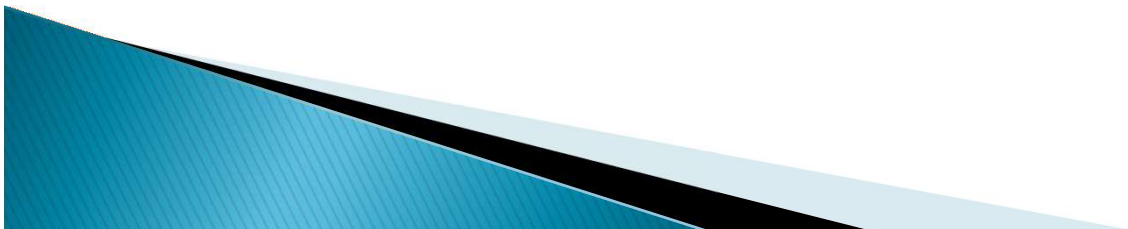
Bayesian Network Learning

□ Structure Learning

- Find a structure of Bayesian Network that best describes the observed data.

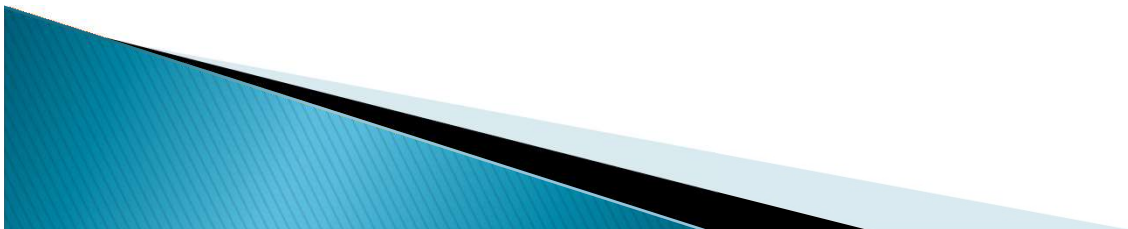
□ Parameter Learning

- Learning the parameters (CPTs) when the structure is known.



Bayes Net Structure Learning

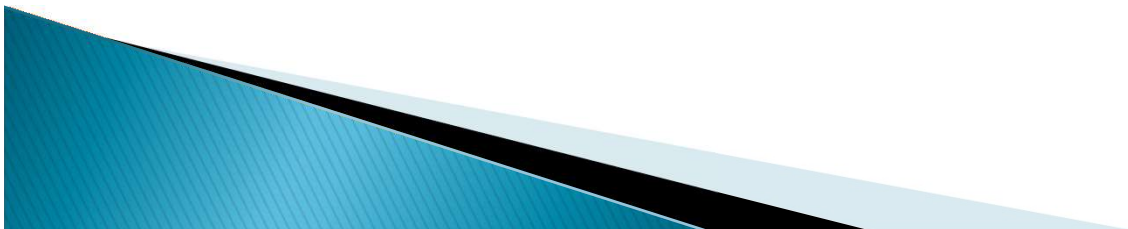
- ❑ Learning the Bayes net structure from data by considering all possible structures is not practical and is shown to be an NP-hard problem [4].
- ❑ Therefore, learning the structure of Bayes nets require sub-optimal heuristic search algorithms.
- ❑ The text books might have assumed that the network structure is already known (like the wetGrass example).



Structure Learning

□ Heuristics:

- i. Score and Search-Based (S&S) approach [5]
 - Learning the network structures
- ii. Constraint-Based (CB) approach [6]
 - Learning the edges composing a structure



Structure Learning (cont.)

