**Homework Week 4**

**Question 1 [50 points]**

**In no more than 500 words, describe the difference between unsupervised and supervised machine learning and provide a real-world problem that could be solved with each.**

* **Points to discuss:**
* **How do we define unsupervised learning? What about supervised learning?**

Unsupervised learning is a branch of machine learning where the algorithm learns patterns and structures in the data without any labelled or pre-classified examples. The goal of unsupervised learning is to explore and extract meaningful insights from the data without explicit guidance. It deals with finding hidden patterns/structures, relationships, or groupings within the dataset according to the structure.

On the other hand, supervised learning is a machine learning paradigm where the algorithm learns from labelled data, meaning the dataset contains input features and corresponding target labels or outcomes. The objective of supervised learning is to build a model that can predict or classify unseen data accurately.

* **For unsupervised learning, discuss both association and clustering. Please provide a short explanation of each of the two.**

Association: Association analysis is a technique used in unsupervised learning to identify interesting relationships or associations between different items in a dataset. It involves discovering co-occurrence patterns or dependencies among variables. For example, in a retail setting, association analysis can identify items frequently purchased together, enabling businesses to make recommendations or optimise product placements.

Clustering: Clustering is another unsupervised learning method that groups similar data points together based on their inherent characteristics. The algorithm automatically identifies clusters or subgroups within the data without any prior knowledge or labels. Clustering can be used for customer segmentation, image segmentation, or anomaly detection, where similar instances need to be grouped together for further analysis or decision-making.

* **For supervised learning, discuss both classification and regression. Please provide a short explanation of each of the two.**

Classification: Classification is a supervised learning technique used for predicting discrete or categorical outcomes. It involves training a model on labelled data to learn the mapping between input features and predefined classes. For instance, email spam detection is a classification problem where the model learns to classify incoming emails as spam or not spam based on features like subject, content, and sender.

Regression: Regression, also a supervised learning approach, is employed to predict continuous numerical values. It aims to find the relationship between input features (x) and a continuous target variable (y). An example of regression is predicting housing prices based on factors like location, square footage, number of bedrooms, etc.

* **What data would we need to be able to apply unsupervised and/or supervised learning?**

For unsupervised learning, we only need the input data without any associated labels or outcomes. Unsupervised algorithms learn patterns and structures solely based on the input features and do not require target labels.

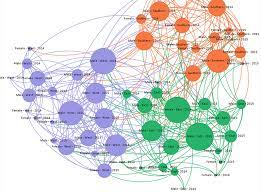
In contrast, supervised learning relies on labelled data, which consists of both input features and corresponding target labels or outcomes. The algorithm learns from this labelled data to make predictions or classifications on unseen instances.

* **What are the limitations of unsupervised learning and what are some methods (if any) we could use to overcome these?**

Unsupervised learning faces challenges such as difficulty in evaluating the accuracy of the learned patterns, subjectivity in interpretation, and the absence of ground truth. To overcome these limitations, methods like cluster validation indices, visualisation techniques, and expert domain knowledge can be utilised. Cluster validation indices help assess the quality of clustering results, visualisation techniques aid in understanding the discovered patterns, and expert domain knowledge provides context and validation.

* **Give examples of real-world problems where you would prefer using unsupervised learning. What about supervised learning?**

Unsupervised Learning: An example of using unsupervised learning is social network analysis, where clustering can be employed to identify communities or groups within a network without any prior knowledge of their existence.



Supervised Learning: In the medical field, supervised learning can be applied to predict the likelihood of a patient developing a certain disease based on their medical history, lifestyle factors, and genetic information. Classification algorithms can be trained using historical data to help healthcare professionals make informed decisions and provide personalised treatment plans.

