Smarter education for smarterplanet
IBM Innovation Centre for Education

Welcome to:

Probabilistic approach to AI

Unit objectives

After completing this unit, you will be able to:

- Understand the basics of probability
- Understand the use of probability as a knowledge representation language
- Gain knowledge on the concept of Probability distributions
- Learn about Bayesian networks as a method for knowledge representation
- Learn about Gaussian Bayesian Networks
- Gain knowledge on Noisy-OR model
- Gain an insight into the tools available on Bayesian Networks

Probability

- Probability theory is the branch of mathematics concerned with probability, the analysis of random phenomena.

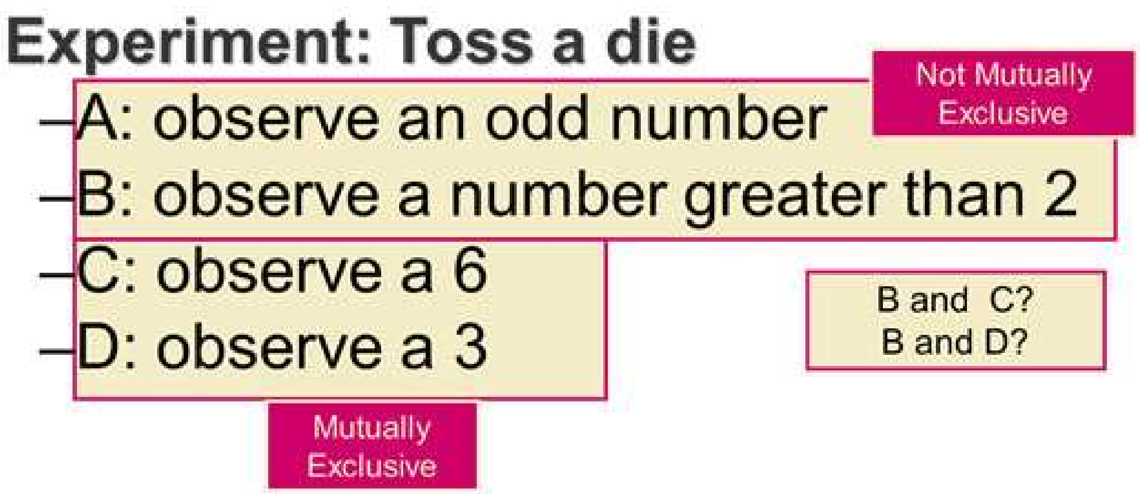
- The central objects of probability theory are random variables, stochastic processes, and events.

- Bayes' theorem is fundamental to Bayesian statistics, and has applications in fields including science, engineering, medicine and law.

- The application of Bayes' theorem to update beliefs is called Bayesian inference.

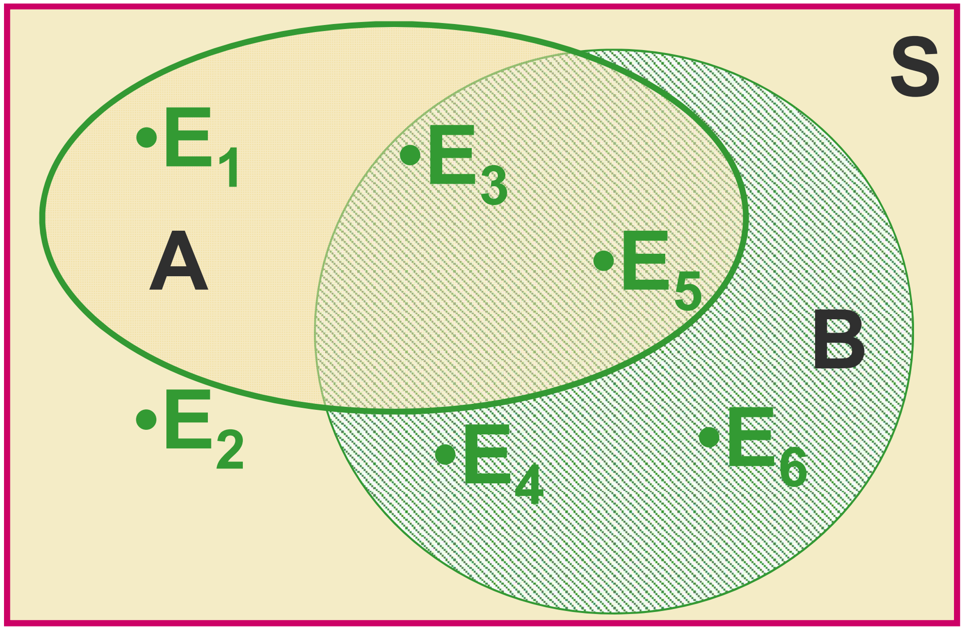
Basic concepts

- An experiment is the process by which an observation (or measurement) is obtained.
- An event is an outcome of an experiment, usually denoted by a capital letter.
  - The basic element to which probability is applied
- When an experiment is performed, a particular event either happens, or it doesn't!
- Experiment: Record an age
  - A: person is 30 years old
  - B: person is older than 65
- Experiment: Toss a die
  - A: observe an odd number
  - B: observe a number greater than 2
- Two events are mutually exclusive if, when one event occurs, the other cannot, and vice versa.



Probability of an event

- An event is a collection of one or more simple events.
- The die toss:
  - A: an odd number
  - B: a number > 2
  - A ={E1, E3, E5}
  - B ={E3, E4, E5, E6}



