DATA MINING

DATA MINING

1. What is **data mining**?
   Exploration and analysis of data to discover valid, useful, and understandable patterns in data. Its functionality is determined by data instead of programmed rules.

2. What **two things** can it be used for?
   Predicting future events and classifying current ones.

3. Describe the 7 main steps in the **data mining process**.
   The main steps are:
   - Generating raw data with the data bank or factory
   - Storing the relevant data in a database (selection and cleaning)
   - Transformation
   - Data mining
   - Interpretation
   - Understanding
   - Integration

4. Describe the two types of **machine learning task**.
   The two types are prediction and classification.

5. Name some **problems** that can be found with data.
   Some problems include:
   - The data could be measured or entered wrong
   - Quantity of data
   - Quality of data
   - Cost of getting the data

6. What kinds of things can be **mined**?
   Relationships between variables such as price affecting sales, and common patterns in categories such as demographic patterns about buying.

7. Name some **techniques** used in data mining.
   - K-nearest neighbour
   - Decision trees
   - Neural networks
   - K-means clustering
   - Market basket analysis
   - Logistic regression
   - ARIMA

8. What is meant by the following terms?
   a. Task – something we want the system to be able to do
   b. Variable – something we can measure that varies
   c. Value – the info assigned to that variable

   d. Data – measurements of values associated with variables

9. What is the difference between **numeric** and **nominal** data?
   Numeric data is continuous or discrete, whereas nominal values are discrete values with no numerical relationship between categories, i.e. pig and bird.

10. What is a **data model**?
    A data model is an abstract model that organises elements of data and how they relate to each other.

11. What is **learning**?
    Learning is the process of using data to build a model capable of performing a given task, usually through training.

12. What is **inference**?
    Inference is the process of drawing conclusions between the models and the data.

DATA PREPARATION

1. Name some points to check when performing **data preparation**.
   Data quantity and quality; how many variables, how complex is the task, is the distribution appropriate (i.e. outliers, balance...).

2. What is the difference between **data quality** and **data quantity**?
   Data quality means if we have good reliable data, and quantity means we have enough data.

3. What is a **frequency distribution**?
   A count of how often each variable contains each value in a data set.

4. What are some features of a **distribution** to look out for?
   Outliers, minority values, data balance, and data entry errors.

5. What is meant by the following terms and how can they affect data mining?
   a. Outliers – values that are smaller or larger than the others, could disrupt the mining process and give misleading results
   b. Minority values – values that occur infrequently in the data – do they appear enough to contribute to the model? Might be worth collecting more data.
   c. Flat & wide variables – can be one or the other – should be excluded

6. How can we ensure **data balance**?
   Try always to maintain balance in the examples i.e. 50-50 for men vs women etc. This is not always possible or necessary but try to strike as best a balance as you can.

7. How can we ensure **data quality**?
   Speaks for itself, cover all the situation, the more good data the better, if you have a good model, do you even need more data?

8. What is **linearity**? How does it relate to data quantity?
   Two variables have a linear relationship if plotting one against the other on a scatter plot produces a straight line.

9. Describe what is meant by **sampling theory**.
   Data will always be a sample from a much larger population, and therefore data you could not collect.

10. What is **noise/variability**? What factors could cause this?
    2 variables with a linear relationship might hover in a cloud around a straight line, could be caused by imperfect measurements or noise, simple randomness, or variability from other factors not being measured.

11. How can we find the right **line**?
    Using mean squared error to calculate the average of the squared errors.

12. What is **Mean Squared Error**?
    The average of the squared errors.

13. Describe what is meant by **learning**.
    Learning is the process of minimising the MSE.

14. How can we perform **learning**?
    We can use linear regression equation solving or iterative searches.

15. What should be done in the event of a **non-linear relationship**?
    In this case we need to capture the nature of the function y=f(x) from the data – hard to tell the difference between random variables from a line and a curve if they do not have much data.

DATA MINING CLASSIFICATION

1. What is **classification**?
   Assigning an object to a certain class based on similarity to previous examples of other objects.

2. What is **certainty**?
   The probability of an object belonging to the class.

3. What is **machine learning**?
   Learning a function (f) that maps input variables (x) to output variables (y).

4. How does an algorithm learn a **target mapping function**?
   From training data.

5. Give some examples of **techniques**.
   Non-parametric, i.e. k-nearest neighbour, mathematical models, i.e., neural networks, and rule-based models, i.e. decision trees.

6. What is the difference between **predictive** and **definitive**?
   With predictive, classification may indicate a propensity to act in a certain way, e.g. a prospect is likely to become a customer; whereas definitive classification may indicate similarity to objects that are members of a certain class.

7. Describe how the **k-nearest neighbour algorithm** works.
   Count the number (k) of other examples that are close, winner is the most common.

8. What is meant by **rule-based**?
   If above or below a certain point that infers it belongs to a group.

9. What is a **decision tree**?
   A rule discovery technique that produces a set of branching decisions ending in classification.

10. Describe the steps and components involved in making **classifications**.
    - Each node = single variable
    - Each branch = value that variable can take
    - To classify an example, start at top of tree and see which variable it represents
    - Follow branch corresponding to value the variable takes
    - Keep going until leaf – now classified

11. What is a **tree structure**?
    A way in which we can arrange a decision tree – we want to make the classification process as fast as possible and optimise the number of correct classifications.

12. Describe an example of a **tree-building algorithm**.
    Divide and conquer works by choosing the variable at the top of the tree, creating a branch for each possible value, and repeating for each branch until there are no more branches to make.

13. Describe how the **ID3 algorithm** works.
    With ID3 we want to split on the variable that gives the greatest information gain. Info can be thought of as a measure of uncertainty.

14. How can we calculate the **information** associated with a single event? Provide the formula.
    The information associated with a single event is |(e) = -log (Pe) where Pe is the probability of the event occurring and log is the base 2 log.

15. What is **entropy**? How can we calculate it?
    Entropy H(x) is a measure of uncertainty in variable x

    Entropy - $H(x) = \sum P(x_i) I(x_i) = -\sum P(x_i) \log(P(x_i))$ **// Log is base 2 log**

    Also, this- $Entropy(S) = -\frac{p}{p+n}\log_2(\frac{p}{p+n}) - \frac{p}{p+n}\log_2(\frac{p}{p+n})$

    Example - For face cards in deck of cards:

    Face = -log (4/13) = 1.7 Not face = -log (9/13) = 0.53

    H = (4/13) *1.7 + (9/13) *0.53 = 0.89 **(The measure of uncertainty)**

16. How does the **entropy** change as the distribution of x becomes **more even**?
    The more even the distribution, the higher the entropy and therefore higher uncertainty becomes.

17. What is **information gain**? How can we calculate it?
    Information cain of input is calculated using

H(Outcome) - H (Outcome | Input)
Number > 9:
Entropy (4F, 1NF) = -(4/5) log (4/5) -(0/5) log (0/5) = 0.2575
Entropy (0F, 8NF) = -(0/8) log (0/8) -(8/8) log (8/8) = 0

Gain(A) = E (Current set)- E (all child set)
Gain (Number > 9) = 0.89 - (5/13*0.2575 + 8/13 * 0) = 0.7910

Whichever variable removes the most uncertainty (higher the gain) becomes the top node, this process continues until all data is accounted for.

DATA VISUALISATION

1. When should **data visualisation** be used?
    - Before starting a data mining project, to understand the problem
    - To guide the data mining project and choice of technique
    - To improve the use of data mining techniques, e.g. choosing number of clusters
    - To show the results of a data mining analysis

2. What is a **scatter plot**?
   A plot that uses cartesian coordinates to display values for usually two variables against one another using dots.

