BIN504 - Lecture XI

Bayesian Inference & Markov Chains

References:
Lee, Chp. 11 & Sec. 10.11
Ewens-Grant, Chps. 4, 7, 11

Including slides by Tolga Can
Rev. 1.2 (Build 20130528191100)

Outline

- Bayesian Inference
  - Bayes' Chain Rule
  - Bayesian Belief Networks
  - Markov Blanket and Conditional Independence
- Markov Chains
  - Irreducibility, Periodicity, and Recurrence
  - Stationary Distribution
  - Random Walks

Refresh: Bayes' Rule

Theorem

Bayes' Theorem states that:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

Likewise:

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Therefore:

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

The Bayes Chain Rule

Corollary

The chain rule follows directly from Bayes' Theorem:

$$\begin{aligned} P(X_1, \ldots, X_N) &= P(X_1 | X_2, \ldots, X_N) P(X_2, \ldots, X_N) \\ &= P(X_1 | X_2, \ldots, X_N) P(X_2 | X_3, \ldots, X_N) P(X_3, \ldots, X_N) \\ &\vdots \\ P(\cap_{k=1}^{n} X_k) &= \prod_{k=1}^{n} \mathrm{P}(X_k \mid \cap_{j=1}^{k-1} X_j) \end{aligned}$$

Notice that since intersection (joint probability) is associative/commutative:

$$\begin{aligned} P(X, Y, Z) &= P(X|Y, Z) P(Y|Z) P(Z) \\ &= P(X|Y, Z) P(Z|Y) P(Y) \\ &= P(Y|X, Z) P(Z|X) P(X) \\ &\vdots \end{aligned}$$

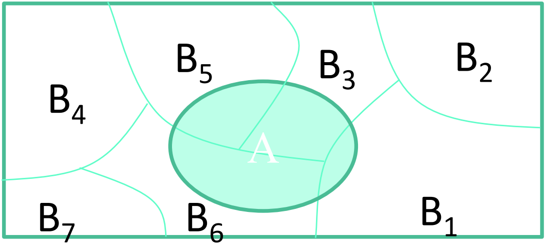
Refresh: Law of Total Probability

Definition

Let $B_1, B_2 \ldots B_k$ be possible values of r.v. B.

$$P(A) = \sum_{j=1}^{k} P(A|B = B_j)P(B = B_j)$$

This is called marginalization: calculating the marginal probability using conditional probability)



Conditional Independence

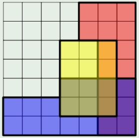
Definition

If given the value of some r.v. Z, also knowing Y does not change our belief about X:

$$X \perp Y | Z$$

we say that X is conditionally independent of Y given Z.

Example



$$Red \perp Blue | Yellow$$

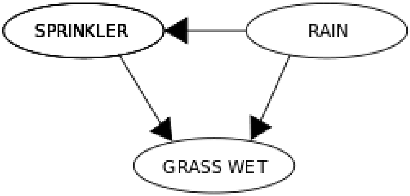
Fig by AzaToth@en.wikipedia

Bayesian Belief Networks

- A Bayesian (Belief) Network is a directed acyclic graph where:
  - Nodes represent random variables
  - Edges represent influence (dependence)
- Each node has a conditional probability table based on its parents.
  - $P(node|parents)$
- This graph represents dependence among variables in a concise fashion
  - Allows us to reason about our belief in certain causes.
    - Using conditional probabilities of symptoms given causes.
  - We can calculate different probabilities using the chain rule and marginalization:
    - joint probabilities of causes and symptoms.
    - posterior probabilities of causes **given** symptoms.

