**Artificial Intelligence**

**Session 9**

1. Probability, Likelihood and Truth
   1. In Propositional Calculus and FOPC, we were able to
      1. write down things that were true
      2. write down rules of inference that were known to hold
      3. applying rules of inference
      4. make explicit any truths which were previously implicit
   2. Showed how proof could be constructed, to show that conclusion followed from a set of facts
      1. But everything is true or false
      2. there’s no way of saying “perhaps” or “maybe”
   3. Refine this by viewing truth values as extreme probabilities
      1. 0 replaces false, 1 replaces true, but now we have all the values in between
2. Using probability for reasoning
   1. In this lecture, we write P(A) to mean “the probability of A”
      1. not Predicate P applied to Object A!
   2. But what does P(A) mean?
      1. P(I will draw an ace of hearts)
      2. P(the coin will come up heads)
      3. P(it will snow tomorrow)
      4. P(the sun will rise tomorrow)
      5. P(the problem is in the third cylinder)
      6. P(the patient has measles)
      7. P(the goalkeeper will dive to the left on this penalty shot)
3. Frequentist interpretation
   1. We can interpret frequencies of events as their probability (see ID3)
      1. Draw a card from a randomly shuffled normal deck
         1. Consisting of 4 suits: 13 spades (s), 13 hearts (h), 13 diamonds (d), and 13 clubs (c)
         2. Each suit (s,h,d,c) consists of: Ace (A), 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack (J), Queen (Q), King (K)
         3. Total number of cards, n = 13 x 4 = 52
      2. The probability that the proposition X = “the card is a heart” is true corresponds to the relative frequency with which we expect to draw a heart
         1. note that, here, each card is equally likely to be drawn
      3. P(X) = |h| / n = 13 / 52 = 0.25
   2. Definition: probability of an event X is number of possible occurrences where X holds divided by total number of possible occurrences
      1. thus, if X always holds, P(X) = n / n = 1 = T in Propositional Calculus/FOPC, conversely for F
4. Definitions & subjective interpretation
   1. Atomic event is an occurrence that can't be made of other events
      1. see Propositional Sentence in Propositional Calculus, Literal/Atomic Sentence in FOPC
   2. Event is a set of atomic events
   3. Set of all possible outcomes of an event is the sample space for that event
   4. Probability of event E in sample space S of equally likely outcomes is ratio of number of elements in E to total number of possible outcomes in S
      1. P(E) = |E| / |S|
   5. Many situations in which there’s no objective frequency interpretation
      1. Just before hang-gliding, I say “there is probability 0.2 that I’ll crash”
      2. You think there’s a 50% chance that you will get a Distinction
   6. P(E) may correspond to degree of subjective belief based on evidence available
      1. Called subjective, evidential or Bayesian interpretation of probability
      2. Related to disposition to gamble at certain odds
      3. It sounds unprincipled, but subjective interpretation has a strong formal basis
5. Axioms of probability
   1. Although there is debate about which interpretation (frequentist or subjectivist) to adopt, there is agreement about underlying maths
   2. Three basic requirements of probability values:
      1. 0 <= P(A) <= 1
      2. P( A union B ) = P( A ) + P( B ) if and only if A and B are mutually exclusive
         1. that means “iff A and B cannot be true at the same time”
      3. P( True ) = 1
   3. This is all we need to define probabilities, all other rules derived from them
   4. Examples:
      1. P( ¬A ) = 1 – P( A )
         1. 1 = P( True ); by axiom 3
         2. P( True ) = P( A union ¬A ); law of excluded middle
         3. P( A ∨ ¬A ) = P( A ) + P( ¬A ); by axiom 2
         4. P( A ) + P( ¬A ) = 1; transitivity of =
      2. P(false) = 0 because
         1. False = ¬ True; by definition
         2. P(False) = 1 – P(True); by P( ¬A ) = 1 – P( A )
6. Another way to look at probability (FIG: Two separate circles (green and pink) enclosed in a rectangle)
   1. Rectangle is the Universe (the Sample Space)
      1. Contains every possible outcome of every atomic event
      2. Has probability 1
   2. A and B are events
      1. Sets in Venn diagram contain cases when they're true
      2. mutually exclusive→ don’t overlap, can't both occur at the same time
   3. Probability represented by area relative to universal set A B
      1. 0 <= P(A) <= 1
         1. set cannot have a negative area; can’t be bigger than Universal Set
         2. nothing can have less than no chance of happening
         3. nothing can have more than absolute certainty of happening
      2. P( A union B ) = P( A ) + P( B ) iff A and B are mutually exclusive
         1. compute respective probabilities by measuring areas and adding
      3. P( True ) = 1
         1. Universal Set is the set of all things that can be true
         2. area of the Universal Set is that of the Universal Set; n / n = 1
      4. P( ¬B ) = 1 – P( B )
         1. B is the green part
         2. ¬B is the yellow part
         3. P( B ) + P( ¬B ) = P( U ) = 1
      5. P( A union B ) = P( A ) + P( B ) - P( A union B )
         1. because intersection area would be included twice otherwise
         2. similarity between connectives and set operators
7. Random variables
   1. Atomic propositions in Propositional Calculus were assigned from { T, F } by an interpretation
   2. Here we think of atomic events as being associated with random variables ranging over a domain of values which are
      1. mutually exclusive
         1. so a variable can only have one value at a time, or only one possibility can be true at a time
      2. exhaustive
         1. all possible values are known
      3. Each possible outcome has an associated probability
   3. Examples
      1. coin toss: { heads, tails }
      2. die roll: { 1, 2, 3, 4, 5, 6 }
      3. has measles: { true, false }
8. Probability distribution of a random variable:
   1. Probability distribution of a random variable is the set of probabilities covering every possible value the variable might take
      1. Uniform distribution: a die where the chance in scoring a particular number from 1 to 6 is 1/6.
      2. Non-uniform distribution: weather predicitions
9. Joint probability distributions
   1. Joint probability distribution enumerates probabilities for all possible combinations of joint outcomes of some random variables
      1. These probabilities are about co-occurrence, not about cause!