3. Which scenarios should **bar charts** be used in, and which for **line charts**?
   Bar charts should be used for frequencies, and line graphs for continuous variables.

4. What is a **boxplot**?
   A boxplot is a method for graphically representing numerical data through their quartiles.

5. Describe **overlap problems** and how **jitter** can help.
   If there is a lot of data plotted around the same point on the graph, then it is hard or impossible to see how much there is at that point. Jitter adds small random amounts to the data to spread them out a little and mitigate overlap.

6. Describe some problems with **dimensions** in visualisation.
   As the number of variables increase it becomes harder to plot things against each other; i.e. if you have 5 differently coloured dots on the graph it could get a little confusing to understand.

7. What is a **correlation matrix**? What is a **correlation coefficient**?
   A correlation matrix plots all pairs of variables; there are scatter plots of each pair lower-left and a histogram of each variable on the diagonal.
   Correlation coefficient is in the upper right along with significance and is the measure of the relationship between two variables.

8. What is **projection**?
   Projection is involved when data comes from a system with more dimensions than can be plotted. Plotting data in fewer dimensions than it contains spoils the plot.

9. How can we solve **projection problems**?

We can represent all dimensions in some way, for example by colour and shape, reduce dimensionality, or using software that is able to represent the model in higher dimensions.

10. Describe what is meant by **parallel coordinates**.
    Parallel coordinates contain many variables at once and identify correlations.

11. Describe what is meant by **dimensionality reduction**.
    Reducing 2 or more related dimensions into a single new dimension that can be plotted against others to allow a deeper relationship to be found, however there is always a loss of info.

12. Give some example techniques of **dimensionality reduction**.
    Principal components reduction, non-linear principal components refraction, and auto association neural net.

13. Name some methods of **visualising data** for the average user.
    Informatics might be the most suitable way to represent data to the average user as it is less of an analysis tool and more of a presentation one.

PREDICTION

1. What is **prediction**?
   Predicting the identity of one thing based purely on the description of another related thing; not necessarily future events.

2. What is the difference between **predicted values** and **classifications**?
   Predicted values are usually continuous while classifications are discrete.

3. Describe some **prediction techniques**.
   - Simple stats models such as linear regression
   - Non-linear stats such as power series
   - Neural networks, RBFs, etc.

4. What is the **mathematical model**?
   A model that learns the relationship between predictors and predicted values.

5. What is **regression analysis**?
   A statistical process for estimating the relationships among variables. Regression models are built from data to predict the average you would expect the variable to have.

6. What is the formula for **regression**?
   $Y1 = bXi + a + Ei$

7. Describe how **simple linear regression** works.
   Mapping one variable onto the mean value of another.
   To find the best values for a and b, SLR uses a method called ordinary least squares. Least squares mean the sum of the squared distance between each data point and its associated data point is minimised, that is:

$$\sum_{i=1}^{n} \varepsilon_i^2$$

8. How can **a** and **b** be found from regression?

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

9. What is **multiple regression**?
   With multiple inputs, general form of linear regression is:

$$y_i = b_0 + x_{i1}b_1 + x_{i2}b_2 + x_{i3}b_3 + \ldots + \varepsilon_i$$

$$Y = Xb + \varepsilon$$

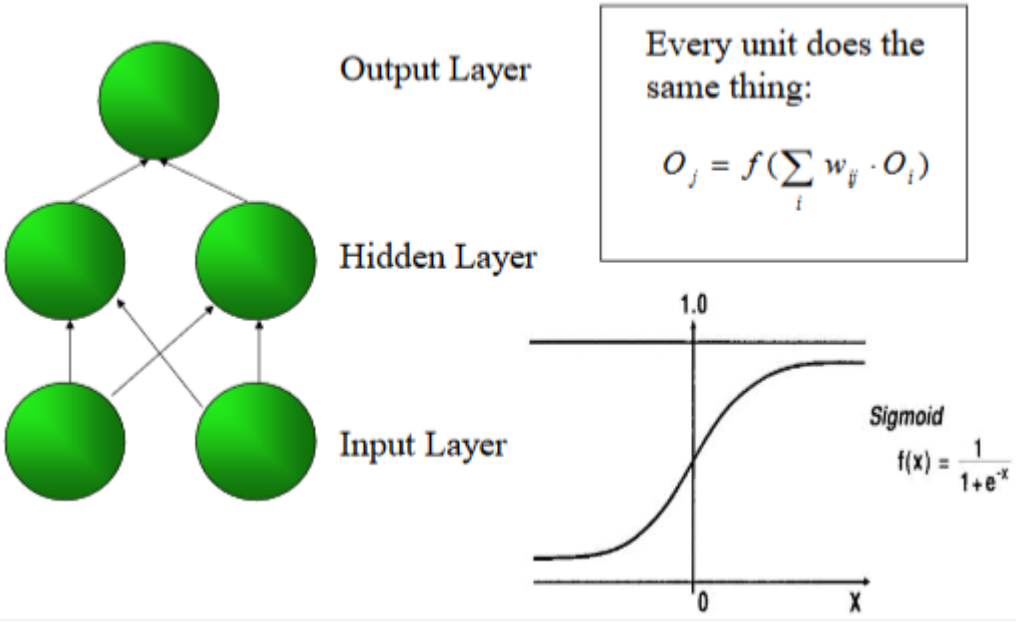
   The parameters in b are calculated as:

$$b = (X^T X)^{-1} X^T Y$$

10. What is a **neural network**? How does it work?
    A certain type of neural network, called a multi-layer perceptron (MLP) can learn a function between our inputs (qualities of a newspaper) and the outcome (Sales)
    It works by building the function out of many small simple functions, joined by weighted connections

11. What is the **MLP structure**?
    Multi layered perceptron uses many small weighted functions joined by weighted connections.
    Starting with random weights error is produced by the difference of the predicted and actual output when training (error).
    Using the weighted input, and an activation function we get an output of a neuron and so on.

12. Provide a **diagram** of the MLP structure.



13. How is **neural network training** performed?
    - Prepare data so that a file contains the predictors and the predicted variables with an example per row
    - Split the data into a test set and a training set
    - Read each row in turn into NN, presenting predictors as inputs and predicted value as the target output
    - Make a prediction and compare value given by NN to the target value
    - Update the weights
    - Present the next example in the file
    - Repeat until error no longer reduces – ideally stop when test error is at its lowest

14. How are **weights** changed in an MLP?
    MLP starts with random weights; each example in the training data is used as an input and the network generates an output. Differences between output and value in the training data is known as the error.

15. Describe the process of **backpropagation**.
    The final output is compared with the desired output, this difference is called the error signal, this gathers error from each neuron and each input node are then modified. This process repeats.

16. What is **deep learning**?
    A host of statistical machine learning techniques that enable the automatic learning of future hierarchies. They are usually based on artificial neural networks.

17. Describe some qualities of a **predictor**.
    The technique should have the ability to make correct predictions on data that is not in the original training data as well as the ability to provide a certainty measure with its predictions.

18. What is **overfitting**?

Occurs when a data mining predictor can capture the structure of the data so well that irrelevant details are picked up when they are not generally true; throws the whole thing off.

