Questions & Answers (ICT 2021)

Part A

What are the different types of Learning/Training models in Machine Learning (ML)? Briefly explain their principles (4 marks)

Indicative answer: ML algorithms can be primarily classified depending on the presence/absence of desired output/s. Supervised learning: The machine learns using labelled data. The model is trained on an existing data set before it starts making decisions with the new data. The desired output can be either continuous or categorical. Unsupervised learning: The machine is trained on unlabelled data and without any proper guidance (i.e. desired output). It automatically infers patterns and relationships in the data (through some type of similarity (by creating clusters). Reinforcement Learning: The model learns through a trial and error method. This kind of learning involves an "agent" that will interact with the environment to create actions and then discover errors or rewards of that action. ***(1 marks for the definition of types and 3 marks for the principles)***

What is the essential difference between classification and clustering? (4 marks)

Indicative answer: Classification is a supervised learning task, which needs training examples consisting of input data and their corresponding class label(s). As a result, the supervised learning process leads to a rule for decision making; i.e., given an unseen input, the classifier predicts its class label. In contrast, clustering is an unsupervised learning task where there are only input data available without knowing their class labels. As a result, the unsupervised learning process would discover the intrinsic structure underlying the input data. ***(2 marks for classification and 2 marks for clustering)***

Summarise the strength(s) and the weakness(es) of K-means clustering. (4 marks)

Indicative answer: The strength of K-means algorithm lies in its computational efficiency and the nature of easy-to-use. In contrast, there are a number of weaknesses: a) requiring the prior knowledge of cluster numbers, *K*, b) sensitive to initialisation, which leads to unwanted solutions, c) sensitive to outliers and noise, which results in an inaccurate partition, and d) inapplicable to categorical data.***(2 marks for strength(s), 2 marks for weakness(es))***

Suppose you are training a decision tree on a dataset that contains k binary features (i.e. attributes). The dataset contains a very large number of examples/samples (N ≫ k). What is the maximum possible depth of the decision tree? How likely is to have such situation in real decision trees cases? Justify your response. (4 marks)

Indicative answer: The maximum depth is k. Each attribute can be used once at most to define a split. In reality, either due to pre-processing analysis (input variables selection) or due to the specific criterion (i.e. entropy for example), not necessary all attributes will be utilised. ***(2 marks for the answer, 2 marks for the justification)***

You are training a Multilayer Perceptron (MLP) neural network for a particular classification task. After, some investigation, your neural network is constructed with 5 input variables, one hidden layer with 12 nodes and one output layer with 3 nodes (the classes). How many network parameters are required to be tuned/trained? Show your detailed calculations. (4 marks)

Indicative answer: (5+1) x 12 + (12+1) x 3 = 111. We have to include also the bias "weights" which is a compulsory component in MLP structure.

Briefly describe the general objective of Association Rules mining. What is the "Apriori Principle"? (4 marks)

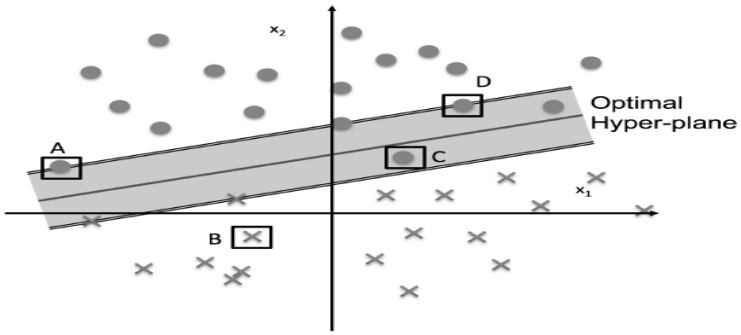
Indicative answer: The objective of association rules mining is to discover interesting relations between objects in large databases. A brute force approach to find frequent itemsets is to form all possible itemsets and check the support value of each one of these. *Apriori principle* helps in making this search efficient. It says

that "Any subset of a frequent itemset must be also frequent", which is an anti-monotone property. ***(2 marks for the objective, 2 marks for the principle)***

