COEN413 (MACHINE LEARNING) 2021/2022 EXAMINATIONS
MODEL SOLUTION AND MARKING SCHEME

QUESTION ONE

1a) Give the Arthur L. Samuel's definition of machine learning.

"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"

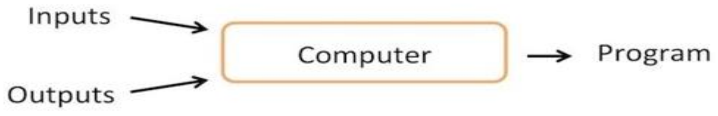
1b) List the six branches of AI

I. Machine learning II. Neural Networks III. Expert systems IV. Robotics V. Fuzzy logic VI. Natural language processing.

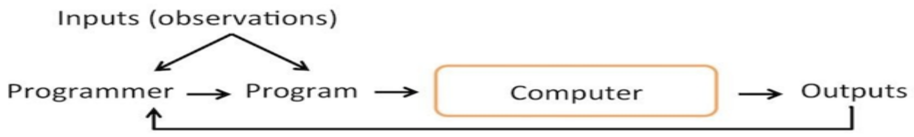
1c) With the aid of block diagram differentiate between normal computer software and machine learning

The difference between normal computer software and machine learning is that a human developer hasn't given codes that instructs the system how to react to situation, instead it is being trained by a large number of data.

Machine Learning



The Traditional Programming Paradigm



1d) What are the steps machine learning follows while building a machine learning model?

I. Gathering data
II. Preparing the data
III. Choosing a model
IV. Training
V. Evaluation
VI. Hyperparameter tuning
VII. Prediction

1e) Why do data be cleaned in machine learning process?

The data collected needs to be cleaned by removing duplicates, incorrect, incomplete, or corrupted data. It needs to be manipulated by data transformation, normalization or standardization, handling missing data and outliers, joining data, feature transformation, and much more. Most commonly used machine learning algorithms require data to be numerical. It is, therefore, necessary to transform any non-numeric features using feature transformation.

QUESTION TWO

2a) Differentiate between Semi-supervised learning & weakly supervised learning

It was found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce a considerable improvement in learning accuracy. **In Semi-supervised learning,** training data is a combination of both labeled and unlabeled data. However, labeled data exists with a very small amount while it consists of a huge amount of unlabeled data. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labeled data. It is why label data is a comparatively, more expensive acquisition than unlabeled data.
In weakly supervised learning, the training labels are noisy, limited, or imprecise; however, these labels are often cheaper to obtain, resulting in larger effective training sets.

2b) Differentiate between self-supervised learning and unsupervised learning

Self-supervised learning is similar to unsupervised learning because both work with data without human added labels.
The difference is that unsupervised learning uses clustering, grouping, and dimensionality reduction, while self-supervised learning draws its own conclusions for regression and classification tasks.

2c) Write five uses of machine learning

Any five.

- Traffic prediction
- Virtual Personal Assistant
- Speech recognition
- Email spam and malware filtering
- Bioinformatics
- Natural Language processing
- Online Transportation
- Social Media Services
- Product Recommendation
- Online Fraud Detection

2d) Write five advantages of machine learning

Any five

a. Fast, Accurate, Efficient.
b. Automation of most applications.
c. Wide range of real-life applications.
d. Enhanced cyber security and spam detection.
e. No human Intervention is needed.
f. Handling multi-dimensional data.

2e) Define data mining and list five types of data

Data mining is the use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data.

The types of data are:

- Numeric data: Each object is a point in a multidimensional space
- Categorical data: Each object is a vector of categorical values
- Set data: Each object is a set of values (with or without counts)
  - Sets can also be represented as binary vectors, or vectors of counts
- Ordered sequences: Each object is an ordered sequence of values.
- Graph data: Web graph and HTML Links

QUESTION THREE

3a) Mention four application areas of artificial intelligence and describe three main branches of machine learning and provide one example for each of the branches **[6 marks]**

The four application areas of machine learning include:

