**Show your working in all calculations.**

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

(6 marks)

*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.* (2 marks) *for definition;* (0.5 marks) *for each correct example (with order defined for ordinal attributes).*

b) Define what an **ROC curve** represents and the type of task that it can be applied to. Describe how you would compute the ROC curve for a logistic regression model.

(6 marks)

*Model Answer:*

*The ROC curve represents the true positive rate ($y$-axis) and false positive rate ($x$-axis) for all possible choices of a threshold in a classification model.* (2 marks) *. The ROC curve can be calculated for two-class classification tasks.* (2 marks) *. To compute the curve for a logistic regression model, its output (which is between 0 and 1, since it represents a probability) is thresholded at a value $p = 0$. The fractional true positive and false positive rates are computed. Then the value of $p$ is increased a little and the results for the model at this threshold are calculated.* (2 marks)

c) Consider models such as decision trees that do not provide accurate class probability estimates. What is the disadvantage of using models of this type in applications where there are costs?

(5 marks)

*Model Answer:*

*A model that does not output probabilities can only predict the class.* (1 mark) *This is equivalent to having a set of probabilities that is 1 for the winning class and 0 for the others.* (2 marks) *Hence the cost matrix has no effect on the prediction (in other words, the costs are ignored).* (2 marks)

d) For each of the following types of problem, describe briefly how it differs from a classification problem and give TWO example applications:

i) regression;

ii) outlier detection.

(8 marks)

*Model Answer:*

*Regression: The target variable is numeric instead of discrete; this means that the error function cannot be misclassification rate but is usually least squares.* (2 marks) *Example applications include predicting financial markets, predicting missing numeric data, weather forecasting etc.* (2 marks)

*Outlier detection. This is an unsupervised problem rather than a supervised one (like classification); often a probability density model is used to model the probability of normal events.* (2 marks) *Example applications include machine condition monitoring and detecting outbreaks of infectious diseases.* (2 marks)

2. This question is concerned with **decision trees**. Consider the following dataset which relates to whether a person plays golf based on the weather conditions.

| Temperature | Humidity | Windy | Play |
|---|---|---|---|
| hot | high | false | no |
| hot | high | true | no |
| hot | high | false | yes |
| mild | high | false | yes |
| cool | normal | false | yes |
| cool | normal | true | no |
| cool | normal | true | yes |
| mild | high | false | no |
| cool | normal | false | yes |
| mild | normal | false | yes |
| mild | normal | true | yes |
| mild | high | true | yes |
| hot | normal | false | yes |
| mild | high | true | no |

a) Write down the equation for the **entropy** $E(p_1, \ldots, p_k)$ of a probability distribution $(p_1, \ldots, p_k)$.

(2 marks)

*Model Answer:*

*The entropy of the probability distribution* $(p_1, \ldots, p_k)$ *is given by the formula*

$$E(p_1, \ldots, p_k) = -\sum_{i=1}^{k} p_i \log_2 p_i.$$

(2 marks)

b) Compute the **information gain** of splitting the data objects using the '*Temperature*' and '*Humidity*' attributes respectively for the weather dataset assuming that '*Play*' represents the class. Show the details of your working. Which attribute of the two is the better one to select for a decision tree?

(14 marks)

*Model Answer:*

*The entropy before any splits is* $E(5/14, 9/14) \approx 0.9403$.

*The Temperature variable has three values: hot, mild, and cool. The value hot occurs 4 times in the dataset, 2 of which have "play=yes" and 2 of which have "play=no". This explains the* $[2, 2]$ *branch. The value mild occurs six times, 4 of which have "play=yes" and 2 have "play=no", which gives rise to the* $[2, 4]$ *branch. Finally, cool occurs four times, once with "play=no" and three times with "play=yes", which gives rise to the* $[1, 3]$ *branch.*

*We can compute the information content of the* $[2, 2]$ *branch as follows:*

$$\begin{aligned} \textit{info}([2,2]) = E(1/2, 1/2) &= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) \\ &= -\frac{1}{2} \times (-1) - \frac{1}{2} \times (-1) = 1. \end{aligned}$$

*The information content of the* $[2, 4]$ *branch is computed by*

$$\begin{aligned} \textit{info}([2,4]) = E(2/6, 4/6) &= -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) \\ &\approx 0.5823 + 0.39 = 0.9183. \end{aligned}$$