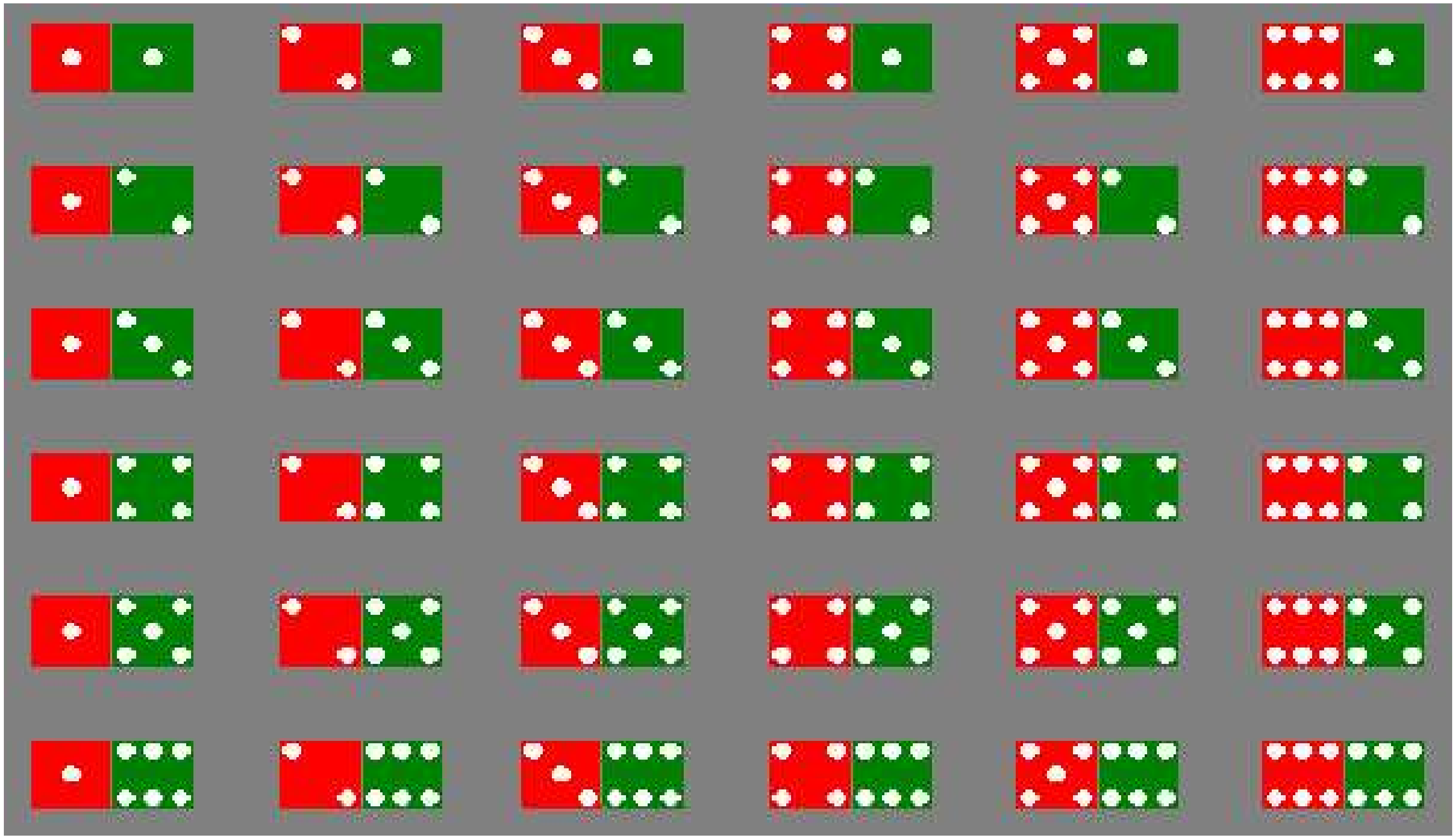
- The probability of an event A measures “how often” A will occur. We write P(A).
- Suppose that an experiment is performed n times. The relative frequency for an event A is

$$\frac{\text{Number of times A occurs}}{n} = \frac{f}{n}$$

- If we let n get infinitely large, $\boxed{P(A) = \lim_{n \to \infty} \frac{f}{n}}$

Example on sample space

- The sample space of throwing a pair of dice is

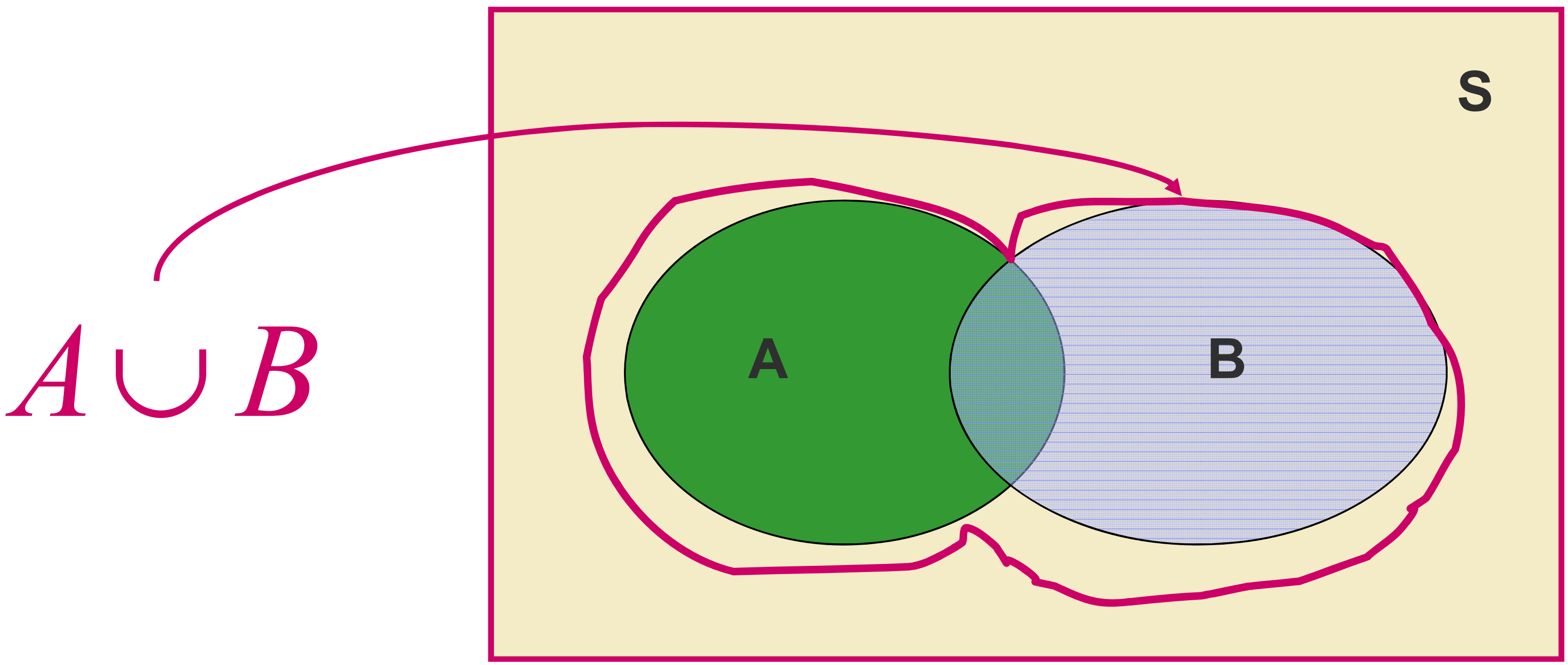


Counting rules

- Sample space of throwing 3 dice has 216 entries, sample space of throwing 4 dice has 1296 entries, …
- At some point, we have to stop listing and start thinking …
- We need some counting rules
- The mn Rule
- If an experiment is performed in two stages, with m ways to accomplish the first stage and n ways to accomplish the second stage, then there are mn ways to accomplish the experiment.
- Rule is extended to k stages, with the number of ways equal to n1 n2 n3 … nk.
- Examples:
  – Toss two coins. The total number of simple events is: 2 x 2 = 4
  – Toss three coins. The total number of simple events is: 2 x 2 x 2 = 8
  – Toss two dice. The total number of simple events is: 6 x 6 = 36
  – Toss three dice. The total number of simple events is: 6 x 6 x 6 = 216
  – Two M&Ms are drawn from a dish containing two red and two blue candies. The total number of simple events is: 4 x 3 = 12

