Lecture 2: Introduction to belief (Bayesian) networks

- Conditional independence
- What is a belief network?
- Independence maps (I-maps)

Recall from last time: Conditional probabilities

- Our probabilistic models will compute and manipulate conditional probabilities.
- Given two random variables X, Y, we denote by $p(X = x|Y = y)$ the probability of X taking value x given that we know that Y is certain to have value y.
- This fits the situation when we observe something and want to make an inference about something related but unobserved:
  - $p(\text{cancer recurs}|\text{tumor measurements})$
  - $p(\text{gene expressed} > 1.3|\text{transcription factor concentrations})$
  - $p(\text{collision to obstacle}|\text{sensor readings})$
  - $p(\text{word uttered}|\text{sound wave})$

Recall from last time: Bayes rule

- Bayes rule is very simple but very important for relating conditional probabilities:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

- Bayes rule is a useful tool for inferring the posterior probability of a hypothesis based on evidence and a prior belief in the probability of different hypotheses.

Using Bayes rule for inference

Often we want to form a hypothesis about the world based on observable variables. Bayes rule is fundamental when viewed in terms of stating the belief given to a hypothesis H given evidence e:

$$p(H|e) = \frac{p(e|H)p(H)}{p(e)}$$

- $p(H|e)$ is sometimes called **posterior probability**
- $p(H)$ is called **prior probability**
- $p(e|H)$ is called **likelihood** of the evidence (data)
- $p(e)$ is just a normalizing constant, that can be computed from the requirement that $\sum_h p(H = h|e) = 1$:

$$p(e) = \sum_h p(e|h)p(h)$$

Sometimes we write $p(H|e) \propto p(e|H)p(H)$

Example: Medical Diagnosis

A doctor knows that pneumonia causes a fever 95% of the time. She knows that if a person is selected randomly from the population, there is a 10^{-7} chance of the person having pneumonia. 1 in 100 people suffer from fever.

You go to the doctor complaining about the **symptom** of having a fever (evidence). What is the probability that pneumonia is the **cause** of this symptom (hypothesis)?

$$p(\text{pneumonia}|\text{fever}) = \frac{p(\text{fever}|\text{pneumonia})p(\text{pneumonia})}{p(\text{fever})} = \frac{0.95 \times 10^{-7}}{0.01}$$

Computing conditional probabilities

- Typically, we are interested in the posterior joint distribution of some **query variables** Y given specific values e for some **evidence variables** E
- Let the **hidden variables** be $Z = X - Y - E$
- If we have a joint probability distribution, we can compute the answer by using the definition of conditional probabilities and marginalizing the hidden variables:

$$p(Y|e) = \frac{p(Y,e)}{p(e)} \propto p(Y,e) = \sum_z p(Y,e,z)$$

- This yields the same big problem as before: the joint distribution is too big to handle

Independence of random variables revisited

- We said that two r.v.'s X and Y are **independent**, denoted $X \perp\!\!\!\perp Y$, if $p(x,y) = p(x)p(y)$.
- But we also know that $p(x,y) = p(x|y)p(y)$.
- Hence, two r.v.'s are independent if and only if:

  $$p(x|y) = p(x) \text{ (and vice versa)}, \forall x \in \Omega_X, y \in \Omega_Y$$

  This means that knowledge about Y does not change the uncertainty about X and vice versa.
- Is there a similar requirement, but less stringent?

Conditional independence

- Two random variables X and Y are **conditionally independent** given Z if:

  $$p(x|y,z) = p(x|z), \forall x, y, z$$

  This means that knowing the value of Y does not change the prediction about X *if the value of Z is known*.
- We denote this by $X \perp\!\!\!\perp Y | Z$.

Example

- Consider the medical diagnosis problem with three random variables: P (patient has pneumonia), F (patient has a fever), C (patient has a cough)
- The full joint distribution has $2^3 - 1$ = 7 independent entries
- If someone has pneumonia, we can assume that the probability of a cough does *not* depend on whether they have a fever:

$$p(C = 1|P = 1, F) = p(C = 1|P = 1) \quad (1)$$

- Same equality holds if the patient does not have pneumonia:

$$p(C = 1|P = 0, F) = p(C = 1|P = 0) \quad (2)$$

- Hence, C and F are *conditionally independent given* P.