|  |  |  |  |
| --- | --- | --- | --- |
| Repairs and traffic on a bridge | | | |
| Road repair | Bad traffic | Probability |
| T | T | 0.3 |
| T | F | 0.2 |
| F | T | 0.1 |
| F | F | 0.4 |

Alternative representation:

|  |  |  |
| --- | --- | --- |
|  | Road repair | ¬ Road repair |
| Bad traffic | 0.3 | 0.1 |
| ¬ Bad traffic | 0.2 | 0.4 |

|  |  |  |  |
| --- | --- | --- | --- |
| Age and weight of humans | | | |
| Age | Weight | Probability |
| Young | Light | 0.5 |
| Young | Heavy | 0.1 |
| Old | Light | 0.1 |
| Old | Heavy | 0.3 |

Alternative representation:

|  |  |  |
| --- | --- | --- |
| Weight/age | Light | Heavy |
| Young | 0.5 | 0.1 |
| Old | 0.1 | 0.3 |

1. Joint and marginal probabilities:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Road repair | ¬ Road repair | Marginal |
| Bad traffic | 0.3 | 0.1 | 0.4 |
| ¬ Bad traffic | 0.2 | 0.4 | 0.6 |
| Marginal | 0.5 | 0.5 | 1 |

1. Joint probabilities allow us to study interaction between component events
   1. P( BadTraffic = true, RoadRepair = false ) = 0.1
   2. P( BadTraffic, ¬ RoadRepair ) = 0.1
   3. P( BadTraffic intersect ¬ RoadRepair ) = 0.1
      1. Can use 'comma' as 'and', & mix Propositional Calculus propositions with variables
2. Marginal probabilities: sum across rows or columns
   1. 0 <= each value in the table <= 1
   2. Values in the whole table should add up to 1
   3. Different outcomes must be mutually exclusive
3. P(BadTraffic ) = P(BadTraffic intersect RoadRepair) + P(BadTraffic intersect ¬ RoadRepair) = 0.3 + 0.1 = 0.4
   1. We can compute more subtle inferences
      1. P( BadTraffic union RoadRepair ) = 0.3 + 0.1 + 0.2 = 0.6
      2. P( RoadRepair → BadTraffic ) = P( ¬RoadRepair union BadTraffic ) = 0.1 + 0.4 + 0.3 = 0.8
      3. This is not causation
4. Prior probability
   1. prior probability or unconditional probability: probability assigned to event in absence of knowledge supporting occurrence
      1. probability of event prior to any evidence
   2. Example: P( Toothache = true ) = 0.1
      1. In absence of other information, 10% chance that patient has toothache
   3. So far, we assumed the world is unchanging
      1. what happens if it can change?
   4. Dynamic probabilistic knowledge base:
      1. P(A): probability of event A, about which we know nothing else
      2. P(A given B): probability of event A given additional information B
      3. Suppose marginal probability of bad traffic is P( T )
      4. If construction work started, probability of bad traffic given construction is P(T given C)
         1. read as “probability of T given C”
         2. sometimes called conditional or posterior probability
5. Conditional or posterior probability:
   1. P(A given B) = (P(A intersect B)) / (P(B)), where P(B) > 0.
   2. P( A given B ) denotes conditional (or posterior) probability of A, given that we know B and B is all we know
      1. If we receive evidence concerning a proposition, prior probabilities are no longer applicable
      2. Need to assess conditional probability of a proposition given available evidence
      3. For this to be different from P(A), A and B would not be independent
   3. Examples:
      1. Interested in probability that stock price of company C will increase, given knowledge about FTSE index from BBC news
         1. If we know FTSE is up, can rule out everything outside green area
         2. Focus attention only on FTSE area  
              
            FIG: Two interlapping circles (“C up” and “FTSE up”) within a rectangle (“FTSE Down”).  
              