- Computer vision: image segmentation, object detection, object recognition
- Natural Language Processing; ChatBot and Conversational Bots
- Robotics; object grasping
- Recommendation System (clustering techniques)
- Anomaly detection
- Temporal Prediction (Time-Series)
- Big Data Analytics
- Gaming

Each correct point takes 0.5 mark (select only 4 = 2 marks)
Branches of machine learning can be categorized into the following: [2.5 marks]

i. **Supervised Learning** is a branch of machine learning that involves building a model from a labeled data: example of supervised learning algorithm; linear regression, logistic regression, KNN, random forest, support vector machines, decision trees, ensemble learning methods, etc. **[1.0 mark]**

ii. **Unsupervised Learning** is a branch of machine learning that involves building a model from unlabeled data with a central goal to find the underlying patterns within the data: example of unsupervised learning algorithm include: clustering (K-means clustering, hierarchal clustering, spectral clustering), Principal Component Analysis, Linear Discriminate Analysis, Factor Analysis, Autoencoder, Generative Adversarial Networks (GAN), etc**. [1.5 mark]**

iii. **Reinforcement Learning** is a branch of machine learning that involves the interaction between an agent and its environment with a central goal to optimize a long-term reward. Example of reinforcement learning algorithm include: Q-learning, SARSA, DQN, DDPG, A3C, A2C, SAC **[1.5 mark]**

3b) Given the following learning examples with input attributes X1 and X2 and desired target attribute D:
Suppose a Gaussian density is used and we have the following data:

Data No.	Size	Gender
1	180	male
2	190	male
3	170	female
4	200	male
5	150	female
6	160	male
7	165	female
8	180	female

i. Compute the mean and variance of both classes. [**4 marks]**

$\mu_M = \frac{180+190+200+160}{4} = 182.5$ **1 mark**

$\mu_F = \frac{170+150+165+180}{4} = 166.25$ **1 mark**

$\sigma_M^2 = \frac{(180-182.5)^2+(190-182.5)^2+(200-182.5)^2+(160-182.5)^2}{4} = 218.75$ **1mark**

$\sigma_F^2 = \frac{(170-166.25)^2+(150-166.25)^2+(165-166.25)^2+(180-166.25)^2}{4} = 117.19$ **1mark**

ii. Suppose an unclassified person has a length of 175 cm. What is the most probable gender of this person according to our model and what is the probability of this prediction? **[5marks]**

$$P(C_i/x) = \frac{P(x/C_i)P(C_i)}{P(x)}$$

Because the decision is independent of $P(x)$ and $P(C_N) = P(C_Y) = 0.5$, we just need to use $P(x|C_i)$ and choose the maximum. For the normal distribution, we have: So we get;

$P(x/C_i) = P(x/\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ **0.5 mark**

$P^M(x = 175/\mu_M = 182.5, \sigma = 14.79) = \frac{1}{14.79\times\sqrt{2\pi}} e^{-\frac{(175-182.5)^2}{2\times 218.75}} = 0.0237$ **1 mark**

$P^F(x = 175/\mu_M = 166.25, \sigma = 10.83) = \frac{1}{10.83\times\sqrt{2\pi}} e^{-\frac{(175-166.25)^2}{2\times 117.19}} = 0.0265$ **1 mark**

So the classification for a male. The probability of this output (certainty);
$P(C_M/x = 175) = \frac{0.0237}{0.0237+0.0265} = 0.4721$ **1 mark**
So the classification for a female. The probability of this output (certainty);
$P(C_F/x = 175) = \frac{0.0265}{0.0237+0.0265} = 0.5278$ **1 mark**

Remark; The gender with length of 175cm is most probable be a female, since the classification probability yielded the highest score. **0.5 mark**

QUESTION FOUR
4a) Explain the main difference between logistic regression and linear regression? **[2 marks]**

A linear regression is a supervised learning algorithm that is used to predict a dependent variable y given a set of independent variable x, such that the dependent variable is continuous.
On the other-hand, **a logistic regression** is a supervised learning algorithm that is used to predict a dependent variable y given a set of independent variable x, such that the dependent variable is categorical. **2 marks**

4b. Differentiate between online gradient descent and batch gradient descent? **[2 marks]**

