Review of Probability Theory

Outline

◇ Uncertainty

◇ Probability basics

◇ Random variables

◇ Probability Distributions

◇ Probabilistic Inference

◇ Independence

◇ Bayes' Rule

From Categorical to Uncertain Beliefs

◇ Intelligent behavior requires knowledge about the world

◇ Propositional logic provides a valuable tool for representing and reasoning with categorical beliefs about the world

◇ Any sentence could be true, false, or unknown

◇ Often, we are uncertain about the state of the world

Wumpus World

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 **B** OK	2,2	3,2	4,2
1,1 OK	2,1 **B** OK	3,1	4,1

Pits are placed randomly, 20% chance per square

◇ Logical agent: choose randomly

◇ Probabilistic agent: 31% probability [1,3] contains a pit, 86% for [2,2]

Representing and Reasoning under Uncertainty

Example of reasoning under uncertainty

$\diamondsuit$ Beliefs:

    -If a patient has lung cancer, there is a 60% chance that an X-ray test will come back positive ; and a 40% percent chance negative.

    -If a patient does not have lung cancer, there is 2% percent chance that an X-ray test will come back positive ; and a 98% percent chance negative.

    -Population cancer rate is $1/1000$

$\diamondsuit$ Observation: X-ray test came back positive.

$\diamondsuit$ Inference task: What is the chance that the patient has lung cancer?

Probability Theory provides a framework for representing and reasoning under uncertainty

Probability basics

Probability deals with chance experiments that have a set of distinct **outcomes**.

    For example, we roll a die and the possible outcomes are 1, 2, 3, 4, 5, 6 corresponding to the side that turns up.

    We toss a coin with possible outcomes H (heads) and T (tails).

The set of all possible outcomes is called the **sample space**, generally denoted by Ω.

    Example: A die is rolled once. Then the sample space for this experiment is the 6-element set $\Omega = \{1, 2, 3, 4, 5, 6\}$.

$\omega \in \Omega$ is a sample point/possible world/atomic event.

    - a complete specification of the state of the world

    - mutually exclusive

    - exhaustive

Probability basics

A probability space or probability model is a sample space
with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.

$\quad 0 \leq P(\omega) \leq 1$
$\quad \Sigma_{\omega} P(\omega) = 1$

For example, unless there is reason to believe the die is loaded, the natural
assumption is that every outcome is equally likely,
$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An event A is any subset of Ω.

$\quad$ Example: The event that the result of the roll is an even number is
$A = \{2, 4, 6\}$.

$$P(A) = \Sigma_{\{\omega \in A\}} P(\omega)$$

E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

Random variables

A random variable is a function from sample points to some range, e.g., the
reals or Booleans

$\quad$ e.g., $Odd(1) = true$.

The set of values that a random variable X can assume is called the **space/domain**
of X.

A probability space P induces a probability distribution for any r.v. X:

$$P(X = x_i) = \Sigma_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

e.g., $P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

Random variables

The set of values that a random variable X can assume is called the **space/domain** of X, $Dm(X)$. The values are mutually exclusive and exhaustive.

Propositional or Boolean random variables (have the domain $\{true, false\}$)

    e.g., $Cavity$ (do I have a cavity?)

Discrete random variables (domain is finite or countably infinite)

    e.g., $Weather$ is one of $\{sunny, rain, cloudy, snow\}$

Continuous random variables (take on values from the real numbers)

    e.g., $Temp$

Random variables

Often in AI applications, random variables are basic elements

    - the sample points are **defined** by the values of a set of random variables

    - A possible world is an assignment of exactly one value to every random variable.

    - the sample space is the (Cartesian product of the) domains of the variables

E.g., if the world consists of only two Boolean variables Cavity and Toothache, then there are 4 distinct atomic events or 4 possible worlds:

$$Cavity = false \wedge Toothache = false$$
$$Cavity = false \wedge Toothache = true$$
$$Cavity = true \wedge Toothache = false$$
$$Cavity = true \wedge Toothache = true$$

Knowledge Representation

Probability theory as a knowledge representation

◇ Basic element: random variables. → propositional symbols in PL

◇ **Propositions** represent the agent's beliefs about the world. (what do we want to talk about?) → logical sentences

◇ Elementary proposition asserts that a random variable has a particular value

    e.g., $Cavity = true$, $Weather = rain$

◇ Complex propositions formed from elementary propositions and standard logical connectives