            P( C up given FTSE up ) = P( C up intersect FTSE up) \* P(FTSE up )
6. Total probability of an event
   1. Event A is composed of those occasions when
      * 1. A and B co-occur
        2. A and ¬B co-occur
      1. Compound events 'A and B' and 'A and ¬B' are mutually exclusive, so probability of A must be sum of these probabilities
   2. Convenient way to calculate P(A) is with
      1. P( A ) = P( A intersect B ) + P( A intersect ¬B) (called marginalisation) = P( A given B )\* P( B ) + P( A given ¬ B ) P( ¬B ) (called conditioning)
         1. produced using conditional probability rule
7. Another example
   1. P( Cavity given Toothache ) = 0.8 is a conditional probability
   2. What happens if we are given other evidence?
      1. P( Cavity given Toothache, Earthquake ) = P( Cavity given Toothache )
         1. because Earthquake information is not relevant
      2. P( Cavity given Toothache, Cavity ) = 1
         1. because non-conditional probability overrides
8. Rules of conditional probability
   1. Product rule
      1. P( A intersect B ) = P( A given B )\*P( B ) = P( B given A )\*P( A )
         1. by rearranging formula defining conditional probability
   2. Chain rule (successive application of product rule)
      1. P( X1 intersect X2 intersect ... intersect Xn ) = P( X1 ) \* product sum of (P( Xi given X1 intersect ... intersect Xi-1 )
      2. Proof
         1. P( X1 intersect ... intersect Xn) = P( X1 intersect ... intersect Xn-1 ) \* P( Xn given X1 intersect ... intersect X\_n-1 )
         2. = P( X1 intersect ... intersect X\_n-2 ) \* P( X\_n-1 given X1 intersect ... intersect X\_n-2 ) \* P( Xn given X1 intersect ... intersect X\_n-1 )
         3. etc.
      3. Example
         1. P( X1 intersect X2 intersect X3 ) = P( X1 )\*P( X2 given X1 )\*P( X3 given X1 intersect X2 )
   3. Note: P( A → B ) is not usually the same as P( B given A )
   4. Conditioning Rule
      1. If we know an exhaustive set of conditional probabilities for an event, we can compute its prior probability
         1. like marginal probability calculation from the table earlier
         2. P( Y ) = sum of (P( Y given Z )\*P( Z ))
         3. This sums all the outcomes of all the variables that condition Y
   5. Conditionalised Product Rule:
      1. P( A intersect B given E ) = P( A given B intersect E ) \* P( B given E ) = P ( B given A intersect E )\*P( A given E )
   6. Bayes Rule**:**
      1. P(A given B) = (P(B given A) \* P(A)) / P(B)
      2. Product rule: P ( A intersect B ) = P( A given B ) \* P( B )
      3. Commutativity of intersect: P( A intersect B ) = P( B intersect A ) = P( B given A ) \* P(A)
      4. Transitivity of: P( B given A ) \* P ( A ) = P( A given B ) \* P( B )
      5. Divide both sides by P( A ) to get the rule
      6. Bayes’ rule allows us to reverse a conditional probability, if we know the probabilities of its components
         1. P( Cause | Effect ) = P( Effect | Cause ) P( Cause ) / P( Effect )
9. Frequentist vs Subjective View of Probability
   1. Nothing about Bayes’ Rule tells you how to interpret the probabilities
      1. Yet Bayes’ Rule is often used with semi-subjective probabilities
   2. Statistical estimates (frequentist interpretation) for probabilities
      1. fine if experiment under consideration can be repeated many times under similar conditions.
   3. Subjective probabilities (degree of belief)
      1. Suppose you want to assign probability that one political party will beat another. Can determine your own personal probability by seeing what kind of bet you’d be willing to make.
      2. Doctor assigns 0.8 probability that a particular patient has meningitis, given that the patient has a stiff neck. May not have statistical data on that one patient for meningitis and stiff necks. But might base subjective assignment of probability on our experiences with patients with a similar medical history.
10. Reasoning with Probabilities
    1. Scenario
       1. Jane goes to the doctor’s for a routine checkup and takes some tests
       2. One test for a rare genetic disease comes back positive
       3. The disease is potentially fatal
    2. Jane looks up wikipedia and learns that
       1. “rare” means P( Disease ) = P( D ) = 1/10,000 = 0.0001
       2. the test is 99% accurate:
          1. a small rate of false positives P( Test = + given ¬ D) = 0.01
          2. no false negatives P( Test = – given D ) = 0
          3. Jane would like to compute the probability that she has the disease and take appropriate action
11. Formulating the situation
    1. The scenario in probabilities
    2. The disease is rare: P( D ) = 0.0001
    3. Reliability of test: P( Test = – given ¬ D ) = 0.99 :  
         
       P(D given Test = +) = P( D intersect Test = + ) / P( Test = + )  
         
       = 1 / ( 1 + 99.99 )  
         