The ethics of how a Machine Learning system is to function is a common thought that arises when we read about all these advancements in this domain. To build however an ML system, we need, among others, lots of data. Unfortunately, the selection/utilisation of data for using it in our ML system generates some ethical and biased issues. Briefly describe some of them. (6 marks)

Indicative answer: The first question is where are we going to get this data? Data could be open-sourced and free. It might be publically available for a price. Or the data might be privately owned by a group of people. Can we use it, without authorisation from its owners? The answer becomes even more ambiguous when we talk about tracking people anonymously, without their consent, to collect data, or even utilise datasets where personal/private information is included. What we do in such cases? Are we allowed to use such information or try to remove/replace it with suitable alternatives ways? Even, if we sorted-out such "ethical" issues, what about the quality of this data? One particular problem in data is that sometimes the sampling quality of this data is not the appropriate one. Describe for example the issue of unbalanced datasets; a characteristic problem occurs in classification problems ***(3 marks for the ethics, 3 marks for the bias)***

The figure below displays the training samples and the learned SVM hyperplane. Which of the four highlighted samples is NOT considered as a support vector? Justify your response. (4 marks)



Indicative answer: Obviously the sample B. Support vector elements are those samples used to create this "region" that separates the two classes. So these elements practically, define the borders on that region. However, in the case of soft-margin case (opposite to the hard-margin case), we allow some samples (like C) which are inside the boundary in order to avoid an overfitting issue.

Briefly describe some of the advantages and disadvantages of Principal Component Analysis (PCA), a classic method used for data analysis. (4 marks)

Indicative answer: Advantages of PCA. ***Removes Correlated Features:*** In a real-world scenario, this is very common that we get many attributes/features in our dataset. We cannot run ours algorithm on all the features as it will reduce the performance of our algorithm. We need to find out the correlation among the features (correlated variables). Finding correlation manually in hundreds/ thousands of attributes is nearly impossible, frustrating and time-consuming. PCA does this very efficiently. ***Improves Algorithm Performance:*** With a large number of attributes, the performance of our algorithm will be drastically degraded. PCA is a very common way to speed up our ML algorithm by getting rid of correlated variables which don't contribute in any decision making. ***Improves Visualization:*** It is rather difficult to visualize and understand the data in high dimensions. PCA transforms a high dimensional data to low dimensional data (say in 2 dimensions) so that it can be visualized easily. **Disadvantages of PCA**. ***Independent variables become less interpretable*****:** After implementing PCA on the dataset, our original features will turn into Principal Components (PCs). PCs are the linear combination of our original features and are not as readable and interpretable as original features. ***Data standardization is must before PCA:*** We need standardize our data before implementing PCA; otherwise PCA will not be able to find the optimal PCs. For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, PCs will be biased towards features with high variance, leading to false results. ***Information Loss:*** Although PCs try to

cover maximum variance among the features in a dataset, if we don't select the number of PCs with care, it may miss some information as compared to the original list of features. (***2 marks for advan. 2 marks for disadv)***

Consider the following confusion table summarising the testing results for iris classification. As, you are aware, the iris data is a classic multi-class benchmark dataset (12 marks)

Confusion Matrix for IRIS dataset		Actual Class		
		Setosa	Versicolor	Virginica
Predicted Class	Setosa	20	0	0
	Versicolor	1	1	1
	Virginica	0	4	16

- **What is the overall classification accuracy? (2 marks)**
- **What is the sensitivity and specificity for each class? (6 marks)**
- **Use the table as an example to explain why confusion matrix is a better way to assess the performance of a classifier than the overall classification accuracy (4 marks)**

Indicative answer:
The overall classification accuracy is: (20+1+16)/(20+1+1+4+1+16) = 86.046%
As this is a multi-class problem, we need to decompose it, so to create "individual" confusion matrices (CM) for each, just to make easier the understanding of the concept. The following CM shows the details.