□ Heuristics:

- i. Score and Search-Based (S&S) approach [4]
 - Learning the network structures
- ii. Constraint-Based (CB) approach [5]
 - Learning the edges composing a structure

*** Independent Feature Model
also called naïve Bayes

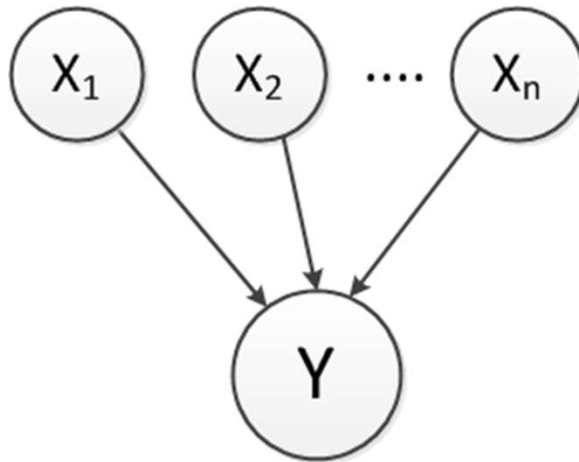
(It assumes the features are conditionally independent)



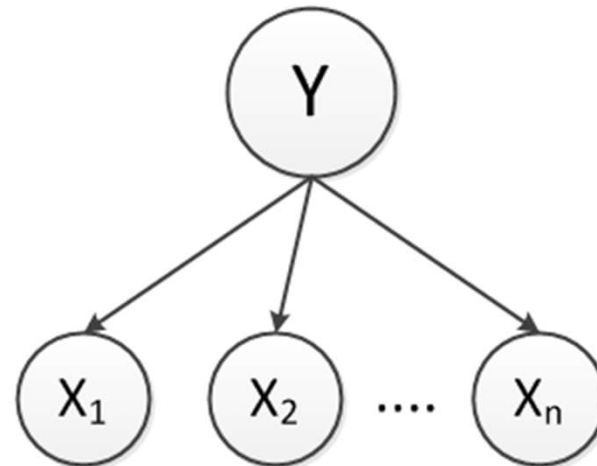
Naïve Bayes Classifier (NBC)

- Which graph represents the NBC structure?
(a) or (b)?

$$\underline{h: X \rightarrow Y}$$



(a)



(b)

Naïve Bayes Classifier (NBC)

Binary classification

$$h: X \rightarrow Y \rightarrow h: X \rightarrow \{+, -\}$$

- We wish to approximate an unknown function $f: X \rightarrow Y$ or equivalently $P(Y|X)$.

$$\begin{aligned} y_{\text{MLE}} &= \operatorname{argmax}_{y \in Y} p(Y = y | \mathbf{X}) \\ &= \operatorname{argmax}_{y \in Y} \frac{p(\mathbf{X} | Y = y) p(Y = y)}{p(\mathbf{X})} \quad (\text{Bayes Rule}) \end{aligned}$$

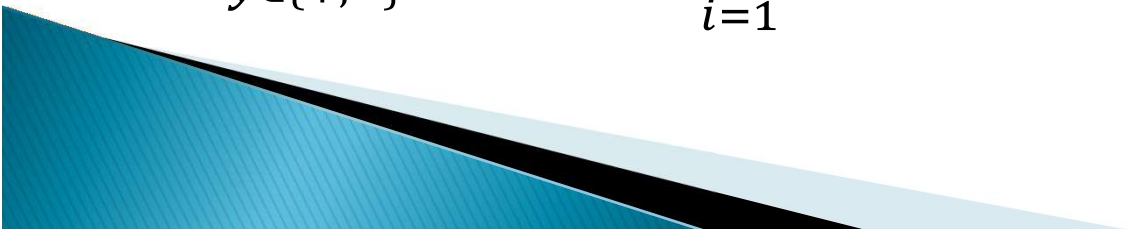
$$= \operatorname{argmax}_{y \in Y} p(\mathbf{X} | Y = y) p(Y = y) = \operatorname{argmax}_{y \in Y} p(\mathbf{X}, Y = y) \quad (\text{Joint Probability})$$