       = 0.0099 (not 0.99!!!)
    4. Chance of false positive: P( Test = + given ¬ D ) = 0.01
    5. No chance of false negative: P( Test = – given D ) = 0
    6. Therefore, Chance of true positive: P( Test = + given D ) = 1
12. Formalising the reasoning
    1. Derivation
       1. Bayes’ rule
          1. P( D given Test = + ) = P( Test = + given D ) \*P( D ) / P( Test= + )
          2. = 1 x 0.0001 / P( Test = + ) = 0.0001 / P( Test = + )
       2. Bayes’ rule •
          1. (¬ D given Test = + ) = P( Test = + given ¬ D) P( ¬ D) / P( Test = + )
          2. = 0.01 x 0.9999 / P( Test = + ) = 0.009999 / P( Test = + )
       3. Excluded middle and arithmetic
          1. 1 = P( D given Test = + ) + P( ¬D given Test = + ) = 0.0001 / P( Test = + ) + 0.009999 / P( Test = + )
          2. = ( 0.0001 + 0.009999 ) / P( Test = + ), therefore, 0.0001 + 0.009999 = 1 x P( Test = + ) = 0.010099
       4. Bayes’ Rule
          1. P( D given Test = + ) = 1 x 0.0001 / 0.010099 = 0.0099
13. Another similar example
    1. Let S be proposition that patient has stiff neck, & M that patient has meningitis
    2. Doctor wants to know P( M given S )
       1. Probabilities from a medical database
       2. P( S given M ) = 0.5
       3. P( M ) = 1 / 50,000 = 0.00002
       4. P( S ) = 1 / 20 = 0.05
    3. Using Bayes’ Rule
       1. P( M given S ) = P( S given M )\*P(M) / P(S) = 0.5 x 0.00002 / 0.05 = 0.0002
    4. If doctor knows this probability, and sees patient with stiff neck, he/she can know how strongly/weakly to lean toward diagnosis of meningitis
14. How practical is Bayes’ rule?
    1. How reasonable is the supposition that you know P( S given M ) but need to calculate P( M given S )?
       1. Why not estimate P( M given S ) directly by sampling?
          1. P( M given S ): This is affected by both the presence/absence of an epidemic and the way meningitis works – complicated.
          2. P( S given M ): Reflects the way meningitis works. Unaffected by epidemic. Unrelated information
          3. P(M): Increases when there is an epidemic. Unrelated information
       2. You can’t separate these two unrelated factors when estimating P( M | S ) by sampling.
15. How practical is Bayes’ rule?
    1. What if prior probability P( S ) in the denominator is too difficult to determine?
       1. Can we avoid direct assessment of the probability of evidence?
          1. That is, the denominator of Bayes’ rule
       2. Yes, with normalisation
          1. As in “Jane” example above
16. Normalisation
    1. Can avoid direct assessment of priors of evidence, such as P( S ) here, by considering an exhaustive set of hypotheses.
       1. P( M given S ) = P( S given M ) \* P( M ) / P( S given M ) \* P( M ) + P( S given ¬M ) \* P( ¬M )
       2. By assessing P( S given ¬M ), we avoid assessing P( S )
17. Can generalise normalisation beyond the P ∨ ¬P case
    1. use a distribution over multiple, mutually exclusive values of a variable
    2. Suppose we wish to compute a posterior distribution over A given B = b, and suppose A has disjoint domain a1, ..., am
    3. Apply Bayes’ rule for each value of A
       1. P( A = a1 | B = b ) = P( B = b | A = a1 ) P( A = a1 ) / P( B = b )
       2. P( A = am | B = b ) = P( B = b | A = am ) P( A = am ) / P( B = b )