◇ What is the **degree of belief** that a given proposition is true (what can we believe?) → probabilities

◇ Degrees of belief are always applied to propositions.

    e.g., $P(Cavity = true \lor Toothache = false) = 0.1$ and $P(Weather = sunny) = 0.72$

Probability Distributions

Probability distribution function (probability mass function) gives probability values for all possible values of a random variable

$P(Weather = sunny) = 0.72, \quad P(Weather = rain) = 0.1$

$P(Weather = cloudy) = 0.08, \quad P(Weather = snow) = 0.1$

$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Probability distribution must satisfy

$$0 \le P(x) \le 1, \quad \sum_{x \in Dm(X)} P(x) = 1$$

Joint Probability Distributions

Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every combination of the values of the set of r.v.s)

$\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

Joint probability distribution must satisfy
$$0 \leq P(\vec{x}) \leq 1, \qquad \sum_{\vec{x} \in Dm(\vec{X})} P(\vec{x}) = 1$$

Knowledge base: joint probability distribution
Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Conditional probability

Prior or unconditional probabilities
e.g., $P(Cavity = true) = 0.1$ and $P(Weather = sunny) = 0.72$
correspond to belief prior to arrival of any (new) evidence

Conditional or posterior probabilities
e.g., $P(cavity|toothache) = 0.8$
i.e., **given that** $toothache$ **is all I know**

If we know more, e.g., $cavity$ is also given, then we have
$P(cavity|toothache, cavity) = 1$

New evidence may be irrelevant, allowing simplification, e.g.,
$P(cavity|toothache, sunny) = P(cavity|toothache) = 0.8$
This kind of inference, sanctioned by domain knowledge, is crucial

(Notation for conditional distributions:
$\mathbf{P}(Cavity|Toothache) =$ 2-element vector of 2-element vectors)

Conditional probability

Definition of conditional probability (a shorthand for $A = a$):

$$P(a|b) = \frac{P(a,b)}{P(b)} \text{ if } P(b) \neq 0$$

Product rule gives an alternative formulation:

$$P(a,b) = P(a|b)P(b) = P(b|a)P(a)$$

A general version holds for whole distributions, e.g.,

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$$

(View as a 4×2 set of equations, **not** matrix mult.)

Chain rule is derived by successive application of product rule:

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1, \ldots, X_{n-1}) \, \mathbf{P}(X_n|X_1, \ldots, X_{n-1})$$
$$= \mathbf{P}(X_1, \ldots, X_{n-2}) \, \mathbf{P}(X_{n-1}|X_1, \ldots, X_{n-2}) \, \mathbf{P}(X_n|X_1, \ldots, X_{n-1})$$
$$= \ldots$$
$$= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$$

Probability for continuous variables

For continuous variables, it is not possible to write out the entire distribution as a table because there are infinitely many values. In fact, it no longer makes sense to talk about the probability that X has a particular value.

Express distribution as a parameterized function of value, called **probability density function**
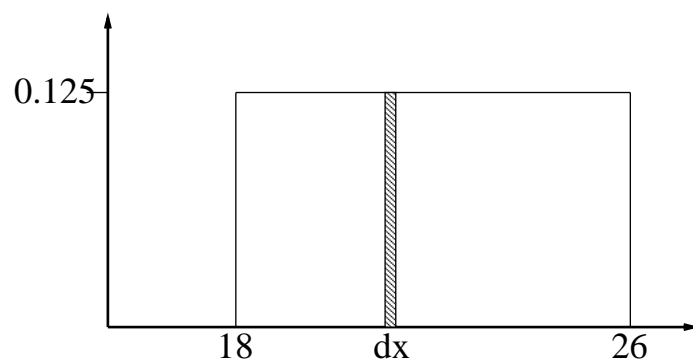
$$P[x \in (a,b)] = \int_a^b p(x)dx$$

$$p(x) \geq 0, \quad \int p(x)dx = 1$$

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

Probability for continuous variables

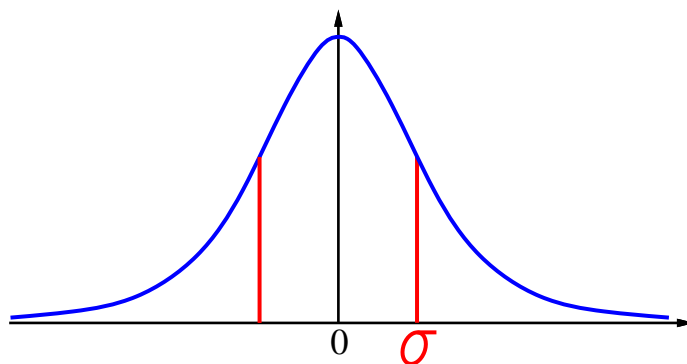$p(x) = U[18, 26](x) =$ uniform density between 18 and 26