Naïve Bayes Classifier (NBC)

$$\mathbf{X} = \langle X_1, X_2, \dots, X_n \rangle, \quad X_i \perp\!\!\!\perp X_j \mid Y \quad (s.t: i, j = 1, 2, \dots, n \text{ \& } i \neq j)$$

$$y_{MLE} = \operatorname{argmax}_{y \in \{+, -\}} p(\mathbf{X}, Y = y) = \operatorname{argmax}_{y \in \{+, -\}} p(\mathbf{X} \mid Y = y) p(Y = y)$$

$$= \operatorname{argmax}_{y \in \{+, -\}} p(X_1 \mid Y = y) p(X_2 \mid Y = y) \dots p(X_n \mid Y = y) p(Y = y)$$

$$= \operatorname{argmax}_{y \in \{+, -\}} p(Y = y) \prod_{i=1}^n p(X_i = x_i \mid Y = y)$$


PlayTennis / Weather Dataset

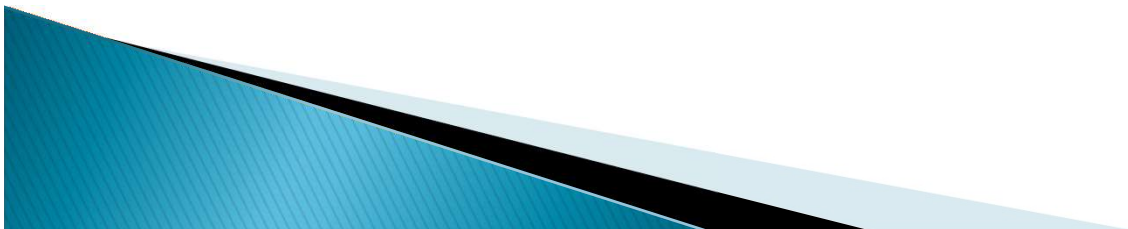
Day	Outlook	Temperature	Humidity	Windy	Play
D1	Sunny	Hot	High	False	No
D2	Sunny	Hot	High	True	No
D3	Overcast	Hot	High	False	Yes
D4	Rainy	Mild	High	False	Yes
D5	Rainy	Cool	Normal	False	Yes
D6	Rainy	Cool	Normal	True	No
D7	Overcast	Cool	Normal	True	Yes
D8	Sunny	Mild	High	False	No
D9	Sunny	Cool	Normal	False	Yes
D10	Rainy	Mild	Normal	False	Yes
D11	Sunny	Mild	Normal	True	Yes
D12	Overcast	Mild	High	True	Yes
D13	Overcast	Hot	Normal	False	Yes
D14	Rainy	Mild	High	True	No

NBC – Example

- Considering the Weather data, what would NBC predict for the following test data?

($\langle O = \text{Rainy}, T = \text{Cool}, H = \text{High}, W = \text{True} \rangle, ?$)

- O: Outlook
- T: Temperature
- H: Humidity
- W: Windy



NBC – Example (cont.)

Test Data: ($< O = \text{Rainy}, T = \text{Cool}, H = \text{High}, W = \text{True} >, ?$)

	Prior Probability	
$p(Y = \text{yes}) = 9/14$		$p(Y = \text{no}) = 5/14$

$$p(\text{rainy}|\text{yes}) = 3/9$$

$$p(\text{rainy}|\text{no}) = 2/5$$

$$p(\text{cool}|\text{yes}) = 3/9$$

$$p(\text{cool}|\text{no}) = 1/5$$

$$p(\text{high}|\text{yes}) = 3/9$$

$$p(\text{high}|\text{no}) = 4/5$$

$$p(\text{true}|\text{yes}) = 3/9$$

$$p(\text{true}|\text{no}) = 3/5$$

NBC – Example (cont.)

Test Data: ($\langle O = \text{Rainy}, T = \text{Cool}, H = \text{High}, W = \text{True} \rangle, ?$)

$$p(Y = \text{yes}) = 9/14$$

$$p(Y = \text{no}) = 5/14$$

$$p(\text{rainy}|\text{yes}) = 3/9$$

$$p(\text{rainy}|\text{no}) = 2/5$$

$$p(\text{cool}|\text{yes}) = 3/9$$

$$p(\text{cool}|\text{no}) = 1/5$$


$$p(\text{high}|\text{yes}) = 3/9$$

$$p(\text{high}|\text{no}) = 4/5$$

$$p(\text{true}|\text{yes}) = 3/9$$

$$p(\text{true}|\text{no}) = 3/5$$

Posterior
Probability



$$p(Y = \text{yes}) \prod_{i=1}^4 p(X_i = x_i | Y = \text{yes}) = 0.0079 < p(Y = \text{no}) \prod_{i=1}^4 p(X_i = x_i | Y = \text{no}) = 0.0137$$

NBC – Example II

- As an another example, what would NBC predict for the following test data?