**Question 2 [50 points]**

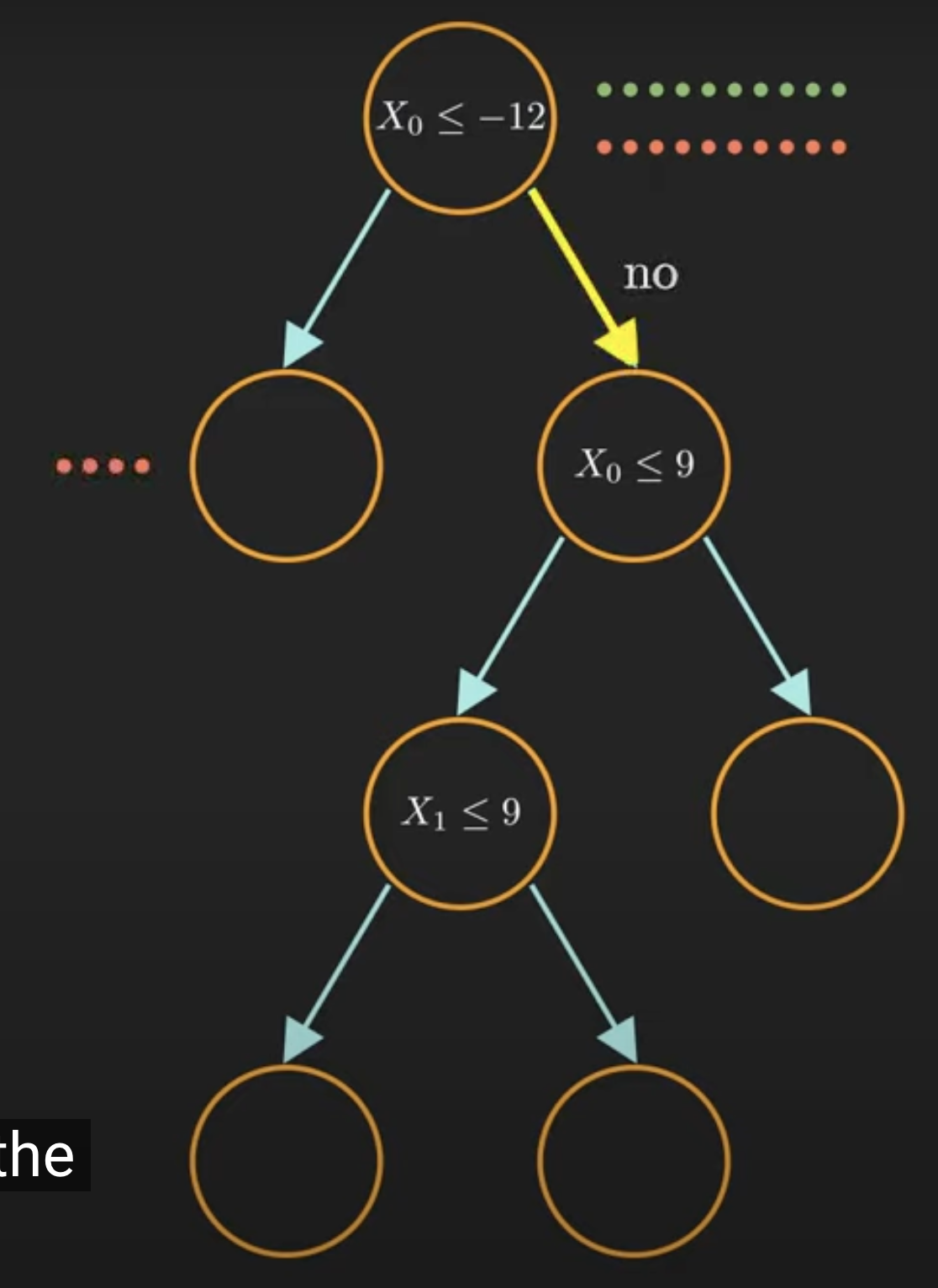
**Having in mind a non-specialised audience, explain a Machine Learning algorithm of your choice which was not covered in the class. The answer should be no longer than 500 words.**

* **For this task you should use the internet and other materials, such as the class slides, to answer the question. We are looking to see that you can describe a relatively complex concept related to Machine Learning using your own words.**
* **You do not have to code the algorithm and a purely theoretical explanation with an example of its application is sufficient.**
* **Points to discuss:**
* **Define the algorithm and mention what it is used for (e.g., unsupervised/supervised learning?)**
* **Can you create a breakdown with steps that someone could follow to implement the algorithm?**
* **Can you think of real-world problems where this algorithm is currently commonly used for? If you cannot find any, can you think of some problems where this algorithm might be useful?**
* **Are there any “limitations” to your algorithm? For example, is there any data-preprocessing that you need to do (e.g., some algorithms cannot work with particular data types or cannot handle missing values).**

I want to introduce the Decision Tree algorithm, which is used for supervised learning tasks.

The implementation of Decision Tree follows several steps:

1. Collect and preprocess the data for training the decision tree, involves handling missing values and categorical values (converting them into numerical representations, e.g. ‘First Class’ -> 1), and splitting the data into training and testing sets (e.g. 75% and 25%)
2. Start with the root node, which represents the entire dataset. Choose an attribute as the root based on a criterion. Split the data based on the chosen attribute into subsets. Repeat this process recursively for each subset until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples per leaf.
3. At each node, select the best attribute to split the data based on a splitting criterion. This criterion measures the quality of a split by evaluating the homogeneity or impurity of the target variable within each subset. Popular criteria include Information Gain, Gini Index, and entropy.



1. Define stopping criteria to control the growth of the decision tree and prevent overfitting. These criteria may include maximum depth, minimum number of samples per leaf, or a minimum improvement in impurity measure after a split.

Real-world problems where Decision Tree are commonly used:

* Disease Diagnosis: Decision Trees can assist in diagnosing diseases based on symptoms, medical history, and test results. By following the decision tree's path, healthcare professionals can identify potential diseases or conditions and recommend further tests or treatments.

Limitations of Decision Trees:

1. Overfitting: Decision Trees tend to create complex trees that can overfit the training data, resulting in poor generalisation on unseen data. Techniques like pruning and setting appropriate stopping criteria can help alleviate this issue.
2. Handling Numerical Data: Decision Trees work well with categorical and binary features but may struggle with continuous numerical data. One approach to overcome this is by discretizing the numerical data into intervals or using algorithms specifically designed for handling continuous attributes, such as Classification and Regression Trees.
3. Sensitivity to Data Variations: Decision Trees can be sensitive to small variations in the data, leading to different tree structures. This can result in instability and reduced robustness. Techniques like ensemble methods (e.g., Random Forests) can be employed to mitigate this issue.