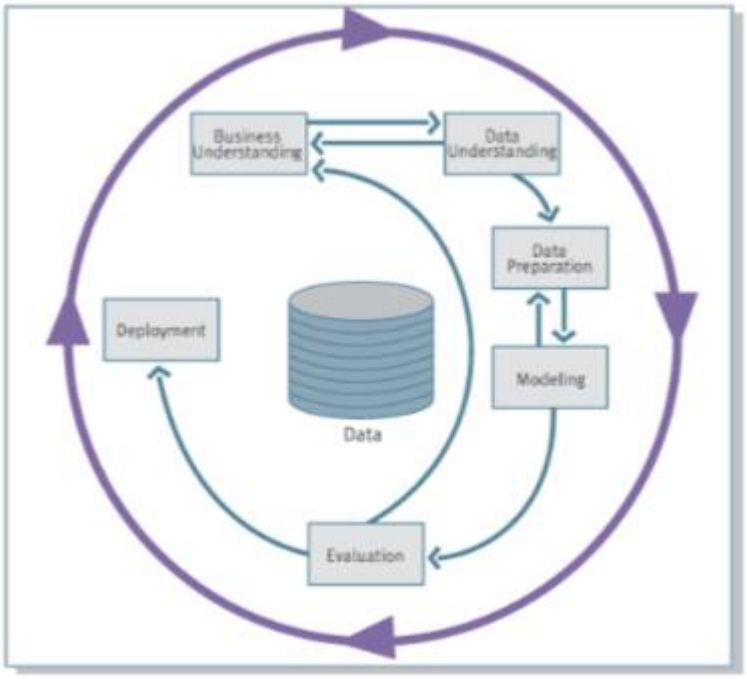
19. What is **data quantity** and **quality**?
    Data quantity – amount of data we have
    Data quality – accuracy and usefulness of the data we have

20. Describe some **advantages and disadvantages** of a neural network.
    Advantages
    - Can cope with non-linear relationships, multiple numeric and discreet variables
    - Able to generalise to data that it has not seen before

    Disadvantages
    - How predictions are gained can be hard to understand by a human user
    - Not easy to ask why answer was given
    - No rules to look at
    - Can make big errors if not trained properly
    - Requires a certain degree of faith

RUNNING A DM PROJECT

1. What is the **CRISP DM Standard**?
   Cross Industry Standard Process for Data Mining



2. Describe the steps involved in **data preparation**.
   - Clean the data
     - Remove rows with missing values
     - Remove rows with obvious data entry errors
     - Recode obvious data entry inconsistencies
     - Remove rows with minority values

3. How can **data quantity** be checked?

- Choose variables to be used for model
- Look at distributions of chosen values
- Look at level of noise in the data
- Decide whether there are enough examples in the data
- Treat unbalanced data

4. What are **error costs**? Why should error costs be considered?
   An error in one direction can cost more than an error in the opposite direction, e.g. better to recommend a blood test based on a false positive than miss an infection due to a false negative.

5. When building models, what should the following be used for? (**prediction** or **classification**)
   a. Neural network – prediction or classification
   b. Decision tree - classification
   c. Rule induction - classification
   d. Regression - prediction
   e. Bayesian – classification

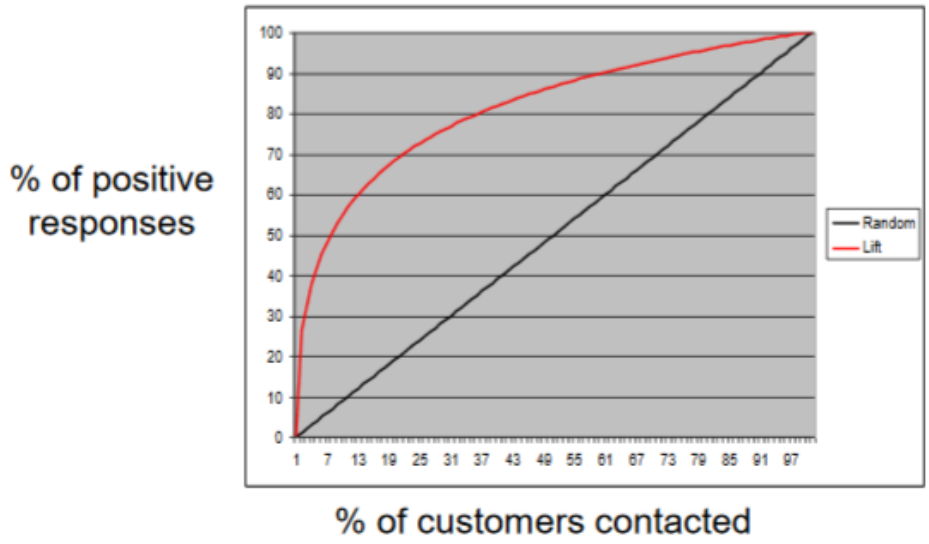
6. How do we **train models**?
   For each technique:
   - Run series of experiments with different parameters, e.g. the number of hidden units in an MLP
   - Each experiment should use ~70% of the data for training & the rest for testing
   - When a good solution is found, use cross validation to verify the result (10-fold good)

7. What is **cross validation**?
   A resampling procedure used to evaluate machine learning models on a limited data set.

8. How is **cross validation** performed?
   - Split data into 10 subsets, then train 10 models – each one using 0 of the 10 subsets as training data and the 10^{th} as test. The score is the average of all 10
   - More accurate representation of how well the data may be modelled, as it reduces the risk of getting a lucky test set

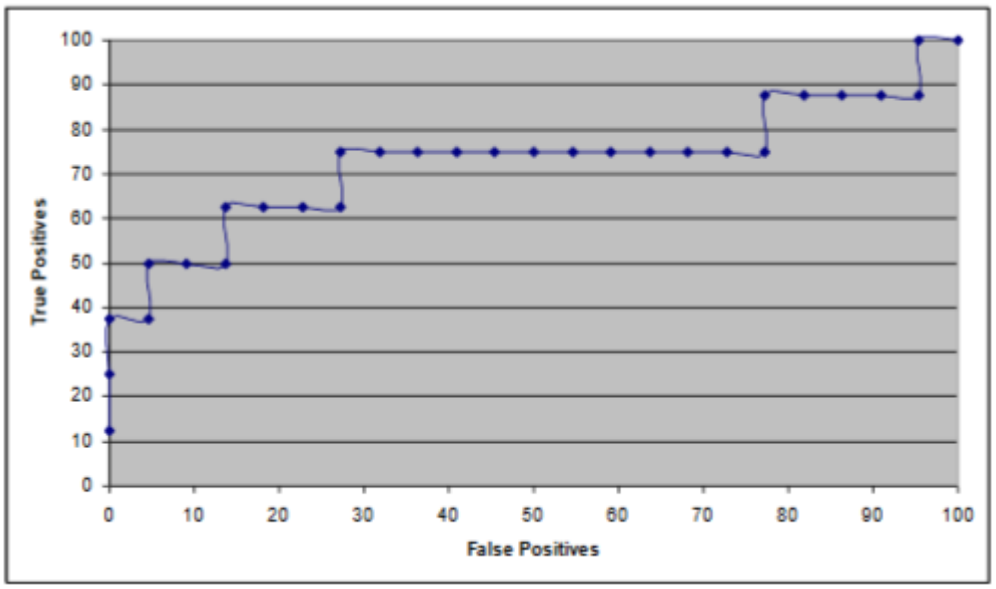
9. Describe some ways in which a model's **accuracy** can be assessed.
   - Mean squared error – not always meaningful
   - Percentage correct - for classification
   - Confusion matrix - for classification

10. What is a **lift curv**e?
    The measure of the performance of a targeting model at predicting or classifying cases as having an enhanced response, measured against a random choice targeting model.