Sample Bayesian Network

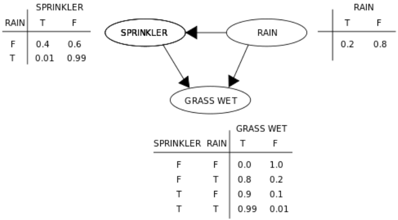
	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Example: Joint Probability



What is the probability that it is raining, that the sprinkler is off, and the grass is wet?

$$P(G, S, R) = P(G|S, R)P(S|R)P(R) \quad \text{(chain rule)}$$

For the required configuration:

$$P(Wet, Off, Yes) = 0.8 \times 0.99 \times 0.2 = 0.1584$$

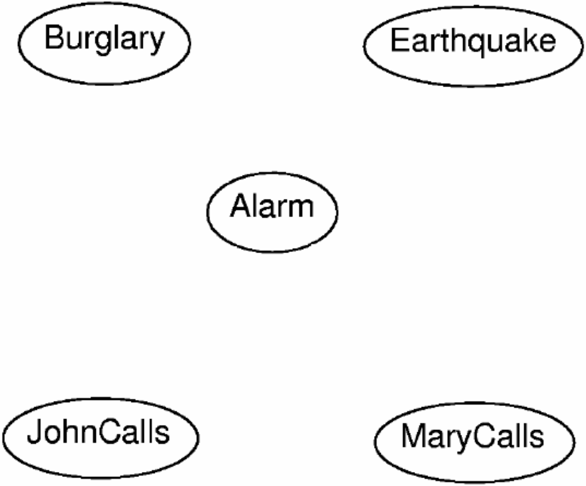
Example: Posterior Probability

What is the probability that it is raining, **given** that the sprinkler is off, and the grass is dry?

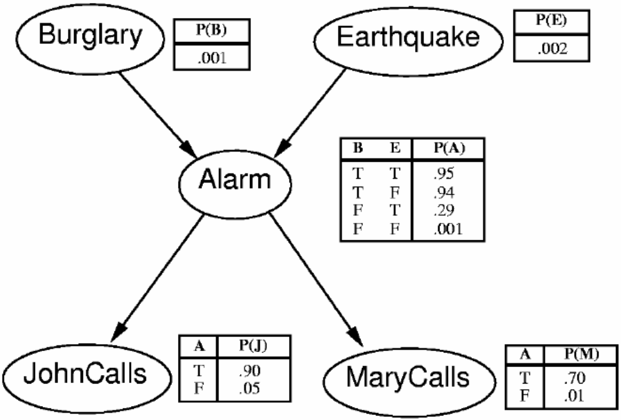
$$\begin{aligned}
P(R|G,S) &= \frac{P(G,S,R)}{P(G,S)} \\
&= \frac{P(G|S,R)P(S|R)P(R)}{P(G|S)P(S)} \\
&= \frac{P(G|S,R)P(S|R)P(R)}{\sum_R P(G|S,R_i)P(R_i)\sum_R P(S|R_i)P(R_i)} \quad \text{(marginalize)} \\
P(Yes|Dry,Off) &= \frac{0.2 \times 0.99 \times 0.2}{(0.2 \times 0.2 + 1.0 \times 0.8)(0.99 \times 0.2 + 0.6 \times 0.8)} \\
&= 0.0695
\end{aligned}$$

Influence

Assume you are at work, have a house alarm and two neighbors (John and Mary):



Influence


Burglary
P(B)
.001
Earthquake
P(E)
.002
Alarm
B E P(A)
T T .95
T F .94
F T .29
F F .001
JohnCalls
A P(J)
T .90
F .05
MaryCalls
A P(M)
T .70
F .01

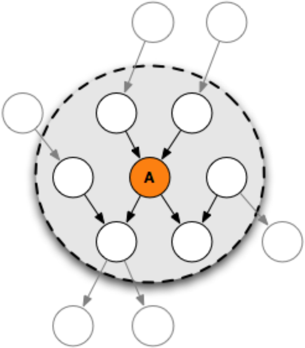
Conditional Independence

- In the previous example, there is a certain conditional independency:
  - Given the alarm state, JohnCalls is independent of MaryCalls, Earthquake or Burglary.

$$P(J|M, A, Q, B) = P(J|A)$$

- Does this mean that an earthquake or burglary does not affect whether John calls?
  - No. John is still affected by burglary or earthquake, but not directly.
  - John calling is conditionally independent of everything else, given the alarm.
  - But John is not absolutely independent of burglary or earthquake.

The Markov Blanket



Definition

The set of nodes ∂A composed of A's parents, children and children's parents is called the Markov Blanket of A

- Given its Markov blanket, a node is conditionally independent of any other nodes in the network.

$$A \perp B | \partial A \ \forall B$$

Plate Notation

For networks with replicated subgraphs we use plate notation, which wraps replicate parts into frames, with number of replicates indicated.