Batch gradient descent computes the gradient of the cost function w.r.t the parameter weights **w** for an entire training data, by updating the weights in a single pass. **1 mark**
Online gradient descent computes the gradients of the cost function w.r.t the weights w, using a single training data sample X_i, by updating the weights for each example in the training data. **1 mark**

4c) What is the core difference between linear neural networks and multilayer perceptron? **[2 marks]**

A linear neural network is a kind of artificial neural network that consist of only input layer and output layer without hidden layers but a multilayer perceptron is a neural network structure containing input layer, hidden layer and output layer. **[2 marks]**

4d) Given a dataset with 3-dimensional Boolean examples $x = (x_A, x_B, x_C)$, train a naïve Bayes classifier to predict the following classification problems:

Attribute A	Attribute B	Attribute C	Classification D
F	T	F	T
F	F	T	T
T	F	F	T
T	F	F	F
F	T	T	F
F	F	T	F

i. What is the predicted probability P(D=T/A=T, B=F, C=T)? [**6 marks]**

P(D=T/A=T, B=F, C=T)

$$P(D=T)=1/2,\ P(D=F)=1/2$$

$$P(A=T|D=T)=1/3, P(A=F|D=T)=2/3$$
$$P(B=T|D=T)=1/3, P(B=F|D=T)=2/3$$
$$P(C=T|D=T)=1/3, P(B=F|D=T)=2/3$$

$$P(A=T|D=F)=1/3, P(A=F|D=F)=2/3$$
$$P(B=T|D=F)=1/3, P(B=F|D=F)=2/3$$
$$P(C=T|D=F)=2/3, P(B=F|D=F)=1/3$$

Each step = 0.25 x 14 = 3.5 marks

$$=\frac{P(A=T,B=F,C=T|D=T)P(D=T)}{P(A=T,B=F,C=T)}$$
$$=\frac{P(A=T,B=F,C=T|D=T)P(D=T)}{P(A=T,B=F,C=T|D=T)P(D=T)+P(A=T,B=F,C=T|D=F)P(D=F)}$$
$$=\frac{\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{2}}=\frac{2}{2+4}=\frac{1}{3}$$

Step 1 = 0.5 mark, Step 2 = 1 mark, and Step 3 = 1 mark

ii. What is the predicted probability P(D=T/B=T)? [**3 marks]**

P(D=T/B=T)

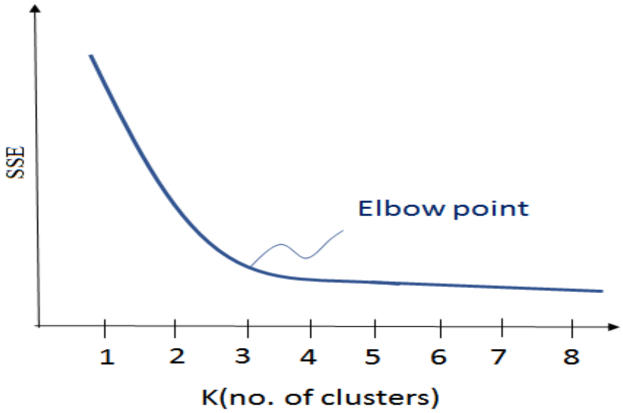
$$=\frac{P(B=T|D=T)P(D=T)}{P(B=T)}$$
$$=\frac{P(B=T|D=T)P(D=T)}{P(B=T|D=T)P(D=T)+P(B=T|D=F)P(D=F)}$$
$$=\frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{2}}=\frac{1}{1+1}=\frac{1}{2}$$

Each step = 1 mark x 3

QUESTION FIVE

5a) Clustering is a technique for finding similarity groups in data, called clusters. I.e., it groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters. ***(3 marks)***

Elbow method: We will run K-Means clustering for a range of K values let's say (K= 1 to 10) and calculate the Sum of Squared Error (SSE). Then plot a line chart for SSE values for each K, if the line chart looks like an arm then the elbow on the arm is the value of K that is the best. ***(3 marks)***



5b) **<u>Iteration-01:</u>**

- We calculate the distance of each point from each of the center of the two clusters.
- The distance is calculated by using the Euclidean distance formula.