Here $p(x)$ is a density; integrates to 1.

$$\lim_{dx \to 0} P(20.5 \le X \le 20.5 + dx)/dx = 0.125$$

Gaussian/Normal density

$p(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$

$$P(|x - \mu| \le \sigma) = 0.68$$

Expected Values

The expectation of the r.v. X (expected value, mean, average)

$$E[X] = \sum_x x P(x)$$

The expectation of function $f(x)$

$$E[f(x)] = \sum_x f(x) P(x)$$

$$E[f(x)] = \int_x f(x) p(x) dx$$

The variance (and the standard deviation) provides a measure of variability around mean value
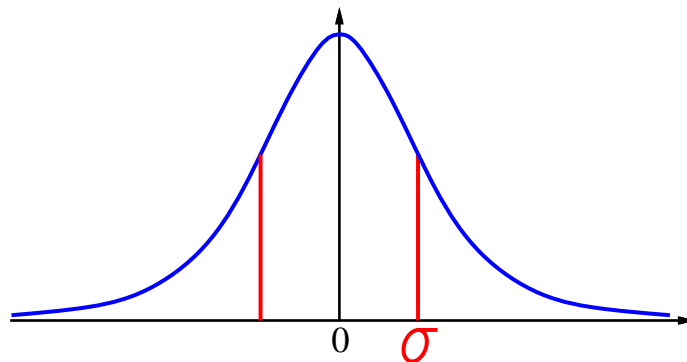
$$var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

The covariance: a measure of statistical dependence between X and Y

$$cov[X, Y] = E[(X - E[X])(Y - E[Y])] = \sum_{x,y} (x - E[X])(y - E[Y]) P(x, y)$$

19

Gaussian/Normal density

$$p(x) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

0   σ

$$E[X] = \mu, \quad var[X] = \sigma^2$$

σ is called the standard deviation

$$P(|x - \mu| \le \sigma) = 0.68$$

20

Inference by enumeration

Start with the joint distribution (as the KB):

(Cavity may cause the dentist's probe to catch)

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

Inference by enumeration

Start with the joint distribution:

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

For any event, sum the atomic events where it is true:

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Probabilistic Inference

One common task is to extract the distribution over a single variable or some subset of variables, called **marginal distribution**

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

$P(\neg toothache) = \ldots = 0.8$

This process is called **marginalization** or **summing out**. For any sets of variables Y and Z

$$P(Y) = \sum_z P(Y, z)$$

A distribution over Y can be obtained by summing out all the other variables from any joint distribution containing Y

A variant of this rule is called **conditioning**

$$P(Y) = \sum_z P(Y|z)P(z)$$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

Can also compute conditional probabilities:

$$P(\neg cavity|toothache) = \frac{P(\neg cavity, toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Normalization

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

Denominator can be viewed as a normalization constant α

$$\mathbf{P}(Cavity|toothache) = \alpha\,\mathbf{P}(Cavity, toothache)$$
$$= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$
$$= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]$$
$$= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle$$

Probabilistic Inference

Let $\mathbf{X}$ be all the variables. Typically, we want
the posterior distribution of the query variables $\mathbf{Y}$
given specific values $\mathbf{e}$ for the evidence variables $\mathbf{E}$

Let the other variables be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the other variables:

$$\mathbf{P}(\mathbf{Y}|\mathbf{E}\!=\!\mathbf{e}) = \Sigma_{\mathbf{h}}\mathbf{P}(\mathbf{Y}, \mathbf{H}\!=\!\mathbf{h}|\mathbf{E}\!=\!\mathbf{e}) = \alpha\Sigma_{\mathbf{h}}\mathbf{P}(\mathbf{Y}, \mathbf{E}\!=\!\mathbf{e}, \mathbf{H}\!=\!\mathbf{h})$$

The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ together exhaust the set of random variables

Probabilistic Inference

In principle, joint distributions can be used to answer any probabilitic queries.

