School of Electrical and Computer Engineering- University of Tehran
Artificial Intelligence Course- 2024
Project 2-Due Date: May 18, 2024, before 4 PM

Part I
Unsupervised Learning

[A]. Explain the mathematical intuition behind Hierarchical Clustering and Density-Based Spatial Clustering (DBSCAN) by providing two examples. One example for each algorithm.
[Length of your recording ≤ 8 minutes]. You should provide your notebook. Use Google Colab.

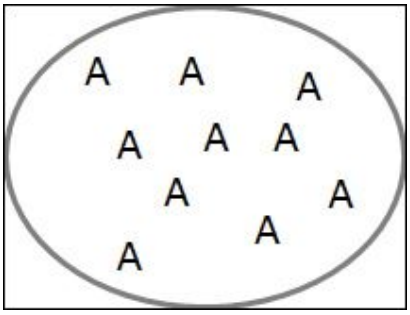
[B]. Use the dataset I provided in the [Project 2 folder] and build three unsupervised models using K-Means, Hierarchical Clustering, and DBSCAN. Compare the performance of each model. You should elucidate the code blocks. In this problem, your objective is to employ clustering algorithms that can successfully group investors based on different factors (such as age, income, and risk tolerance) that portfolio managers can further use to standardize portfolio allocation and rebalance strategies across the clusters, making the investment management process faster and more effective. The dataset has the following features.

AGE: There are six age categories, where 1 represents an age less than 35 and 6 represents an age more than 75.
EDUC: There are four education categories, where 1 represents no high school, and 4 represents a college degree.
MARRIED: There are two categories to represent marital status, where 1 represents married, and 2 represents unmarried.
OCCU: This represents the occupation category. A value of 1 represents managerial status, and 4 represents unemployed.
KIDS: Number of children.
WSAVED: This represents the individual's spending versus income, split into three categories. For example, 1 represents spending exceeded income.
NWCAT:This represents the net worth category. There are five categories, where 1 represents net worth less than the 25th percentile, and 5 represents net worth more than the 90th percentile.
INCCL: This represents the income category. There are five categories, where 1 represents income less than 10,000 USD, and 5 represents income more than 100,000 USD.
RISK: This represents the willingness to take risk on a scale of 1 to 4, where 1 represents the highest level of willingness to take risk.
LIFECYCL:This lifecycle variable approximates a person's ability to take on risk. There are six categories for increasing the level of ability to take risk. A value of 1 represents "age under 55, not married, and no kids," and a value of 6 represents "age over 55 and not working."
HHOUSES: This is a flag indicating whether the individual is a homeowner. A value of 1 (0) implies the individual does (does not) own a home.
SPENDMOR: This represents higher spending preference if assets are appreciated on a scale of 1 to 5.
[Length of your recording ≤ 8 minutes]. You should provide your notebook. Use Google Colab.

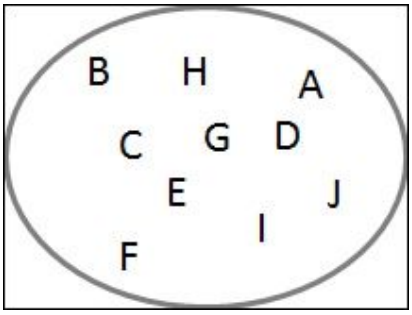
Part II

Supervised Learning

In the lecture, I explained the concepts of entropy and information gain for building a decision tree. Entropy originated in thermodynamics as a measure of molecular disorder. Entropy approaches zero when molecules are still well-ordered. It later spread to various domains, including machine learning. We use it in the process of building a Decision Tree. In essence, entropy measures the homogeneity of a dataset. Imagine a dataset with ten observations with one attribute, as shown in the following diagram:



The value of this attribute is A for the ten observations. This dataset is entirely homogenous, and it is easy to predict the value of the next observation. The entropy in a dataset that is completely homogenous is zero. On the other hand, imagine a similar dataset, but in this dataset, each observation has a different value, as shown in the following diagram:



This dataset is very heterogeneous, and predicting the following observation is hard. In this dataset, the entropy is higher. Another measurement used to split a decision tree is called Gini Impurity Index. The Gini Impurity Index is another way we can measure the diversity in a dataset. This means that if we have a dataset in which all elements are similar, the dataset has a low Gini Impurity Index, and if all elements are different, it has a large Gini Impurity Index.

[A]. Elucidate the mathematical intuition of the Gini Impurity Index by providing an example. You should demonstrate your work mathematically and by using codes. [Length of your recording $\leq$ 5 minutes]. You should provide your notebook. Use Google Colab.

