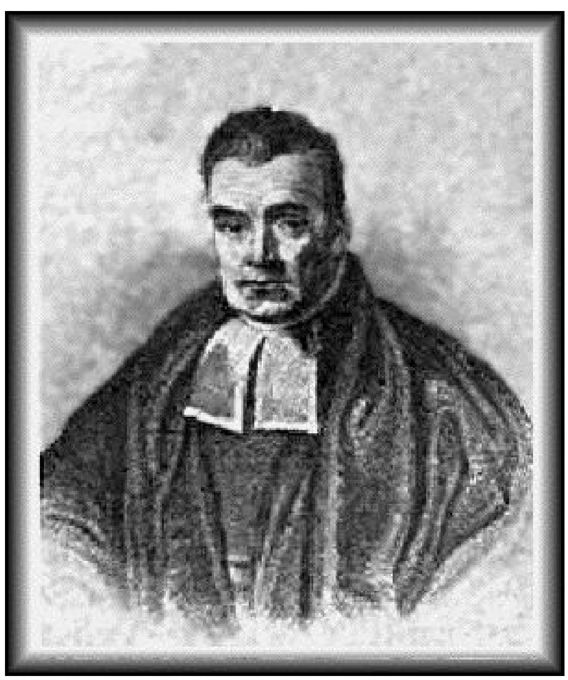
Lecture 2: Conditional independence. Belief networks

- Conditional probability and Bayes rule
- Independence of random variables
- Using Bayes rule for inference
- Conditional independence
- Bayes nets: a graphical representation for conditional independence

Rev. Thomas Bayes (1706-1761)



Conditional probability

The basic statements in the Bayesian framework talk about **conditional probabilities**. $p(A|B)$ is the belief in event A given that event B is known with absolute certainty:

$$p(A|B) = \frac{p(A \cap B)}{p(B)} \text{ if } p(B) \neq 0$$

Note that we can use either the set intersection or the logical "and" notation ($p(A \wedge B)$, or $p(A, B)$).

The **product rule** gives an alternative formulation:

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

Bayes rule

Bayes rule is another alternative formulation of the product rule:

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

The **complete probability formula** states that:

$$p(A) = p(A|B)p(B) + p(A|\neg B)p(\neg B)$$

or more generally,

$$p(A) = \sum_i p(A|b_i)p(b_i),$$

where b_i form a set of exhaustive and mutually exclusive events.

Chain rule

Chain rule is derived by successive application of product rule:

$$
\begin{aligned}
& p(X_1, \ldots, X_n) = \\
= \;& p(X_1, \ldots, X_{n-1}) p(X_n | X_1, \ldots, X_{n-1}) \\
= \;& p(X_1, \ldots, X_{n-2}) p(X_{n-1} | X_1, \ldots, X_{n-2}) p(X_n | X_1, \ldots, X_{n-1}) \\
= \;& \ldots \\
= \;& \prod_{i=1}^{n} p(X_i | X_1, \ldots, X_{i-1})
\end{aligned}
$$

Simpson's paradox (Pearl)

The following table describes the effectiveness of a certain drug on a population:

	Male		Female		Overall	
	Recovered	Died	Recovered	Died	Recovered	Died
Drug used	15	40	90	50	105	90
No drug	20	40	20	10	40	50

Good news: the ratio of recovery for the whole population increases from 40/50 to 105/90

But the ratio of recovery decreases for both males and females!

Simpson's paradox (2)

The paradox lies in ignoring the context in which the results are given.