11. What is a **cumulative gains chart**?
    Evaluation curve that assesses the performance of the model and compares the results with the random pick. Shows percentage of targets reached when considering a certain percentage of the population with the highest probability to be target according to the model.

12. What is a **ROC curve**?
    Tells you how many false positives and true positives you would get for each possible threshold; the threshold for a positive is varied from 0 to 1 and false and true positives counted for each.



CLUSTERING

1. Describe the elements involved in **supervised learning**.
    - F(x) – true function (usually not known)
    - D – training sample (x, f(x))
    - G(x) – model learned from D
    - Goal – E[(f(x) – g(x)) ^2] is near zero for future samples

2. What are the main differences between **supervised** and **unsupervised learning**?

Supervised learning has a true function and uses a labelled data set, whereas unsupervised has no true function and an unlabelled data set.

3. What is **clustering**? How does it work, and is it **supervised** or **unsupervised**?
   Clustering is a type of unsupervised learning:
   - What we have – Data set D and a similarity/distance metric
   - What we need to do – find partitioning of data, or groups of similar/close items
   - An illumination – find natural grouping of instances given un-labelled data

4. Name some potential applications of **clustering**.
   Marketing, image segmentation and natural language processing.

5. What is **similarity**? How can it be used?
   Similarity just refers to what it means; how similar are groups of similar customers, products, etc. Can include things like numeric data, i.e. Manhattan and Euclidean distance as well as categorical data, or both.

6. What is the difference between **Manhattan** and **Euclidean distance**?
   Manhattan is along the corridor and up the stairs, whereas Euclidean is "as the crow flies", so with a triangle length 4 and width 3, the Euclidean distance is 5 (sqrt(4^2 + 3^2)) and the Manhattan is 7.

7. What is **mean clustering**? How can we calculate the mean average of data set size S?
   An approach to clustering numeric data based on picking several mean values, one for each cluster.
   $$\bar{x} = \frac{\sum_{S} x}{S}$$

8. How does a clustering algorithm **work**?
   By separating out the data and calculating the means.

9. How can we **minimise the total distance between data points and the means to which they are assigned**?
   $$\arg\min(S) \sum_{i=1}^{k} \sum_{x_j \in s} \left\| x_j - m_i \right\|^2$$

10. Describe how the **k-means clustering algorithm works**.
    The goal is to minimise the sum of square of distance from all data points to their means
    The algorithm:
    - Pick K different points from the data and assume they are the cluster centres
      - Random, first K, K separated points
    - Repeat until stabilisation:
      - Assign each point to closest cluster centre
      - Generate new cluster centres by calculating and move the centroids.

11. What are the **disadvantages**?

- Only measures the mean for each cluster so tells you nothing of the shape – must assume the cluster is round, but they rarely are.
- Need to know K before starting
- Assumes all distances are equally important

12. How does a **hierarchical clustering algorithm** work? What can it be used for?
    - Start with the same number of clusters as you have data points
    - Repeat until reached number of clusters (or everything in one cluster)
        - Join the two most similar clusters together
        - Calculate the new centre

13. What is a **minimum spanning tree**? How can it be used?
    A minimum spanning tree looks for clusters withing clusters; cluster 1, the root, is the whole data set, that splits into a small number of subsets – each subset splits into 0 or more subsets.

    Cluster qualities

    - Population size – how many data points in that cluster
    - Variance/range – how far from centre does most data lie

14. What is a **dependency rule**?
    Predict occurrence of some items based on occurrences of other items.

15. What is a **frequent itemset**?
    An itemset whose support is >= a minsup threshold.

16. What is meant by the **support** and **support count**?
    - **Support count -** Frequency of item occurrence
    - **Support -** Fraction of transactions the contain and itemset

17. How do **rule evaluation metrics** work? What is meant by **support** and **confidence**, and how can they be calculated?
    - Support - % of transactions that contain both X and Y
    - Confidence - % (of transactions that contain X), that also contain Y

    Find all itemsets with support >= 2

Number	Items	Set	Support	Set	Support	Set	Support
1	a, b, c	{a}	3	{a, b}	3	{a, b, c}	2
2	b, c, d, e	{b}	4	{a, c}	2	{b, c, d}	1
3	c, d	{c}	4	{a, d}	1		
4	a, b, d	{d}	3	{b, c}	3		
5	a, b, c	{e}	1	{b, d}	2		
				{c, d}	2		

F = {{a},{b},{c},{d},{a,b},{a,c},{b,c},{b,d},{c,d},{a,b,c}}

Find minimum support 40% and minimum confidence 70%

Rules	Confidence	Set	Support
ac=>b (2/2) *100	100%	{}	5
a=>b (3/3) *100	100%	{a}	3
b=>a (3/4) *100	75%	{b}	4
b=>c (3/4) *100	75%	{c}	4
c=>b (3/4) *100	75%	{d}	3
{} =>b (4/5) *100	80%	{a, b}	3
{} =>c (4/5) *100	80%	{a, c}	2
		{b, c}	3
		{b, d}	2
		{c, d}	2
		{a, b, c}	2

18. Is X → Y the same as Y → X?
    No.

19. How can we mine **association rules**?

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Rules of itemset: {Milk, Diaper, Beer}

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- Computationally expensive as all the above rules are **binary partitions** of the same itemset: {Milk, Diaper, Beer}
- Rules originating from the same itemset have **identical support** but have different confidence

20. What steps are involved in **finding the rules**?
    Apriori algorithm works as follows:
    - Find all the acceptable itemsets – Support
    - Use them to generate acceptable rules – Confidence

Step 1 – Generate Itemsets

1. Find all the acceptable itemsets of size 1
2. Use the items from step 1 to generate all itemsets of size two and count their support. Keep those that are supported.
3. Repeat for increasingly large itemsets until none of the current size are supported

Example

- With a minimum support of 20%
  - Bread = 40%: Keep
  - Milk = 60%: Keep
  - Porcini = 2%: Discard
- {Bread, Milk} = 30%: Keep
- {Bread, Milk, Sardines} = 15%: Discard
- These are NOT rules yet! Just itemsets

Step 2: Generate Rules

- Generate every combination from the acceptable itemsets:

$$X \rightarrow Y \text{ where } X \cap Y = \text{Empty}$$

- That is, where nothing in X appears in Y, and vice-versa.

Example

- {Bread} → {Milk} is good
- {Bread, Milk} → {Coffee} is good
- {Bread} → {Bread, Milk} is not allowed

Finally

- Discard all the rules that have a confidence score lower than some pre-defined target
- Remember, confidence is the percentage of baskets that contain both parts of the rule

TIME SERIES FORECASTING

1. What is a **time series**?
   A sequence of values or events where the next event is determined by the events that precede it. The next step in a time series may be determined by 1 or more of the previous steps.

2. What is meant by the **order** of the time series?
   The number of steps is known as the order of the time series.

3. Describe the **anatomy** of a time series.
   TS reflects the process being measured, and the process has certain components that affect its behaviour.

4. What is meant by the following terms?
   a. Level – average value of the time series – if same throughout length, series is stationary
   b. Trend – function of time or previous values – process that produces values to get continually larger (or smaller) over time is said to have a trend (or to be non-stationary)
   c. Seasonality – any period that repeats through the data

d. Cycles – smooth undulations of a process; cycles often add together to produce complex waveforms

5. What is the difference between **cycles** and **seasonality**?
   Cyclic patterns are non-fixed fluctuations. Seasonal changes are due to very specific times.

