Show your working in all calculations.

1. This question is related to **Classification**.

   There are two closely related cultivars of the cassava plant *Manihot esculenta* that grow in South America: one that is edible without any special processing, and another that presents a high content of hydrogen cyanide (HCN), which can be lethal to humans if not properly treated before consumption. Laboratory testing is expensive, and you are trying to investigate whether it is possible to classify plants based on their physical characteristics.

   In the samples available to you, some characteristics of the plants have been measured, and each specimen has been identified as poisonous or not. This data is shown in the table below. Based on this information and your knowledge of data mining, answer the following questions.

Sample	RedBranches	Smooth	PinkSkin	Bitter	Poisonous
A	1	0	0	1	1
B	0	0	0	1	0
C	1	0	1	1	0
D	0	1	0	0	0
E	0	1	1	1	1
F	1	1	0	0	0
G	0	0	1	0	0
H	1	1	1	1	1

   a) Estimate the following **conditional probabilities**:

   - $P(RedBranches = 1 \mid Poisonous = 1)$
   - $P(RedBranches = 1 \mid Poisonous = 0)$
   - $P(Smooth = 1 \mid Poisonous = 1)$
   - $P(Smooth = 1 \mid Poisonous = 0)$
   - $P(PinkSkin = 0 \mid Poisonous = 1)$
   - $P(PinkSkin = 0 \mid Poisonous = 0)$
   - $P(Bitter = 1 \mid Poisonous = 1)$
   - $P(Bitter = 1 \mid Poisonous = 0)$

Model Answer:

- $P(RedBranches = 1 \mid Poisonous = 1) = 2/3 = 0.67$
- $P(RedBranches = 1 \mid Poisonous = 0) = 2/5 = 0.40$
- $P(Smooth = 1 \mid Poisonous = 1) = 2/3 = 0.66$
- $P(Smooth = 1 \mid Poisonous = 0) = 2/5 = 0.40$
- $P(PinkSkin = 0 \mid Poisonous = 1) = 1/3 = 0.33$
- $P(PinkSkin = 0 \mid Poisonous = 0) = 3/5 = 0.60$
- $P(Bitter = 1 \mid Poisonous = 1) = 3/3 = 1.00$
- $P(Bitter = 1 \mid Poisonous = 0) = 2/5 = 0.40$

   b) Write down the **Bayes' theorem** and explain briefly what it means.

Model Answer:

- *Let E denote evidence and H denote hypothesis. Then Bayes' theorem can be stated as:*

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}.$$

- *Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence. $P(H)$, the prior, is the initial degree of belief in H. $P(H|E)$, the posterior, is the degree of belief having accounted for E. The quotient $P(E|H)/P(E)$ represents the support E provides for H.*

c) Use the conditional probabilities estimated in item (a) to predict the value of ($Poisonous$) for a test sample $(RedBranches = 1; Smooth = 1; PinkSkin = 0; Bitter = 1)$ using the naïve Bayes approach.

Model Answer:

Let $E = (RedBranches = 1; Smooth = 1; PinkSkin = 0; Bitter = 1)$ be our evidence.
We need to compare $P(Poisonous = 1|E)$ against $P(Poisonous = 0|E)$, where:

$$P(Poisonous = X|E) = \frac{P(E|Poisonous = X) \times P(Poisonous = X)}{P(E)}$$

Since the denominator $P(E)$ is the same for both $Poisonous = 1$ and $Poisonous = 0$, it can be ignored in the comparisons and does not need to be computed.

***

The prior probabilities are:

$$P(Poisonous = 0) = 5/8 = 0.625$$
$$P(Poisonous = 1) = 3/8 = 0.375$$

***

Under the independence assumption of the Naïve Bayes classifier, we have that:

$$P(E|Poisonous = X) = P(RedBranches = 1|Poisonous = X) \times P(Smooth = 1|Poisonous = X) \times P(PinkSkin = 0|Poisonous = X) \times P(Bitter = 1|Poisonous = X)$$

This allows us to calculate the likelihoods as:

$$P(E|Poisonous = 0) = 2/5 \times 2/5 \times 3/5 \times 2/5 = 24/625 = 0.038$$
$$P(E|Poisonous = 1) = 2/3 \times 2/3 \times 1/3 \times 3/3 = 4/27 = 0.148$$

***

We can now calculate:

$$P(E|Poisonous = 0) \times P(Poisonous = 0) = 0.038 \times 0.625 = 0.024$$
$$P(E|Poisonous = 1) \times P(Poisonous = 1) = 0.148 \times 0.375 = 0.056$$

Given these results, a Naïve Bayes classifier would suggest that the new sample should receive a value of $Poisonous = 1$.

d) In the context of Bayesian classification, what is the problem that could arise from having an estimated (prior) conditional probability of zero? What could be done to prevent this problem?