Event relations

- The beauty of using events, rather than simple events, is that we can combine events to make other events using logical operations: and, or and not.
- The union of two events, A and B, is the event that either A or B or both occur when the experiment is performed. We write: AUB



Conditional probabilities

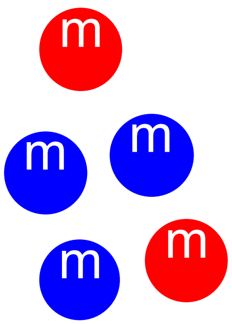
- The probability that A occurs, given that event B has occurred is called the conditional probability of A given B and is defined as:

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

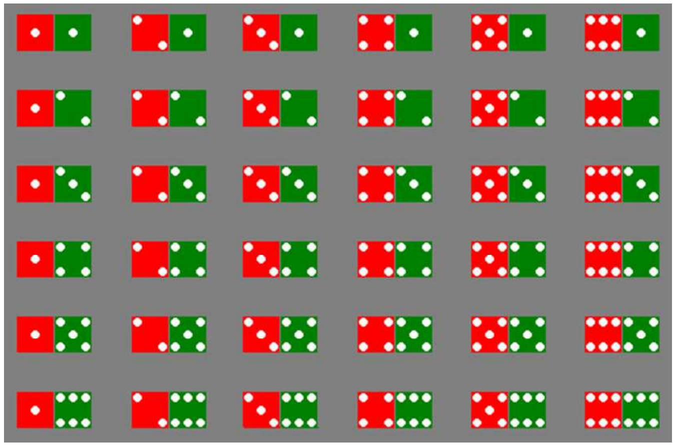
- A bowl contains five M&Ms, two red and three blue. Randomly select two candies, and define
  - A: second candy is red.
  - B: first candy is blue.

  P(A|B) =P(2^{nd} red|1^{st} blue)= 2/4 = 1/2

  P(A|not B) = P(2^{nd} red|1^{st} red) = 1/4



- Toss a pair of fair dice. Define
  - A: add to 3
  - B: add to 6



- P(A|B) = P(A and B)/P(B) = 0/36/5/6 = 0

Defining independence

- We can redefine independence in terms of conditional probabilities:

- Two events A and B are independent if and only if

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

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

Otherwise, they are dependent.

- For any two events, A and B, the probability that both A and B occur is
  - $P(A \cap B) = P(A)\,P(B \text{ given that } A \text{ occurred}) = P(A)P(B|A)$

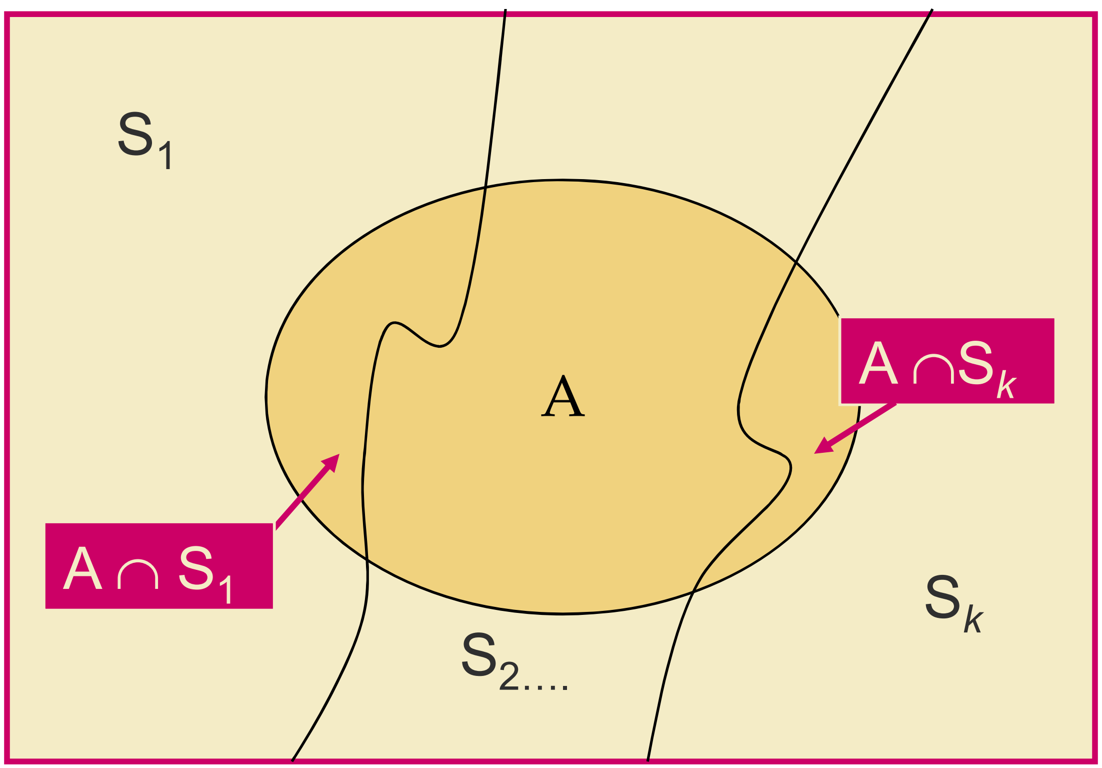
- If the events A and B are independent, then the probability that both A and B occur is
  - $P(A \cap B) = P(A)\,P(B)$

The law of total probability

- Let S_1 , S_2 , S_3 ,..., S_k be mutually exclusive and exhaustive events (that is, one and only one must happen). Then the probability of any event A can be written as
- $P(A) = P(A \cap S_1) + P(A \cap S_2) + \ldots + P(A \cap S_k)$

  $= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) + \ldots + P(S_k)P(A|S_k)$
- $P(A) = P(A \cap S_1) + P(A \cap S_2) + \ldots + P(A \cap S_k)$

  $= P(S_1)P(A|S_1) + P(S_2)P(A|S_2) + \ldots + P(S_k)P(A|S_k)$



Bayes' rule

- Let S_1 , S_2 , S_3 ,..., S_k be mutually exclusive and exhaustive events with prior probabilities $P(S_1)$, $P(S_2)$,…,$P(S_k)$. If an event A occurs, the posterior probability of S_i, given that A occurred is

$$P(S_i \mid A) = \frac{P(S_i)P(A \mid S_i)}{\sum P(S_i)P(A \mid S_i)} \text{ for } i = 1, 2, \ldots k$$

- From a previous example, we know that 49% of the population are female. Of the female patients, 8% are high risk for heart attack, while 12% of the male patients are high risk. A single person is selected at random and found to be high risk. What is the probability that it is a male?