If we derive correct conditional probabilities based on this data (assuming 50% males in the population) we get:

$$p(\text{recovery} \mid \text{drug}) = \frac{1}{2}\frac{15}{15+40} + \frac{1}{2}\frac{90}{90+50} \approx 0.46$$

$$p(\text{recovery} \mid \text{no drug}) = \frac{1}{2}\frac{20}{20+40} + \frac{1}{2}\frac{20}{20+10} = 0.5$$

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**
- $p(e)$ is just a normalizing constant, that can be computed from the requirement that $p(H|e) + p(\neg H|e) = 1$:

$$p(e) = p(e|H)p(H) + p(e|\neg H)p(\neg H)$$

Sometimes we write $p(H|e) = \alpha p(e|H)p(H)$

Example: Medical Diagnosis

A doctor knows that SARS 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 SARS. 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 meningitis is the **cause** of this symptom (hypothesis)?

Let S be SARS, F be fever:

$$p(S|F) = \frac{p(F|S)p(S)}{p(F)} = \frac{0.95 \times 10^{-7}}{0.01} = 0.95 \times 10^{-5}$$

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 "summing out" the hidden variables:

$$p(Y|e) = \alpha p(Y, e) = \alpha \sum_z p(Y, e, z)$$

Big problem: the joint distribution is too big to handle!

Example

Suppose we consider medical diagnosis, and there are 100 different symptoms and test results that the doctor could consider. A patient comes in complaining of fever, dry cough and chest pains. The doctor wants to compute the probability of SARS.

- The probability table has $>= 2^{100}$ entries!
- For computing the probability of a disease, we have to sum out over 97 hidden variables!

Independence of random variables

Two random variables X and Y are **independent**, denoted $X \perp\!\!\!\perp Y$, if knowledge about X does not change the uncertainty about Y and vice versa.

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

or equivalently, $p(x, y) = p(x)p(y)$ If n Boolean variables are independent, the whole joint distribution can be computed as:

$$p(x_1, \ldots x_n) = \prod_i p(x_i)$$

Only n numbers are needed to specify the joint, instead of 2^n
But absolute independence is a very strong requirement, seldom met

Conditional independence

Two 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 SARS diagnosis problem with three random variables: S, F, C (patient has a cough)

The full joint distribution has $2^3 - 1$ = 7 independent entries

If someone has SARS, we can assume that, the probability of a cough does **not** depend on whether they have a fever:

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

I.e., C is *conditionally independent* of F given S

Same independence hold if the patient does not have SARS.

$$p(C|\neg S,F) = p(C|\neg S) \quad (2)$$

Example (continued)

Full joint distribution can now be written as:

$$\begin{aligned} & p(C,F,S) = \\ = \quad & p(C,F|S)p(S) \\ = \quad & p(C|S)p(F|S)p(S) \end{aligned}$$

I.e., 2 + 2 + 1 = 5 independent numbers (equations 1 and 2 remove two numbers)

Much more important savings happen if the system has lots of variables!

Naive Bayesian model

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

- Let $s_1, \ldots s_n$ 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:

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

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 $s_1, \ldots s_n$ the probability of D is:

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

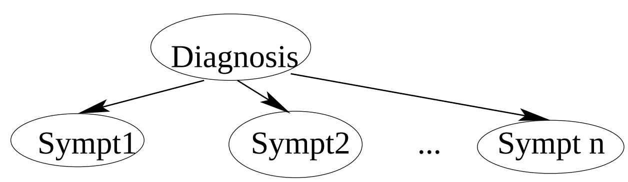
What happens if a new symptom s_{n+1} appears?

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

An even nicer formula can be obtained by taking logs:

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

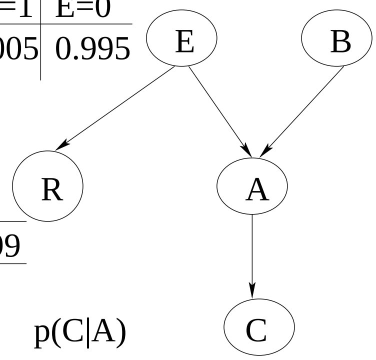
A graphical representation of the naive Bayesian model



- The nodes represent random variables
- The arcs represent “influences”

This is a simple Bayes network!

A Bayes net example



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