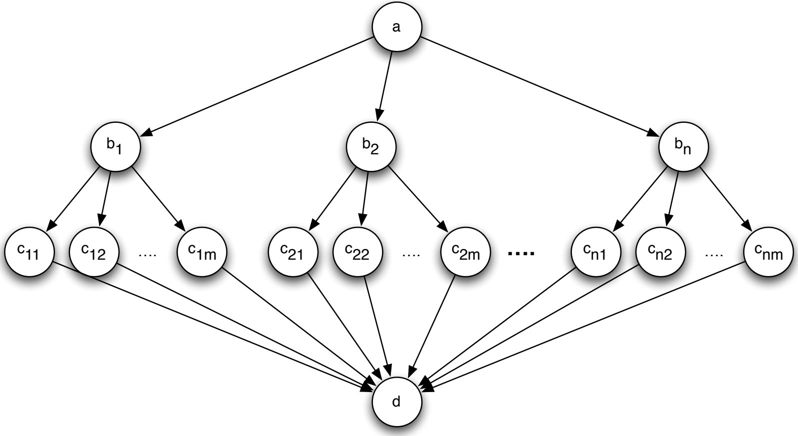
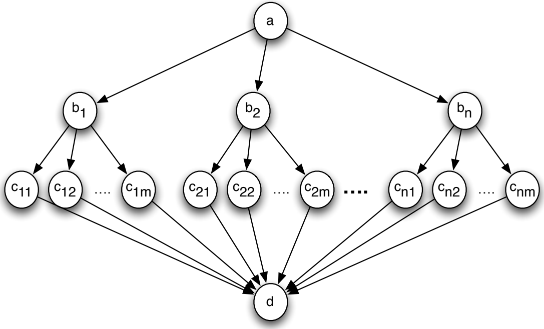
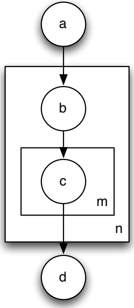


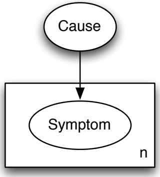
Plate Notation (2)





Naïve Bayes Model

- Normally, some symptom (evidence) variables are dependent on each other as well as being dependent on the causes.
- However, as the graph gets deeper, the conditional probability tables become intractably large.
- The Naïve Bayes Model assumes independence of symptoms given the cause:



- Naïve Bayes is a special case of Bayesian Network that is very easy to do inference on.

Naïve Bayes Model (2)

$$P(C, S_1, \dots, S_n) = P(S_1|C)P(S_2|C, S_1) \dots P(S_n|C, S_1, \dots, S_{n-1})$$

- Since all symptoms (S) are conditionally independent given the cause (C), the joint probability reduces to:

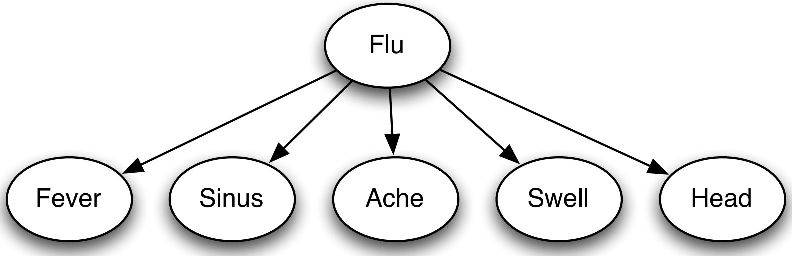
$$P(C, S_1, \dots, S_n) = P(C) \prod_{i=1}^{n} P(S_i|C)$$

- Likewise, the posterior is reduced to:

$$P(C|S_1, \dots, S_n) = \frac{P(C, S_1, \dots, S_n)}{\prod_{i=1}^{n} P(S_i)} = \frac{P(C, S_1, \dots, S_n)}{\prod_{i=1}^{n} \sum_C P(S_i|C)}$$

Naïve Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	Y	N	Y	Y
Y	M	Y	N	N	Y
?	M	Y	N	N	N