$(\langle O = \text{Rainy}, T = \text{Cool}, H = \text{High}, W = \text{False} \rangle, ?)$

- O: Outlook
- T: Temperature
- H: Humidity
- W: Windy



NBC – Example II (cont.)

Test Data: ($\langle O = \text{Rainy}, T = \text{Cool}, H = \text{High}, W = \text{False} \rangle, ?$)

$$p(Y = \text{yes}) = 9/14$$

$$p(Y = \text{no}) = 5/14$$

$$p(\text{rainy}|\text{yes}) = 3/9$$

$$p(\text{rainy}|\text{no}) = 2/5$$

$$p(\text{cool}|\text{yes}) = 3/9$$

$$p(\text{cool}|\text{no}) = 1/5$$

$$p(\text{high}|\text{yes}) = 3/9$$

$$p(\text{high}|\text{no}) = 4/5$$

$$p(\text{false}|\text{yes}) = 6/9$$

$$p(\text{false}|\text{no}) = 2/5$$

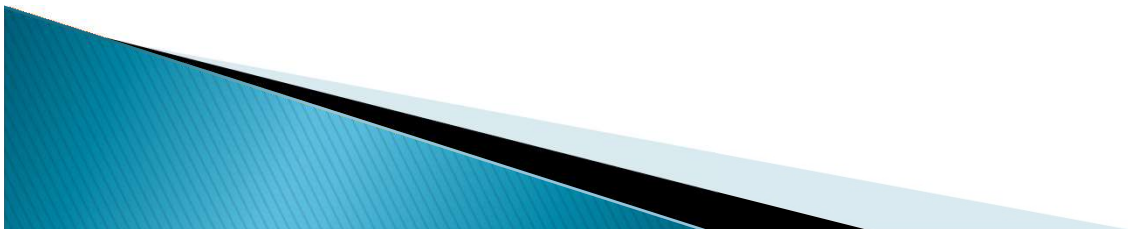
$$p(Y = \text{yes}) \prod_{i=1}^4 p(X_i = x_i | Y = \text{yes}) = 0.0159 > p(Y = \text{no}) \prod_{i=1}^4 p(X_i = x_i | Y = \text{no}) = 0.0091$$

NBC – Example III

- As the third example, what would NBC predict for this test data?

($\langle O = \textit{Overcast}, T = \textit{Cool}, H = \textit{High}, W = \textit{False} \rangle, ?$)

- O: Outlook
- T: Temperature
- H: Humidity
- W: Windy



NBC – Example III (cont.)

Test Data: ($O = \text{Overcast}, T = \text{Cool}, H = \text{High}, W = \text{False}$), ?

$$p(Y = \text{yes}) = 9/14$$

$$p(Y = \text{no}) = 5/14$$

$$p(\text{overcast}|\text{yes}) = 4/9$$

$$p(\text{overcast}|\text{no}) = 0.0$$

...


...

- The zero probability, i.e. $p(\text{overcast}|\text{no}) = 0$, is problematic, because it will wipe out all the information in the other probabilities. Solution?
 - (Laplace) Smoothing

Laplace Smoothing

- Handling insufficient data
- Smoothed likelihood:

$$p(X_i = x_{ik} | Y = y_j) = \frac{\text{count}(X_i = x_{ik}, Y = y_j) + l}{\text{count}(Y = y_j) + l \times k}$$

- Weather data: ($i = 1..4, j = 1..2$)
 - l : Laplace smoothing strength
 - k : number of possible values for X_i
- 

