CSC384 Summer 2018

Week 7 - Probability Review

Ilir Dema

University of Toronto

Jul 9, 2018

Overview

Reasoning Under Uncertainty

- Chapter 13 (R&N, 3rd edition)
  - Basic background on probability from the point of view of AI.
- Chapter 14 (R&N, 3rd edition)
  - Bayesian Networks, exact reasoning in Bayes Nets as well as approximate reasoning.

UNCERTAINTY

- For the most part we have dealt with deterministic actions.
  - If you are in state S_1 and you execute action A you will always arrive at a particular state S_2.
- When there is a fixed initial state S_0, we will know exactly what state we are in after executing a sequence of deterministic actions (yours and the actions of the other agents).
- These assumptions are sensible in some domains
- But in many domains they are not true.
  - We have already seen some modeling of uncertainty in Expectimax search where we were not sure what our opponent would do.
  - But the actions were still deterministic- we just didn't know which action was executed.

Uncertainty

- We might not know exactly what state we start off in
  - E.g., we can't see our opponents cards in a poker game
  - We don't know what a patient's ailment is.
- We might not know all of the effects of an action
  - The action might have a random component, like rolling dice.
  - We might not know all of the long term effects of a drug.
  - We might not know the status of a road when we choose the action of driving down it.
- In many applications we cannot ignore this uncertainty.
  - In some domains we can (e.g., build a schedule with some slack in it to account for delays).

UNCERTAINTY

- In such domains we still need to act, but we can't act solely on the basis of known true facts. We have to "gamble"
- E.g., we don't know for certain what the traffic will be like on a trip to the airport.
- But how do we gamble rationally?
  - If we must arrive at the airport at 9pm on a week night we could "safely" leave for the airport 1/2 hour before.
    - Some probability of the trip taking longer, but the probability is low.
  - If we must arrive at the airport at 4:30pm on Friday we most likely need 1 hour or more to get to the airport.
    - Relatively high probability of it taking 1.5 hours.
- Acting rationally under uncertainty typically corresponds to maximizing one's expected utility.
  - various reason for doing this.

Maximizing Expected Utility

- Don't know what state arises from your actions due to uncertainty.
- But if you know (or can estimate) the probability you are in each of these different states (i.e., the probability distribution) you can compute the expected utility and take the actions that leads to a distribution with highest expected utility.

Maximizing Expected Utility

- Probabilities of different outcomes.

Event	Go to Bloor St.	Go to Queen St.
Find Ice Cream	0.5	0.2
Find Donuts	0.4	0.1
Find Live Music	0.1	0.7

- Your utility of different outcomes.

Event	Utility
Ice Cream	10
Donuts	5
Music	20

Maximizing Expected Utility

- Expected utility of different actions.

Event	Go to Bloor St.	Go to Queen St.
Find Ice Cream	0.5*10	0.2*10
Find Donuts	0.4*5	0.1*5
Find Live Music	0.1*20	0.7*20
Utility	9.0	16.5

- Maximize Expected Utility would say that you should “Go to Queen Street”
- But it would recommend going to Bloor if you liked ice cream and donuts more than live music.

UNCERTAINTY

- To use the principle of maximizing expected utility we must have the probabilities
- So we need mechanisms for representing and reason about probabilities.
- We also need mechanisms for finding out utilities or preferences. This also is an active area of research.

Sets

Definition

1. A **set** is a collection of distinct objects called **elements** that does not contain itself as an element.
2. There exists a relation denoted $\in$ such that for any object x and for any set S, the statement $x \in S$ is a predicate.
3. A predicate $P(x)$ such that $S = \{x|P(x)\}$ is a set is said **specifies** the set S.

Does every predicate specify a set?

Consider $P(x) : x$ is a set. Let $S = \{x|P(x)\}$. What happens?