    4. Adding these up, and noting that sum of(P( A = ai | B = b )) = 1
    5. 1 / P( B = b ) = 1 / sum of(P ( B = b | A = ai ) P( A = ai ))
    6. This quantity is known as the normalisation factor, denoted by alpha
18. How practical is Bayes’ rule?
    1. Normalisation is one way to make Bayes’ rule more practically applicable, but there’s another
       1. Issue is that Bayes’ rule requires us to estimate too many probabilities before we can apply it
       2. Want to estimate one conditional probability; need to know
          1. another conditional probability
          2. two prior probabilities
       3. Why are we willing to make this effort?
          1. Because other probabilities are typically easier to obtain by observation
       4. Can we reduce number of probabilities needed to apply Bayes’ Rule?
          1. There is a technique for doing this when there are multiple sources of evidence and we can use the idea of conditional independence
19. Independence revisited
    1. Unconditional independence
       1. A and B are independent iff
          1. P( A given B ) = P( A )
          2. P( B given A ) = P( B )
          3. P( A intersect B ) = P( A ) P( B )
       2. Conditional independence
          1. independence holds, given a common cause
          2. A and B are independent given C iff
             1. P( A given B intersect C ) = P( A given C )
             2. P( A intersect B given C ) = P( A given C ) P( B given C )
             3. Often, pairs of variables are not independent, but are conditionally independent
20. Bayesian Belief Networks (BBNs)
    1. When there are many variables in a model, full probability distribution is too complex to compute (2n entries for n variables)
    2. Conditional independence allows us to reduce complexity to manageable proportions for many real-world problems
    3. Bayesian Belief Network (BBN) models dependencies between variables in an intuitive graphical format
    4. Belief networks are now standard technology for expert systems
    5. Domain experts generally report it is not too hard to interpret the links and fill in the requisite probabilities
    6. Some (e.g., Pathfinder IV) seem to be outperforming the experts who designed them!
21. Bayesian Belief Networks (BBN)
    1. A Bayesian Belief Network is a graph for which the following properties hold:
       1. A set of random variables makes up the nodes of the network
          1. Variables may be discrete or continuous
          2. Each node is annotated with quantitative probability information.
       2. A set of directed links or arrows connects pairs of nodes
          1. If there is an arrow from node X to node Y, then X is said to be a parent of Y
       3. Each node X has a conditional probability distribution P(X given Parents(X)) that quantifies the effect of the parents on the node
       4. Note there can be more than one parent
       5. The graph has no directed cycles (so it is a directed, acyclic graph, or DAG)
22. BBN Example:
    1. You have an alarm system which is quite reliable at detecting burglars, but it sometimes goes off when there is an earthquake
    2. You have two neighbours who agree to call you if they hear the alarm: John is usually home, but sometimes he thinks the telephone is the alarm; Mary is out more often than John