*Finally, the information content of the* $[1, 3]$ *branch is computed by*

$$\begin{aligned} \textit{info}([1,3]) = E(1/4, 3/4) &= -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) \\ &\approx 0.5 + 0.3113 = 0.8113. \end{aligned}$$

*To compute the information of the Temperature attribute, we form a weighted sum of these entropy terms, where the weights are given by the fraction of examples that satisfy the corresponding branches: 4/14 for hot, 6/14 for mile, and 4/14 for cool. Thus the information of Temperature is*

$$\begin{aligned} \textit{info}([2,2],[2,4],[1,3]) &= (4/14) \times 1 + (6/14) \times 0.9183 + (4/14) \times 0.8113 \\ &\approx 0.9111, \end{aligned}$$

*which corresponds to an information gain of approximately* $0.9403 - 0.9111 \approx 0.0292$ *bits.*

*By a similar calculation, the information of Humidity is*

$$\begin{aligned} \textit{info}([4,3],[1,6]) &= (1/2) \times 0.9852 + (1/2) \times 0.5917 \\ &\approx 0.7884, \end{aligned}$$

*which corresponds to an information gain of approximately* $0.9403 - 0.7884 \approx 0.1519$ *bits. Thus Humidity is the better attribute to choose since it gives the greater information gain.*

( (5 marks) for each entropy calculation (so (10 marks) in total for this aspect), (2 marks) for the pair of information calculations, and (2 marks) for final choice (based on whatever calculations the student has achieved).)

---

c) What is the purpose of pruning a decision tree? Explain what is meant by each of **pre-pruning** and **post-pruning** of a decision tree. What is the main advantage of post-pruning over pre-pruning?

(9 marks)

---

*Model Answer:*

*The purpose of pruning is to remove those parts of the true that are just modelling noise and hence to improve generalisation* (2 marks) *. Pre-pruning of a decision tree takes place during tree construction: a condition, which has been selected using information gain or some similar measure, is only added to the tree if the split it creates is statistically significant* (2 marks) *. Post-pruning of a decision tree is the process of removing branches or other structures from the tree that are not statistically significant at some determined significance level* (2 marks) *. The main advantage of post-pruning over pre-pruning*

*is that post-pruning allows the tree to create complex structure before determining its significance. Pre-pruning may not allow complex conditions on multiple attributes to be formed since combinations of fewer attributes may never be deemed signficant* (3 marks) .

3. This question is concerned with the **Naïve Bayes** classifier.

a) Write down the **Bayes' theorem** and explain briefly what it means. (5 marks)

*Model Answer:*

*Let $E$ denote evidence and $H$ denote hypothesis, the Bayes' theorem is:*

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

(3 marks)

*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 $H$.* (2 marks)

b) Refer to the data contained in the following table:

| **Magazine Promotion** | **Watch Promotion** | **Credit Card Insurance** | **Sex** | **Life Insurance Promotion** |
|---|---|---|---|---|
| Yes | No | No | Male | No |
| Yes | Yes | Yes | Female | Yes |
| No | No | No | Male | Yes |
| Yes | Yes | Yes | Male | Yes |
| Yes | No | No | Female | Yes |
| No | No | No | Female | No |
| Yes | Yes | Yes | Male | Yes |
| No | No | No | Male | No |
| Yes | No | Yes | Male | No |
| Yes | Yes | No | Female | No |

Copy the table presented below to your answer book. Fill in the counts and probabilities in the table. The output attribute is '*life insurance promotion*'.

(5 marks)

| | Magazine Promotion | | Watch Promotion | | Credit Card Insurance | | Sex | |
|---|---|---|---|---|---|---|---|---|
| Life Insurance Promotion | Yes | No | Yes | No | Yes | No | Male | Female |
| Counts | | | | | | | | |
| Yes | | | | | | | | |
| No | | | | | | | | |
| Probabilities | | | | | | | | |
| Yes | | | | | | | | |
| No | | | | | | | | |

*Model Answer:*

| | *Magazine Promotion* | | *Watch Promotion* | | *Credit Card Insurance* | | *Sex* | |
|---|---|---|---|---|---|---|---|---|
| *Life Insurance Promotion* | *Yes* | *No* | *Yes* | *No* | *Yes* | *No* | *Male* | *Female* |
| *Counts* | | | | | | | | |
| *Yes* | *4* | *1* | *3* | *2* | *3* | *2* | *3* | *2* |
| *No* | *3* | *2* | *1* | *4* | *1* | *4* | *3* | *2* |
| *Probabilities* | | | | | | | | |
| *Yes* | *4/5* | *1/5* | *3/5* | *2/5* | *3/5* | *2/5* | *3/5* | *2/5* |
| *No* | *3/5* | *2/5* | *1/5* | *4/5* | *1/5* | *4/5* | *3/5* | *2/5* |

