**Data Mining**

**Q1:**  
Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late. What is all the excitement about? The world is data rich and knowledge poor. This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world applications of knowledge discovery, and current and future research directions in the field.

**Q2:**

It means that the how distant values are dispersed from mean. That can be achieved by standard deviation. Mode, mean and median are also included.

In data mining, it helps to collect and how values are lying are in which range.

**Q.3:**

Value of k:

Optimal value of k can be determined empirically. There no specific rule to find the optimal value of k. however to find the optimal value of K we try differing values of k on given dataset we then choose the value of k which gives better result on that dataset.

To classify the given instance:

To classify the given un-seen instance and assign a class we need the following steps of calculations.

1. **Normalization:** normalize the given dataset by subtracting the minimum value from all the data values of an attribute in the given training set, and then divide all the data value on max-value of that attribute.
2. **Calculate distance:** calculate distance of the unseen instance to with all the instance in the training set by measuring distance of each attribute distance + incorporate the weight in the distance formula as (C>A>B=D)
3. **Take K neighbor:** take **K**-nearest instances. Based on minimum distance.
4. **Assign the class:** assign the class of the most frequent instance among the k-nearest to unseen-instance

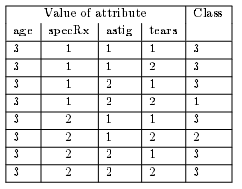
**Q4**

**Information Gain:**

Information Gain = *Estart− Enew* then

The ‘***entropy method***’ of attribute selection is to choose to split on the attribute that gives the. the one that maximizes the value of Information Gain.

For example



The information gain from splitting on attribute *age* is

1*.*3261 *−* 1*.*2867 = 0*.*0394 bits

**Q5 Differentiate between:-**

**Conditional Probability and Posterior probability**

Conditional Probability and Posterior probability is same thing:

Conditional Probability is calculated this as e.g:

* The number of times class = on time and season = winter (in the same instance), divided by the number of times the season is winter, which comes to 2/6 = 0.33.
* The probability of an event occurring if we know that an attribute has a particular value (or that several variables have particular values) is called the *conditional probability* of the event occurring and is written as, e.g.
* *P*(class = on time | season = winter).
* ‘the probability that the class is *on time* given that the season is *winter*

***Supervised and Unsupervised learning***

Data mining using labelled data is known as ***supervised learning***. E.g classification

Data mining of unlabelled data is known as ***unsupervised learning***. E.g Assosiation rule

**Interval Scaled And Ratio Scaled variables:**

Interval-scaled variables are variables that take numerical values which are measured at equal intervals from a zero point or origin.

Two well-known examples of interval-scaled variables are the ***Fahrenheit*** and ***Celsius*** temperature scales.

**Ratio Scaled Variables:**

Ratio-scaled variables are similar to interval-scaled variables except that the zero point does reflect the absence of the measured characteristic

For example Kelvin temperature and molecular weight.

In the former case the zero value corresponds to the lowest possible temperature ‘absolute zero’, so a temperature of 20 degrees Kelvin is twice one of 10 degrees Kelvin.

A weight of 10 kg is twice one of 5 kg, a price of 100 Rupees is twice a price of 50 Rupees etc.

**Deduction and Induction:**

The conclusion is shown to follow necessarily from the truth of the premises, for example

All Men Are Mortal

John is a Man

🡺Therefore John is Mortal

If the first two statements (the *premises*) are true, then the conclusion must be true. This type of reasoning is entirely reliable but in practice rules that are 100% certain (such as ‘all men are mortal’) are often not available.

This is a process of generalization based on repeated observations.

After many observations of *x* and *y* occurring together, learn the rule

**if** *x* **then** *y*

For example, if I see 1,000 dogs with four legs I might reasonably conclude that “if *x* is a dog then *x* has 4 legs” (or more simply “all dogs have four legs”).This is *induction*.