Example (continued)

- The joint distribution can now be written as:

$$p(C, P, F) = p(C|P, F)p(F|P)p(P) = p(C|P)p(F|P)p(P)$$

- Hence, the joint can be described using $2 + 2 + 1 = 5$ numbers instead of 7
- Much more important savings happen with more variables

Naive Bayesian model

A common assumption in early diagnosis is that the symptoms are independent of each other given the disease

- Let $s_1, \dots s_n$ be the symptoms exhibited by a patient (e.g. fever, headache etc)
- Let D be the patient's disease
- Then by using the naive Bayes assumption, we get:

$$p(D, s_1, \dots s_n) = p(D)p(s_1|D) \cdots p(s_n|D)$$

- The conditional probability of the disease given the symptoms:

$$p(D|s_1, \dots s_n) = \frac{p(D, s_1, \dots s_n)}{p(s_1, \dots s_n)} \propto p(D)p(s_1|D) \cdots p(s_n|D)$$

because the denominator is just a normalization constant.

Recursive Bayesian updating

- The naive Bayes assumption allows also for a very nice, incremental updating of beliefs as more evidence is gathered
- Suppose that after knowing symptoms $s_1, \dots s_n$ the probability of D is:

$$p(D, s_1 \dots s_n) = p(D) \prod_{i=1}^{n} p(s_i|D)$$

- What happens if a new symptom s_{n+1} appears?

Recursive Bayesian updating

- The naive Bayes assumption allows also for a very nice, incremental updating of beliefs as more evidence is gathered
- Suppose that after knowing symptoms $s_1, \dots s_n$ the probability of D is:

$$p(D, s_1 \dots s_n) = p(D) \prod_{i=1}^{n} p(s_i|D)$$

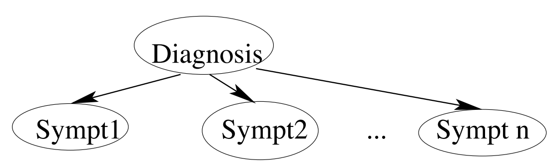
- What happens if a new symptom s_{n+1} appears?

$$p(D, s_1 \dots s_n, s_{n+1}) = p(D) \prod_{i=1}^{n+1} p(s_i|D) = p(D, s_1 \dots s_n) p(s_{n+1}|D)$$

- An even nicer formula can be obtained by taking logs:

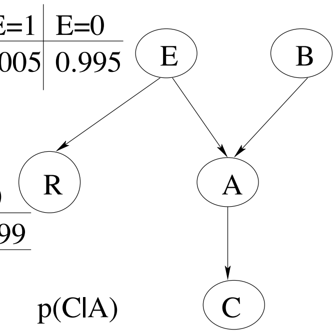
$$\log p(D, s_1 \dots s_n, s_{n+1}) = \log p(D, s_1 \dots s_n) + \log p(s_{n+1}|D)$$

A graphical representation of the naive Bayesian model



- The nodes represent random variables
- The arcs represent "influences"
- The **lack of arcs** represents **conditional independence** relationships

More generally: Bayesian networks



p(E)

E=1	E=0
0.005	0.995

p(B)

B=1	B=0
0.01	0.99

p(R|E)

	R=1	R=0
E=0	0.0001	0.9999
E=1	0.65	0.35

p(A|B,E)

	A=1	A=0
B=0,E=0	0.001	0.999
B=0,E=1	0.3	0.7
B=1,E=0	0.8	0.2
B=1,E=1	0.95	0.05

p(C|A)

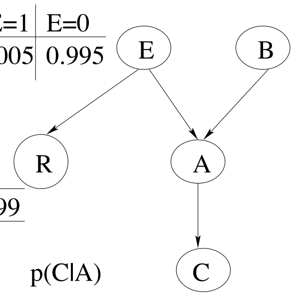
	C=1	C=0
A=0	0.05	0.95
A=1	0.7	0.3

Bayesian networks are a graphical representation of conditional independence relations, using graphs.

A graphical representation for probabilistic models

- Suppose the world is described by a set of r.v.'s $X_1, \ldots X_n$
- Let us define a *directed acyclic graph* such that each node i corresponds to an r.v. X_i
- Since this is a one-to-one mapping, we will use X_i to denote both the node in the graph and the corresponding r.v.
- Let X_{π_i} be the set of parents for node X_i in the graph
- We associate with each node the conditional probability distribution of the r.v. X_i given its parents: $p(X_i|X_{\pi_i})$.

Example: A Bayesian (belief) network



p(E)

E=1	E=0
0.005	0.995

p(B)

B=1	B=0
0.01	0.99

p(R|E)

	R=1	R=0
E=0	0.0001	0.9999
E=1	0.65	0.35

p(A|B,E)

	A=1	A=0
B=0,E=0	0.001	0.999
B=0,E=1	0.3	0.7
B=1,E=0	0.8	0.2
B=1,E=1	0.95	0.05

p(C|A)

	C=1	C=0
A=0	0.05	0.95
A=1	0.7	0.3

- The nodes represent random variables
- The arcs represent “influences”
- At each node, we have a conditional probability distribution (CPD) for the corresponding variable *given its parents*

Factorization

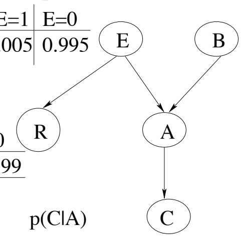
Let G be a DAG over variables $X_1, \ldots, X_n$. We say that a joint probability distribution p **factorizes according to** G if p can be expressed as a product:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i})$$

The individual factors $p(x_i | x_{\pi_i})$ are called **local probabilistic models** or **conditional probability distributions (CPD)**.