Alarm:

|  |  |  |
| --- | --- | --- |
| B | E | P(A) |
| T | T | 0.95 |
| T | F | 0.94 |
| F | T | 0.29 |
| F | F | 0.001 |

Burglary:

|  |
| --- |
| P(B) |
| 0.001 |

Earthquake:

|  |
| --- |
| P(E) |
| 0.002 |

MaryCalls:

|  |  |
| --- | --- |
| A | P(M) |
| T | 0.70 |
| F | 0.01 |

JohnCalls:

|  |  |
| --- | --- |
| A | P(M) |
| T | 0.90 |
| F | 0.05 |

* 1. The tables show conditional probabilities P(A given B, E), P(J given A) and P(M given A) and priors P(B) and P(E)
     1. All variables are binary, so we only need to show one value, for P(x) = true
  2. Note that the full joint probability table for 5 binary values has 25=32 entries; here we can specify the full distribution with 10 values
     1. Assumptions: John and Mary do not notice the burglar directly, nor earthquakes, nor do they confer before calling
     2. e.g. P(j,m,a,¬b,¬e) = P(j given a) \* P(m given a) \* P(a given ¬b intersect ¬e) \* P(¬b) \* P(¬e)
     3. = 0.90 \* 0.70 \* 0.001 \* 0.999 \* 0.998 = 0.000628
  3. Any query about the domain can be answered by summing entries from the full joint distribution
  4. In more complex networks, various methods of exact and approximate inference are available

1. Summary
   1. Probability theory enables use of varying degrees of likelihood to represent uncertainty
   2. A (joint) probability distribution completely describes a (set of) random variable(s)
   3. Conditional probabilities let us calculate probabilities relative to priors that we know
   4. Bayes’ rule is helpful in relating conditional probabilities and priors
   5. Independence assumptions let us make intractable problems tractable
   6. Belief networks are now standard technology for expert systems