CSC 520-001/601

Introduction to Artificial Intelligence

Fall 2015

[Back to Moodle] [Back to Course at a Glance]

[Section Index] [Previous] [Next]

Outline

1. Probability Review
2. Using a Joint Distribution
3. Types of Probabilistic Inference
4. Bayes Nets
5. Testing for Conditional Independence
6. Probabilistic Reasoning with Bayes Nets

Probability Review

Assume: all variables A, B, C, D, ... are conventionally Boolean, so that all variables are either true or false.

Notation

P(A) -- the **probability** of some event A. Note that in the Boolean case this involves two values: P(A true) and P(A false), sometimes written P(A=t) and P(A=f). (Of course, we don't need both values because we can compute one from the other.)

Frequentist Semantics	The theoretical proportion of times an event occurs under repeated trials (dice, coin flip, etc.)
Bayesian Semantics	The strength of belief that an event has occurred, will occur, etc.

P(A,B) -- the **joint** probability of two events A and B. This involves 2^2 values: P(A true and B true), P(A true and B false), P(A false and B true), and P(A false and B false). Again, we need only three of these, because we can compute the other one from them, because the values in a discrete joint distribution must sum to 1.

P(A | B) -- **conditional** probability, the probability of A given B, defined by P(A | B) == P(A,B)/P(B). This also involves 4 values, but we need only two of these, because we can compute the other two from them. The values in a conditional distribution do not have to sum to 1, but pairs of them will in the discete case.

Statistics uses terms such as maximum likelihood, etc. The **likelihood** of B given A, written L(B | A), is just P(A | B).

Σ_B P(A,B) -- the sum of the joint probabilities, summed over all values taken on by B -- **marginal** probability or "marginalization" or "marginalizing out B"
Σ_B P(A,B) == P(A)
See example below.

Basic Stuff

P(A), summed over all values A can take on, must sum to 1.

The joint distribution P(A,B), summed over all combinations of values A and B can take on, must sum to 1. The same is true of all joint distributions.

P(A,B) = P(A | B) * P(B) -- sometimes called the Fundamental Law, computes a joint prob as the product of a conditional and something else

P(A | B)P(B) = P(A,B) = P(B,A) = P(B | A)P(A) so P(A | B)P(B) = P(B | A)P(A)
This is **Bayes' Theorem**.

A and B are **independent** iff P(A | B) = P(A). By the Fundamental Law, if A and B are independent, P(A,B) = P(A)P(B). Independence is a symmetric relation, i.e., if A is independent of B, then B is independent of A.

If we conditionalize all terms by the same variable, equalities remain true.

	Unconditional	Conditionalized by C
Axiom	$P(A) + P(\neg A) = 1$	$P(A\|C) + P(\neg A\|C) = 1$
The Fundamental Law	$P(A,B) = P(A\|B)P(B)$	$P(A,B\|C) = P(A\|B,C)P(B\|C)$
Bayes Theorem	$P(B\|A)P(A) = P(A\|B)P(B)$	$P(B\|A,C)P(A\|C) = P(A\|B,C)P(B\|C)$
Independence	A, B independent iff $P(A\|B) = P(A)$	A, B conditionally independent given C iff $P(A\|B,C) = P(A\|C)$
Independence Version 2	A, B independent iff $P(A,B) = P(A)P(B)$	A, B conditionally independent given C iff $P(A,B\|C) = P(A\|C)P(B\|C)$
Marginalization	$\Sigma_B\ P(A,B) = P(A)$	$\Sigma_B\ P(A,B\|C) = P(A\|C)$
Expectation	$E[f] = \Sigma_X\ f(X)\ P(X)$	$E_X[f\|C] = \Sigma_X\ f(X)P(X\|C)$

Independence and conditional independence are different concepts. Neither independence nor conditional independence implies the other.

Conditional independence is often notated as $(X \perp Y \mid Z)$, meaning X is conditionally independent of Y, given Z.

X, Y, and Z can all be *sets* of variables, not just a single variable.

Note that conditional independence is commutative: these three sentences are equivalent:

- $(A \perp B \mid C)$ -- "A is conditionally independent of B given C."
- $(B \perp A \mid C)$ -- "B is conditionally independent of A given C."
- "A and B are conditionally independent given C."

Using a Joint Distribution

With boolean vars, a table of joint probabilities looks a lot like a truth table in Propositional Logic, only with numbers. A joint table just gives a probability for each boolean combination of the variables. The entries in a joint table must sum to 1.

