DATA MINING

DATA MINING

1. What is **data mining**?
2. What **two things** can it be used for?
3. Describe the 7 main steps in the **data mining process**.
4. Describe the two types of **machine learning task**.
5. Name some **problems** that can be found with data.
6. What kinds of things can be **mined**?
7. Name some **techniques** used in data mining.
8. What is meant by the following terms?
    a. Task
    b. Variable
    c. Value
    d. Data
9. What is the difference between **numeric** and **nominal** data?
10. What is a **data model**?
11. What is **learning**?
12. What is **inference**?

DATA PREPARATION

1. Name some points to check when performing **data preparation**.
2. What is the difference between **data quality** and **data quantity**?
3. What is a **frequency distribution**?
4. What are some features of a **distribution** to look out for?
5. What is meant by the following terms and how can they affect data mining?
    a. Outliers
    b. Minority values
    c. Flat & wide variables
6. How can we ensure **data balance**?
7. How can we ensure **data quality**?
8. What is **linearity**? How does it relate to data quantity?
9. Describe what is meant by **sampling theory**.
10. What is **noise/variability**? What factors could cause this?
11. How can we find the right **line**?
12. What is **Mean Squared Error**?
13. Describe what is meant by **learning**.
14. How can we perform **learning**?
15. What should be done in the event of a **non-linear relationship**?

DATA MINING CLASSIFICATION

1. What is **classification**?
2. What is **certainty**?
3. What is **machine learning**?
4. How does an algorithm learn a **target mapping function**?
5. Give some examples of **techniques**.
6. What is the difference between **predictive** and **definitive**?

7. Describe how the **k-nearest neighbour algorithm** works.
8. What is meant by **rule-based**?
9. What is a **decision tree**?
10. Describe the steps and components involved in making **classifications**.
11. What is a **tree structure**?
12. Describe an example of a **tree-building algorithm**.
13. Describe how the **ID3 algorithm** works.
14. How can we calculate the **information** associated with a single event? Provide the formula.
15. What is **entropy**? How can we calculate it?
16. How does the **entropy** change as the distribution of x becomes **more even**?
17. What is **information gain**? How can we calculate it?

DATA VISUALISATION

1. When should **data visualisation** be used?
2. What is a **scatter plot**?
3. Which scenarios should **bar charts** be used in, and which for **line charts**?
4. What is a **boxplot**?
5. Describe **overlap problems** and how **jitter** can help.
6. Describe some problems with **dimensions** in visualisation.
7. What is a **correlation matrix**? What is a **correlation coefficient**?
8. What is **projection**?
9. How can we solve **projection problems**?
10. Describe what is meant by **parallel coordinates**.
11. Describe what is meant by **dimensionality reduction**.
12. Give some example techniques of **dimensionality reduction**.
13. Name some methods of **visualising data** for the average user.

PREDICTION

1. What is **prediction**?
2. What is the difference between **predicted values** and **classifications**?
3. Describe some **prediction techniques**.
4. What is the **mathematical model**?
5. What is **regression analysis**?
6. What is the formula for **regression**?
7. Describe how **simple linear regression** works.
8. How can **a** and **b** be found from regression?
9. What is **multiple regression**?
10. What is a **neural network**? How does it work?
11. What is the **MLP structure**?
12. Provide a **diagram** of the MLP structure.
13. How is **neural network training** performed?
14. How are **weights** changed in an MLP?
15. Describe the process of **backpropagation**.
16. What is **deep learning**?
17. Describe some qualities of a **predictor**.
18. What is **overfitting**?
19. What is **data quantity** and **quality**?
20. Describe some **advantages and disadvantages** of a neural network.

RUNNING & DM PROJECT

1. What is the **CRISP DM Standard**?
2. Describe the steps involved in **data preparation**.
3. How can **data quality** be checked?
4. What are **error costs**? Why should error costs be considered?
5. When building models, what should the following be used for? (**prediction** or **classification**)
    a. Neural network
    b. Decision tree
    c. Rule induction
    d. Regression
    e. Bayesian
6. How do we **train models**?
7. What is **cross validation**?
8. How is **cross validation** performed?
9. Describe some ways in which a model's **accuracy** can be assessed.
10. What is a **lift curv**e?
11. What is a **cumulative gains chart**?
12. What is a **ROC curve**?

CLUSTERING

1. Describe the elements involved in **supervised learning**.
2. What are the main differences between **supervised** and **unsupervised learning**?
3. What is **clustering**? How does it work, and is it **supervised** or **unsupervised**?
4. Name some potential applications of **clustering**.
5. What is **similarity**? How can it be used?
6. What is the difference between **Manhattan** and **Euclidean distance**?
7. What is **mean clustering**? How can we calculate the mean average of data set size S?
8. How does a clustering algorithm **work**?
9. How can we **minimise the total distance between data points and the means to which they are assigned**?
10. Describe how the **k-means clustering algorithm works**.
11. What are the **advantages** and **disadvantages**?
12. How does a **hierarchical clustering algorithm** work? What can it be used for?
13. What is a **minimum spanning tree**? How can it be used?
14. What is a **dependency rule**?
15. What is a **frequent itemset**?
16. What is meant by the **support** and **support count**?
17. How do **rule evaluation metrics** work? What is meant by **support** and **confidence**, and how can they be calculated?
18. Is X → Y the same as Y → X?
19. How can we mine **association rules**?
20. What steps are involved in **finding the rules**?