		Actual Class	
		Yes	Not
Predicted Class	Yes	TP	FP
	Not	FN	TN

Confusion Matrix for IRIS dataset		Actual Class	
		Setosa	Not
Predicted Class	Setosa	20	0
	Not	1	22

Confusion Matrix for IRIS dataset		Actual Class	
		Versicolor	Not
Predicted Class	Versicolor	1	2
	Not	4	36

Confusion Matrix for IRIS dataset		Actual Class	
		Virginica	Not
Predicted Class	Virginica	16	4
	Not	1	22

Sensitivity: TP/(TP+FN)
Specificity: TN/(TN+FP)

For setosa: Sensitivity: 20/21 = 95.23%, specificity: 22/22 = 100%
For Versicolor: Sensitivity: 1/5= 20%, specificity: 36/38 = 94.73%
For Virginica: Sensitivity: 16/17 = 94.11%, specificity: 22/26 = 84.61%

Classification accuracy can be unreliable when assessing unbalanced data. In the above example, the overall classification accuracy of 86% shows the classifier does well in general. It does not reflect the poor performance when classifying the Versicolor class, which has small sample number. The confusion matrix shows a rather poor performance of classifying the Versicolour class, which also leads to low sensitivity.

Part B

Question B-1

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. Association Rules are widely used to analyse retail basket or transaction data. You have been given the following transaction database that consists of items (a, b, c, d &e) bought in a store by customers.

TID	Items
1	a,b,d,e
2	b,c,d
3	a,b,d,e
4	a,c,d,e
5	b,c,d,e
6	b,d,e
7	c,d
8	a,b,c
9	b,d
10	a,d,e

Find all the closed frequent itemsets which are not maximal, along with their support, for a minsupp threshold of 0.3. **Procedure:** Define first, all the frequent itemsets (10 marks), then all the closed frequent itemsets (10 marks) and finally all the closed frequent itemsets which are not maximal (10 marks). Show all steps/results of your work and justify any decision you have taken in your analysis.

Marks: 30

Indicative answer:

We check first the 1-size itemsets

a 5/10 = 0.5
b 0.7
c 0.5
d 0.9
e 0.6

Everything is above min support so they are candidates for the creation of 2-size itemsets

ab 0.3
ac 0.2 reject, due to less 0.3
ad 0.4
ae 0.4
bc 0.3
bd 0.6
be 0.4
cd 0.4
ce 0.2 reject, due to less 0.3
de 0.6

8 2-itemsets are suitable; they will create the 3-size itemsets in the next stage

Abc reject as it contains ac
Abd
Abe
Acd reject due to ac
Ace reject due to ac
Ade

Bcd
Bce reject due to ce
Bde
Cde reject due to ce

Thus, only 5 3-itemsets are suitable candidates. We need to calculate their support

Abd 0.2 reject, due to less 0.3
Abe 0.2 reject, due to less 0.3
Ade 0.4
Bcd 0.2 reject, due to less 0.3
Bde 0.4

These 2 3-itemsets are ok for the creation of 4-itemsets

Abcd reject as it includes ac
Abce reject as it includes bc
Abde reject as it includes abd
Acde 0.1 reject, less 0.3
Acde 0.1 reject, less 0.3
So, finally we have 5 1-itemsets, 8 2-itemsets and 2 3-itemsets (all frequent)

(10 marks for frequent itermsets: 3 marks for correct lists of candidates in various stages, 3 marks for current lists of final chosen ones in various stages and 4 marks for correct list of removed itemsets with a justification)

In summary, these are the following frequent itemsets

	items	support
[1]	{c}	0.5
[2]	{a}	0.5
[3]	{e}	0.6
[4]	{b}	0.7
[5]	{d}	0.9
[6]	{b,c}	0.3
[7]	{c,d}	0.4
[8]	{a,e}	0.4
[9]	{a,b}	0.3
[10]	{a,d}	0.4
[11]	{b,e}	0.4
[12]	{d,e}	0.6
[13]	{b,d}	0.6
[14]	{a,d,e}	0.4
[15]	{b,d,e}	0.4

Based on definitions:

- **Closed Frequent Itemset:** An itemset is closed if none of its immediate supersets has the same support as that of the itemset.
- **Maximal frequent itemset:** The definition says that an itemset is maximal frequent if none of its immediate supersets is frequent

