Reasoning under uncertainty

Probability Review I

CSC384

March 14, 2018

Reasoning under uncertainty

- Thanks to Sonya Allin, Faheim Bacchus, and Sheila McIlraith for the slides.
- This material is covered in chapters 13 and 14. Chapter 13 gives some basic background on probability from the point of view of AI. Chapter 14 talks about Bayesian Networks, exact reasoning in Bayes Nets as well as approximate reasoning.

Outline

1. Announcements
2. Probability over finite sets
3. Probability over feature vectors
4. Conditional probability
5. Independence
6. Conditional independence

Announcements

- Assignment 3 (Part I) is due March 16 (Friday).
- Drop deadline is March 14 (today!)
- Midterms can be picked up in BA4208 (undergrad office) Monday-Friday 10:00am-12:00pm and 2:00pm-4:00pm
- Assignment 4 (5%)
- Quiz (5%)

Assignment 4

- Consists of 5 modules: Probability review, Introduction to Bayes Nets, Independence and Dependence, More Bayes Nets, Variable elimination
- Google Forms
- First 5 attempts are free
- Each module worth 20 pts

Assignment 4 - Help sessions

- March 27, 11:00am (PT266)
- March 29, 4:00pm (PT378)
- March 30, 11:00am (PT378)
- April 2, 11:00am (PT378)
- April 3, 12:00pm (PT378)
- April 3, 4:00pm (PT378)

Life in an uncertain world

- To date, we have seen many algorithms that will predictably move you from state to state to state (or from states of knowledge to states of knowledge).
  - We have viewed actions as being deterministic. e.g. If we are in state S_1 and we execute action A, we will arrive at state S_2
  - With deterministic actions, after executing any sequence of actions, we know exactly what state we will have arrived at, e.g., we always know what state we are in.
- These assumptions are sensible in some domains... but in many domains they are not true.
  - Weve basically ignored the fact that the world is NOT always predictable:
    i.e., if you are in state S1 and you execute action A, you may not always arrive at state S2!

Life in an uncertain world

- We might not know the effects of an action:
  - The action might have a random component, like rolling dice.
  - We might not know the long term effects of a drug.
  - We might not know the status of a road when we choose to drive down it.
- We might not know what state we are in:
  - e.g., we cant see our opponents cards in a poker game; we can only see our opponents face.
  - We may not know what a patients ailment is; we can only see symptoms.
- We still need to act, but we cant act solely on the basis of known true facts. We have to gamble.

Life in an uncertain world

- 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 0.5 hours before.
  There is 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 will most likely need 1 hour or more to get to the airport.
  There is a high probability of it taking 1.5 hours.

Life in an uncertain world

- To act rationally under uncertainty we must be able to evaluate how likely certain things are based on what we know.
- By weighing likelihoods of events we can develop mechanisms for acting rationally under uncertainty.
- We will measure likelihoods of events using probability.

Probability over finite sets

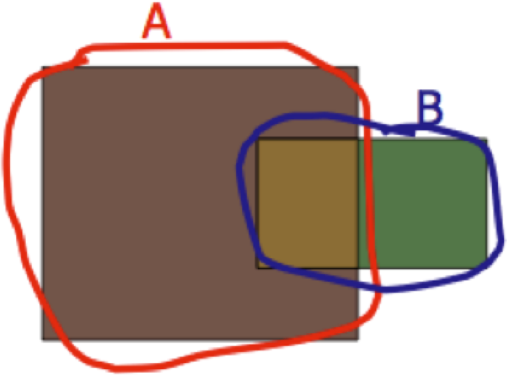
- A probability distribution is a function that is defined over a set of atomic events U. U represents the universe of events.
    - An atomic event is an event which contains only a single outcome in the sample space.
- Probability assigns a value to each event that is in the range [0,1].
- Probability assigns a value to a set of events (e.g. set A) by summing the probabilities of the members of that set.
  $P(A) = \sum_{a \in A} P(a)$