Define H: high risk F: female M: male

$$P(M \mid H) = \frac{P(M)P(H \mid M)}{P(M)P(H \mid M) + P(F)P(H \mid F)}$$
$$= \frac{.51(.12)}{.51(.12) + .49(.08)} = .61$$

Examples

- Suppose a rare disease infects one out of every 1000 people in a population. And suppose that there is a good, but not perfect, test for this disease: if a person has the disease, the test comes back positive 99% of the time. On the other hand, the test also produces some false positives: 2% of uninfected people are also test positive. And someone just tested positive. What are his chances of having this disease?

- Define
  - A: has the disease
  - B: test positive

We know:

$P(A) = .001$ $P(A^c) = .999$

$P(B|A) = .99$ $P(B|A^c) = .02$

- We want to know P(A|B)=?

$$P(A \mid B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A^c)P(B|A^c)}$$

$$= \frac{.001 \times .99}{.001 \times .99 + .999 \times .02} = .0472$$

Random variables

- A quantitative variable x is a random variable if the value that it assumes, corresponding to the outcome of an experiment is a chance or random event.
- Random variables can be discrete or continuous.
- Examples:
  - x = CAT score for a randomly selected student
  - x = number of people in a room at a randomly selected time of day
  - x = number on the upper face of a randomly tossed die
- Probability Distributions for Discrete Random Variables
  - The probability distribution for a discrete random variable x resembles the relative frequency distributions. It is a graph, table or formula that gives the possible values of x and the probability p(x) associated with each value.

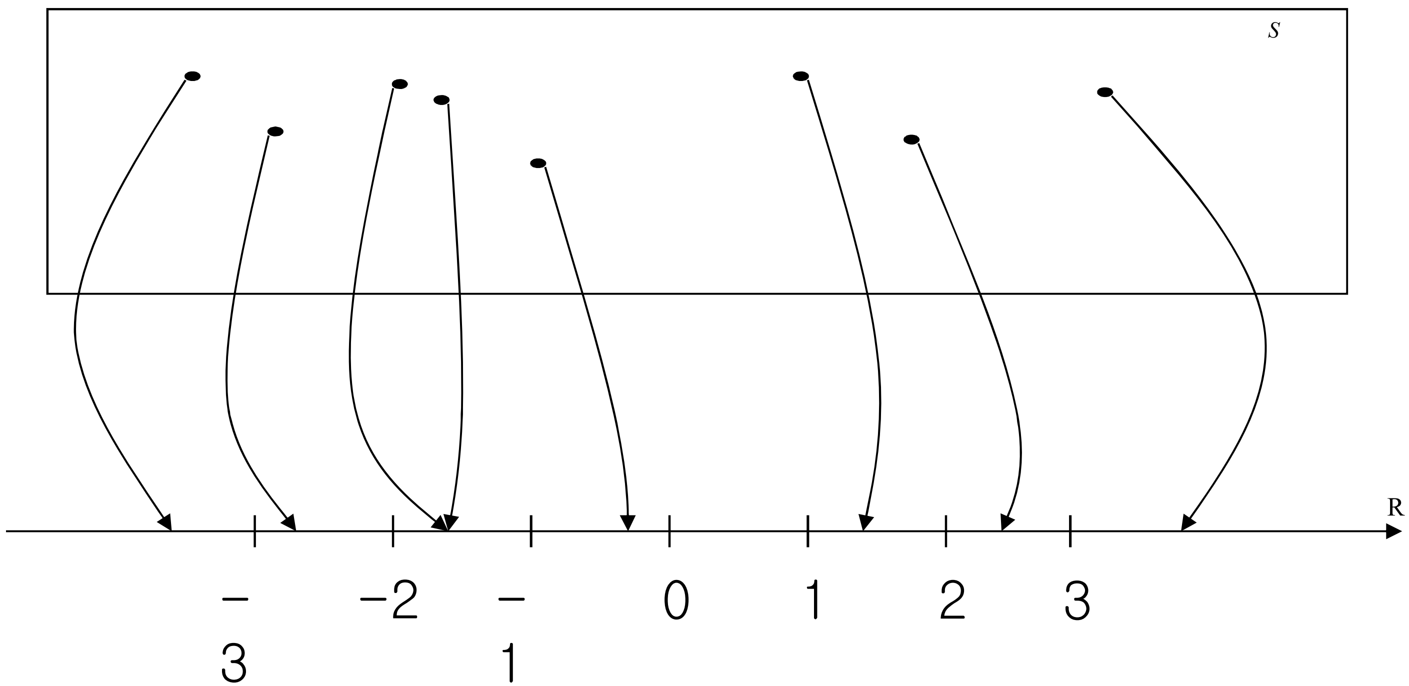
We must have

$$0 \leq p(x) \leq 1 \text{ and } \sum p(x) = 1$$

Discrete random variable

- Random variable
  - A numerical value to each outcome of a particular experiment

- Example 1 : Machine Breakdowns
  - Sample space : $S = \{electrical, mechanical, misuse\}$
  - Each of these failures may be associated with a repair cost
  - State space : $\{50, 200, 350\}$
  - Cost is a random variable : 50, 200, and 350



Probability distributions

- Probability distributions can be used to describe the population.
- Shape: Symmetric, skewed, mound-shaped…
- Outliers: unusual or unlikely measurements
- Center and spread: mean and standard deviation.
- A population mean is called m and a population standard deviation is called s.
- Let x be a discrete random variable with probability distribution p(x). Then the mean, variance and standard deviation of x are given as:

$$\text{Mean} : \mu = \sum x p(x)$$

$$\text{Variance} : \sigma^2 = \sum (x-\mu)^2 p(x)$$

$$\text{Standard deviation} : \sigma = \sqrt{\sigma^2}$$

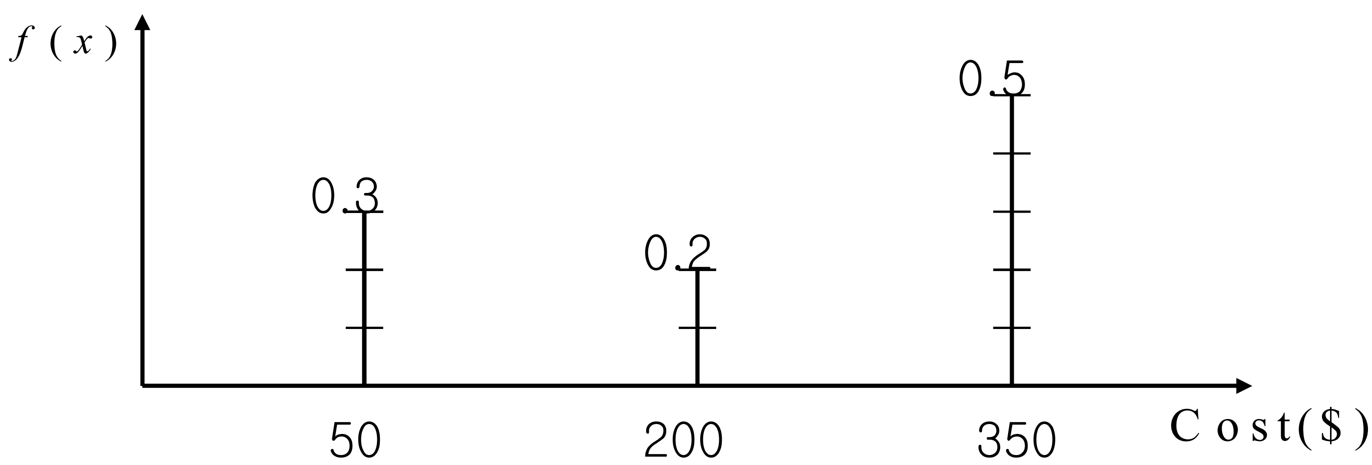
Probability mass function

- Probability Mass Function (p.m.f.)
  - A set of probability value p_i assigned to each of the values taken by the discrete random variable x_i

$$0 \le p_i \le 1 \qquad \sum_i p_i = 1$$

  - Probability : $P(X = x_i) = p_i$

- Example 1: Machine Breakdown
  - P (cost=50)=0.3, P (cost=200)=0.2, P (cost=350)=0.5
  - 0.3 + 0.2 + 0.5 =1



Probability density function

IBM ICE (Innovation Centre for Education)

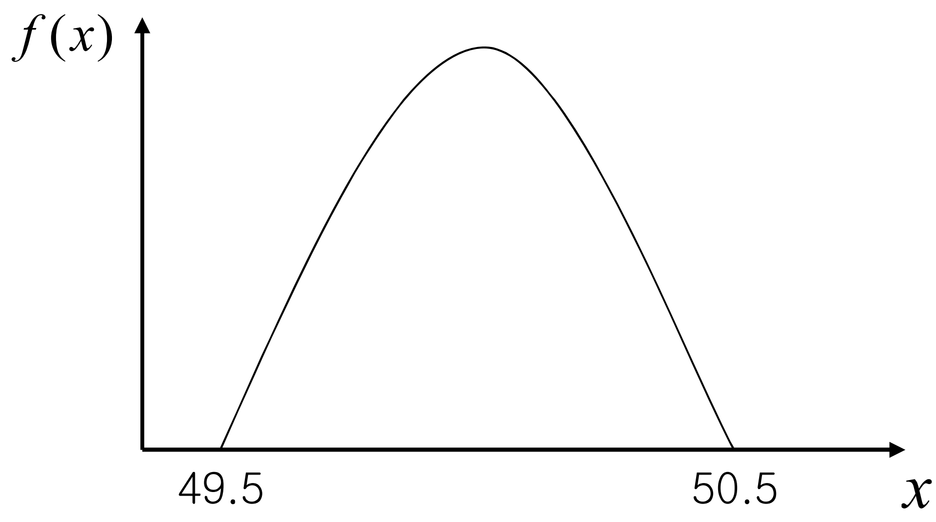
- Probability Density Function (p.d.f.)
  - Probabilistic properties of a continuous random variable

$$f(x) \geq 0$$

$$\int_{statespace} f(x)dx = 1$$

- Example 14
  - Suppose that the diameter of a metal cylinder has a p.d.f

$$f(x) = 1.5 - 6(x-50.2)^2 \quad \text{for} \quad 49.5 \leq x \leq 50.5$$
$$f(x) = 0, \quad \text{elsewhere}$$



Expectations of random variables

- Expectation of a discrete random variable with p.m.f : $P(X = x_i) = p_i$

$$E(X) = \sum_i p_i x_i$$

- Expectation of a continuous random variable with p.d.f $f(x)$

$$E(X) = \int_{\text{state space}} x f(x) dx$$

- The expected value of a random variable is also called the mean of the random variable
- Example: (discrete random variable)
  - The expected repair cost is $E(\text{cost}) = (\$50 \times 0.3) + (\$200 \times 0.2) + (\$350 \times 0.5) = \$230$
- Example: (continuous random variable)
  - The expected diameter of a metal cylinder is $E(X) = \int_{49.5}^{50.5} x(1.5 - 6(x - 50.0)^2) dx$

$$E(x) = \int_{-0.5}^{0.5} (y + 50)(1.5 - 6y^2) dy$$

$$= \int_{-0.5}^{0.5} (-6y^3 - 300y^2 + 1.5y + 75) dy$$

$$= [-3y^4 / 2 - 100y^3 + 0.75y^2 + 75y]_{-0.5}^{0.5}$$

$$= [25.09375] - [-24.90625] = 50.0$$

Medians of random variables

Median

– Information about the “middle” value of the random variable $F(x) = 0.5$

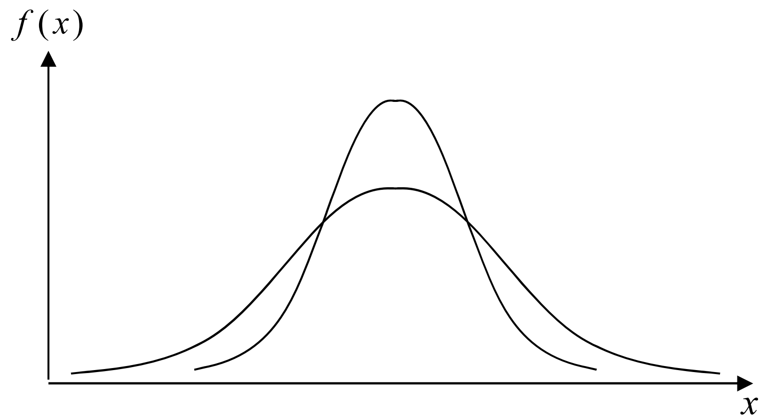
- Symmetric Random Variable
  – If a continuous random variable is symmetric about a point μ, then both the median and the expectation of the random variable are equal to μ.
- Example: $$F(x) = 1.5x - 2(x - 50.0)^3 - 74.5 = 0.5$$

$$x = 50.0$$

The variance of a random variable

- Variance(σ^2)
  - A positive quantity that measures the spread of the distribution of the random variable about its mean value
  - Larger values of the variance indicate that the distribution is more spread out
- Standard Deviation σ
  - The positive square root of the variance, denoted by σ

$$\begin{aligned}\mathrm{Var}(X) &= E((X - E(X))^2) \\ &= E(X^2 - 2XE(X) + (E(X))^2) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2\end{aligned}$$