Bayesian network definition

A Bayesian network is a DAG G over variables $X_1, \ldots, X_n$, together with a distribution p that factorizes over G. p is specified as the set of conditional probability distributions associated with G's nodes.



p(E)

E=1	E=0
0.005	0.995

p(B)

B=1	B=0
0.01	0.99

p(R|E)

	R=1	R=0
E=0	0.0001	0.9999
E=1	0.65	0.35

p(A|B,E)

	A=1	A=0
B=0,E=0	0.001	0.999
B=0,E=1	0.3	0.7
B=1,E=0	0.8	0.2
B=1,E=1	0.95	0.05

p(C|A)

	C=1	C=0
A=0	0.05	0.95
A=1	0.7	0.3

.

Using a Bayes net for reasoning (1)

- Computing any entry in the joint probability table is easy because of the factorization property:

$$
\begin{aligned}
&p(B=1, E=0, A=1, C=1, R=0) \\
&\quad = p(B=1)p(E=0)p(A=1|B=1,E=0)p(C=1|A=1)p(R=0|E=0) \\
&\quad = 0.01 \cdot 0.995 \cdot 0.8 \cdot 0.7 \cdot 0.9999 \approx 0.0056
\end{aligned}
$$

- Computing marginal probabilities is also easy.
  E.g. What is the probability that a neighbor calls?

$$p(C=1) = \sum_{e,b,r,a} p(C=1, e, b, r, a) = \ldots$$

Using a Bayes net for reasoning (2)

- One might want to compute the conditional probability of a variable given evidence that is "upstream" from it in the graph
- E.g. What is the probability of a call in case of a burglary?

$$p(C=1|B=1) = \frac{p(C=1,B=1)}{p(B=1)} = \frac{\sum_{e,r,a} p(C=1,B=1,e,r,a)}{\sum_{c,e,r,a} p(c,B=1,e,r,a)}$$

- This is called **causal reasoning** or **prediction**

Using a Bayes net for reasoning (3)

- We might have some evidence and need an explanation for it. In this case, we compute a conditional probability based on evidence that is "downstream" in the graph
- E.g. Suppose we got a call. What is the probability of a burglary? What is the probability of an earthquake?

$$p(B=1|C=1) = \frac{p(C=1|B=1)p(B=1)}{p(C=1)} = \ldots$$

$$p(E=1|C=1) = \frac{p(C=1|E=1)p(E=1)}{p(C=1)} = \ldots$$

- This is **evidential reasoning** or **explanation**.

Using a Bayes net for reasoning (4)

- Suppose that you now gather more evidence, e.g. the radio announces an earthquake. What happens to the probabilities?

$$p(E=1|C=1,R=1) \gg p(E=1|C=1) \text{ and}$$
$$p(B=1|C=1,R=1) \ll p(B=1|C=1)$$

- This is called **explaining away**

I-Maps

A directed acyclic graph (DAG) G whose nodes represent random variables $X_1, \ldots, X_n$ is an **I-map (independence map)** of a distribution p if p satisfies the independence assumptions:

$$X_i \perp\!\!\!\perp \text{Nondescendents}(X_i)|X_{\pi_i}, \forall i = 1, \ldots n$$

where X_{π_i} are the parents of X_i

Example

Consider all possible DAG structures over 2 variables. Which graph is an I-map for the following distribution?

x	y	$p(x, y)$
0	0	0.08
0	1	0.32
1	0	0.32
1	1	0.28

What about the following distribution?

x	y	$p(x, y)$
0	0	0.08
0	1	0.12
1	0	0.32
1	1	0.48

Example (continued)

- In the first example, X and Y are not independent, so the only I-maps are the graphs $X \rightarrow Y$ and $Y \rightarrow X$, which assume no independence
- In the second example, we have $p(X = 0) = 0.2$, $p(Y = 0) = 0.4$, and and for all entries $p(x, y) = p(x)p(y)$
- Hence, $X \perp\!\!\!\perp Y$, and there are three I-maps for the distribution: the graph in which X and Y are not connected, and both graphs above.
- Note that independence maps may have extra arcs!