CSE 591 - FALL 03.

Chitta Baral

Department of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287-5406 USA
chitta@asu.edu

http://www.public.asu.edu/∼cbaral/cse571-f99/

October 12, 2003

PROBABILITY, BAYES NETS AND CAUSALITY

Basic Concepts in probability theory

- 3 basic axioms of probability calculus in the Bayesian formalism
  - $0 \leq P(A) \leq 1$
  - $P(\text{Sure proposition}) = 1$
  - $P(A \vee B) = P(A) + P(B)$, if A and B are mutually exclusive.
    * $P(A) = P(A, B) + P(A, \neg B)$
      ($P(A, B)$ is short for $P(A \wedge B)$)
    * If B_i, $i = 1, 2, \ldots, n$ is a set of exhaustive and mutually exclusive propositions (called a partition or a variable), then

$$P(A) = \sum_i P(A, B_i)$$

- Basic expression in Bayesian formalism
  - Conditional probabilities of the form $P(A|B)$

– means: belief in A under the assumption that B is known with absolute certainty.

– $P(A|B) = P(A)$ – A and B are independent.

– $P(A|B, C) = P(A|C)$ – A and B are conditionally independent given C.

– Dawid's notation: $(A \amalg B|C)$

– Bayesian philosophers see the conditional relationship as more basic than that of joint events.
  $P(A \wedge B) = P(A|B)P(B)$

Bayesian Networks

- Goal:
  - to provide convenient means of expressing substantive assumptions
  - to facilitate economical representations of joint probability functions
  - to facilitate efficient inferences from observations
- Idea: Directed acyclic graphs is used to represent causal or temporal relationship
- Basic decomposition scheme
  - $P(A \wedge B) = P(A|B)P(B)$
  - $P(x_1, x_2, x_3) = P(x_1 \wedge x_2 \wedge x_3) = P(x_1|x_2, x_3)P(x_2 \wedge x_3) = P(x_1|x_2, x_3)P(x_2|x_3)P(x_3)$

– In general,

$$P(x_1, \ldots, x_n) = \prod_j P(x_j | x_1, \ldots, x_{j-1})$$

* Let $PA_j \subseteq \{x_1, \ldots, x_{j-1}\}$, such that x_j is independent of $\{x_1, \ldots, x_{j-1}\} \setminus PA_j$ once we know the value of PA_j.
* We can then write $P(x_j | x_1, \ldots, x_{j-1}) = P(x_j | pa_j)$
* If PA_j is a minimal set of predecessors of X_j that renders X_j independent of all its other predecessors, then PA_j is said to be **Markovian parents** of X_j.

- Markov factorization: If a probability function P admits the factorization

$$P(x_1, \ldots, x_n) = \prod_j P(x_j | parents_j)$$

relative to a DAG G, we say G represents P, that G and P are compatible, or P is Markov relative to G.

Inference with Bayesian Networks

- Prediction and abduction
  - x – a set of observations
  - y – a set of variables deemed important for prediction or diagnosis
  - Need to compute $P(y|x)$.
  - $$P(y|x) = \frac{p(y,x)}{p(x)} = \frac{\Sigma_s P(y,x,s)}{\Sigma_{y,s} P(y,x,s)}$$
- An example:
  - The Network
    - $P(tampering) = 0.02$; $P(fire) = 0.01$
    - Directed Edges: $(tampering, alarm)$, $(fire, alarm)$, $(fire, smoke)$, $(alarm, leaving)$, $(leaving, report)$

- local probability distributions:
  $P(alarm|fire, tampering) = 0.5;$
  $P(alarm|fire, \neg tampering) = 0.99;$
  $P(alarm|\neg fire, tampering) = 0.85;$
  $P(alarm|\neg fire, \neg tampering) = 0.0001.$
  $P(smoke|fire) = 0.9; P(smoke, \neg fire) = 0.01.$
  $P(leaving|alarm) = 0.88; P(leaving|\neg alarm) = 0.001.$
  $P(report|leaving) = 0.75; P(report|\neg leaving) = 0.01.$