   Average length - Cycles are longer than that of seasonal changes.
   Magnitude - Cycles tend to be far more variable than seasonal changes.

6. Describe some **techniques** for time series forecasting.
   Different techniques that use different components of a time series – cannot find trends on cyclic data for example.
   Simple
   - Predict the next step will be
     - The same as the previous one
     - The average of the last few
     - A weighted average of the last few

   Level
   - A process that operates at a fixed level might never leave that level – can use ARMA models to predict how quickly the process moves back to its level after being pushed off it by a shock

7. What is an **ARMA model**? How do they work? How does **trend** factor in?
   Auto-regressive moving average, AR affects previous values, MA how shock(variations) affect the future values. Regression can be used to find the trend.

8. What is **ARIMA**? How does it work?
   Auto-regressive integrated moving average, same as ARMA but it incorporates trend.

9. In the context of **ARIMA**, what it the difference between a **linear** and **non-linear** trend?
   Trend can be linear - growth by constant factor or non-linear - the rate of growth changes over time too.

10. What is a **growth rate**?
    The rate at which the trend line grows, could be 3x, x^2, ab^x, etc.

11. What is **seasonality**?
    Additive or multiplicative factors, i.e. summer 4 degrees warmer than winter, or December sales are three times as high as April sales, respectively.

12. What is **auto-correlation**?
    The degree of correlation between the values of the same variables across different observations in the data. Can also auto-correlate with the value two steps before, then three, and so on. High correlations should indicate you need to look for seasonal effects.

13. Describe what the difference is between **Fourier Transform** and **Recurrent Neural Networks**.
    Fourier Transform is good for taking any signal and decomposing it into a set of sine waves, whereas a recurrent neural network is good at finding cyclical components in a time series.

14. What are some problems with **time series forecasting**?
    - All techniques can appear to work even if time series is random

- Predictable time series can look random to the eye
- Strict tests needed to establish whether predictions are better than guess work
- Longer term trends are harder to capture

15. What are **time intervals**?
    - A system is said to be 'temporally dependent' if each step is predicted by previous ones
    - Many time series need to be measured at fixed time intervals to make sense
    - Many series are not measurable at fixed intervals and some do not depend on a fixed interval to be predictive

16. Why is **certainty** difficult?
    Unless system completely closed, future steps will be affected not only by previous steps but by outside forces too.
    Any part of the series you cannot account for is called the residual.

REASONING SYSTEMS

RULE BASED SYSTEMS

1. What is an **expert** (**rule-based)** system?
   A rule-based system aims to capture the knowledge of a human expert in a domain and embody it within a software system.

2. Describe the **5 main components** of a rule-based system.
   A rule-based system has the following:
   - Knowledge base – rules embodying expert knowledge about problem domain
   - Database – contains set of known facts about problem domain
   - Inference engine – carries out reasoning process, using rules and facts
   - Explanation facilities – provides info to users about reasoning steps being followed
   - User interface – communication between the user and the system

3. What is an **expert system shell**?
   An expert system shell is a program or framework that can be customised to create an RBS known as shells or rule engines. A shell usually provides an inference engine, explanation facility and infrastructure for populating the knowledge base and the rules.

4. Describe the **two main approaches** to **deductive inference**.
   Forward chaining uses rules like "If A and B then C" to increase facts in KB; the problem with this approach is that different rules might be valid at different times, so we need to identify (match) them and decide which one to use (Conflict resolution) – this one is useful for RBS with no specific goal.
   Backward chaining uses rules like "A if B and C" to increase KB facts; the system tries to justify all sub rules that satisfy the goal until an answer is found – might lead to many dead ends before a solution,

REASONING ABOUT UNCERTAINTY

1. Name some reasons why **uncertainty** can be introduced to a rule-based system.
   Some reasons include the information not being reliable, being incomplete, the language use being imprecise, conflicting information, or approximated information.

2. What is a **certainty**, or **confidence factor**?

A measure which represents a degree of confidence that some condition is true.

3. Provide the **two main equations** for **combining certainty factors** upon joining rules together.
   CF (A and B) = min (CF(A), CF(B))
   CF (A or B) = max (CF(A), CF(B))

   Or is max, and is min
   If B or C Then A @ 60 / If D and E Then B @ 100 / If G or H Then C @75 (find A)
   CF(D) = 80 / CF(E) = 50% / CF(F) = 90% / CF(G) = 20 / CF(H) = 80%
   CF(C) = CF (G or H) x75% = max (20%, 80%) x75% = 80%x75% = 60%
   CF(B) = CF (D and E and F) x 100% = min (80%,50%,90%) x 100% = 50%
   CF(A) = CF (B or C) x 60% = max (50, 60%) x 60% = 60% x 60% = 36%

   <For multiple independent rule conclusions>
   CF(A) = CF1 + CF2(100-CF1)/100
   1st rule A = 36% 2nd rule A = 75% so… CF(A) = CF1+CF2(100/CF1)/100
   CF(A) = 35+75(100-36)/100 = 84%

4. Given the statement "If P then Q @ n", what is the **uncertainty factor** of Q?
   CF(Q) – CF(P). n

5. What is the difference between **certainty factors** and **probability**?
   There is no rigorous mathematical foundation for certainty factors; they are often altered with the design, and there is no strict way of evaluating the certainty of a rule and/or fact.

6. What assumptions do the **older rule-based systems**, using **Bayesian probability**, make?
   Older systems assumed that the hypothesis is mutually exclusive and exhaustive, and that pieces of evidence are conditionally independent.

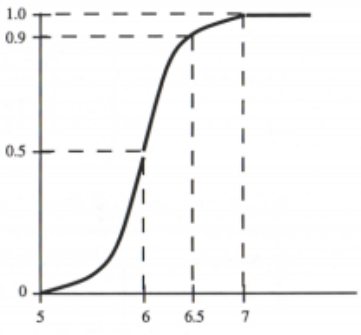
7. Why is it usually **inaccurate**? What model did this **evolve** into?
   Usually inaccurate as the assumptions above are mostly never true; model evolved into belief networks that give promising results.

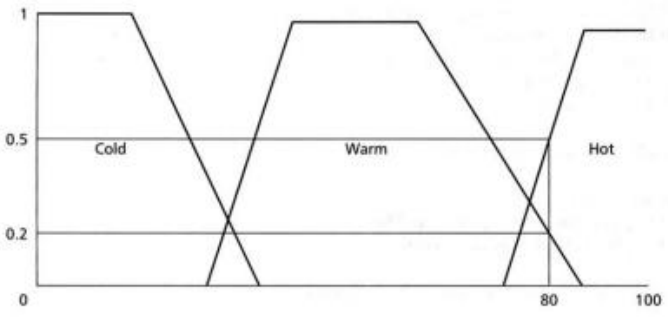
FUZZY LOGIC

1. What is **fuzzy logic**?
   An alternative method to Boolean logic for determining the value of a property – instead of true or false a percentage of how the current condition fits is used, from 0 to 1.

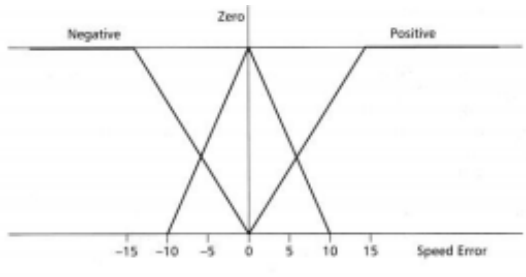
2. What is the purpose of **fuzzy set theory** and **fuzzy logic**?
   Fuzzy set theory and fuzzy logic provide a precise and mathematical basis for uncertainty reasoning.