*(question continues on next page. . . )*

*(Question 3 continued. . . )*

c) Use the completed table in Part (b) together with the Naïve Bayes classifier to determine the value of life insurance promotion for the following instance:

Magazine Promotion = Yes
Watch Promotion = Yes
Credit Card Insurance = Yes
Sex = Female
Life Insurance Promotion = ?

(9 marks)

*Model Answer:*

*Essentially compare* $P(L = yes|M = yes, W = yes, C = yes, S = female)$ *and* $P(L = no|M = yes, W = yes, C = yes, S = female)$

$$P(L = yes|M = yes, W = yes, C = yes, S = female) = \frac{P(M = yes, W = yes, C = yes, S = female|L = yes) \times P(L = yes)}{P(M = yes, W = yes, C = yes, S = female)} \quad (1)$$

*Since the denominator* $P(M = yes, W = yes, C = yes, S = female)$ *is the same for* $P(L = yes|Atributes)$ *and* $P(L = no|Atributes)$*, it can be ignored.*

(2 marks)

$$\begin{aligned} &P(M = yes, W = yes, C = yes, S = female|L = yes) \times P(L = yes) \\ &= P(M = yes|L = yes) \times P(W = yes|L = yes) \times P(C = yes|L = yes) \times P(S = female|L = yes) \\ &\quad \times P(L = yes) \\ &= (4/5) \times (3/5) \times (3/5) \times (2/5) \times (1/2) = 0.0576 \end{aligned}$$

(3 marks)

*Similarly,*

$$\begin{aligned} &P(M = yes, W = yes, C = yes, S = female|L = no) \times P(L = no) \\ &= P(M = yes|L = no) \times P(W = yes|L = no) \times P(C = yes|L = no) \times P(S = female|L = no) \\ &\quad \times P(L = no) \\ &= (3/5) \times (1/5) \times (1/5) \times (2/5) \times (1/2) = 0.0048 \end{aligned}$$

(3 marks)

$P(L = yes|M = yes, W = yes, C = yes, S = female) > P(L = no|M = yes, W = yes, C = yes, S = female)$*, therefore, life insurance promotion is YES.* (1 mark)

d) Repeat Part (c), but assume that the gender of the customer is unknown.

(4 marks)

*Model Answer:*

$$P(M = yes, W = yes, C = yes|L = yes) \times P(L = yes)$$
$$= P(M = yes|L = yes) \times P(W = yes|L = yes) \times P(C = yes|L = yes) \times P(L = yes)$$
$$= (4/5) \times (3/5) \times (3/5) \times (1/2) = 0.144$$

(2 marks)

$$P(M = yes, W = yes, C = yes|L = no) \times P(L = no)$$
$$= P(M = yes|L = no) \times P(W = yes|L = no) \times P(C = yes|L = no) \times P(L = no)$$
$$= (3/5) \times (1/5) \times (1/5) \times (1/2) = 0.012$$

$P(L = yes|M = yes, W = yes, C = yes) > P(L = no|M = yes, W = yes, C = yes)$, *therefore, life insurance promotion is YES.* (2 marks)

e) Comment on the results obtained in Part (c) and Part (d). (2 marks)

*Model Answer:*

*Naïve Bayes assumes that attributes are independent. Hence, if some attribute is unknown, we simply omit this attribute from the calculation of posterior probabilities of class given attributes. In the example here,* $P(S = Female|L = Yes)$ *and* $P(S = Female|L = No)$ *are the same. Therefore, we still get the same classification result (Life Insurance Promotion = Yes) for the given instance regardless whether the gender of the customer is known or not.*

4. a) For the following **cost matrix**:

| Actual \ Predicted | a | b | c |
|---|---|---|---|
| a | 0 | 10 | 5 |
| b | 50 | 0 | 30 |
| c | 20 | 10 | 0 |

If a classification model predicts a with probability 0.4, b with probability 0.3 and c with probability 0.3, what is the optimal (i.e. **lowest cost**) decision? (7 marks)