Two distribution with identical mean values but different variances

- Example:

$$\begin{aligned}\mathrm{Var}(X) &= E((X - E(X))^2) = \sum_i p_i (x_i - E(X))^2 \\ &= 0.3(50-230)^2 + 0.2(200-230)^2 + 0.5(350-230)^2 \\ &= 17{,}100 = \sigma^2\end{aligned}$$

$$\sigma = \sqrt{17{,}100} = 130.77$$

Chebyshev's inequality

- Chebyshev's Inequality
  - If a random variable has a mean μ and a variance σ^2, then

$$P(\mu - c\sigma \le X \le \mu + c\sigma) \ge 1 - \frac{1}{c^2}$$

$$\text{for} \quad c \ge 1$$

  - For example, taking c=2 gives

- Proof:

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) \ge 1 - \frac{1}{2^2} = 0.75$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \ge \int_{|x-\mu|>c\sigma} (x-\mu)^2 f(x)dx \ge c^2\sigma^2 \int_{|x-\mu|>c\sigma} f(x)dx.$$

$$\Rightarrow P(|x-\mu| > c\sigma) \le 1/c^2 \qquad \Rightarrow P(|x-\mu| \le c\sigma) = 1 - P(|x-\mu| > c\sigma) \ge 1 - 1/c^2$$

Example:

Suppose we have sampled the weights of dogs in the local animal shelter and found that our sample has mean of 20 pounds with standard deviation of 3 pounds. With the use of Chebyshev's inequality, we know that at least 75% of the dogs that we sampled have weights that are two standard deviations from the mean. Two times the standard deviation gives us 2 x 3 = 6. Subtract and add this from the mean of 20. This tells us that 75% of the dogs have weight from 14 pounds to 26 pounds.

Quantiles of random variables

- Quantiles of Random variables
  - The pth quantile of a random variable X is F(x) = p
  - A probability of p that the random variable takes a value less than the pth quantile
- Upper quantile
  - The 75th percentile of the distribution
- Lower quantile
  - The 25th percentile of the distribution
- Inter-quantile range
  - The distance between the two quantiles
- Example:

$$F(x) = 1.5x - 2(x - 50.0)^3 - 74.5 \text{ for } 49.5 \le x \le 50.5$$

  - Upper quantile:

$$F(x) = 0.75 \quad x = 50.17$$

  - Lower quantile:

$$F(x) = 0.25 \quad x = 49.83$$

  - Interquantile range:

$$50.17 - 49.83 = 0.34$$

Jointly distributed random variables

- Joint Probability Distributions
  - Discrete: $P(X = x_i, Y = y_j) = p_{ij} \geq 0$

    $$\text{satisfying} \quad \sum_i \sum_j p_{ij} = 1$$

  - Continuous: $f(x,y) \geq 0 \ \text{satisfying} \ \iint_{\text{state space}} f(x,y)dxdx = 1$

- Joint Cumulative Distribution Function
  - Discrete: $F(x,y) = P(X \leq x_i, Y \leq y_j)$ $\quad F(x,y) = \sum_{i:x_i \leq x} \sum_{j:y_j \leq y} p_{ij}$
  - Continuous: $F(x,y) = \int_{w=-\infty}^{x} \int_{z=-\infty}^{y} f(w,z)dzdw$

Marginal probability distributions

- Marginal probability distribution
  - Obtained by summing or integrating the joint probability distribution over the values of the other random variable
  - Discrete $P(X = i) = p_{i+} = \sum_{j} p_{ij}$
  - Continuous $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$
- Example:
  - Marginal p.m.f of X: $P(X = 1) = \sum_{j=1}^{3} p_{1j} = 0.12 + 0.08 + 0.01 = 0.21$
  - Marginal pmf of Y: $P(Y = 1) = \sum_{i=1}^{4} p_{i1} = 0.12 + 0.08 + 0.07 + 0.05 = 0.32$
- Joint PDF: f(x, y) $f_X(x) = \int f(x, y)dy$
- Marginal pdf's of X & Y: $f_Y(y) = \int f(x, y)dx$

Independence and covariance

- Two random variables X and Y are said to be independent if
  - Discrete: $p_{ij} = p_{i+}p_{+j}$ for all values i of X and j of Y
  - Continuous: $f(x,y) = f_X(x) f_Y(y)$ *for all x and y*
- Covariance $\mathrm{Cov}(X,Y) = E((X - E(X))(Y - E(Y)))$

$$= E(XY) - E(X)E(Y)$$

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E((X-E(X))(Y-E(Y))) \\ &= E(XY - XE(Y) - E(X)Y + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

- May take any positive or negative numbers.
- Independent random variables have a covariance of zero

Bayesian networks

- Bayesian nets (BN) are a network-based framework
- BN are different from other knowledge-based systems tools
- BN are different from other probabilistic analysis tools
- Knowledge structure
- Computational architecture

Merits of Bayesian networks

- Handling of Incomplete Data Sets
- Learning about Causal Networks
- Facilitating the combination of domain knowledge and data
- Efficient and principled approach for avoiding the over fitting of data
- Bayesian Probability : the degree of belief in that event
- Classical Probability : true or physical probability of an event

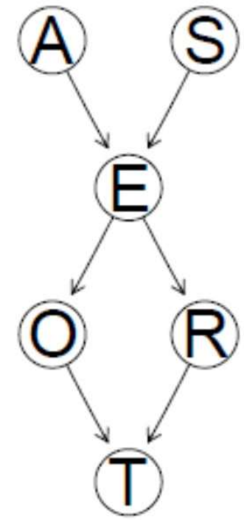
Construction of a Bayesian network

The approach is based on the following observations:

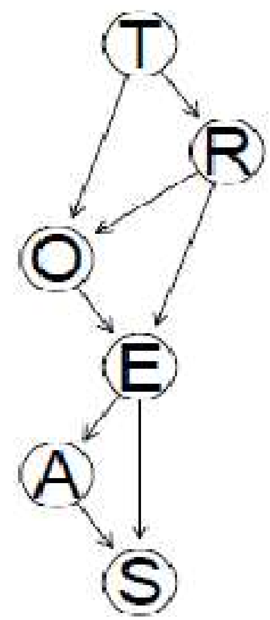
- People can often readily assert causal relationships among the variables
- Casual relations typically correspond to assertions of conditional dependence
- To construct a Bayesian Network we simply draw arcs for a given set of variables from the cause variables to their immediate effects.
- In the final step we determine the local probability distributions.