NBC Application

□ Spam Filtering



Naïve Bayes → Hidden Markov Model

- Previously, we saw that the naïve Bayes model provides a joint distribution as follows:

$$p(\mathbf{X}, Y) = \prod_{i=1}^n p(X_i|Y) p(Y)$$

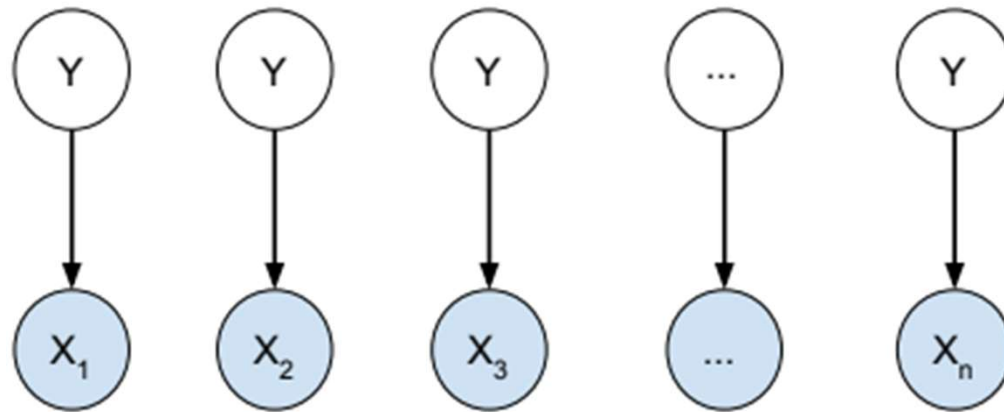
- As an extension of naïve Bayes for sequential data, the hidden Markov model (HMM) adds state transition $P(Y_i|Y_{i-1})$. Therefore, HMM provides a joint distribution as follows:

$$p(\mathbf{X}, \mathbf{Y}) = p(X_1|Y_1)p(Y_1) \prod_{i=2}^n p(X_i|Y_i) p(Y_i|Y_{i-1})$$