Fundamental operations on sets

- In what follows we will assume we are working over a set U, often called *universe*.
- A predicate defined over U, specifies a set V. Then the implication $x \in V \implies x \in U$ is true for all $x \in V$ and false for all $x \notin V$. We say V is a subset of U, write $V \subseteq U$.
- Let $V, W \subseteq U$. The predicate $P(x) : x \in V \vee x \in W$ specifies a subset of U, called *union* $V \cup W$.
- The predicate $Q(x) : x \in V \wedge x \in W$ specifies *the intersection* $V \cap W$.
- For any $V \subseteq U$, the predicate $x \notin V$ specifies the *complement* V^c of V wrt U.
- A set S such that $x \in S$ is always false is called *empty* set, denoted $\emptyset$.
- Exercise: Prove that $U^c = \emptyset$.

FINITE SETS.

- A set S is called **finite** if $S = \emptyset$ or if there exist $n \in \mathbb{N}$ and a bijection $f : \{1, \ldots, n\} \to S$. If so, n is called *cardinality* of S, denoted $|S|$. The cardinality of empty set is zero.
- A set is infinite if it is not finite.
- By definition, we will call "event" any subset of U.
- If $A \subseteq U$ and $|A| = 1$, then A will be called "atomic event".

Probability over finite sets.

- Probability is a function defined over subsets of U (the universe of events), satisfying the following axioms:
  - For all $A \subseteq U$, $0 \leq P(A) \leq 1$.
  - For all $A, B \subseteq U$, $A \cap B = \emptyset \implies P(A \cup B) = P(A) + P(B)$.
  - $P(U) = 1$.
- Using P we may define a function $p : U \to [0,1]$ as follows:

  For all $x \in U$, $p(x) = P(\{x\})$.

Exercises

1. Any subset of a finite set is a finite set.
2. For any $A \subseteq U$, $P(A) = \sum_{x \in A} p(x)$.
3. $P(\emptyset) = 0$.

Additive law for probabilities.

Theorem. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof.

Recall $A \setminus B = A \cap B^c$.

$A = A \setminus B \cup (A \cap B), A \setminus B \cap (A \cap B) = \emptyset$.

Then $P(A) = P(A \setminus B) + P(A \cap B)$.

Similarly $P(B) = P(B \setminus A) + P(B \cap A)$.

Then
$P(A) + P(B) = P((A \setminus B) \cup (A \cap B) \cup (B \setminus A)) + P(A \cap B)$.

The result follows.

Probability over Feature Vectors.

- Like CSPs, we have
  - a set of variables $V_1, \ldots, V_n$.
  - a finite domain (set of values) for each variable $Dom[V_i]$.
- Each variable represents a different feature of the world that we might be interested in knowing.
- Each different total assignment to these variables will be an atomic event.
- So there are $\Pi_{i=1}^{n} |Dom[V_i]|$ assignments.
- Example
  - $V_i, Dom[V_i] = \{1, 2, 3\}, i = 1, 2$
  - There are 9 atomic events in this feature vector space.
- How many atomic events for $V_i \in \{0, 1\}, i = 1, \ldots, n$?

Probability over Feature Vectors.

- By fixing values of some of the variables, say $V_1 = a_1, \ldots, V_k = a_k$, we can restrict the set of atomic events to
$$(a_1, \ldots, a_k, x_{k+1}, \ldots, x_n)$$
where $x_i \in D[V_i]$, we have shortened *Dom* to D.
- The, we can write

$$P(V_1 = a_1, \ldots, V_k = a_k)$$
$$= \sum_{x_{k+1} \in D[V_{k+1}], \ldots, x_n \in D[V_n]} P(V_1 = a_1, \ldots, V_k = a_k, V_{k+1} = x_{k+1}, \ldots, V_n = x_n)$$

PROBABILITY OVER FEATURE VECTORS

- Problem:
  - There are an exponential number of atomic probabilities to specify. (can't get all that data)
  - Computing, e.g, $P(V_1 = a)$ would requires summing up an exponential number of items. (even if we have the data, we cannot compute efficiently with it)
- AI techniques for dealing with these two problems involve:
  - Using knowledge of conditional independence to simplify the problem and reduce the data and computational requirements.
  - Using approximation techniques after we can simplify with conditional independence. (Many approximation methods rely on having distributions structured by independence).