Representation in Bayesian networks

- A prognostic view of the survey as a BN is given here:
  - the blocks in the experimental design on top (e.g. stuff from the registry office);
  - the variables of interest in the middle (e.g. socio-economic indicators);
  - the object of the survey at the bottom (e.g. means of transport).
- Variables that can be thought as “causes" are on above variables that can be considered their “effect", and confounders are on above everything else.

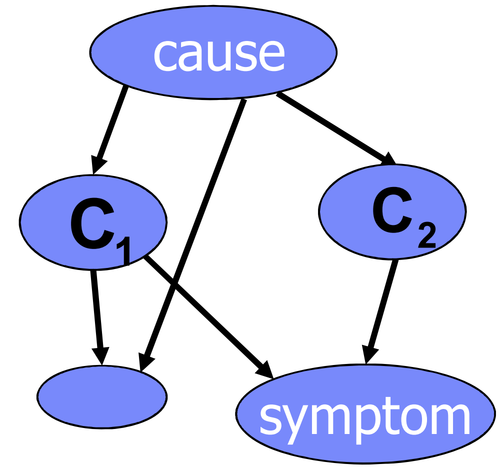


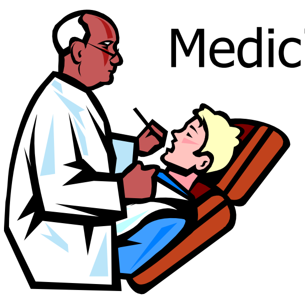
- A diagnostic view of the survey as a BN:
  - Encodes the same dependence relationships as the prognostic view but is laid out to have “effects” on top and “causes" at the bottom.
- Depending on the phenomenon and the goals of the survey, one may have a graph that makes more sense than the other; but they are equivalent for any subsequent inference. For discrete BNs, one representation may have fewer parameters than the other.



Benefits of Bayesian networks

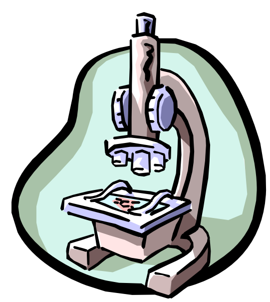
- Diagnosis: P(cause | symptom)=?
- Prediction: P(symptom | cause)=?
- Classification: maxclass P(class | data)
- Decision-making (given a cost function)



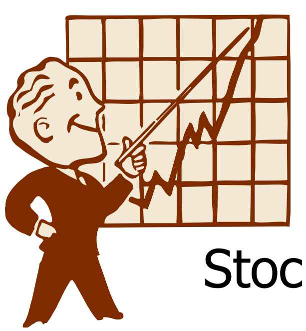


Medicine

Speech recognition



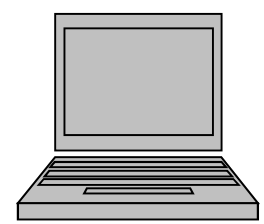
Bio-informatics



Stock market

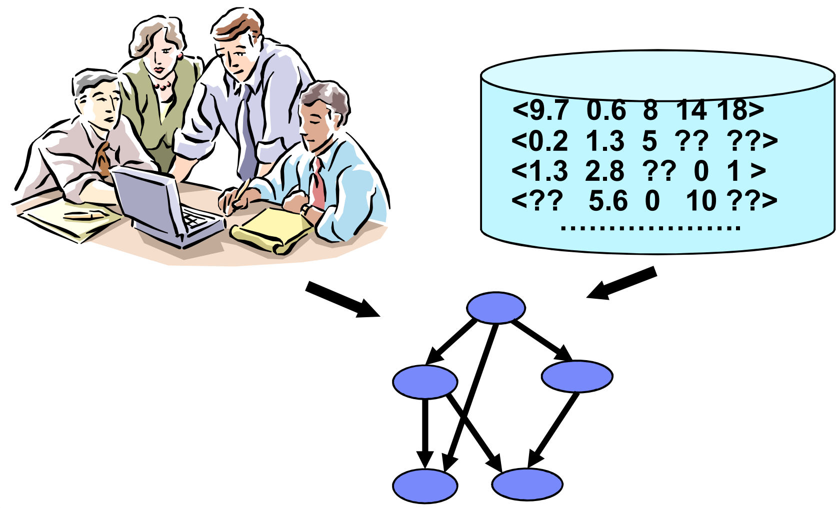
Text Classification

Computer troubleshooting



Why learn Bayesian networks?

- Combining domain expert knowledge with data

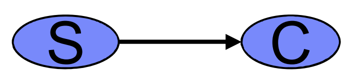


- Efficient representation and inference

- Incremental learning

- Handling missing data: **<1.3 2.8 ?? 0 1 >**

- Learning causal relationships:



Constructing Bayesian networks

- Choose an ordering of variables $X_1, \ldots, X_n$
- For i = 1 to n
  - add X_i to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that

$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$$

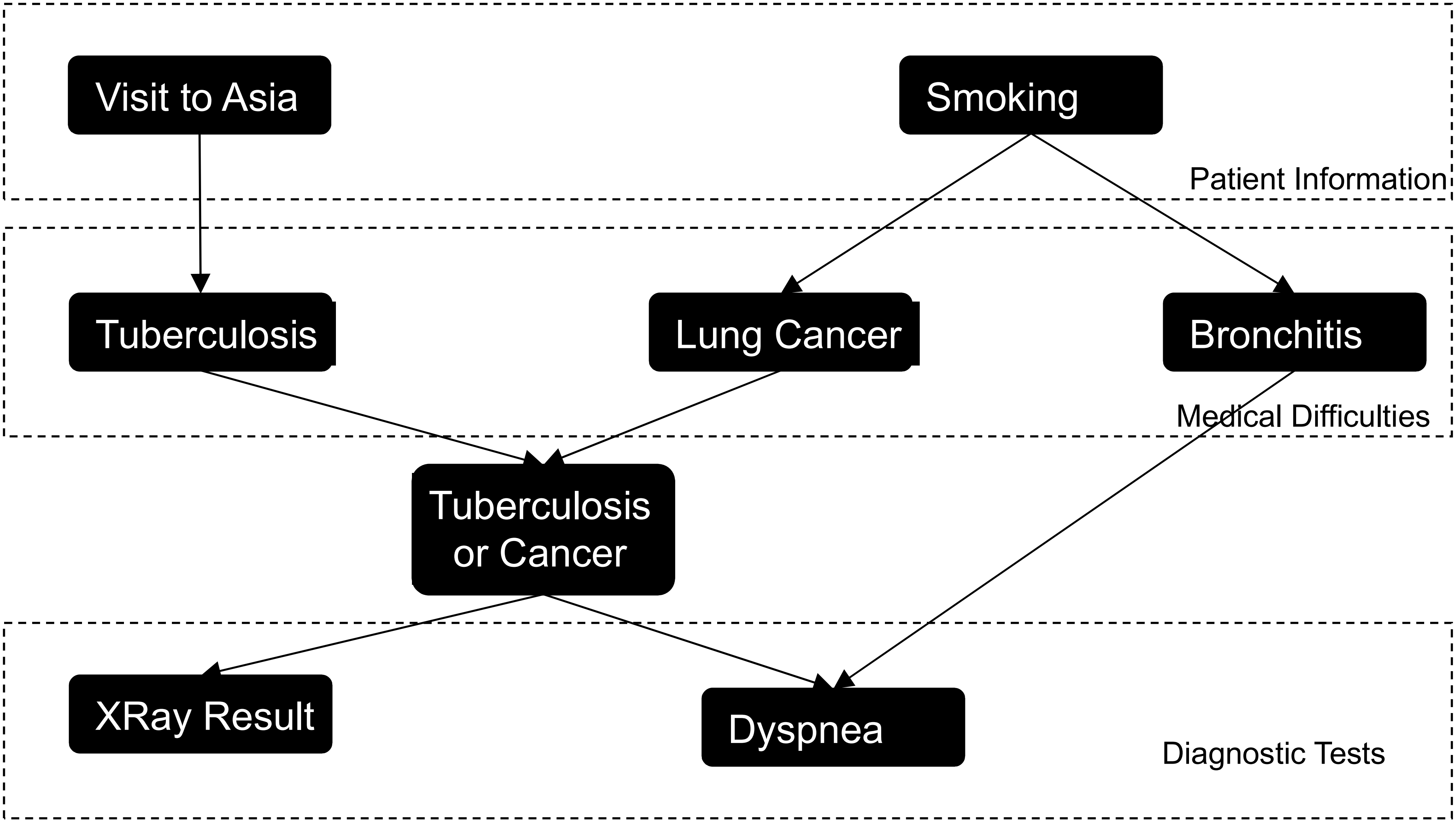
- This choice of parents guarantees:

$\mathbf{P}(X_1, \ldots, X_n) = \pi_{i=1} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$

(chain rule)

(by construction) $= \pi_{i=1} \mathbf{P}(X_i \mid Parents(X_i))$

Example from medical diagnostics



- Network represents a knowledge structure that models the relationship between medical difficulties, their causes and effects, patient information and diagnostic tests

Software for Bayesian networks

- BUGS: http://www.mrc-bsu.cam.ac.uk/bugs
  - parameter learning, hierarchical models, MCMC
- Hugin: http://www.hugin.dk
  - Inference and model construction
- xBaies: http://www.city.ac.uk/~rgc
  - chain graphs, discrete only
- Bayesian Knowledge Discoverer: http://kmi.open.ac.uk/projects/bkd
  - commercial
- MIM: http://inet.uni-c.dk/~edwards/miminfo.html
- BAYDA: http://www.cs.Helsinki.FI/research/cosco
  - classification
- BN Power Constructor: BN PowerConstructor
- Microsoft Research: WinMine (http://research.microsoft.com/~dmax/WinMine/Tooldoc.htm)

Gaussian Bayesian networks

- Gaussian Bayesian Networks is a BN all of whose variables are continuous and all of the CPDs are linear Gaussians.

- Linear Gaussian Bayesian networks are an alternative representation to multivariate Gaussian distributions.

Theorem:

- Let Y be a linear Gaussian of parents X1,.. Xk :
  - P(Y|x) ~ N(β0+βTx; σ2)

- Assume X1,..Xk are jointly Gaussian with distribution N(μ,Σ).

- Then distribution of Y is a normal distribution p(Y)=N(μY;σY2) where
  - μY=β0+βTμ,
  - σ2Y=σ2+βTΣβ
- The joint distribution over { X,Y } is a normal distribution where Cov(Xi ; Y)=Σj βjΣi,j

- It follows that if B is a linear Gaussian Bayesian network then it defines a joint Gaussian distribution.

Linear Gaussian BN to joint Gaussian

- Bayesian Network:

- CPDs:
  - p(X1)=N(1 ; 4)
  - p(X2|X1) =N(0.5X1-3.5 ; 4)
  - p(X3|X2)=N(-X2+1 ; 3)

- From theorem, compute joint Gaussian p(X1, X2, X3):
  - μ2=0.5μ1-3.5=0.5-3.5=-3
  - μ3=(-1)μ2+1=(-1)(-3)+1=4
  - Σ22=4+(1/2)24=5
  - Σ33=3+(-1)25=8
  - Σ12=(1/2)4=2
  - Σ23=(-1)Σ22=-5 Σ13=(-1)Σ12=-2

$$p(X_1, X_2, X_3) = N(\mu, \Sigma) \quad \mu = \begin{bmatrix} 1 \\ -3 \\ 4 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{bmatrix}$$

Theorem: From Gaussian to Bayesian networks

- Let χ={X1,..,Xn} and let p be a joint Gaussian distribution over χ

- Given any ordering X1,..,Xn over χ we can construct a Bayesian network graph G and a Bayesian network B over G such that
  - PaXi G ≤ {X1,..Xi-1}
  - The CPD of Xi in B is a linear Gaussian of its parents
  - G is a minimal I-map for p

- Given Gaussian

$$p(X_1, X_2, X_3) = N(\mu, \Sigma) \quad \mu = \begin{bmatrix} 1 \\ -4.5 \\ 8.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{bmatrix}$$

- Ordering X1, X2, X3

- Want to build linear Gaussian Bayes net:

$$X_1 \to X_2 \to X_3$$

- We will show that the CPDs are
  - p(X1)=N(1 ; 4)
  - p(X2|X1) =N(0.5X1-5 ; 4)
  - p(X3|X2)=N(-X2+4 ; 3)

Noisy OR-Gate model

- In BN, Conditional Probability Tables (CPTs) should be defined to measure the relationships between variables.
- Difficult to quantify the CPTs due to the complexity.
- Solutions to this problem is the Noisy-OR (NOR) gate, which can be attributed.
- Traditional NOR can only deal with the binary variables.
- The theory of belief functions, also called Dempster – Shafer Theory (DST), offers a mathematical framework for modeling uncertainty and imprecise information.
- Belief functions are widely employed in various fields, such as data classification, data clustering, social network analysis and statistical estimation.

Promedas: A clinical diagnostic decision support system

- Promedas is a medical patient-specific clinical diagnostic decision support systems based on graphical probabilistic models.

- Promedas aims to improve the quality and efficiency of the diagnostic process, while reducing its costs at the same time.

- A clinical diagnostic decision support system (CDDSS), called Promedas is a computer program that contains all relevant knowledge about a certain medical (sub)domain and generates patient specific diagnostic advice in the form of a list of likely diagnoses and suggestions for additional laboratory tests that are expected to be particularly informative to establish or rule out any of the diagnoses considered.

Organization of PROMEDAS development