**Artificial Intelligence**

**Session 10**

1. Probabilistic modelling
   1. A linear equation – (perhaps) simplest model
      1. y = beta\*T^x
         1. E.g., y - the price of a house, and x - are a series of factors that affect this price, e.g., the location, the number of bedrooms, the age of the house, etc.
   2. Probability distribution for modelling the world
      1. p(x, y)
      2. We could ask lot of questions
         1. What is the probability that house prices will rise over the next five years?
         2. Given that the house costs $100,000, what is the probability that it has three bedrooms?
   3. The probabilistic aspect of modelling is very important
      1. The world itself is stochastic
      2. We need to assess the confidence of our predictions
2. Joint probabilities
   1. A complete probability model is a single joint probability distribution over all propositions/variables in the domain
      1. P(X1 , X2 , ..., Xi , ...)
   2. A particular instance of the world has the probability
      1. P(X1=x1 intersect X2=x2 intersect ... intersect Xi=xi intersect ...) = p

|  |  |  |
| --- | --- | --- |
|  | ¬ WetGrass | WetGrass |
| ¬ Raining | 0.8 | 0.04 |
| Raining | 0.01 | 0.15 |

* 1. Rather than stating knowledge as
     1. Raining ⟹ WetGrass
  2. We can state it as
     1. P(Raining, WetGrass) = 0.15
     2. P(Raining, ¬WetGrass) = 0.01
     3. P(¬Raining, WetGrass) = 0.04
     4. P(¬Raining, ¬WetGrass) = 0.8