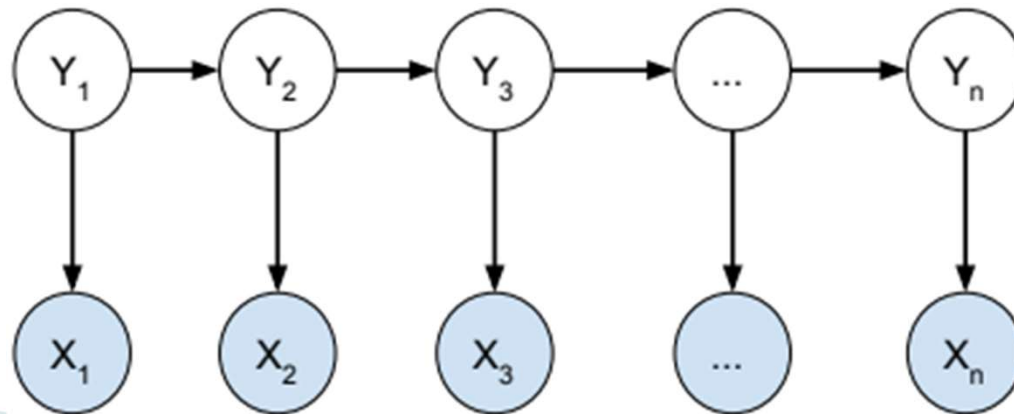
Naïve Bayes vs. HMM

□ Graphical representation

Naïve Bayes

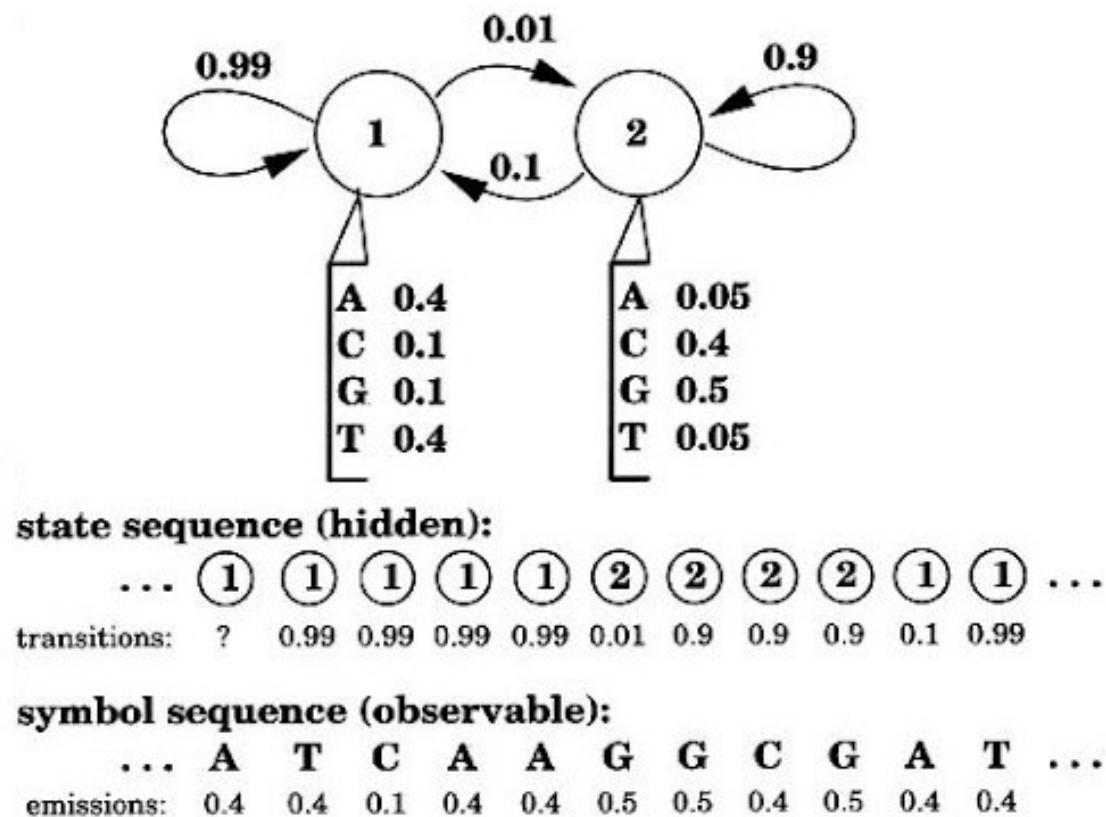


Hidden Markov Model



HMM – State Machine

- (i) Prior Probabilities
- (ii) State Transition Probabilities
- (iii) Emission Probabilities



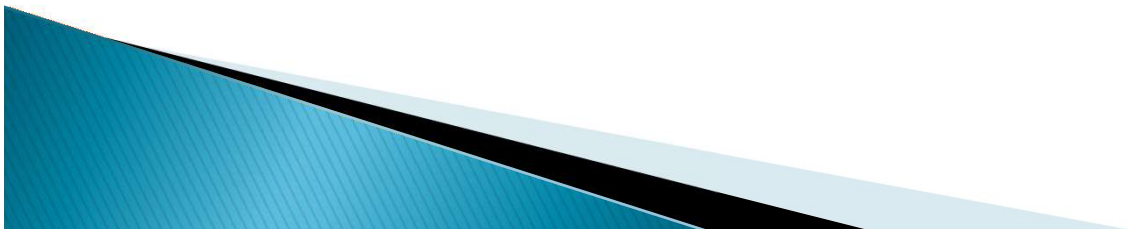
Three Basic HMM Problems (1 / 3)

- Given an observation sequence X and a model M_i , how do we efficiently compute $p(X | M_i)$, i.e the probability of the observation sequence, given the model?
- Solution: The Forward–Backward algorithm.



Three Basic HMM Problems (2/3)

- Given the observation sequence X and the model M_i , how do we choose a corresponding state sequence Y that best explains the observations?
- Solution: The Viterbi algorithm.