– Different kinds of inferences

- Diagnostic inferences: $P(fire|report)$
- Causal inferences (prediction): $P(leaving|tampering)$
- Intercausal inferences: $P(fire|alarm, tampering)$
- Mixed inferences: $P(alarm|report, fire)$

– An illustration:

$P(tampering|report, smoke)$
$= \frac{P(tampering, report, smoke)}{P(report, smoke)}$

$$= \frac{\Sigma_{leaving,alarm,fire} P(tampering{=}T, report{=}T, smoke{=}T, leaving, alarm, fire)}{\Sigma_{tampering,leaving,alarm,fire} P(report{=}T, smoke{=}T, tampering, leaving, alarm, fire)}$$

- Let us compute the denominator D first.

  $\Sigma_{tampering,leaving,alarm,fire}\, P(tampering)\ P(fire)$
  $P(smoke = T|fire)\ P(alarm|tampering, fire)$
  $P(leaving|alarm)\ P(report = T|leaving)$
  $= \Sigma_{tampering,leaving,alarm}\, P(tampering)\ P(leaving|alarm)$
  $P(report = T|leaving)\ \Sigma_{fire}\, P(fire)\ P(smoke = T|fire)$
  $P(alarm|tampering, fire)$

- Let $f_1(alarm, tampering) = \Sigma_{fire}\, P(fire)\ P(smoke = T|fire)$
  $P(alarm|tampering, fire)$

  Now let us compute $f_1(alarm = T, tampering = T)$
  $= \Sigma_{fire}\, P(fire)\ P(smoke = T|fire)$
  $P(alarm = T|tampering = T, fire)$
  $= P(fire = T)\ P(smoke = T|fire = T)$
  $P(alarm = T|tampering = T, fire = T) +$
  $P(fire = F)\ P(smoke = T|fire = F)$
  $P(alarm = T|tampering = T, fire = F)$

$= 0.01 \times 0.9 \times 0.5 + 0.99 \times 0.01 \times 0.85$
Similarly, we can also compute $f_1(alarm = T, tampering = F)$, $f_1(alarm = F, tampering = T)$ and $f_1(alarm = F, tampering = F)$.

* We can now write the denominator as:
$\Sigma_{tampering,leaving,alarm}\, P(tampering)\ P(leaving|alarm)\ P(report = T|leaving)\ f_1(alarm, tampering)$
$= \Sigma_{tampering,leaving}\, P(tampering)\ P(report = T|leaving)\ \Sigma_{alarm}\ P(leaving|alarm)\ f_1(alarm, tampering)$
Let us denote $\Sigma_{alarm}\ P(leaving|alarm)\ f_1(alarm, tampering)$ by $f_2(leaving, tampering)$. We can compute it as we compute f_1
* The denominator can now be written as:
$= \Sigma_{tampering,leaving}\, P(tampering)\ P(report = T|leaving)\ f_2(leaving, tampering)$
$= \Sigma_{tampering}\, P(tampering)\ \Sigma_{leaving}\ P(report = T|leaving)\ f_2(leaving, tampering)$
Let us denote $\Sigma_{leaving}\ P(report = T|leaving)$

$f_2(leaving, tampering)$ by $f_3(tampering)$ and compute it like the other f_is.

* The denominator can now be written as: $\Sigma_{tampering} P(tampering)\ f_3(tampering)$

- Main Issues and challenges
  - Computing the conditional probabilities efficiently
  - Inference in general networks in NP-hard
  - Many efficient algorithms are defined for particular kind of networks (say for trees).
    * Algorithm based on message passing architecture for trees.
    * Join-tree propagation
    * Cutset conditioning
    * Hybrid combinations of the above two
    * Approximation methods: stochastic simulation.

Causal Bayesian Networks

- Motivation
  - A joint distributions tells us how probable events are and how probabilities would change with subsequent observations.
  - A causal model also tells us how these probabilities would change as a result of external interventions.

    Such a change can not be deduced from a join distribution even if fully specified.
- Importance
  - Difference between **observing** the alarm is on, and **turning** the alarm on.
  - $P(fire|alarm) > 0.01$.