If we know the joint distribution, we can answer queries numerically by marginalizing, i.e., using some rows but not others.

Here's a joint table over 3 vars E, A, and T (source unspecified for the moment -- any set of 8 values summing to 1 is a legal discrete joint distribution over 3 boolean vars). The row numbers are for reference only.

Row	E	A	T	P(E,A,T)
1	t	t	t	.001
2	t	t	f	.003
3	t	f	t	.014
4	t	f	f	.006
5	f	t	t	.008
6	f	t	f	.002
7	f	f	t	.093
8	f	f	f	.873
				1.000

1. $P(E) = P(E = t) = \Sigma_{A,T}\, P(E,A,T) = .001 + .003 + .014 + .006 = .024$ -- rows for which E is t
2. $P(A) = .001 + .003 + .008 + .002 = .014$
3. $P(T) = .001 + .014 + .008 + .093 = .116$
4. $P(\neg T) = P(T = f) = 1 - P(T) = .884 = .003 + .006 + .002 + .873$
5. $P(A,T) = \Sigma_{E}\, P(E,A,T) =$

A	T	P(A,T)
t	t	.001 + .008 = .009

t	f	.003 + .002 = .005
f	t	.014 + .093 = .107
f	f	.006 + .873 = .879
		1.000

This is a joint table, so its entries must sum to 1.

6. P(A | T) = P(A,T)/P(T) =

A	**T**	**P(A | T)**
t	t	.009/.116 = .078
t	f	.005/.884 = .006
f	t	.107/.116 = .922
f	f	.879/.884 = .994

This is a conditional prob table, so its entries need not sum to 1, although P(A | T) + P(~A | T) = 1 etc. still. Thus rows 1 and 3 above sum to 1, and so do rows 2 and 4.

7. Similarly for any other probability involving these three variables.

Two things are true about joint distributions in probabilistic inference in AI systems:

1. **If** we know the full joint distribution over a set of random variables, we can answer any query about those variables, by using a combination of marginalization and Bayes Theorem.
2. **We never want to compute the full joint distribution!** Doing so is far too expensive, because the size of a distribution is exponential in the number of variables in it. In fact, in complex cases there may not be any easily computed form for the joint distribution at all.

Types of Probabilistic Inference

Here is Bayes Theorem again:
P(A | B)P(B) = P(A,B) = P(B | A)P(A)
or just
P(A | B)P(B) = P(B | A)P(A)

We are interested in using Bayes Theorem to reason about hypotheses or explanations or diagnoses, given evidence or data or observations.

In place of of A, let's use the variable H to represent a hypothesis or explanation or diagnosis. In place of of B, let's use the variable D to represent the evidence/data/observations we have observed or collected.

So now Bayes Theorem becomes: P(H | D)P(D) = P(D | H)P(H)

- **Bayesian Model Averaging:**

  We want to use the evidence D as input to produce a hypothesis H as output, so Bayes Theorem can be written as

  P(H | D) = P(D | H)P(H)/P(D)

  which tells us how to find the probability of a hypothesis, given the data available. To find P(D), we can use $P(D) = \sum_h P(D,H) = \sum_h P(D \mid h)P(h)$, marginalizing over all hypotheses.

Such an approach, which is also known as Bayesian Prediction, requires that we know three of the terms of Bayes Theorem to compute the fourth. This can actually be a difficult requirement, because this compact notation can represent a lot of probability values. For example, if we wish to consider 10 hypotheses and 100 sets of evidence, then this use of Bayes Theorem will require 10x100=1,000 conditional probabilities, 10 prior probabilities for the hypotheses, and 100 prior probabilities for the evidence, and will compute 10x100=1,000 values, the probability of each hypothesis for each data set. This is often far more than we really need.

We can simplify Bayes Theorem in two successive ways.

- **Maximum A Posteriori (MAP):**

  Usually we are not interested not in every probability for every possible hypothesis for every possible data set, but only the hypothesis of highest probability. Since the probability of our data doesn't depend on any hypothesis, we can eliminate the P(D) term and rewrite Bayes Theorem as

  $P(H \mid D) = \alpha_1 P(D \mid H)P(H)$, where α_1 is a constant of proportionality.

  We can use this to get a ranking of hypotheses according to their relative probabilities. This approach will give us what is known as the Maximum A Posteriori (MAP) hypothesis. In MAP prediction, we use only the "best" hypothesis, which we take as the one of greatest posterior probability. We can use the hypothesis priors to penalize complexity: the larger, more complex the model, the lower its prior.