Model Answer:

An estimated conditional probability of zero would mean that the posterior probability would never be able to receive any value different from zero, regardless of the evidence. This is a problem because rare events can have initial conditional estimates equal to zero, but it does not necessarily mean that they never occur, only that they have been so far unobserved.

A common approach to prevent this problem is to use the Laplace estimator, which adds 1 to all counts when calculating the probabilities table. In this way no estimated conditional probability is ever equal to zero, and the Bayesian approach can be able to change the support for certain hypotheses regarding such rare events as new evidence becomes available.

Show your working in all calculations.

2. This question is related to **Clustering**.

a) Describe briefly the main objective of Cluster Analysis.

Model Answer:

The objective of cluster analysis is to find groups of objects in a data set such that objects in a group will be similar (or related) to one another and different from (or unrelated to) objects in other groups.

b) Explain the difference between *partitional clustering* amd *hierarchical clustering* approaches.

Model Answer:

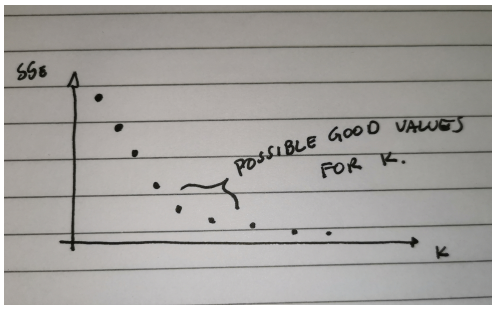
Partitional clustering *is the division of data objects into non- overlapping subsets (clusters), such that each data object is in exactly one subset* ***Hierarchical clustering*** *generates nested clusters, organised as a hierarchical tree, so that clusters can have sub-clusters and each node in the tree (except for the leaf nodes) is the union of its children nodes, with the root node being the cluster containing all objects*

c) Describe briefly and illustrate graphically the *elbow method* for selecting the number of clusters in k-means clustering.

Model Answer:

The elbow method is a very simple approach for selecting the value of k. It essentially consists of:

- *Iteratively increasing number of clusters*
- *Observe total SS_E - SS_E gains tend to flatten after a certain point*
- *Select a k with good tradeoff between model complexity (number of clusters) and SS_E*



d) Discuss the basic DIFFERENCE between the agglomerative and divisive hierarchical clustering algorithms. Explain also the difference between single (MIN) and complete (MAX) linkage in the context of hierarchical clustering.

Model Answer:

Agglomerative methods start with each object as and individual cluster and then incrementally build larger clusters by merging clusters with smallest dissimilarity Divisive methods, on the other hand, start with all points belonging to one cluster and then iteratively split the clusters In single linkage the similarity of two clusters is based on the two most similar (closest) points in the different clusters, while complete linkage, also known as MAX linkage, computes cluster similarity based on the two least similar (most distant) points in the different clusters

Show your working in all calculations.

3. This question is related to **Regression**.

   Consider the samples below, related to a regression problem.

x	y
1	2
3	5
5	6
7	9

   a) Estimate the linear regression line for these points based on the ordinary least squares formulas.[1]

Model Answer:

The calculation is straightforward. The means can be calculated as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = (1+3+5+7)/4 = 4$$
$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = (2+5+6+9)/4 = 5.5$$

The quantities required for the calculation of $\hat{\beta}_1$ are

x	*y*	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	*2*	*-3*	*-3.5*	*10.5*	*9*
3	*5*	*-1*	*-0.5*	*0.5*	*1*
5	*6*	*1*	*0.5*	*0.5*	*1*
7	*9*	*3*	*3.5*	*10.5*	*9*
			SUMS:	*22*	*20*

Based on these values, we can easily calculate

$$\hat{\beta}_1 = \frac{22}{20} = 1.1$$
$$\hat{\beta}_0 = 5.5 - 1.1 \times 4 = 1.1$$