    But $P(fire|do(alarm = T)) = P(fire) = 0.01$

- Causal networks can predict the effect of actions. (Simple joint distributions can not.)
- Stability and autonomy
  - Autonomy: It is possible to change one parent child relationship in the network without changing the others.
  - Stability: One can predict the effect of external interventions with minimum of extra information.
  - Autonomy and intervention: Instead of specifying a new probability function for each of the many possible interventions, we specify merely the immediate changes implied by the intervention. Because of autonomy, the change is local.
- Definition: Causal Bayesian network

  Let $P(v)$ be a probability distribution on a set V of variables, and let $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ which sets any subset X of variables to constants x. Denote by $P*$ the set of all interventional distributions $P_x(v)$, $X \subseteq V$, including $P(v)$ which represents no intervention. A DAG G is said to be a **causal Bayesian network** compatible with $P*$ iff the following three conditions hold for every $P_x \in P*$.

1. $P_x(v)$ is Markov relative to G.
2. $P_x(v_i) = 1$, for all $V_i \in X$, whenever v_i is consistent with $X = x$.
3. $P_x(v_i|pa_i) = P(v_i|pa_i)$ for all $V_i \notin X$, whenever pa_i is consistent with $X = x$.

- Properties:
  - for all v consistent with x:

$$P_x(v) = \prod_{\{i|V_i \notin X\}} P(v_i|pa_i)$$

  - For all i, $P(v_i|pa_i) = P_{pa_i}(v_i)$
    (The above ensures, conditional probabilities with respect to parents, corresponds to causal effects.)

- For all i, and for every subset S of variables disjoint of $\{V_i, PA_i\}$ we have: $P_{pa_i,s}(v_i) = P_{pa_i}(v_i)$
  (Expresses invariance of causality)

- Causal relationship is more stable than probabilistic relationships.
  - Causal relationship remains unaltered as long as no change has taken place in the environment, even when our knowledge about the environment undergoes change.
    * $(season, sprinkler)$, $(season, rain)$, $(sprinkler, wet)$, $(rain, wet)$, $(wet, slippery)$.
    * S_1 – Turning the sprinkler on would not affect rain
    * S_2 – The state of the sprinkler is independent of the state of the rain.
    * S_2 changes from false to true when we learn what season it is.
    * Given that we know the season, S_2 changes from true to false once we observe that the pavement is wet.

* S_1 remains true regardless of what we learn or know about the season or the pavement.
* Falling barometer predicts rain, does not explain it.

Functional Causal Models

- Two views of non-determinism
  - Laplace's (1814) conception of natural phenomena:
    Nature's laws are deterministic, and randomness surfaces merely due to our ignorance of the underlying boundary condition.
  - Modern (quantum mechanical) conception of physics:
    All relationships are inherently stochastic.
- Why Pearl's book uses Laplace's conception of causality
  - besides the fact that it is used in genetics, econometrics and social sciences
  - It is more general.
    * Every stochastic model can be emulated by many functional relationships (with stochastic inputs), but not the other way round;

    * Functional relationships can only be approximated as a limiting case, using stochastic models.
  - Laplacian conception is more in tune with human intuition.
  - Certain important concepts can only be defined in Laplacian framework (i.e., they can not be defined in terms of purely stochastic models.)
    * the probability that event B occurred *due to* event A.
    * the probability that event B would have been different if it were not for event A
      (they are called counterfactuals)

- (Functional) causal model:
  A causal model is a triple $M = \langle U, V, F \rangle$ where
  - U is a set of background (or exogenous, or error) variables, that are determined by factors outside the model.
  - V is a set $\{V_1, \ldots, V_n\}$ of variables, that are determined by the variables in $U \cup V$.

  – F is a set of functions $\{f_1, \ldots, f_n\}$ giving rise to a set of structural equations of the form: $x_i = f_i(pa_i, u_i)$, $i = 1, \ldots, n$

- Types of queries that can be answered using functional causal models
  – **Prediction**: Would the pavement be slippery if we *find* the sprinkler off?
  – **Interventions**: Would the pavement be slippery if we *make sure* that the sprinkler is off?
  – **Counterfactuals**: Would the pavement be slippery *had* the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?

- Prediction using Markovian causal models:
  – Causal diagram: A graph obtained by having edges from each member of PA_i to X_i.
  – If the causal diagram is acyclic then the corresponding model is called semi-Markovian.