Naïve Bayes Example

Flu	Fever	Sinus	Ache	Swell	Head
Y	L	Y	Y	Y	N
N	M	N	N	N	N
Y	H	Y	N	Y	Y
Y	M	Y	N	N	Y
?	M	Y	N	N	N

$$P(Flu = Y, F = M, S = Y, A = N, Sw = N, H = N) =$$

$$0.75 \times 0.33 \times 1 \times 0.66 \times 0.33 \times 0.33 = 0.0178$$

$$P(Flu = Y | F = M, S = Y, A = N, Sw = N, H = N) =$$

$$\frac{0.0178}{P(F = M, S = Y, A = N, Sw = N, H = N)}$$

$$= \frac{0.0178}{0.5 \times 0.75 \times 0.75 \times 0.5 \times 0.5} = 0.253$$

Markov Chains

- A stochastic process $\mathbf{X}$
  - is a sequence of r.v.'s $\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(i)}, \dots\}$ which represent the state of the process at different points i
  - typically i's represent different time points
  - each $x^{(i)}$ can have one of a finite set of values $\mathbf{s} = \{s_1, s_2, \dots s_m\}$
    - Called states of the process
- A stochastic process $\mathbf{X}$ is called a Markov Chain if the state of $x^{(i+1)}$ is conditionally independent of all other points given the state of $x^{(i)}$:

$$P(x^{(i+1)}|x^{(i)}, x^{(i-1)}, \dots, x^{(1)}) = P(x^{(i+1)}|x^{(i)})$$

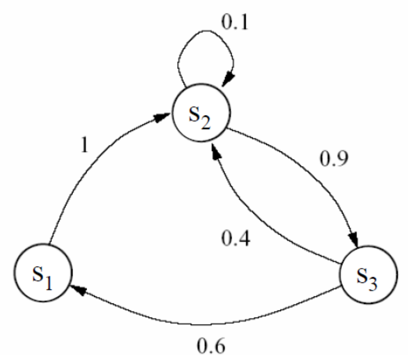
  - The probability distribution of the next state depends only on the state before it.
  - Called a homogenous Markov chain if $P(x^{(i+1)}|x^{(i)}$ does not depend on i

Transition Matrix

- When the Markov chain is homogenous, the probability distribution of the next state can be represented with a transition matrix **T** s.t.:

$$T_{jk} = P(x^{(i+1)} = s_k | x^{(i)} = s_j)$$

- For example:



$$\mathbf{T} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{pmatrix}$$

Multiple Step Transitions

- The probability of transitioning from state i to state j in two steps would be:

  $$T_{ij}^{(2)} = \sum_k T_{ik} T_{kj}$$

  so $\mathbf{T}^{(2)}$ would be:

  $$\mathbf{T}^{(2)} = \mathbf{T} \times \mathbf{T} = \mathbf{T}^2$$

  the multiplication of the transition matrix with itself.
- This can be generalized for any number of steps:

  $$\mathbf{T}^{(n)} = \mathbf{T}^n$$

  is the n-step transition matrix.

Accessibility and Irreducibility

- A state j is "accessible" from a state i ($s_i \rightarrow s_j$) if a system started in state i has a non-zero probability of transitioning into state j at some point.
- $s_i \rightarrow s_j$ if for some $n \geq 0$:

$$T_{ij}^{(n)} > 0$$

- States i and j "communicate" if both $s_i \rightarrow s_j$ and $s_j \rightarrow s_i$
- s_i is "essential" if all states j that are accessible from s_i also communicate with s_i
- A Markov chain is called irreducible if all of its states are essential.
  - In other words, all states are accessible from all other states.

Periodicity

- Some states can be periodic. If a point 0, we are at s_i, the period is the minimum number of steps required to get back to s_i:

$$period(s_i) = \gcd\{n : T_{ii}^{(n)} > 0\}$$

- If the period is 1, the state is called aperiodic.
- If all states in a Markov chain are aperiodic, the chain itself is aperiodic.

Example

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.3 & 0.7 \\ 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \end{pmatrix}$$

is periodic. All states have $period = 2$

Absorbing and Transient States

- A state, s_i, is called an absorbing state if

$$T_{ii} = 1 \quad \text{and} \quad T_{ij} = 0 \quad \text{for } i \neq j$$

  In other words, once transitioned, and absorbing state will never be left.