[B]. Using Entropy or the Gini Impurity Index in most decision tree applications leads to similar results. That is said, the Gini Impurity Index is slightly faster to compute. However, when they differ, the Gini Impurity Index tends to isolate the most frequent class in its own branch of the tree, while entropy tends to produce slightly more balanced trees. Provide an example showing that using entropy tends to produce somewhat more balanced trees than the Gini Impurity Index.

You should demonstrate your work by using codes. [Length of your recording ≤ 5 minutes]. You should provide your dataset and notebook. Use Google Colab.

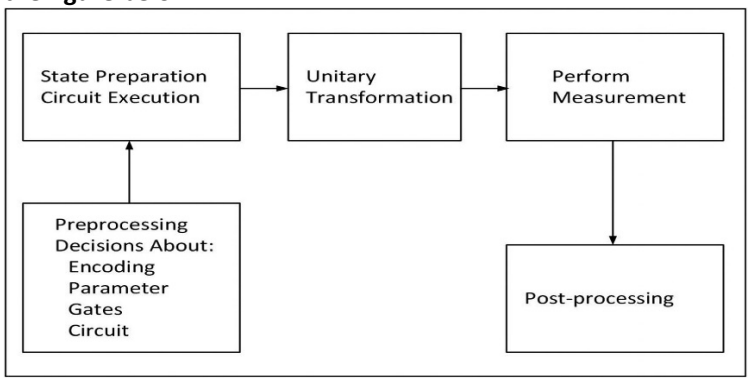
[C]. Use the dataset I provided for Assignment 1 in the [Assignment 1-Dataset] folder and build a Decision Tree and Random Forest Models. Compute the classification report for each model and compare your results. You should explain the codes for Decision Tree and Random Forest Models. [Length of your recording ≤ 8 minutes]. You should provide your notebook. Use Google Colab.

[D]. In the lecture, I provided an example of Decision Tree and Random Forest Models for classification problems. Explain the decision tree algorithm for solving regression problems. [Length of your recording ≤ 5 minutes]. Use the Boston housing dataset to build a Decision Tree and Random Forest Regression model (Note: For this part, you only need to provide your notebook, no recording is required). You should provide your notebook. Use Google Colab.

Part III

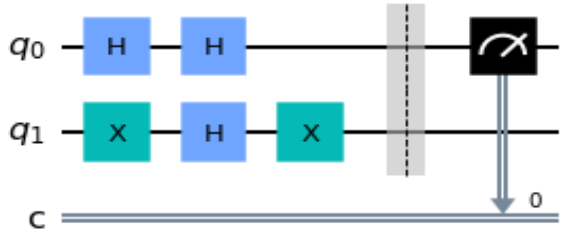
Quantum Computing and Quantum Neural Networks

[A]. In the lecture, I explained that in order to load our classical data into quantum states, various methods have been used, such as Basis Encoding, Amplitude Encoding, and Angle Encoding. See the figure below.



Angle Encoding was explained during the lecture. Elucidate the mathematical intuition for Basis Encoding and Amplitude Encoding by providing two examples. Use Qiskit(IBM) to demonstrate each method. [Length of your recording ≤ 8 minutes]. You should provide your notebook. Use Google Colab.

[B]. Explain the following Quantum Circuit and use Qiskit to create it. [Length of your recording ≤ 5 minutes]. You should provide your notebook. Use Google Colab.



[C]. During the lecture, I explained a Quantum Neural Network(QNN) model and provided a notebook. You are required to elucidate each block of codes in the notebook. [Length of your recording $\leq$ 8 minutes]. In your third project, you will be asked to build a QNN model

Important Note

1. For this project, you should record a video. You should show your face. Your voice must be clear. It is your responsibility to make sure your video is working. You should put your video and notebook in Google Drive. Your folder should have your [student ID number, first name, and last name]. Only one video should be submitted for all parts of this project.
2. Your notebook should be well structured, and before each code block, you should explain the codes of the next block. You should follow the format of the notebooks I presented in the lectures.
3. You should make your Google link open access and submit your video to the auxiliary TA of our course, Mr Behzad Mohasel Afshari, via e-mail on May 18, 2024, before 4 PM. His e-mail for this course is ai.2024.cs@gmail.com.
4. You should use Google Colab. Insert all the slides into your Google Colab environment, then record all parts of this project.
5. You should submit the following two items
A. Your Google Colab notebook.
B. Your video file. The extension of your video file should be .mp4. Other formats will not be accepted.
C. Your dataset for the parts your elucidation is augmented with examples.
6. Your video file and notebook should be saved as [student ID number, first name, and last name]. Your total video file should not be more than 60 minutes. Your recording must be in English.
7. You should ensure that your codes and mathematical algorithms are flawless.

If further elucidation is warranted, please don't hesitate to contact me.