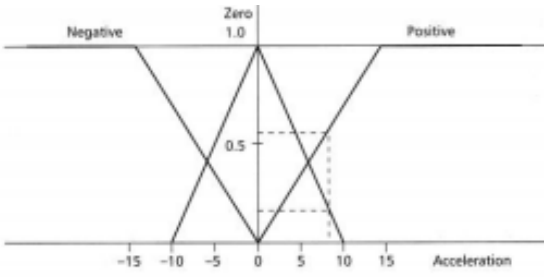
3. What is a **fuzzy set**?
   A fuzzy set can be shown in graphs, allowing calculation of corresponding degrees given the measurement. Actual measurement is defined as the opposite of fuzzy.



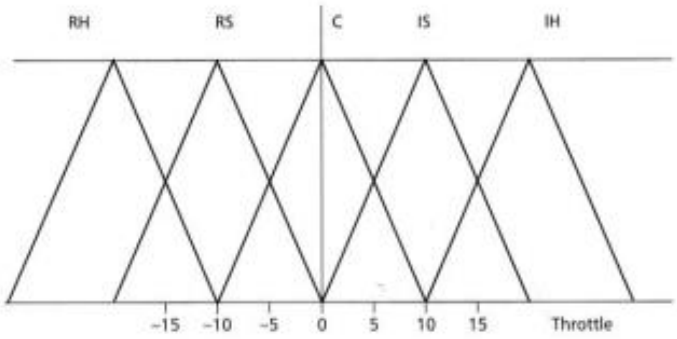


The best place to use fuzzy logic is in automatic control systems such as washing machines, air conditioners etc. They can allow for higher efficiency and smoother operation is these types of devices.





Example of plotting singular values I.e. acceleration or temperature (flat top or triangular).



Used when we have multiple values, i.e. more than three. (The tops could be flat, also the ends could be constant) The graphs depend more on what we are measuring here.

4. What is the opposite of **fuzzy**?
   The opposite of fuzzy is crisp.

5. How can they be used in **rule-based systems**?
   They can be used in rule-based systems to define how much something fits a certain quality; successful in automatic control systems like washing machines.

6. What is **defuzzification**?
   The process of producing quantifiable results in crisp logic, given fuzzy sets and corresponding membership degrees.

7. Name a common **defuzzification technique**.
   Centre of gravity (or centroid).

8. What is a **hedge**, and how are they used?
   A qualifying word provided using a mathematical definition, i.e. very, slightly, somewhat.

CASE-BASED REASONING

1. What is **case-based reasoning**?
   Relies on the analogy between a problem and a previous one, looking for the most similar problem that has been solved in the past, and applying the same solution; does not use inference and knowledge is stored as a record of cases.

2. Describe the **3 steps** towards creating a **CBR system**.
   The steps involved are identifying the attributes, identifying cases, and comparing a new case with the record of cases.

3. What **4 things** should the system do?
   - Retrieve the most similar case(s) from library
   - Reuse the case(s) – attempt to solve the problem
   - Revise proposed solution if necessary
   - Retain the new solution as part of a new case

4. Describe how CBRs use **similarity**.
   CBRs often use nearest-neighbour to find the most similar case, using the distance of two points in n-dimensional space.

5. What type of values is the **nearest-neighbour approach** effective with?
   Only effective with numerical values.

   i.e. if have three values age = 30 hair darkness = 8 eye darkness = 8 form vector3 (30,8,8) our case is (27,7,7) use Euclidean distance $\sqrt{[(x1-x2)^2+(y1-y2)^2+(z1-z2)^2]}$ = 9+1+1 = **square root of 11** or **3.317**
   Issues with the approach - All attributes are measured with numerical codes. For say colours, we would be a taxonomy, or table of colours.

6. What methods can be used to deal with the **other type** of values?
   With nominal values, we can use tables, ordered lists, taxonomies, etc.

7. What is **taxonomy-based comparison**?
   Like a decision tree but for computing nominal values that can be classified into groups.

8. What is **adaptation**?
   Once we have a best match we still need to adapt, this may involve eliciting more info or some judgement or expertise. At this point it could be carried out by a separate expert system or a human

expert. A classification or identification CBR system may not need this stage. Always used for design or planning however.

9. Why should **newly created cases** not be directly added to the case-base?
   If new cases are directly added, then the case base will become progressively degraded – should first be analysed and accepted.

10. Describe some **advantages** and **disadvantages** of case-based reasoning.
    Advantages
    - Knowledge based on previous data
    - Casual relationships do not need identification
    - Less formal logic needed
    - Adding new knowledge is simple
    - No need for general rules
    - Most people reason by analogy

    Disadvantages
    - All attribute values must be found
    - Must decide on algorithm to measure similarity
    - Must make effective adaptation strategy

11. When is it suitable to use **case-based reasoning**?
    - There are records of previously solved cases
    - Historical cases are used to solve new ones
    - Human experts tend to talk about examples rather than rules
    - Problem domain is not well-defined or well-understood
    - Experience is as important as theoretical knowledge

BAYESIAN BELIEF NETWORKS

UNCERTAINTY & PROBABILITY

1. What is a **decision support system**?
   A computer-based system that collects organises and analyses data.

2. What is a **Bayesian network**?
   A way of describing the relationships between causes and effects.

3. What is the meaning of **nodes** and **arcs** in a Bayesian network?
   Nodes are random variables, and arcs are casual or influential relations.

4. What is **conditional probability**, and how is it used in the context of Bayesian networks?
   Conditional probability is the likelihood of that outcome from the combinations of parent nodes.

5. Name some **applications** of Bayesian networks.
   Finding strategies to solve tasks under uncertainty, supporting decision making, fault diagnosis, etc.

6. Describe how Bayesian networks can be used to **model** and **reason** about **uncertainty**.
   Bayesian networks require probability to quantify uncertainty…

7. Which **two tools** are needed to implement BNs?
   Probability theory and propositional logic.

8. What is meant by the following terms?
   a. Uncertainty – lack of certainty; impossible to exactly describe outcome or state.
   b. Measurement of uncertainty – set of possible states/outcomes where probabilities are assigned to each possible state outcome.
   c. Risk – state of uncertainty where some possible outcomes have an undesired effect or significant loss.
   d. Measure of risk – set of measured uncertainties where some possible are losses, and the magnitude of those losses.

9. Name some **methods** of dealing with uncertainty.
   Confidence factors, fuzzy logic, and Bayesian theory.

10. What is **propositional logic**?
    Study of the ways of joining and/or modifying entire statements to form more complex ones.

11. How is it implemented?
    In order to implement it, many things are needed:
    - Logic – study of principles of correct reasoning
    - Statement – declarative statement, i.e. snow is white
    - Operator connectives – not, and, or also used
    - Sample space W – set of all possible outcomes
    - w in W - the possible outputs
    - Probability model – samples space with an assignment P(W) for each possible outcome (P(W) is a number in [0,1]; sum of all P(W) = 1)

12. What is a **random variable**?
    A function that associates a numerical value to each outcome of an experiment.

13. Describe the difference between **discrete** and **continuous** variables.
    A discrete variable is one which is obtained by counting, a continuous variable is obtained by measuring; for example, time is continuous but an age in years is discrete.