  MAP reasoning eliminates the need to worry about the probability of the data itself, but at the cost of no longer being able to discover the full probability distribution over hypotheses.

- **Maximum Likelihood (ML):**

  We can simplify Bayes Theorem even further. Suppose we are willing to assume that every hypothesis is equally probable. Then we no longer need the P(H) term, and can rewrite Bayes Theorem as

  $P(H \mid D) = \alpha_2 P(D \mid H)$, where α_2 again is a (different) constant of proportionality.

  Again, this will give us a ranking of hypotheses according to their relative probabilities. Because $P(D \mid H) = L(H \mid D)$, the likelihood of hypothesis H given data D, this simplest of all uses of Bayes Theorem will give us the Maximum Likelihood (ML) hypothesis.

  ML reasoning eliminates the need to worry about the probability distribution over hypotheses, but at the cost of no longer being able to prefer simpler hypotheses over more complex or larger ones, because in ML all hypotheses are considered equally probable.

Bayes Nets

The Chain Rule

For any set of n boolean variables, e.g. {A,B,C,D,E}, n= 5, choose an arbitrary ordering, e.g., ABCDE.

Then the joint distribution will be given by
P(A,B,C,D,E) = P(A)P(B | A)P(C | A,B)P(D | A,B,C)P(E | A,B,C,D) where each term represents the probability of a variable, conditioned on all the variables which precede it in the ordering.

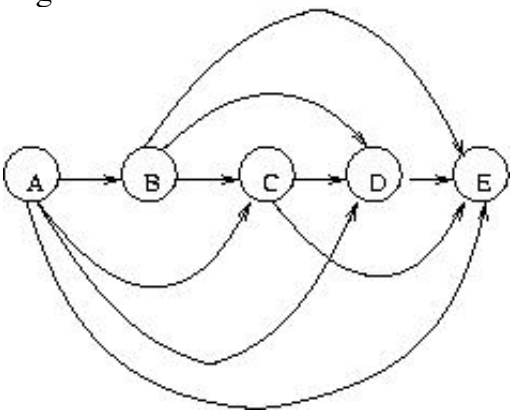
This general result is known in probability as the **Chain Rule** and is independent of the ordering, so

P(A,B,C,D,E) = P(A)P(B | A)P(C | A,B)P(D | A,B,C)P(E | A,B,C,D)
= P(E)P(D | E)P(C | D,E)P(B | C,D,E)P(A | B,C,D,E)
= P(D)P(C | D)P(A | D,C)P(E | D,C,A)P(B | D,C,A,E)
etc. for all the (5! = 120) possible orderings of {A,B,C,D,E}.

This is not saying that all these formulas are equally easy to compute.

A Graphical Representation

For n = 5, we can construct a graphical representation of the chain rule with this graph on 5 nodes and 10 edges:



This graph contains one node for each variable, and an edge to each node from the variables on which its probability is conditioned. Or, to go from the graph to the distribution,
$P(A,...,E) = \prod_i P(i \mid \text{parents}(i))$, where parents(i) means all parents of node i in the graph.

We call these graphs **Bayes Nets** or Bayesian Belief Networks. Some authors use the more general term **Probabilistic Graphical Models**, of which Bayes Nets are but one special case. They also bear a strong resemblance to Markov Random Fields, Causal Networks, and Influence Diagrams, but if you don't know what those are, don't worry.

A Bayes Net is a graphical representation of the **minimal** set of conditional probabilities over a set of variables necessary to (re)produce the joint distribution.

Because of the chain rule, the 5-node, 10-edge graph is the densest possible Bayes Net on 5 nodes.

The Impact of Conditional Independence

Now suppose some of our 5 variables are conditionally independent. Specifically, suppose:

1. C is conditionally independent of B given A,
2. D is conditionally independent of B given A and C, and

3. E is conditionally independent of (A, B, and D) given C.

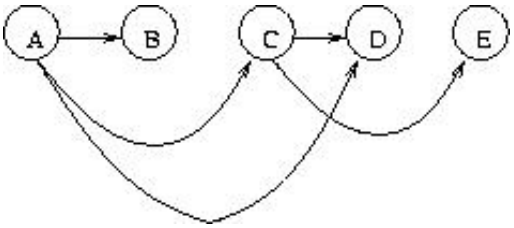
From the definition of conditional independence:

	Statement	written	means
1.	C is conditionally independent of B given A.	$(C \perp B \mid A)$	$P(C \mid A,B) = P(C \mid A)$
2.	D is conditionally independent of B given A and C.	$(D \perp B \mid A, C)$	$P(D \mid A,B,C) = P(D \mid A,C)$
3.	E is conditionally independent of A, B, and D given C.	$(E \perp A, B, D \mid C)$	$P(E \mid A,B,C,D) = P(E \mid C)$

Now we can substitute simpler terms for more complex terms in the Chain Rule expression:

$$P(A,B,C,D,E) = P(A)\ P(B \mid A)\ P(C \mid A,B)\ P(D \mid A,B,C)\ P(E \mid A,B,C,D)$$
$$= P(A)\ P(B \mid A)\ P(C \mid A)\ \ P(D \mid A,C)\ \ P(E \mid C)$$

and the corresponding Bayes Net based on $P(A,...,E) = \prod_i P(i \mid \text{parents}(i))$ looks like this:



where 5 edges are gone because of conditional independence.

Moreover, even for Boolean variables the naive original Chain Rule expression contains 15 variable instances, and so requires multiplication over 2^{15}=32,768 values, while the version with conditional independence requires 2^{10}=1,024, a 32-fold savings.

With some experience we can identify conditional independence from a Bayes Net, by considering the edges which are NOT in the graph.

Because

1. The size of a conditional or joint probability table is exponential in the number of variables,
2. Variables that are conditionally independent of a table don't have to be in it,
3. We can answer a query correctly by computing only part of the joint distribution if there's conditional independence in the data,
4. The complexity of inference using a Bayes Net turns out to be proportional to the number of walks (paths without regard to edge direction) in it,
5. Real problems are usually characterized by lots of conditional independence,

any reduction in edges because of conditional independence is a good thing. It is from this that Bayes Nets derive their usability.

Testing for Conditional Independence

How can we tell if two random variables are conditionally independent given one or more other variables? Go back to the definition of conditional independence and use the definition of conditional

probability derived from Bayes Theorem:
(A ⊥ B | C)
iff P(A | B, C) = P(A | C)
iff P(A, B, C)/P(B, C) = P(A, C)/P(C).

If we have a table of values, we simply test these ratios for equality.

As an example, let's test for (E ⊥ A | T) in the tables above:
(E ⊥ A | T)
iff P(E | A, T) = P(E | T)
iff P(E, A, T)/P(A, T) = P(E, T)/P(T)
iff .001/(.001 + .008) = (.001 + .014)/(.001 + .014 + .008 + .093)
iff .001/.009 = .015/.116
iff .111 = .129 which is false, so the conditional independence is not true.

Probabilistic Reasoning with Bayes Nets

In general, probabilistic reasoning:

1. Represents objects, concepts, etc. as random variables.
2. Models knowledge (or beliefs) about the value of such a variable with a probability distribution across its possible values.
3. Models relationships among variables with conditional probability distributions. This includes the logical relationships we've already dealt with, including facts, implications, rules, actions, etc.
4. Modifies these distributions as knowledge/evidence is updated.
5. Provides answers to queries in the form of probabilities, although in particular systems responses can be in the form of the highest probability outcome, a rank ordering, the n most probable outcomes, etc. with or without the actual probability values.

In particular, a Bayes Net consists of:

1. A directed acyclic graph (DAG) with
2. A variable labeling each node, and
3. For each node, a probability distribution giving the conditional probability of a variable given the values of its parents.

When new evidence comes in, inference in Bayes Nets:

1. Updates the distributions of the appropriate variables. (See the section on Incorporating Evidence later.)
2. Propagates this update along the network to (potentially) all other nodes, (potentially) causing updates to their own distributions.

Some Nice Results

Because conditional independence is reflected in a Bayes Net by the **absence** of edges, there are some handy results relating graph structure to independence.

RESULT 1: Every node is conditionally independent of the rest of the network given the values of its:

1. parents,
2. children, and
3. spouses (other parents of its children).

Equivalently, a node is conditionally independent of its non-descendants, given its parents.

COROLLARY 1: Children are conditionally independent of their grandparents, given their parents.

COROLLARY 2: Children are independent of each other, given their parents.

COROLLARY 3: Parents are **dependent** on each other, given their children.

These corollaries correspond to simple subgraph patterns you can look for in any Bayes net.

One Application: Spam Filtering

[Original] [Improved]

[Section Index] [Previous] [Next]

Top of Page