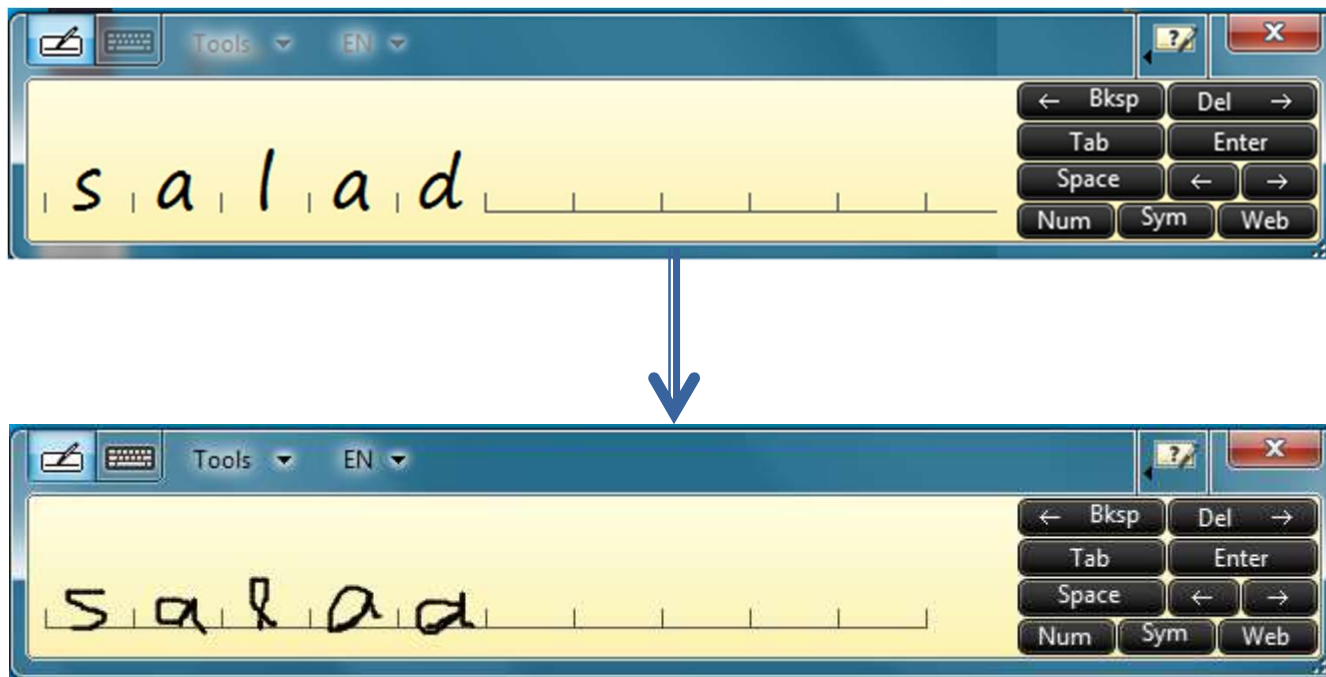
Three Basic HMM Problems (3 / 3)

- How do we adjust the parameters of each model to maximize $p(\mathbf{X} | M_i)$? In other words, how do we train each model so that the model works as well as it can?
- Solution: The Baum–Welch algorithm.