14. Give the **Probability Mass Function** of x.

    PMF of x $f(x) = \frac{p_j}{O}$ f(x) = Pj when $x = x_j$

15. Describe the difference between **Probability Density Function** and **Probability Mass Function**.
    PDF does not define a probability but probability density – to obtain probability we must integrate it in an interval.
    PMF gives true probability – does not need to be integrated to obtain probability.

16. Describe how to **calculate probabilities** when using **OR**, and when using **AND**.

    Example - Probability of picking an ace or king from a deck of cards.

    P (A v K) = (4+4)/52 = 8/52 = 2/13 = 0.15 = 15%

P (A v K) =P(A)+P(K)= (4/52) +(4/52)

Probability of picking a heart or an ace.

P (H or A) = (13+3)/52 this is not equal to P(H)+P(A)= (4+13)/52

P (A or B) = P (A v B) = P(A)+P(B) - P (A ^ B) therefore…

= (4+13-1)/52 = **0.308 or 30.8%**

Combination example - Probability of rolling a 6 and flipping a head on a coin.

P (roll a 6) = 1/6 P (flip a head) = 1/2 together = (1/6) *(1/2) = 1/12

Rolling a 6 or flipping a head = 1-(5/6) *(1/2) = 7/12

Conditional probability - Probability of B given A **= P(B|A)**

P (A and B) = P(A)*P(B|A) or P(A^B) =P(A)*P(B|A) **Order matters here.**

Smoker

	Yes	No	Total
Male	19	41	60
Female	12	28	40
Total	31	69	100

What is the probability of randomly picking a male who smokes?

19/100 = 0.19 why P (male ^ yes) = 0.19

Having picked a male what is the probability he smokes?

19/60 = 0.3167 why P (smoker | male) = P (male ^ smoker) / p(male) = 19 / 60

What is the probability of randomly selecting a smoker that is male?

19/31 = 0.6129 why P (male | smoker) = P (smoker ^ male) / p(smoker) = 19/31

Remember if the events are independent * else use conditional rule.

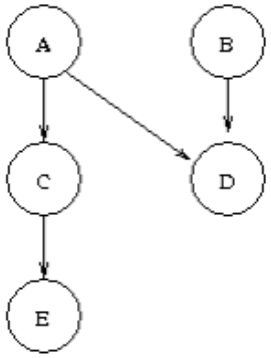
BAYESIAN CLASSIFICATION

1. What is **Bayes' Theorem**?

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Simple example if it cloudy it is rainy, P(C|R) = 1.0 what are the chances it rains if it is cloudy? Rain = 50% Cloudy = 80%

Therefore - P(R|C) = P(R)P(C|R)/P(C) = 0.5*1.0/0.8 = 0.625 62.5% chance to rain if cloudy.
More complex…



A and B are (absolutely) independent.

C is independent of B given A.

D is independent of C given A and B.

E is independent of A, B and D given C.

Suppose

Prob (A=T) = 0.3

Prob (B=T) = 0.6

Prob (C=T|A=T) = 0.8

Prob (C=T|A=F) = 0.4

Prob (D=T|A=T, B=T) = 0.7

Prob (D=T|A=T, B=F) = 0.8

Prob (D=T|A=F, B=T) = 0.1

Prob (D=T|A=F, B=F) = 0.2

Prob (E=T|C=T) = 0.7

Prob (E=T|C=F) = 0.2

If A=T is 0.3 A=F is 0.7

If B=T is 0.6 B=F is 0.4

Solve Prob(D=T)

= P(D=T|A=T,B=T) + P(D=T|A=T,B=F) + P(D=T|A=F,B=T) + P(D=T|A=F,B=F)

= P(D=T|A=T,B=T)P(A=T,B=T) + P(D=T|A=T,B=F)P(A=T,B=F)

+P(D=T|A=F,B=T)(A=F,B=T) + P(D=T|A=F,B=F)P(A=F,B=F)

A and B are independent so…

= P(D=T|A=T,B=T)P(A=T)P(B=T) + P(D=T|A=T,B=F)P(A=T)P(B=F) +

P(D=T|A=F,B=T)P(A=F)P(B=T) + P(D=T|A=F,B=F)P(A=F)P(B=F)

Sub values…

= (0.7 * 0.3 *0.6) + (0.8*0.3*0.4) + (0.1*0.7*0.6) +(0.2*0.7*0.4) = 0.32

2. What are **Bayesian classifiers**?
   Statistical classifiers based on Bayes' theorem. They can predict the probability that a sample is a member of a particular class.

3. Give an **example** of a Bayesian classifier.
   The Naïve Bayesian Classifier is a simple example; based on an independence assumption.

4. Describe some **features** of Bayesian classifiers.
   Performs well even with high-dimensional data points and/or many data points. They also have a comparable performance to decision trees or neural network classifiers; also assumes values given for one variable are not influenced by values given to another,

5. Why **use** Bayesian classifiers?
   No classification method has been found to be superior over all others in every case; methods can be compared based on training time, scalability, accuracy, and interpretability of the results. They can be useful in marketing products and can learn info about customers, such as how likely they are to do x, for example.

BAYESIAN NETWORKS

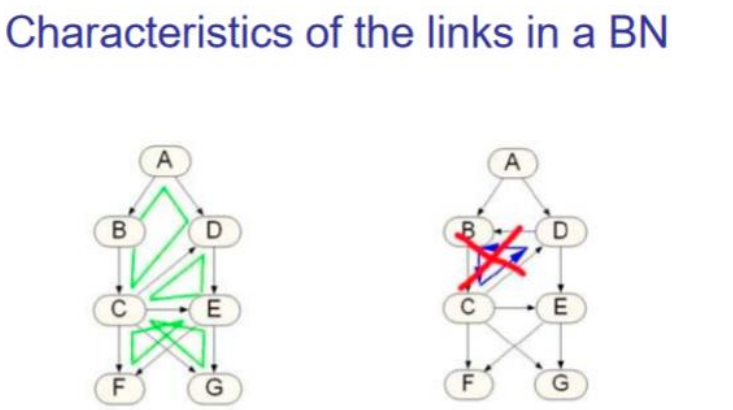
Again, some overlap, but important to know.

1. What is a **Bayesian network**?
   A network-based framework for representing and analysis models involving uncertainty. Used for intelligent decision aids, data fusion, intelligent diagnostic aids, and data mining.

2. Why are Bayesian networks of **growing interest**?
   They have a growing number of uses such as in dementia diagnosis and cancer case symptom modelling and are different from other knowledge-based systems tools and probabilistic analysis

tools.

3. Draw an example **diagram** of a simple Bayesian network, and describe **how it works**.



- The direction of the link arrows roughly corresponds to "causality"
- The nodes higher up in the diagram tend to influence those below
- The links may from loops, but they may not form cycles
- This does not limits the modeling power of the network. We need to be careful when building BNs
- Avoiding cycles makes possible very fast update algorithms.

4. Describe what is meant by **decision theory**.
   Decision theory is the science of decision making and wanting to maximise utility.

5. What is a **decision network**?
   A decision network produces the best decision for the user.

6. Name some **uses** of Bayesian networks.
   Bayesian networks can be used in any situation where modelling an uncertain reality is involved – decision support is involved whenever helpful to make decisions that maximise changes of a desirable outcome.
   Can also be used in diagnosis, prediction, risk assessment and sensor fusion.