Let's try to find the closed frequent itemsets

items		**support**
[1]	{c}	0.5
[2]	{a}	0.5
[3]	{b}	0.7

[4] {d} 0.9
[5] {b,c} 0.3
[6] {c,d} 0.4
[7] {a,b} 0.3
[8] {d,e} 0.6
[9] {b,d} 0.6
[10] {a,d,e} 0.4
[11] {b,d,e} 0.4

There are 4 itemsets from the frequent list (e, ae, ad, be) that are not closed, so they have been omitted from the closed list. So, "e" for example has been excluded as "de" has also 0.6. "ae" the same due to "ade", "ad" due to "ade" and "be" is also omitted due to "bde" (the same support).

(10 marks: 5 marks for correct final lists of closed ones and 5 marks for justification, especially those which have been rejected)

But we are interested in those closed itemsets which are not maximal ones. Let's try to find those which are maximal ones and remove them for the closed list. Then, the remaining ones will be the required answer.

The maximal ones are:

items support
[1] {b,c} 0.3
[2] {c,d} 0.4
[3] {a,b} 0.3
[4] {a,d,e} 0.4
[5] {b,d,e} 0.4

Thus, if we remove them from the above list of closed ones (11 itemsets), we have the final result.

[1] {c} 0.5
[2] {a} 0.5
[3] {b} 0.7
[4] {d} 0.9
[8] {d,e} 0.6
[9] {b,d} 0.6

(10 marks: 5 marks for finding the correct closed maximal ones and 5 marks for the correct closed non-maximal ones)

Question B-2

Complete the tasks based on provided observation table.

row	Weather in Athens	Mood
1	Overcast	good
2	Rain	good
3	Rain	bad
4	Sunny	good
5	Sunny	good
6	Overcast	bad
7	Overcast	good
8	Rain	bad
9	Sunny	good

10	Overcast	bad
11	Sunny	bad
12	Rain	good
13	Sunny	good
14	Sunny	bad
15	Rain	bad
16	Rain	bad

Frequency Table			
Class Weather	**Good**	**Bad**	
Overcast			4/16=0.25
Rain			
Sunny			
Total	8		
	8/16=0.5		

Likelihood Table – P(weather| mood)		
Class Weather	**Good**	**Bad**
Overcast		
Rain		
Sunny		

- **Complete the Frequency and Likelihood tables (5 marks for each table)**
- **What is Naive Bayes Prediction for the mood when weather is Rain and Sunny respectively? (5 marks for each case)**

Perform and show all calculations.

Marks: 20

Indicative answer:

Frequency Table			
Class Weather	**Good**	**Bad**	
Overcast	2	2	4/16=0.25
Rain	2	4	6/16=0.375
Sunny	4	2	6/16=0.375
Total	8	8	
	8/16=0.5	8/16=0.5	

Likelihood Table – P(weather| mood)		
Class Weather	**Good**	**Bad**
Overcast	2/8=0.25	2/8=0.25
Rain	2/8=0.25	4/8=0.5
Sunny	4/8=0.5	2/8=0.25

For simplicity, we write P(Good|Rain) instead of P(Mood=good | weather= rain)

The naïve bayes formula tells us:

P(mood|weather) = P(weather|mood) x P(mood) / P(weather)

P(weather) could be omitted as exists in both sides of the (in)equality. Since the value is positive it would be omitted by multiplying to the both sides of the (in) equality without changing the (in)equality

For the rain case:

P(rain|good) x P(good) = 0.25 x 0.5 = 0.125 < P(rain|bad) x P(bad) = 0.5 x0.5 = 0.25

P(weather) could be omitted as exists in both sides of the (in)equality. Since the value is positive it would be omitted by multiplying to the both sides of the (in)equality without changing the (in)equality. Thus, when it is rain, it is more likely to have bad mood (take the higher value)

For the sunny case:

P(sunny|good) x P(good) = 0.5 x 0.5 = 0.25 > P(sunny|bad) x P(bad) = 0.25 x0.5 = 0.125

Thus, when it is sunny, it is more likely to have good mood.