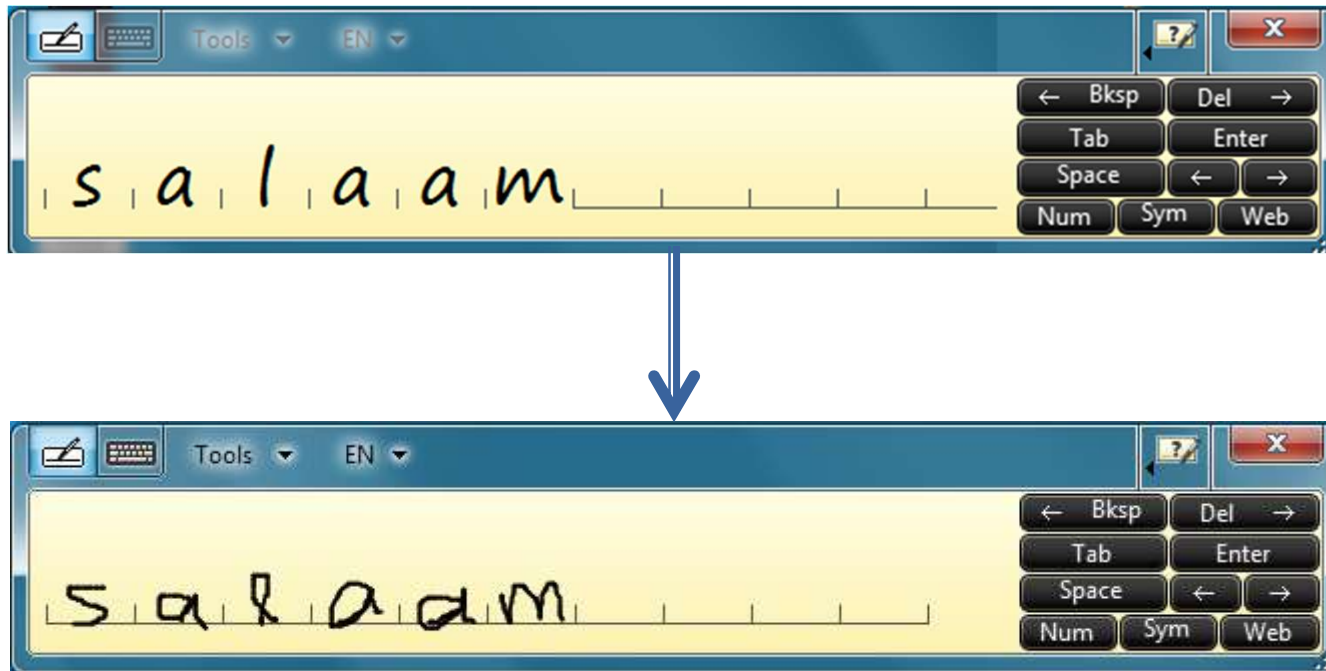
HMM Application

□ Tablet PC



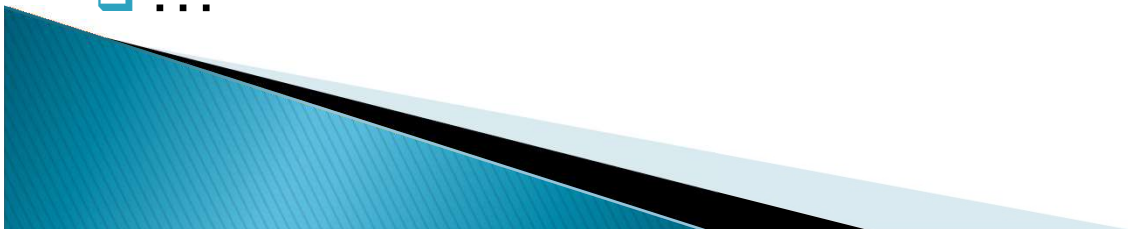
HMM Application (cont.)

□ Tablet PC



Further Reading

- ❑ Dynamic/Temporal Bayesian Networks (DBNs)
- ❑ Latent Dirichlet Allocation (LDA)
- ❑ TAN Classifier
 - TAN short for Tree Augmented Naïve Bayes
- ❑ Kalman Filter
- ❑ ...



References

1. Zoubin Ghahramani, [Learning Dynamic Bayesian Networks](#), 1997.
2. Kevin Murphy, [PhD Thesis](#), Dynamic Bayesian Networks: Representation, Inference and Learning, UC Berkeley, 2002.
3. Probabilistic Graphical Models, Principles and Techniques, Daphne Koller and Nir Friedman, 2009.
4. Lawrence Rabiner, [A tutorial on hidden Markov models and selected applications in speech recognition](#), 1988.
5. Learning equivalence classes of Bayesian-network structures, DM. Chickering, Journal of Machine Learning Research, 2:445–498, 2002.
6. Being Bayesian about Bayesian Network structure: A Bayesian approach to structure discovery in Bayesian Networks, N. Friedman and D. Koller, Machine Learning, 50:95–125, 2003.