7. What is meant by, in the context of Bayesian networks:
   a. Top nodes – predispositions which influence likelihood
   b. Second/third layer – internal conditions and failure states.
   c. Last layer – nodes of observables
   d. Links – correspond to causation

8. Name some **properties** of Bayesian networks.
   Probabilities need not be exact to be useful; approximate probabilities, even subjective, give good results. BNs are quite robust to imperfect knowledge and casual conditional probabilities are easier to estimate than reverse since people are better at estimating probabilities in the forward direction.

BAYESIAN NETWORKS 2

1. Describe the process of **building a Bayesian network**.
   - Step 1: collect info
     - List info given

- Determine info we can deduce from info given
- Convert info into BN
  - Determine nodes
  - Determine relationships between nodes
  - Convert info into network

2. Understand how to **use** a Bayesian network.

What we know

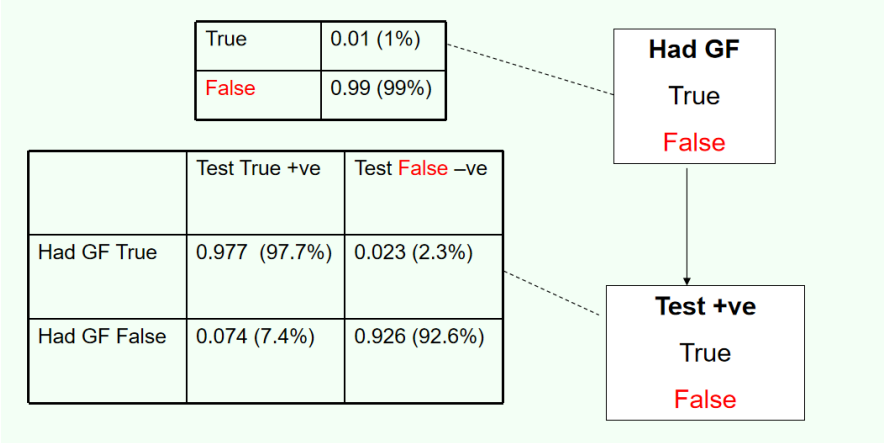
This gives us the following information:

P(had GF) = 0.01
P(not had GF) = 0.99

P(+ve test | had GF) = 0.977
P(-ve test | had GF) = 0.023

P(+ve test | not had GF) = 0.074
P(-ve test | not had GF) = 0.926

Convert to BBN



Use this to further elicit information

What is the probability of a +ve test result?

P (A n B) = P(B|A) P(A)

P (+ve test result n had GF) = P (had GF | +ve test) * P (+ve test)

P (+ve test result n not had GF) = P (not had GF | +ve test) * P (+ve test)

Stitch to…

P (had GF n +ve test result) = 0.977 * 0.01 = 0.00977

P (Not had GF n +ve test result) = 0.074 *0.01 = 0.07326

P (+ve test) = P (+ve test n had GF) + P (+ve test n not had GF)

=0.00977 + 0.07326 = **0.08303**

What is the probability that person has GF given a positive result?

P(A|B) =P(B|A) P(A)/P(B)

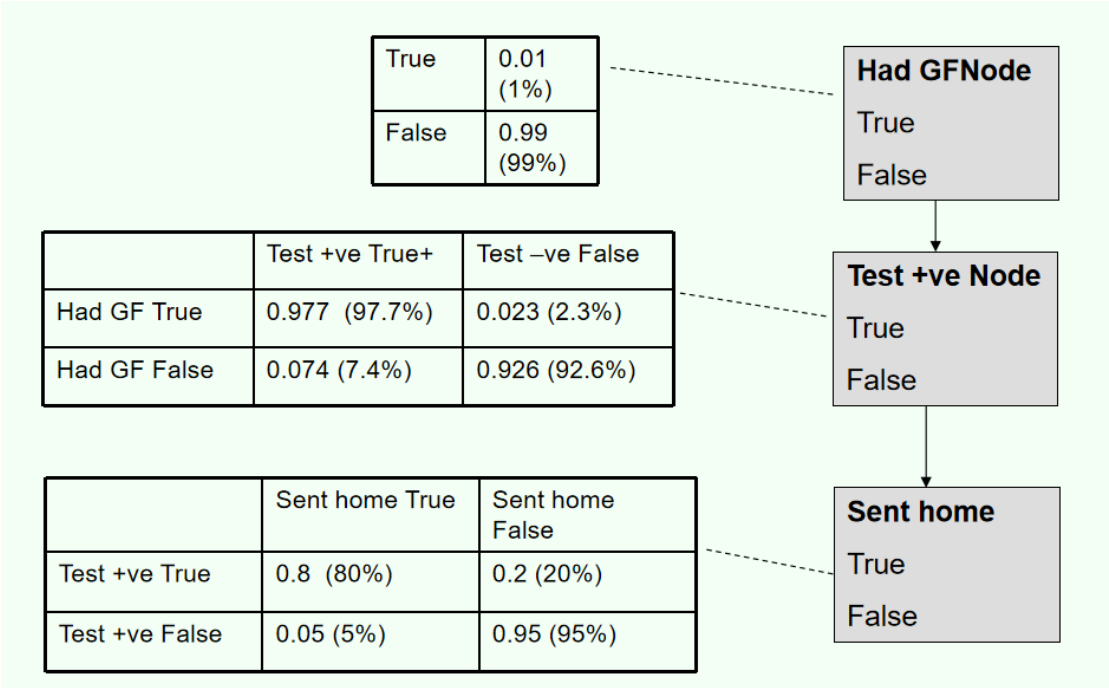
So…

P (has had GF | +ve test) = P (+ve test | has had GF) *P (has had GF)/P (+ve test)

= 0.977 * 0.01 / 0.08303 = **0.118**

How do we add this information into our network?

Also, we are told 80% of students testing positive are sent home. 5% are sent home for other reasons. Easily deduce the other side of the argument.



3. What is the difference between a **parent node** and a **child node**?

A parent node represents a node with no arrows leading into it. Probability info is given or can be deduced, and the probabilities of different states a node can have must sum to 1 (or 100%).
A child node represents a node with at least one arrow leading into it. Probability info is dictated by the values of parent(s) nodes feeding into the child node, and like parent nodes, the probabilities of different states a node can have must sum to 1 (or 100%).

BAYESIAN NETWORKS 3

1. What **formula** should be used in the following situations:
    a. Knowing the **parent** info; want to find **child** node info – use conditional probability P(AnB) = P(B|A) P(A)
    b. Knowing **child** info; want to find **parent** node info – use Bayes' Theorem P(A|B) = (P(B|A) P(A)) / P(B)
    c. If have more than one **parent** to a node (remember parents are **independent**) – P (Parent A n Parent B) = P (Parent A) * (Parent B)

2. What are the **qualitative** and **quantitative** parts of a Bayesian network?
   Qualitative – directed acyclic graph; nodes and edges
   Quantitative – set of conditional probability distributions

3. Describe what is meant by **posterior probabilities**.
   The probability of any event given any evidence.

4. Why is it important to learn Bayesian networks?
   Conditional independencies and graphical language capture the structure of many real-world distributions, the graph structure provides much inside into domain and allows knowledge discovery, and the learned model can be used for many tasks.

5. Describe some methods for learning the **structure** of a Bayesian network.
    - From an expert – knowledge acquisition bottleneck (expensive process and often not an expert available)
    - From data – data is cheap (the amount of available info grows rapidly; learning allows us to construct models from raw data)
    - Learning from experience
    - Learning using trees
    - Heuristic methods
    - Bayesian inference
    - Scoring methods