  * the values of X variables will be uniquely determined by the U variables.
  * The joint distribution $P(x_1, \ldots, x_n)$ is determined uniquely by the distribution $P(u)$ of the error variables.
- If in addition the error terms are mutually independent, the model is called *Markovian*.
- Theorem (Pearl and Verma): Every Markovian causal model M induces a distribution $P(x_1, \ldots, x_n)$ that satisfies the Markov condition relative to the causal diagram G associated with M, that is each variable X_i is independent on all its non-descendants, given its parents PA_i in G.
- Theorem (Drudgel and Simon): For every Bayesian network G characterized by a distribution P, there exists a function model that generates a distribution identical to P.
- Advantages of doing prediction using causal-functional specification over the probabilistic specification
  * When organizing knowledge using Markov causal models reliable

assertions about conditional independence can be made without assessing numerical probabilities. (They come later when writing what f exactly is and what the $P(u_i)$'s are.)

* Functional specification is often more meaningful, natural and yields a smaller number of parameters.
* Judgemental assumptions of conditional independence of observable quantities are simplified, and made more reliable, when cast directly as judgments about the presence or absence of *unobserved* common causes. (Instead of judging whether each variable is independent of all its nondescendants, given its parents, we need to judge whether the parent set contains *all* relevant immediate causes, namely whether two omitted factors (say U_i and U_j) share a common cause.
* When some conditions in the environment undergo change, it is simpler to reassess (judgmentally) or reestimate (statistically) the model parameters knowing that the change is local, affecting just a few parameters, than reestimating the whole model from

scratch.

- Interventions and causal effects in functional models.
  - Submodels of causal models:
    Let M be a causal model,X be a set of variables in V, and x be a particular realization of X. A submodel M_x of M is the causal model $M_x = \langle U, V, F_x \rangle$, where
    $F_x = \{f_i : V_i \notin X\} \cup \{X = x\}$.
  - Effects of actions on a causal model: The effect of action $do(X = x)$ on a causal model M is given by the submodel M_x.
  - Effects of actions on other variables: The potential response (or value) of a variable Y in V after an action $do(X = x)$ denoted by $Y_x(u)$ is the solution for Y using the set of equations F_x.
  - Advantages over stochastic models
    * The analysis of interventions can be directly extended to cyclic models.
      $(demand = f(price, income, u_1); price = f'(demand, cost, u_2))$

  * Analysis of causal effects in non-Markovian models will be greatly simplified using functional models.
    (Because: There are infinitely many conditional probabilities $P(x|pa_i)$, but only finite number of functions $x_i = f_i(pa_i, u_i)$, among discrete variables X_i and PA_i.)

- **Counterfactuals**
  - Why we can not use causal Bayes nets.
    * Counterfactuals involve dealing with both actions and observations. (Effect of a drug on a patient with certain symptoms.)
    * The observations alter the conditional probabilities.
  - An example illustrating the inadequacy of using causal Bayes nets.
    * X denotes a treatment.
    * $Y = 0$ means recovery and $Y = 1$ means death.
    * Q: A certain patient Joe, took the treatment and died. Our question is whether Joe's death occurred *due* to the treatment.

I.e., What is the probability that Joe (or any patient for that matter) , who died under treatment ($x = 1, y = 1$) would have recovered ($y = 0$) had he not been treated ($x = 0$).

* An extreme case: 50% of the patients recover and 50% die in both the treatment and the control groups. (assume sample size to be infinite.) I.e. $P(y|x) = \frac{1}{2}$.
* Bayes net 0: edge-less, with $P(y, x) = 0.25$, for all x and y
* Functional model 1: $x = u_1$, $y = u_2$, with $P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2}$.
* Functional model 2: $x = u_1$, $y = xu_2 + (1 - x)(1 - u_2)$, with $P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2}$.
* Both functional model 1 and 2 correspond to the same joint probability $P(y, x) = 0.25$, for all x and y. But will give different answers.