The following illustration shows the calculation of distance between point A(2, 2) and each of the center of the two clusters-

Calculating Distance Between A(2, 2) and C1(2, 2)-

P(A, C1)
= sqrt [$(x2 – x1)^2 + (y2 – y1)^2$]
= sqrt [$(2 – 2)^2 + (2 – 2)^2$]
= sqrt [0 + 0]
= 0

Calculating Distance Between A(2, 2) and C2(1, 1)-

P(A, C2)
= sqrt [$(x2 – x1)^2 + (y2 – y1)^2$]
= sqrt [$(1 – 2)^2 + (1 – 2)^2$]
= sqrt [1 + 1]
= sqrt [2]
= 1.41

In the similar manner, we calculate the distance of other points from each of the center of the two clusters.

Next,

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Given Points	**Distance from center (2, 2) of Cluster-01**	**Distance from center (1, 1) of Cluster-02**	**Point belongs to Cluster**	
A(2, 2)	0	1.41	C1	***(1 mark)***
B(3, 2)	1	2.24	C1	***(1 mark)***
C(1, 1)	1.41	0	C2	***(1 mark)***
D(3, 1)	1.41	2	C1	***(1 mark)***
E(1.5, 0.5)	1.58	0.71	C2	***(1 mark)***

From here, New clusters are-

Cluster-01:

First cluster contains points-

- A(2, 2)
- B(3, 2)
- D(3, 1) ***(1 mark)***

Cluster-02:

Second cluster contains points-

- C(1, 1)
- E(1.5, 0.5) ***(1 mark)***

Now,

- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

Center of Cluster-01

= ((2 + 3 + 3)/3, (2 + 2 + 1)/3)

= (2.67, 1.67) ***(1 mark)***

For Cluster-02:

Center of Cluster-02

= ((1 + 1.5)/2, (1 + 0.5)/2)

= (1.25, 0.75) ***(1 mark)***

This is completion of Iteration-01.

QUESTION 6

6a) K should be less than n. This is because dimensionality reduction seeks to reduce data from higher to a lower dimensional dataset ***(2 marks)***

k=2 or k=3. This is because, we can plot 2D or 3D data but don't have ways to visualize higher dimensional data ***(2 marks)***

6b) (i)

$\bar{x} = 4.17 \quad \bar{y} = 3.83$

Point	X	Y	X – X̄	Y - Ȳ
A	1	1	-3.17	-2.83
B	2	1	-2.17	-2.83
C	4	5	-0.17	1.17
D	5	5	0.83	1.17
E	5	6	0.83	2.17
F	8	5	3.83	1.17

$C = \begin{pmatrix} 5.139 & 3.694 \\ 3.694 & 4.139 \end{pmatrix}$ Positive cov_{ij} values → x and y values increase together in dataset ***(3 marks)***

6b) (ii)

$$C - \lambda \cdot E = \begin{pmatrix} 5.139-\lambda & 3.694 \\ 3.694 & 4.139-\lambda \end{pmatrix}$$ Where E is identity matrix

The characteristic polynom is the determinant. The roots of the function, that appears if you set the polynom equals zero, are the eigenvalues

$$\det(C - \lambda \cdot E) = (5.139-\lambda)(4.139-\lambda) - (3.694)^2 \Rightarrow \lambda_1 = 8.367$$
$$= \lambda^2 - 9.278\lambda + 7.620 \qquad \lambda_2 = 0.911$$

(3 marks)

6b(iii)

Eigenvector v_1 with highest eigenvalue is our principal component

$$\begin{pmatrix} 5.139 & 3.694 \\ 3.694 & 4.139 \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = 8.367 \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$$

$$\Rightarrow v_1 = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} -0.753 \\ -0.658 \end{pmatrix}$$

(2 marks)

6c) Three (3) applications of PCA are:

Data compression so that it takes up less computer memory/disk space
Dimensionality reduction so as to speed up learning algorithm
Data visualization of high dimensional data ***(3 marks)***