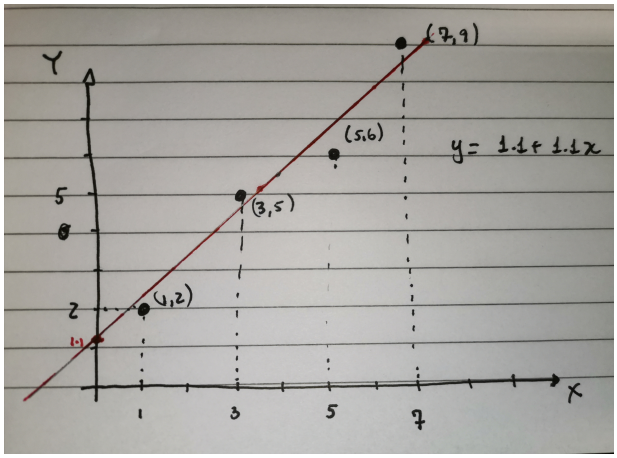
Which results in the linear model $y = 1.1 + 1.1x = 1.1(x+1)$.

[1] The least-squares estimators for the simple linear regression coefficients are calculated as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

b) Draw a scatter plot of the points given in the table, and add the regression line obtained above to your plot.

Model Answer:



c) Based on your model, what would be the expected value of y for $x = 20$? Please discuss your result.

Model Answer:

The model suggests $y = 1.1 + 20 \times 1.1 = 23.1$. This prediction may or may not be accurate, since the value of x considered lies considerably far from those used in the model fitting. If the linear relationship is expected to hold up to (or beyond) $x = 20$ (based, e.g., on domain-specific knowledge), then we could expect this prediction to be valid, otherwise it should be taken with a grain of salt.

Show your working in all calculations.

5. This question is related to **Anomaly/Outlier Detection** and **Data Properties and Preprocessing**

   a) Briefly describe TWO possible applications of outlier detection: explain what the application is, what an outlier would indicate in that context and why it would be important to detect.

Model Answer:

- *Fraud detection: outliers would consist of deviations from normal purchasing behaviour in a given dataset, and could indicate e.g. that a card is stolen. Important to prevent/minimise financial losses.*
- *Intrusion detection in networks: outliers could be unusual traffic, communication behaviour and other anomalies. Important for detecting system intrusions.*
- *Public health: detect, for instance, anomalous occurrences of certain diseases which can indicate low vaccination coverage or introduction of new strains. Important for evidence-guided policy making in public health.*
- *Ecosystem monitoring: unusual patterns, e.g., in species counts can suggest environmental change or some other disturbance in the system. Important for research, environmental protection, detection of newly introduced pests etc.*

   b) What is the best distance (or similarity) measure for each of the following applications?

      i) measure the dissimilarity between two dogs based on 7 numeric attributes.

      ii) compare similar diseases with a set of medical tests that show results as either positive or negative.

      iii) find similar documents in a plagiarism checking system.

Model Answer:

i) Observations composed of few numerical attributes - Euclidean distance

ii) Observations composed of binary vectors - Jaccard's coefficient seems the best one, assuming that only the positive results carry relevant information.

iii) After feature extraction (eg. tf-idf) the feature space is likely to be very large and very sparse. Cosine similarity is the most recommended.

   c) For the following two vectors, $p = [1, 1, 0, 0, 0, 0, 1, 0, 0, 0]$ and $q = [0, 1, 0, 0, 0, 0, 1, 0, 1, 0]$, compute the following similarities:

- Simple Matching Similarity
- Jaccard Similarity
- Cosine Similarity

Model Answer:

$$SMC = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}} = \frac{8}{10} = 0.8$$
$$JS = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{2}{4} = 0.5$$
$$CS = cos(p \angle q) = \frac{(p^T q)}{\|p\| \, \|q\|} = \frac{2}{1.73 \times 1.73} = 0.67$$

d) Explain the difference between **ordinal** and **continuous** attributes. Give TWO examples of each type of attribute.

Model Answer:

An ordinal attribute has a finite (and usually small) set of possible values on which a (total) order can be defined. Continuous variables are measured on a scale of real numbers. Exam grades and the Beaufort scale are ordinal variables; wind speed and the price of shares in BT are both continuous variables.

END OF EXAMINATION PAPER