1. How can a joint-distribution help us?
   1. A principled way to model the world
      1. Handles missing data/variables
      2. Easy to incorporate prior knowledge
   2. With the joint distribution, we can do anything we want
      1. Design classifiers: y = maximise argument of P(Y = y given X)
      2. We can make Bayesian inference (probabilistic reasoning)
         1. Sherlock Holmes: P(Murderer | Observed Evidence)
         2. Doctor: P(Disease | Symptoms), P(Effect | Treatment)
         3. Parenting:
            1. P(Dirty Diaper, Hungry, Lonely given 5 a.m., Baby crying)
            2. P(Baby crying at 5 a.m. given feeding at 2 a.m.)
            3. P(Baby crying at 5 a.m. given feeding at 1 a.m.)
2. Modeling the world with probability distribution
   1. Example: Author attribution as in a paper
      1. Variables: Word 1, Word 2, Word 3, ..., Word N
      2. 15 authors in total: {Dickens, Shakespeare, Jane Austen, Tolkien, George RR. Martin, …, Douglas Adams}
      3. A vocabulary of size 3000
   2. Questions:
      1. What is the dimension(s) of the joint distribution?
      2. How many free parameters are needed to represent this distribution?
   3. Answers: 3000^N × 15 and 3000^N × 15 – 1   
      Note: Exponential in the number of words per document...
3. Statistical Independences
   1. (Marginal / absolute) Independence
   2. X and Y are independent iff
      1. P(X, Y) = P(X)\*P(Y) [by definition]
      2. P(X given Y) = P(X) : Since P(X | Y) = P(X, Y)/P(Y) = P(X) P(Y)/P(Y)
   3. If X and Y are (conditionally) independent given Z, then
      1. P(X given Y, Z) = P(X given Z)
      2. Example: P(WetGrass given Season, Rain) = P(WetGrass given Rain)
4. Example of Conditional Independence
   1. In practice, conditional independence is more common than marginal independence.
      1. P(Final exam grade given Weather) != P(Final exam grade)
         1. i.e., they are not independent
      2. P(Final exam grade given Weather, Effort) = P(Final exam grade given Effort)
         1. But they are conditionally independent given Effort
5. Benefit of conditional independence
   1. If some variables are conditionally independent, the joint probability can be specified with many fewer than 2^{N}-1 numbers (or 3^{N}-1, or 10^{N}-1, or...)
      1. For example: (for binary variables W, X, Y, Z)
         1. P(W,X,Y,Z) = P(W) P(X given W)\*P(Y given W,X)\*P(Z given W,X,Y)
      2. 1 + 2 + 4 + 8 = 15 numbers to specify
         1. But if Y and W are independent given X, and Z is independent of W and X given Y, then P(W,X,Y,Z)=P(W)\*P(X given W)\*P(Y given X)\*P(Z given Y)
            1. 1 + 2 + 2 + 2 = 7 numbers
      3. This is often the case in real problems.
6. (Example continued) Modelling the world with probability distribution
   1. Example: Author attribution as in a paper
      1. Variables: Word 1, Word 2, Word 3, ..., Word N
      2. 15 authors in total: {Dickens, Shakespeare, Jane Austen, Tolkien, George RR. Martin, …, Douglas Adams}
      3. A vocabulary of size 3000
   2. In addition, assume that: Word 1, …, Word N are mutually independent given Author
      1. P(Word1, …, Word N given Author) = P(Word 1 given Author) \* ... \* P(Word N given Author)
   3. Question:
      1. What are the dimensions of each factor?
      2. How many “free parameters” are needed in total? • Answer: 14 + 2999 \* 15 \* N. ( Linear in N)
7. Quiz time: Representing a joint Probability
   1. Joint probability: P(X1 , X2 , ..., XN )
      1. Defines the probability for any possible state of the world
      2. Let the variables be binary. How many numbers (“free parameters”) does it take to define the joint distribution?
         1. Defined by 2^{N}-1 parameters
      3. If the variables are independent, then
         1. P(X1 , X2 , ..., XN ) = P(X1)\*P(X2) ... P(XN)
         2. How many numbers does it take to define the joint distribution?
            1. Just N parameters!
8. Key points so far
   1. Representing a joint-distribution
      1. Number of parameters exponential in the number of variables
      2. Calculating marginals and conditionals from the joint-distribution.
   2. Conditional independences and factorization of joint distributions
      1. Saves parameters, often exponential improvements
9. Graphical models
   1. Graphical models come out of the marriage of graph theory and probability theory.
   2. Directed graph 🡪 Bayesian Networks / Belief Networks
   3. Undirected Graph 🡪 Markov Random Fields.
10. Graphical models
    1. Used as a modelling tool. Many applications!
11. Example: Image denoising
    1. Given an image corrupted by noise
    2. Restore it based on our probabilistic model of what images look like
    3. A graphical model: p(original image given noisy image)
    4. Observing the noisy image, we can sample or use inference to predict the original image
12. Example: Translation
    1. A training set of text that were transcribed in both English and Chinese
    2. A probabilistic model p(y given x)
       1. y - generated English sentence
       2. x - corresponding Chinese sentence
13. Example: Speech recognition
    1. Given a (joint) model of speech signals and language (text)
    2. Infer spoken words from audio signals.
       1. y - generated English sentence
       2. x - corresponding Chinese sentence
14. Graphical models
    1. A particular factorization of a joint distribution
       1. P(X,Y ,Z) = P(X)\*P(Y given X)\*P(Z given Y)
    2. A collection of conditional independences
       1. { X uptick Z given Y, ... }
15. Bayesian networks
    1. Bayesian networks
       1. A data structure (depicted as a graph) that represents the dependence among variables and allows us to concisely specify the joint probability distribution
       2. The graph itself is known as an “influence diagram”
    2. A belief network is a directed acyclic graph where:
       1. The nodes represent the set of random variables (one node per random variable)
       2. Arcs between nodes represent influence, or causality
          1. A link from node X to node Y means that X “directly influences” Y
       3. Each node has a conditional probability table (CPT) that defines P(node | parents).
16. Example
    1. Random variables X and Y
       1. X – It is raining
       2. Y – The grass is wet
    2. X has an effect on Y or Y is a symptom of X
    3. Draw two nodes and link them
    4. Define the CPT for each node
       1. P(X) and P(Y given X)
    5. Typical use: we observe Y and we want to query P(X given Y)
       1. Y is an evidence variable
       2. X is a query variable
17. Bayesian networks represent the joint probability
    1. The joint probability function can be calculated directly from the network
       1. It’s the product of the CPTs of all the nodes
       2. P(var1 , ... , varN ) = product of P(var\_i given Parents(var\_i))
       3. P(X,Y) = P(X)\*P(Y given X)
       4. P(X,Y,Z) = P(X)\*P(Y)\*P(Z given X,Y)
18. Bayesian networks
    1. Step 1: Choose variables in the environments, represent them as nodes.
    2. Step 2: Connect the variables by inspecting the “direct influence”: cause-effect
    3. Step 3: Fill in the probabilities in the CPTs.
19. Example: Modelling with Bayesian networks
    1. I’m at work and my neighbor John called to say my home alarm is ringing, but my neighbor Mary didn’t call. The alarm is sometimes triggered by minor earthquakes. Was there a burglar at my house?
    2. Random (boolean) variables:
       1. JohnCalls, MaryCalls, Earthquake, Burglar, Alarm
    3. The belief net shows the causal links
    4. This defines the joint probability
       1. P(JohnCalls, MaryCalls, Earthquake, Burglar, Alarm)
    5. What do we want to know? P(B given J, ¬M)
    6. Example:

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

* 1. Read the joint pf from the graph:
     1. Read the joint pf from the graph
        1. P(J, M, A, B, E) = P(B) \* P(E) \* P(A given B,E) \* P(J given A) \* P(M given A)
     2. Plug in the desired values
        1. P(J, ¬M, A, B, ¬E) = P(B) \* P(¬E) \* P(A given B,¬E) \* P(J given A) \* P(¬M given A)
     3. How about P(B given J, ¬M) ? Remember, this means P(B=true given J=true, M=false)
     4. Calculate P(B | J, ¬M)
        1. 𝑃(𝐵 given 𝐽, ¬𝑀) = 𝑃(𝐵,𝐽, ¬𝑀) / 𝑃(𝐽, ¬𝑀)

1. Key points so far
   1. Bayesian network as a modelling tool
   2. By inspecting the cause-effect relationships, we can draw directed edges based on our domain knowledge
   3. The product of the CPTs give the joint distribution
      1. We can calculate P(A given B) for any A and B
      2. The factorization makes it computationally more tractable. What else can we get?
2. Example: Conditional independence
   1. Conditional independence is seen here
      1. P(JohnCalls given MaryCalls, Alarm, Earthquake, Burglary) = P(JohnCalls given Alarm)
      2. So JohnCalls is independent of MaryCalls, Earthquake, and Burglary, given Alarm
   2. Does this mean that an earthquake or a burglary do not influence whether or not John calls?
      1. No, but the influence is already accounted for in the Alarm variable
      2. JohnCalls is conditionally independent of Earthquake, but not absolutely independent of it
      3. This conclusion is independent to values of CPTs!
3. Question
   1. If X and Y are independent, are they therefore independent given any variable(s)?
      1. I.e., if P(X, Y) = P(X) \* P(Y) [ i.e., if P(X given Y) = P(X)], can we conclude that P(X given Y, Z) = P(X given Z)?
   2. The answer is no, and here’s a counter example:
      1. P(X given Y) = P(X)
      2. P(X given Y, Z) != P(X given Z)
      3. Note: Even though Z is a deterministic function of X and Y, it is still a random variable with a probability distribution
      4. Again: This conclusion is independent to values of CPTs!
4. Is there a general way that we can answer questions about conditional independences by just inspecting the graphs? Turns out the answer is “Yes!”  
     