TIME SERIES FORECASTING

1. What is a **time series**?
2. What is meant by the **order** of the time series?
3. Describe the **anatomy** of a time series.
4. What is meant by the following terms?

   a. Level?
   b. Trend?
   c. Seasonality?
   d. Cycles
5. What is the difference between **cycles** and **seasonality**?
6. Describe some **techniques** for time series forecasting.
7. What is an **ARMA model**? How do they work? How does **trend** factor in?
8. What is **ARIMA**? How does it work?
9. In the context of **ARIMA**, what it the difference between a **linear** and **non-linear** trend?
10. What is a **growth rate**?
11. What is **seasonality**?
12. What is **auto-correlation**?
13. Describe what the difference is between **Fourier Transform** and **Recurrent Neural Networks**.
14. What are some problems with **time series forecasting**?
15. What are **time intervals**?
16. Why is **certainty** difficult?

REASONING SYSTEMS

RULE BASED SYSTEMS

1. What is an **expert** (**rule-based)** system?
2. Describe the **5 main components** of a rule-based system.
3. What is an **expert system shell**?
4. Describe the **two main approaches** to **deductive inference**.

REASONING ABOUT UNCERTAINTY

1. Name some reasons why **uncertainty** can be introduced to a rule-based system.
2. What is a **certainty**, or **confidence factor**?
3. Provide the **two main equations** for **combining certainty factors** upon joining rules together.
4. Given the statement “If P then Q @ n”, what is the **uncertainty factor** of Q?
5. What is the difference between **certainty factors** and **probability**?
6. What assumptions do the **older rule-based systems**, using **Bayesian probability**, make?
7. Why is it usually **inaccurate**? What model did this **evolve** into?

FUZZY LOGIC

1. What is **fuzzy logic**?
2. What is the purpose of **fuzzy set theory** and **fuzzy logic**?
3. What is a **fuzzy set**?
4. What is the opposite of **fuzzy**?
5. How can they be used in **rule-based systems**?
6. What is **defuzzification**?
7. Name a common **defuzzification technique**.
8. What is a **hedge**, and how are they used?

CASE-BASED REASONING

1. What is **case-based reasoning**?
2. Describe the **3 steps** towards creating a **CBR system**.

3. What **4 things** should the system do?
4. Describe how CBRs use **similarity**.
5. What type of values is the **nearest-neighbour approach** effective with?
6. What methods can be used to deal with the **other type** of values?
7. What is **taxonomy-based comparison**?
8. What is **adaptation**?
9. Why should **newly created cases** not be directly added to the case-base?
10. Describe some **advantages** and **disadvantages** of case-based reasoning.
11. When is it suitable to use **case-based reasoning**?

BAYESIAN BELIEF NETWORKS

UNCERTAINTY & PROBABILITY

1. What is a **decision support system**?
2. What is a **Bayesian network**?
3. What is the meaning of **nodes** and **arcs** in a Bayesian network?
4. What is **conditional probability**, and how is it used in the context of Bayesian networks?
5. Name some **applications** of Bayesian networks.
6. Describe how Bayesian networks can be used to **model** and **reason** about **uncertainty**.
7. Which **two tools** are needed to implement BNs?
8. What is meant by the following terms?
    a. Uncertainty
    b. Measurement of uncertainty
    c. Risk
    d. Measure of risk
9. Name some **methods** of dealing with uncertainty.
10. What is **propositional logic**?
11. How is it implemented?
12. What is a **random variable**?
13. Describe the difference between **discrete** and **continuous** variables.
14. Give the **Probability Mass Function** of x.
15. Describe the difference between **Probability Density Function** and **Probability Mass Function**.
16. Describe how to **calculate probabilities** when using **OR**, and when using **AND**.

BAYESIAN CLASSIFICATION

1. What is **Bayes' Theorem**?
2. What are **Bayesian classifiers**?
3. Give an **example** of a Bayesian classifier.
4. Describe some **features** of Bayesian classifiers.
5. Why **use** Bayesian classifiers?

BAYESIAN NETWORKS

Again, some overlap, but important to know.

1. What is a **Bayesian network**?
2. Why are Bayesian networks of **growing interest**?
3. Draw an example **diagram** of a simple Bayesian network, and describe **how it works**.
4. Describe what is meant by **decision theory**.
5. What is a **decision network**?

6. Name some **uses** of Bayesian networks.
7. What is meant by, in the context of Bayesian networks:
    a. Top nodes
    b. Second/third layer
    c. Last layer
    d. Links
8. Name some **properties** of Bayesian networks.

BAYESIAN NETWORKS 2

1. Describe the process of **building a Bayesian network**.
2. Understand how to **use** a Bayesian network.
3. What is the difference between a **parent node** and a **child node**?

BAYESIAN NETWORKS 3

1. What **formula** should be used in the following situations:
    a. Knowing the **parent** info; want to find **child** node info
    b. Knowing **child** info; want to find **parent** node info
    c. If have more than one **parent** to a node (remember parents are **independent**)
2. What are the **qualitative** and **quantitative** parts of a Bayesian network?
3. Describe what is meant by **posterior probabilities**.
4. Why is it important to learn Bayesian networks?
5. Describe some methods for learning the **structure** of a Bayesian network.