- The Hitting Time of a state is

$$T_i = \inf\{n \geq 1 : x^{(n)} = i | x^{(0)} = i\}$$

  the probability distribution of the first return time to s_i

- If $P(T_i < \infty) < 1$, or if there is a chance that we never return to s_i, the state is called Transient, otherwise it is called Recurrent.
- The Mean Recurrence Time: $M_i = E[T_i]$

Example

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0.3 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix}$$

- State 1 is absorbing
- State 2 is transient
- State 3 is transient
- State 4 is neither absorbing nor transient
- State 5 likewise
- States 4 and 5 are called ergodic
  - Neither transient nor periodic.

Stationary Distribution

- Let π_i be the probability that a homogenous Markov chain is at s_i at some arbitrary time t
- In other words, the π_j **is the probability that we will catch the process at s_i if we took a snapshot at a random time.**
- The definition of π_i is recursive. The probability that the process will be at s_i at time t is a function of where it is likely to be at $t - 1$

$$\pi_i = \sum_{s_j \in S} \pi_j T_{ji}$$

- The vector $\pi = \{\pi_1, \pi_2, \ldots \pi_k\}$ is called the stationary distribution of the Markov Chain if all $\pi_i > 0$ and $\sum \pi_i = 1$

Finding the Stationary Distribution

- Since
$$\pi_i = \sum_{s_j \in S} \pi_j T_{ji}$$
multiplying T with π should give π.
- In other words,
$$\pi = \pi \mathbf{T}$$
- Again, we get the eigenproblem.
- π is the normalized left eigenvector of $\mathbf{T}$ which has an eigenvalue $= 1$
  - Let the left eigenvector with eigenvalue 1 of $\mathbf{T}$ be $\mathbf{v}$
  - $\pi = \frac{\mathbf{v}}{\sum v_i}$

An Easier Way

- Another way to look at the problem is: "Where will I end up if I transition infinite times?"

$$\lim_{n \to \infty} \mathbf{T}^n$$

- For many irreducible and aperiodic Markov chains:

$$\lim_{n \to \infty} \mathbf{T}^n = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} \times \pi$$

- In other words, if you multiply **T** with itself many times, all of its rows will converge to π

Example

- Let's assume the DNA sequence of human chromosome 22 is a Markov Chain.
- So, the probability of the next nucleotide in the sequence depends only on the current one.
- The state space is $S = \{s_1 = A, s_2 = T, s_3 = C, s_4 = G\}$
- The transition matrix is:

$$\mathbf{T} = \begin{pmatrix} 0.6 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.5 & 0.1 \\ 0.1 & 0.3 & 0.1 & 0.5 \end{pmatrix}$$

- It is obvious that the chain is irreducible, aperiodic, and recurrent.
- What are the proportions of A, T, C, and G in the chromosome?

$$\mathbf{T} = \begin{pmatrix} 0.6 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.5 & 0.1 \\ 0.1 & 0.3 & 0.1 & 0.5 \end{pmatrix}$$

$$\mathbf{T}^4 = \begin{pmatrix} 0.2908 & 0.3182 & 0.2286 & 0.1624 \\ 0.2151 & 0.4326 & 0.1899 & 0.1624 \\ 0.2538 & 0.3569 & 0.2269 & 0.1624 \\ 0.2151 & 0.4070 & 0.1899 & 0.1880 \end{pmatrix}$$

$$\mathbf{T}^8 = \begin{pmatrix} 0.24596 & 0.37787 & 0.20961 & 0.16656 \\ 0.23873 & 0.38946 & 0.20525 & 0.16656 \\ 0.24309 & 0.38223 & 0.20812 & 0.16656 \\ 0.23873 & 0.38880 & 0.20525 & 0.16721 \end{pmatrix}$$

$$\mathbf{T}^{16} = \begin{pmatrix} 0.24142 & 0.38494 & 0.20692 & 0.16667 \\ 0.24135 & 0.38510 & 0.20688 & 0.16667 \\ 0.24140 & 0.38503 & 0.20691 & 0.16667 \\ 0.24135 & 0.38510 & 0.20688 & 0.16667 \end{pmatrix}$$

A: 24.1% **T:** 38.5% **C:** 20.7% **G:** 16.7%