   Directed Acyclic Graph 🡪 By traversing the DAG 🡪 Factorization of Joint Distribution 🡪 By definition 🡪 The ser of conditional independences that hold all CPTs 🡪 By d-separation Bayes Ball alg.
5. d-separation:
   1. Intuition: the graph and the edges controls the information flow, if there is no path that the information can flow from one-node to another, we say these two nodes are independent.
   2. “Shading” denotes “observing” or ”conditioning on” that variables.
6. d-separation in three canonical graphs:
   1. X uptick Z given Y
      1. “X and Z are d-separated by the observation of Y.”
   2. X uptick Z given Y
      1. “X and Z are d-separated by the observation of Y.”
   3. X uptick Z given Y
      1. “X and Z are d-separated by NOT observing Y nor any descendants of Y.”
7. Examples:
   1. P(W given R, G) = P(W given G)
   2. P(T given C, F) = P(T given F)
   3. P(W given I, M) != P(W given M) \* P(W given I) = P(W)
   4. FIG: A graph with Nodes:
      1. A connected to B and E
      2. B connected to F and C
      3. C connected to F
      4. D connected to C
      5. E connected to null
      6. F connected to null
      7. Show that {A} is d-separated from {F} given {B}
         1. Every path from A to F must pass through either the segment →B→C or A→B→F. Each of these is blocked given the vertex B. As a result, A is d-separated from F given B.
      8. Is {B} d-separated from {D} given {F}?
         1. No. There are two paths from B to D, B→C←D and B→F←C←D. Neither of these paths is blocked when F is in the separating set.
      9. Is {B} d-separated from {D} given {C}?
         1. No. Again, there are two paths from B to D, B→C←D and B→F←C←D. The second is blocked on the segment F←C←D when C is in the separating set as this is a sequential path. However, the first is not blocked as it is a convergent path and C is in the separating set.
8. Markov Random Fields (MRFs)
   1. Bayesian networks model clear dependencies, often causal dependencies. Bayesian networks are acyclic.
   2. How can we model mutual and cyclic dependencies?
   3. Example (economy):
      1. demand and supply determine the price
      2. high price fosters supply
      3. low price fosters demand
9. Motivation
   1. Example (physics): modeling ferromagnetism in statistical mechanics
      1. a grid of magnetic dipoles in a volume
      2. every dipole causes a force on its neighbors
      3. every dipole is forced by its neighbors
   2. The dipoles might change their orientation. Every configuration of the magnetic dipole field can be characterized by its energy. The probability of a certain configuration depends on its energy: high energy configurations are less probable, low energy configurations are more probable.
10. Markov Random Fields (MRFs)
    1. A Markov random field is an undirected graphical model
       1. Undirected graph 𝐺=(𝑉,𝐸)
       2. Each node represents a random variable
       3. Links indicate an explicitly modeled stochastic dependence
       4. Nonnegative potential function or “factor” associated with cliques, 𝐶, of the graph
       5. Nonnegative potential functions represent interactions and need not correspond to conditional probabilities (may not even sum to one)
11. Clique
    1. A clique of size k is a subset C of k nodes of the MRF so that for each pair X, Y in C with X != Y holds that X and Y are connected by an edge.
    2. Example: the MRF on the right
       1. One clique of size 3
          1. {A, B, C}
       2. Four cliques of size 2
          1. {A, B}, {A, C}, {B, C}, {C, D}
       3. Four cliques of size 1
          1. {A}, {B}, {C}, {D}
12. Joint probability distribution
    1. For every clique 𝑐 in 𝐶 in MRF, we specify a potential function
    2. Omega(c): C 🡪 R\_{> 0}
       1. large values of Omega(c) indicate that a certain configuration of the random variables in the clique is more probable
       2. small values of Omega(c) indicate that a certain configuration of the random variables in the clique is less probable
    3. Joint probability distribution of a MRF is defined over cliques in the graph:
       1. P(x1, …, xN) = (1/Z) \* product sum of (Omega(c))
       2. Z = sum of(product sum of (Omega(c)))
          1. Z: normalizing constant, often called the partition function
13. Potential function
    1. Potential functions are usually given in terms of Gibbs/Boltzmann distributions
       1. Omega(c) = e^{E\*c(c)}
       2. 𝐸\_𝐶 ∶ C → R is an “energy function”
          1. Large energy means low probability
          2. Small energy means high probability
       3. But there are other choices for potential function too.
14. Example
    1. Let us model the food preferences of a group of four persons: Antonia, Ben, Charles, and Damaris. They might choose between pasta, fish, and meat
       1. Ben likes meat and pasta but hates fish
       2. Antonia, Ben, and Charles prefer to choose the same
       3. Charles is vegetarian
       4. Damaris prefers to choose something else than all the other
    2. Create an MRF on the blackboard that models the food preferences of the four persons and assign potential functions to the cliques.
    3. One way to model the food preference task
    4. Random variables A, B, C, D model Antonias, Bens, Charles, and Damaris’ choice. Discrete variables with values 1=pasta, 2=fish, 3=meat
    5. Energy functions which are relevant (all others are constant):
       1. E\_{B}(b) = 0 if b in {1,3}; 100 if b = 2
       2. E\_{A,D}(a,d) = 0 if a != d; 10 if a = d
       3. E\_{A,B,C}(a,b,c) = 0 if a = b = c; 30 otherwise
       4. E\_{B,D}(b,d) = 0 if b != d; 10 if b = d
       5. E\_{C}(c) = 0 if c = 1; 50 of c = 2; 200 if c = 3
       6. E\_{C,D}(c,d) = 0 if c != d; 10 if c = d
    6. Given a graph 𝐺=(𝑉,𝐸), express the following as probability distributions that factorize over 𝐺
    7. Uniform distribution over independent sets
    8. Number the vertices of G from 1 to |V|
       1. Represent each possible subset of V as a binary vector, x, whose i-th component is equal to 1 if the i-th vertex is selected to be in the subset and 0 otherwise.
       2. The uniform distribution over independent sets of G as:  
            
          P(x1, …, x\_|V|) = (1/Z) \* product sum of (Nabla\_{xi+xj <= 1})
       3. Nabla\_{xi+xj <= 1} is an indicator function, equal to 1 if xi + xj <= 1, 0 otherwise.
       4. The total number of independent sets in the graph.
       5. Which means that, with this choice of potential functions, p is the uniform probability distribution over independent sets of G.
       6. The probability of any vector x′ whose corresponding subset is not an independent set should be zero.
15. Independence assertions
    1. Vertex A is graph separated from vertex B given vertex C if every path from A to B contains C.
    2. In general, the set of vertices X in V is graph separated from a set of vertices Y in V given a set Z in V if every path from a vertex x in X to a vertex y in Y contains at least one vertex from the set Z.
    3. If X in V is graph separated from a set of vertices Y in V given a set Z in V, then X uptick Y given Z
       1. Each variable is independent of all of its non-neighbours given its neighbours
       2. All paths leaving a single variable must pass through some neighbours
    4. Is D graph separated from A given B and E
       1. Yes, as every path from D to A must pass through the neighbors of A, namely either B or E.
    5. Is C graph separated from A given B and D
       1. No, the path C−E−A does not contain B or D.
16. BNs vs. MRFs:

|  |  |  |
| --- | --- | --- |
| **Property** | **Bayesian Networks** | **Markov Random Fields** |
| Factorisation | Conditional distributions | Potential functions |
| Distribution | Product of conditional distributions | Normalized product of potentials |
| Cycles | Directed not allowed | Allowed |
| Partition function | 1 | Potentially NP-hard to compute |
| Independence test | d-separation | Graph separation |

1. Moralization
   1. Every Bayesian network can be converted into an MRF with some possible loss of independence information
      1. Remove the direction of all arrows in the network
      2. If A and B are parents of C in the Bayesian network, we add an edge between A and B in the MRF
   2. This procedure is called “moralization” because it “marries” the parents of every node
2. Factorisations
   1. Many factorizations over the same graph may represent the same joint distribution
      1. Some are better than others (e.g., they more compactly represent the distribution)
      2. Simply looking at the graph is not enough to understand which specific factorization is being assumed
3. Factor Graphs
   1. Factor graphs are used to explicitly represent a given factorization over a given graph
      1. Variable nodes that model random variables
      2. Factor nodes that model a probabilistic relationship between variable nodes. Each factor node is assigned with a potential function
      3. Variable nodes and factor nodes are connected by undirected links.
      4. Note: Not a different model, but rather different way to visualize an MRF
   2. For each MRF we can create a factor graph as follows:
      1. The set of variable nodes is taken from the nodes of the MRF
      2. for each non-constant potential function omega\_c
      3. Create a new factor node f
      4. Connect f with all variable nodes in clique C
      5. Assign the potential function omega\_c to f
      6. Hence, the joint probability of the MRF is equal to the Gibbs distribution over the sum of all factor potentials
   3. P(A,B,C) = (1/Z) \* omega\_AB (A,B) \* omega\_BC (B,C) \* omega\_AC (A,C)
4. Special types of MRFs
   1. MRFs are very general and can be used for many purposes. Some models have been shown to be very useful.
      1. Potts model. Useful for image segmentation and noise removal
      2. Conditional random fields. Useful for image segmentation
      3. Boltzmann machine. Useful for unsupervised and supervised learning
      4. Markov logic networks. Useful for logic inference on noisy data
5. Conditional Random Fields (CRFs)
   1. Undirected graphical models that represent conditional probability distributions P(𝑌 given 𝑋 )
   2. Potentials can depend on both X and Y.
6. Conditional Random Fields (CRFs)
   1. CRFs often assume that the potentials are log-linear functions
      1. Omega\_c(xc, yc) = exp(w^T \* f\_c(xc, yc))
      2. f\_c is referred to as a feature vector and w is some vector of feature weights
   2. The feature weights are typically learned from data
   3. CRFs don’t require us to model the full joint distribution (which may not be possible anyhow)
7. Example: Binary image segmentation
   1. Label the pixels of an image as belonging to the foreground or background
   2. +/- correspond to foreground/background
   3. Interaction between neighbouring pixels in the image depends on how similar the pixels are
      1. Similar pixels should preference having the same spin (i.e., being in the same part of the image)
   4. This can be modelled as a CRF where the image information (e.g., pixel colours) is observed, but the segmentation is unobserved
   5. Because the model is conditional, we don’t need to describe the joint probability distribution of (natural) images and their foreground/background segmentations
   6. CRFs will be particularly important when we want to learn graphical models from observed data
8. Example: Low Density Parity Check Codes
   1. Want to send a message across a noisy channel in which bits can be flipped with some probability –use error correcting codes
   2. Omega\_A, Omega\_B, Omega\_C are all parity check constraints: they equal one if their input contains an even number of ones and zero otherwise
   3. Phi\_𝑖(𝑥𝑖 , 𝑦𝑖 ) = 𝑝(𝑦𝑖 given 𝑥𝑖 ), the probability that the 𝑖-th bit was flipped during transmission
   4. The parity check constraints enforce that the 𝑦’s can only be one of a few possible codewords: 000000, 001011, 010101, 011110, 100110, 101101, 110011, 111000
   5. Decoding the message that was sent is equivalent to computing the most likely codeword under the joint probability distribution.
   6. Most likely codeword is given by maximum a posteriori (MAP) inference
   7. Maximise argument of (p(y given x))
   8. Do we need to compute the partition function for MAP inference?