*Model Answer:*

*The cost matrix gives the cost of each decision that is made; each row gives the cost of classifying an example that belongs to a given class to each of the other classes.*

*In the example, we assume that the predicted probabilities are correct.*

*Classifying the example as Class 'a' has an expected cost of* $50 \times 0.3 + 20 \times 0.3 = 21$. (2 marks)

*Classifying the example as Class 'b' has an expected cost of* $10 \times 0.4 + 10 \times 0.3 = 7$. (2 marks)

*Classifying the example as Class 'c' has an expected cost of* $5 \times 0.4 + 30 \times 0.3 = 11$. (2 marks)

*Thus the lowest cost decision is to classify the example as Class 'b' even though the class with the highest probability is 'a'.* (1 mark)

b) Compare and contrast **K-means** and **agglomerative clustering**. Make a list of similarities and differences between the two approaches. (8 marks)

*Model Answer:*

*Similarities* (2 marks)

- *Both are used for unsupervised clustering.*
- *Both use similarity measurement to determine whether to join 2 objects or 2 clusters.*

*Differences*
*K-means (any three points below)* (3 marks)

- *K-means is a partitioning method where it aims to construct a partition of n objects into a set of k clusters.*
- *It computes seed points as the centroids (mean points) of the clusters of the current partition and reassign each object to the cluster with the nearest seed point. This procedure repeats until no more new assignment can be found.*
- *K-Means does not guarantee to find global minimum SSE. Instead, it finds local minimum.*
- *Invoking algorithm using variety of initial cluster centers improves probability of achieving global minimum*
- *Potential problem for applying k-Means is that analyst needs to have a priori knowledge of k.*

*Agglomerative clustering* (3 marks)

- *Agglomerative clustering creates clusters through recursive combining existing clusters.*
- *Each object is initialised to become own cluster. At each iteration two closest clusters aggregated together. Eventually, all records combined into single cluster.*
- *This method does not require the number of clusters k as an input, but needs a termination condition.*

c) Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item $\{a\}$ is 25%, the support for item $\{b\}$ is 90% and the support for itemset $\{a, b\}$ is 20%. What is the **confidence** of the association rule $\{a\} \rightarrow \{b\}$?

(3 marks)

*Model Answer:*

$$\textit{confidence}(\{a\} \rightarrow \{b\}) = \frac{\textit{Frequency of the itemset}\{a,b\}}{\textit{Frequency of the itemset}\ \{a\}} = \frac{0.2}{0.25} = 0.8 \quad (2)$$

(3 marks)

d) A doctor can run a test for the horrible disease *Examophobia*. The test has two possible outcomes: positive and negative.

It is known that among all students, if Examophobia is present, the test comes out positive 80% of the time, and negative 20% of the time. If Examophobia is not present, the test comes out positive 1% of the time, negative 99%.

Among the general student population, Examophobia is known to occur in 35% of all students.

A student enters the clinic and tests positive for the disease. What is the probability they really have Examophobia?

(7 marks)

*Model Answer:*

*We have the following information:*

- *if Examophobia is present, the test comes out positive 80% of the time, and negative 20% of the time; i.e.,* $p(positive|Examophobia) = 0.8$, $p(negative|Examophobia) = 0.2$ (1 mark)
- *If Examophobia is not present, the test comes out positive 1% of the time, negative 99%, i.e.,* $p(positive|not\ Examophobia) = 0.01$, $p(negative|not\ Examophobia) = 0.99$ (1 mark)
- *Examophobia is known to occur in 35% of all students, i.e.,* $p(Examophobia) = 0.35$ *and* $p(not\ Examophobia) = 0.65$ (1 mark)

*We need to calculate the probability of a student really has Examophobia if he was tested positive for the disease, i.e.,* $p(Examophobia|positive)$.

*Based on the Bayes Theorem,*

$$
\begin{aligned}
&p(Examophobia|positive) \\
&= \frac{p(positive|Examophobia) \times p(Examophobia)}{p(positive)} \\
&= \frac{p(positive|Examophobia) \times p(Examophobia)}{p(positive|Examophobia) \times p(Examophobia) + p(positive|not\ Examophobia) \times p(not\ Examophobia)} \\
&= \frac{0.8 \times 0.35}{0.8 \times 0.35 + 0.01 \times 0.65} = 0.977
\end{aligned}
$$

(4 marks)