Obvious problems:
   1) Worst-case time complexity $O(d^n)$ where d is the largest arity
   2) Space complexity $O(d^n)$ to store the joint distribution
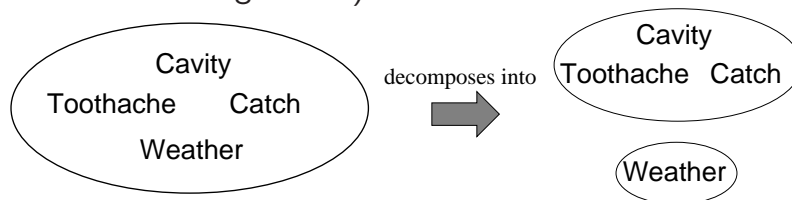   3) How to find the numbers for $O(d^n)$ entries???

Use domain knowledge to relieve this problem: one type of knowledge is independence relationships

Independence

Two sets of variables A and B are independent iff
$$\mathbf{P}(A|B) = \mathbf{P}(A) \quad \text{or} \quad \mathbf{P}(B|A) = \mathbf{P}(B) \quad \text{or} \quad \mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$$
(for all possible value assignments)

$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$
$$= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

32 entries reduced to 12; for n independent biased coins, $2^n \to n$

Independence assertions are usually based on knowledge of the domain. They can help in reducing the size of the domain representation and the complexity of the inference. Unfortunately absolute independences are rare. E.g., Dentistry involves hundreds of diseases/symptoms, all of which interrelated

Conditional independence

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) $P(catch|toothache, cavity) = P(catch|cavity)$

The same independence holds if I haven't got a cavity:

(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ is conditionally independent of $Toothache$ given $Cavity$:

$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

Conditional independence

X is independent of Y given Z, denoted $I(X, Z, Y)$, iff

$$P(x|y, z) = P(x|z), \forall x \in Dm(X), y \in Dm(Y), z \in Dm(Z)$$

Equivalent statements:

$$P(Y|X, Z) = P(Y|Z), \quad P(X, Y|Z) = P(X|Z)P(Y|Z)$$

$X_1, \ldots, X_n$ are mutually independent given Z if

$$P(X_1, \ldots, X_n|Z) = \prod_i P(X_i|Z)$$

Conditional independence contd.

$\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries

Write out full joint distribution using chain rule:

$\mathbf{P}(Toothache, Catch, Cavity)$
$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch, Cavity)$
$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$
$= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$

I.e., $2 + 2 + 1 = 5$ independent numbers (equations 1 and 2 remove 2)

In some cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.

Conditional independence assertations are much more commonly available than absolute independence assertations

Conditional independence is one of the most basic and robust form of knowledge about uncertain environments.

Bayes' Rule

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$\implies \text{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

or in distribution form

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X|Y)\mathbf{P}(Y)$$

$(1/P(X)$ is often viewed as a normalization constant $\alpha)$

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Plays a central role in machine learning

Example use of Bayes' Rule

A patient takes a X-ray test, comes back positive for lung cancer

Test accuracy: false negative rate of .4
$$P(test = +|cancer = yes) = .6$$
false positive rate of .02:
$$P(test = +|cancer = no) = .02$$

Population cancer rate .001
$$P(cancer = yes) = .001.$$

Answer

$$P(yes|+) = .6 * .001/(.6 * .001 + .02 * .999) = .029$$

Bayes' Rule and conditional independence

$\mathbf{P}(Cavity|toothache, catch)$
$= \alpha \, \mathbf{P}(toothache, catch|Cavity)\mathbf{P}(Cavity)$
$= \alpha \, \mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)$

This is an example of a naive Bayes model:

$$\mathbf{P}(Cause|Effect_1, \ldots, Effect_n) = \alpha\mathbf{P}(Cause)\prod_i\mathbf{P}(Effect_i|Cause)$$

Total number of parameters is **linear** in n

Wumpus World

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 **B** OK	2,2	3,2	4,2
1,1 OK	2,1 **B** OK	3,1	4,1

Pits are placed randomly, 20% chance per square

$$P(p_{1,3}|\neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}, \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}) = .31$$

$$P(p_{2,2}|\neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}, \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}) = .86$$

Summary

Probability is a rigorous formalism for uncertain knowledge

Joint probability distribution specifies probability of every atomic event

Queries can be answered by summing over atomic events

For nontrivial domains, we must find a way to reduce the joint size

Independence and conditional independence provide the tools