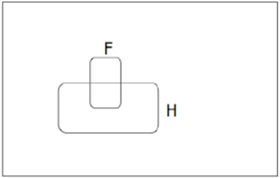
The idea of conditional probability

$P(A|B)$ is fraction of the worlds in which B is true that also have A true. Let U, F, H be the following sets:

$$U = \{x | Person(x)\},$$

$$H = \{x | Person(x) \text{ and } x \text{ has a headache}\},$$

$$F = \{x | Person(x) \text{ and } x \text{ has flu}\}$$



$$P(H) = 1/10$$
$$P(F) = 1/40$$
$$P(H|F) = 1/2$$

It is rare to have a headache, it is even rarer to have a flu, but if one has a flu, chances are 50%-50% one has a headache as well.

Conditional Probabilities

1. Conditional probabilities capture conditional information, i.e. information about the influence of any one variable's value on the probability of others.
2. Conditional probabilities are essential for both representing and reasoning with probabilistic information.

Definition

Let A, B events such that $P(A) > 0$. Then conditional probability of B w.r.t A is called the number

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

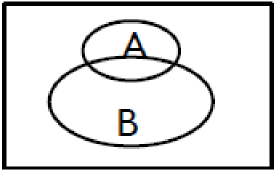
Conditional Probability Properties.

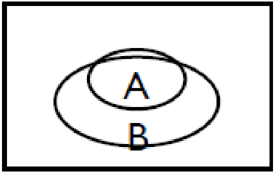
A conditional probability function is simply the restriction of the function P over a subset $A \subseteq U$. The following properties hold (proof - exercise):

- $P(A|A) = 1$
- $0 \leq P(B|A) \leq 1$
- $P(C \cup B|A) = P(C|A) + P(B|A) - P(C \cap B|A)$

THE IDEA OF INDEPENDENCE

It could be that the likelihood of B given A si identical to the likelihood of B in U.
Alternatively, the likelihood of B given A could be very different that its likelihood in U.





In the first case we say B is **independent** of A. The second case, we have B is dependent on A.

Independence

Definition

- A and B are independent if

$$P(B|A) = P(B)$$

- A and B are dependent if

$$P(B|A) \neq P(B)$$

Implications of independence

Say that we have picked an element from the entire set of events. Then we find out that this element is a member of the set A (i.e. it has some specific feature like $V_1 = a$).

- Does this tell us anything more about how likely it is that the element is also in set B (i.e. that it has some other specific feature like $V_2 = b$)?
- If B is independent of A then we have learned nothing new about the likelihood of the element being a member of B.

Implications of independence

Say we have already learned that the randomly picked element in set A (i.e. has $V_1 = a$). To determine whether or not the element is also in set B (i.e. has $V_2 = b$) we may want to use:

- $P(B|A)$, which gives us the likelihood of B given A.
- Now say we learn that the element also is in set C (i.e. has $V_3 = c$). Does this give us more information about B-ness?
- $P(B|A \cap C)$ gives us the likelihood of B given A and C.

Conditional independence

- If $P(B|A \cap C) = P(B|A)$, we have not gained any additional information from knowing that the element is also a member of the set C.
- In this case we say that B is conditionally independent of C given A.
- That is, once we know A, additionally knowing C is irrelevant (it will give us no more information as to the truth of B).
- Note we could have $P(B|C) \neq P(B)$. But once we learn A, C becomes irrelevant.

Computational Impact

Theorem: A, B independent $\implies P(A \cap B) = P(A)P(B)$.

Proof.
Independence says $P(B|A) = P(B)$. But $P(B|A) = \frac{P(B \cap A)}{P(A)}$. The result follows.