Axioms of probability

- $P(U) = 1$, i.e., sum of probabilities over all events is 1.
- $P(A) = [0, 1]$
- $P(\{\}) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  NB: if $A \cap B = \{\}$ then $P(A \cup B) = P(A) + P(B)$

Axioms visualized

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



Notation

- A $\vee$ B: the set of events with either property A or B, i.e. the set A$\cup$B
- A $\wedge$ B: the set of events with both property A and B, i.e. the set A$\cap$B *
- $\leftarrow$ A: the set of events that do not have property A: the set U-A (i.e., the complement of A w.r.t. the universe of events U)

* With probabilities, commas will also be used to denote intersection, i.e. P(A$\cap B) = P(A, B)$

Probability over feature vectors

- We will model sets of events in our universe as vectors of feature values. Like CSPs, we have:
  1. a set of variables $V_1, V_2, ..., V_n$
  2. a finite domain of values for each variable, $Dom[V_1], Dom[V_2], ..., Dom[V_n]$.
- The universe of events U will be the set of all vectors of values for the variables:
  $\langle d_1, d_2, ..., d_n \rangle : d_i \in Dom[V_i]$

Note that we often use lowercase letters to denote values of an event that is sampled from the universe of all events.

Probability over feature vectors

- This event space has size $\prod_i |Dom[V_i]|$, i.e., the product of the domain sizes.
- e.g., if $|Dom[V_i]| = 2$ we have 2^n distinct atomic events (Exponential!)
- We have 2^n resulting probabilities; our probability functions are discrete (finite).

Probability over feature vectors

- Asserting that some variables have particular values allows us to specify a useful subsets of U.
- e.g.
  $\{V_1 = a\}$ is the set of all events where $V_1 = a$
  $\{V_1 = a, V3 = d\}$ is the set of all events where $V1 = a$ and $V_3 = d$
- e.g.
  $P(\{V_1 = a\}) = \sum_{x \in Dom[V3]} P(V_1 = a, V_3 = x)$

Probability over feature vectors

- If we have the probability of every atomic event (i.e. a full instantiation of the variables), we can compute the probability of any other set.
- e.g. if $V_1 = a$ is the set of all events where $V_1 = a$, then:

$$P(V_1 = a) = \sum_{x_2 \in Dom[V_2]} \sum_{x_3 \in Dom[V_3]} \sum_{x_4 \in Dom[V_4]} \dots \sum_{x_n \in Dom[V_n]}$$

$$P(V_1 = a, V_2 = x_2, V_3 = x_3, V_4 = x_4, \dots, V_n = x_n)$$

- This is called **summing out variables** (or marginalizing your distribution).

Probability over feature vectors

- Problem:There is an exponential number of atomic probabilities to specify.
    - To compute probabilities of subsets, we have to sum up an exponential number of items.
- Thankfully, to evaluate the probability of sets containing a particular subset of variable assignments we can do much better. Improvements come from the use of probabilistic independence, especially conditional independence.

Conditional probability

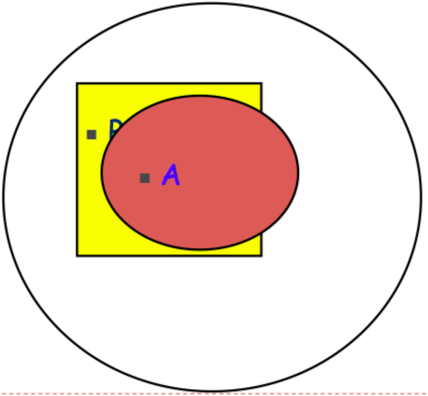
- Before we get to conditional independence, we need to define the meaning of conditional probabilities.
- These capture conditional information, i.e. information about the influence of any one variables value on the probability of others.
- Conditional probabilities are essential for both representing and reasoning with probabilistic information.