* Answering Q using model 1 and model 2

y	u_2	x	$P_{model1}(y\|u_2, x)$	$P_{model2}(y\|u_2, x)$
0	0	0	0.25	0
0	0	1	0.25	0.25
$\underline{0}$	$\underline{1}$	$\underline{0}$	$\underline{0}$	$\underline{0.25}$
0	1	1	0	0
1	0	0	0	0.25
$\underline{1}$	$\underline{0}$	$\underline{1}$	0	0
1	1	0	0.25	0
$\underline{1}$	$\underline{1}$	$\underline{1}$	0.25	0.25

* Using model 1 the answer to Q would be 0.
  Intuitively: the treatment has no effect. 50% die and 50% recover.
* Using model 2 the answer to Q would be 1.
  Intuitively, the treatment kills 50% of the people and cures the other 50%.

- Answering counter-factual queries using functional models.
  - Counterfactual: Let Y be a variable in V in the causal model $M = \langle U, V, F \rangle$. The counterfactual sentence "The value that Y would have obtained, had X been x" is interpreted as denoting the potential response $Y_x(u)$.
  - Probabilistic causal model: Is a pair $\langle M, P(u) \rangle$, where M is a causal model and $P(u)$ is a probability function defined over the domain of U.
    - $P(y) = P(Y = y) = \Sigma_{\{u|Y(u)=y\}} P(u)$
    - $P(Y_x = y) = \Sigma_{\{u|Y_x(u)=y\}} P(u)$
    - $P(Y_x = y, X = x') = \Sigma_{\{u|Y_x(u)=y \& X(u)=x'\}} P(u)$
    - $P(Y_x = y, Y_{x'} = y') = \Sigma_{\{u|Y_x(u)=y \& Y_{x'}(u)=y'\}} P(u)$
  - One purpose of counter-factuals: We want to show that the event $X = x$ was *the cause* of the event $Y = y$.
  - So we ask the question: What is the probability that Y would not be equal to y had X not been equal to x?

– To answer the above we need to evaluate $P(Y_{x'} = y' | X = x, Y = y)$

– Given M, a three step procedure to evaluate the conditional probability $P(B_A|e)$ of a counter factual sentence "If it were A then B,", given evidence e.
(e is $X = x$ and $Y = y$. A is $X \neq x$.)

* Abduction: Update $P(u)$ by the evidence e, to obtain $P(u|e)$.
(explain the past (U) in light of the current evidence e.)
* Action: Modify M by the action $do(A)$ to obtain M_A.
(minimally bend the course of history, to comply with the hypothetical condition $X \neq x$)
* Prediction: Use the modified model $\langle M_A, P(u|e) \rangle$ to compute the probability of B.
(predicting the future (Y) on the basis of the above 2 steps.)

Evaluating Counter-factuals: an example

- The Causal relationship in a 2-man firing squad:
  - Nodes
    - U : Court orders the execution.
    - C : Captain gives a signal.
    - A : Rifleman-A shoots.
    - B : Rifleman-B shoots.
    - D : Prisoner dies.
  - Edges: $(U, C), (C, A), (C, B), (A, D), (B, D)$.
- Logical structural equations
  - $C \Leftrightarrow U$
  - $A \Leftrightarrow C$
  - $B \Leftrightarrow C$

- $D \Leftrightarrow A \vee B$

- Questions that we want to answer:
  - (prediction) : If the rifleman did not shoot, the prisoner would be alive.
  - (abduction) : If the prisoner is alive, then the captain did not signal.
  - (transduction) : If rifleman-A shot, then B shot as well.
  - (action) : If the captain gave no signal and rifleman-A decides to shoot, the prisoner will dies and B will not shoot.
  - (counter-factual) : If the prisoner is dead, then even if A were not to have shot, the prisoner would still be dead.

- Probabilistic analysis: a modification of the story
  - There is a probability $P(u = 1) = p$ that the court has ordered the execution.
  - Rifleman-A has a probability q of pulling the trigger out of nervousness. $(w = 1)$

- Rifleman-A's nervousness is independent of U.
- We wish to compute the probability that the prisoner would be alive if A were not to have shot, given that the prisoner is in fact dead.
- The solution steps:
  * (abduction) : $P(u, w|D)$
  * (action) :
  * (prediction) :