Conditional Independence

Theorem. $P(B \cap C|A) = P(B|A)P(C|A)$.
Proof.
Conditional independence says $P(B|C \cap A) = P(B|A)$. Rewrite as $\frac{P(B \cap C \cap A)}{P(C \cap A)} = \frac{P(B \cap A)}{P(A)}$. Multiply both sides by $P(C \cap A)$. Divide by $P(A)$ and the result follows.

Computational impact

Summarizing:

- So if we have conditional independence we get to break up this computation:

$$P(B \cap C|A) = P(B|A)P(C|A)$$

- And we get to ignore some information:

$$P(B|A \cap C) = P(B|A)$$

Partitions

Definition

A finite collection $B_1, \ldots, B_k$ of subsets of U forms a **partition** of U if:

$$B_i \cap B_j = \emptyset, i \neq j$$

$$\bigcup_{i=1}^{k} B_i = U$$

In probabilities:

$$P(B_i \cap B_j) = 0, i \neq j$$

$$P(\bigcup_{i=1}^{k} B_i) = 1$$

Applications

Let $B_1, \ldots, B_k$ be a partition of U. Then for all $A \subset U$, the collection

$$A \cap B_1, \ldots, A \cap B_k$$

forms a partition for A (proof - exercise).
it follows:

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_k)$$
$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k)$$

Often we know $P(A|B_i)$, so we can compute $P(A)$ by summing across the sets B_i.

Bayes Theorem

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

Proof

The probability of two events A and B happening, $P(A \cap B)$ is the probability of A, $P(A)$, times the probability of B that A has occurred, $P(B|A)$.

$P(A \cap B) = P(A)P(B|A)$.

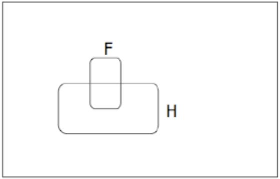
Similarly, $P(B \cap A) = P(B)P(A|B)$.

- The result follows.

BACK TO THE FLU

$P(A|B)$ is fraction of the worlds in which B is true that also have A true. Let U, F, H be the following sets:

$$U = \{x | Person(x)\},$$
$$H = \{x | Person(x) \text{ and } x \text{ has a headache}\},$$
$$F = \{x | Person(x) \text{ and } x \text{ has flu}\}$$



$P(H) = 1/10$
$P(F) = 1/40$
$P(H|F) = 1/2$

If one has the flu, what is the probability the person has a headache?

$$P(F|H) = P(F)\frac{P(H|F)}{P(H)} = \frac{1}{8}$$

EXAMPLE

Say a person goes to the doctor with a fever. To simplify the matter, assume the set of possible diseases in the town is {cold, flu, malaria}. To prescribe malaria medication, need to know $P(malaria|fever)$.

- $P(malaria|fever)$ is not easy to asses, nor does it it reflect the underlying casual mechanism (i.e. malaria causes fever).
- $P(malaria|fever)$ also is not stable; a malaria epidemic may skew the number significantly.
- Bayes to the rescue:

$$P(malaria|fever) = P(fever|malaria)P(malaria)/P(fever)$$

Example

- $P(malaria)$: this is the prior probability of malaria, i.e., before one exhibits fever, and with it we can account for other factors (epidemic, travel to a malaria risk zone, etc.), e.g., the center for disease control keeps track of the rates of various diseases.
- $P(fever|malaria)$: This is the probability a patient with malaria exhibits a fever. Again this kind of information is available from people who study malaria and its effects.
- $P(fever)$: typically not known, but it can be computed.

EXAMPLE

- Find a set of mutually exclusive and exhaustive causes of fever. (We have it: M=malaria, C=cold, F=flu). Also denote fever by S (=symptom)
- That is we assume $P(M \cap F) = P(M \cap C) = P(F \cap C) = 0, P(M \cup C \cup F) = 1$.
- From know sources can compute $P(S|M)P(M), P(S|C)P(C), P(S|F)P(F)$.
- That is $P(S) = P(S|M)P(M) + P(S|C)P(C) + P(S|F)P(F)$.
- Finally, $P(M|S) = \frac{P(S|M)P(M)}{P(S)}$.