Conditional probability

- Say that A is a set of events such that $P(A) > 0$.
- Then one can define a conditional probability w.r.t. the event A:

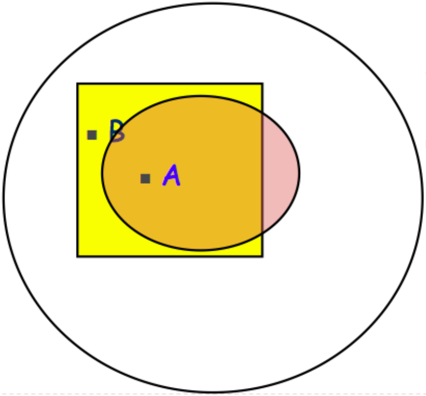
$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Conditional probability - visualized



- B covers about 30% of the entire space, but covers over 80% of A.

Conditional probability - visualized



- B's probability in the new universe A is 0.8.

Conditional probability - visualized

- Conditioning on A, corresponds to restricting ones attention to the events in A.
- We now consider A to be the whole set of events (a new universe of events): $P(A|A) = 1$.
- Then we assign all other sets a probability by taking the probability mass that lives in A ($P(B \wedge A)$), and normalizing it to the range [0,1] by dividing by $P(A)$.

Conditional probability - Axioms of probability

- A conditional probability is a probability function, but now over a subset of events in the universe instead of over the entire universe.
- The axioms hold:
  - $P(A|A) = 1$
  - $P(B|A) \in [0, 1]$
  - $P(C \cup B|A) = P(C|A) + P(B|A) - P(C \cap B|A)$

Independence

- It could be that the density of B on A is identical to its density on the entire set.
    - Density: pick an element at random from the entire set. How likely is it that the picked element is in the set B?
- Alternately the density of B on A could be much different that its density on the whole space.
- In the first case we say that B is independent of A. While in the second case B is dependent on A.

Independence

- Formally, A and B are independent if:

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

- A and B are dependent if:

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

- Implication: 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 V1= 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.

Independence

- E.g., we have a feature vector, we dont know which one. We then find out that it contains the feature $V_1 = a$.
  - i.e., we know that the vector is a member of the set $\{V_1 = a\}$.
- Does this tell us anything about whether or not $V_2 = a$, $V_3 = c$, ..., etc?
- This depends on whether or not these features are independent/dependent of $V_1 = a$.

Conditional independence

- Say we have already learned that the randomly picked element has property A.
- We want to know whether or not the element has property B:
  - $P(B|A)$ expresses the probability of this being true.
- Now we learn that the element also has property C. Does this give us more information about B?
  - $P(B|A \wedge C)$ expresses the probability of this being true under the additional information.

Conditional independence

- If $P(B|A \wedge C) = P(B|A)$, then 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 of independence

- We will see in more detail how independence allows us to speed up computation. The key to understanding how we can speed up computation is to note that if A and B are independent then:

$$P(A \wedge B) = P(B) * P(A)$$

Computational impact of independence

- We will see in more detail how independence allows us to speed up computation. The key to understanding how we can speed up computation is to note that if A and B are independent then:

$$P(A \wedge B) = P(B) * P(A)$$

- Proof:
  $P(B|A) = P(B)$ (defn of independence)
  $P(A \wedge B)/P(A) = P(B)$
  $P(A \wedge B) = P(B) * P(A)$

Conditional independence

- Similar savings can be gained from conditional independence.

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

means we can break up $P(B \wedge C|A)$ efficiently, as it implies $P(B \wedge C|A) = P(B|A) * P(C|A)$

Conditional independence

- Similar savings can be gained from conditional independence.

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

  means we can break up $P(B \wedge C|A)$ efficiently, as it implies $P(B \wedge C|A) = P(B|A) * P(C|A)$
- Proof:
  $P(B|C \wedge A) = P(B|A)$ (defn of conditional independence)
  $P(B \wedge C \wedge A)/P(C \wedge A) = P(B \wedge A)/P(A)$
  $P(B \wedge C \wedge A)/P(A) =$
  $P(C \wedge A)/P(A) * P(B \wedge A)/P(A) P(B \wedge C|A) = P(B|A) * P(C|A)$