UNIVERSITI MALAYA

FAKULTI SAINS KOMPUTER DAN TEKNOLOGI MAKLUMAT

*Faculty of Computer Science and Information Technology*

# WQD7012 APPLIED MACHINE LEARNING

## OCCURRENCE 3

## SEMESTER 2 SESSION 2024/2025

## *TUTORIAL 1*

**NAME:**

BARKAVI A/P P CHEVEN (23093149)

**LECTURER'S NAME:**

DR RIYAZ AHAMED ARIYALURAN HABEEB MOHAMED

# WQD7006: Machine Learning for Data Science

## *Tutorial 1*

1. A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E. Suppose we feed a learning algorithm a lot of historical weather data, and have it learned to predict weather. In this setting, what is T?
   *In this weather prediction scenario:*
   *T (Task): Predicting the weather*
   *E (Experience): Historical weather data*
   *P (Performance measure): Accuracy of weather predictions*
   *The task (T) is specifically predicting the weather based on historical patterns.*

2. Suppose you are working on weather prediction, and your weather station makes one of three predictions for each day's weather: Sunny, Cloudy or Rainy.
   You'd like to use a learning algorithm to predict tomorrow's weather. Would you treat this as a classification or a regression problem? Why?
   *This is a classification problem because the output variable (weather prediction) is categorical with discrete classes (Sunny, Cloudy, or Rainy). Henceforth, we're assigning the weather to one of several predefined categories. In this case, we're not predicting a continuous numerical value*

3. Suppose you are working on stock market prediction. You would like to predict whether or not a certain company will win a patent infringement lawsuit (by training on data of companies that had to defend against similar lawsuits). Would you treat this as a classification or a regression problem? Why?
   *This is a classification problem because the output is binary (win or lose the lawsuit). Here, we're predicting which of two discrete categories the outcome falls into and eventually not predicting a continuous value but rather a discrete outcome.*

4. Some of the problems below are best addressed using a supervised learning algorithm, and the others with an unsupervised learning algorithm. Which of the following would you apply supervised learning to? And which of them would you apply to an unsupervised learning algorithm. In each case, assume some appropriate data set is available for your algorithm to learn from. Justify your answers.

- Given genetic (DNA) data from a person, predict the odds of him/her developing diabetes over the next 10 years.
  *This is a supervised learning problem. Justification: This requires labelled training data where we know both the genetic data (features) and whether those individuals developed diabetes (target variable). To be more profound, we're trying to predict a specific outcome based on historical examples.*

- Given data on how 1000 medical patients respond to an experimental drug (such as effectiveness of the treatment, side effects, etc.), discover whether there are different categories or "types" of patients in terms of how they respond to the drug, and if so what these categories are.
  *This is an unsupervised learning problem. Justification: In this case, we're trying to discover hidden patterns or groupings in the data without predefined categories. This is a clustering problem where we want to find natural groupings of patients based on their responses.*

- Have a computer examine an audio clip of a piece of music and classify whether or not there are vocals (i.e., a human voice singing) in that audio clip, or if it is a clip of only musical instruments (and no vocals).

  *This is a supervised learning problem, specifically a binary classification task. Justification: We need to classify audio clips into two distinct categories: "contains vocals" or "instrumental only". This requires labelled training data where we have examples of both vocal and instrumental-only clips. The algorithm would learn to identify patterns in audio features (like spectral characteristics, frequency distributions, etc.) that distinguish human voices from instruments. Therefore, the output is a discrete category (vocals present or not), not a continuous value*

- Given a large dataset of medical records from patients suffering from heart disease, try to learn whether there might be different clusters of such patients for which we might tailor separate treatments.
  *This is an unsupervised learning problem, specifically a clustering task. Justification: Here, eventually we're trying to discover natural groupings or patterns within the heart disease patient data. Therefore, we don't have predefined categories or labels for these potential patient clusters. The ultimate goal is to identify if distinct subgroups exist based on similarities in their medical records. These discovered clusters could then inform personalized treatment approaches as we're not predicting a specific outcome but rather exploring the inherent structure in the data*

5. **Which** of these is a reasonable definition of Machine Learning?

    - Machine learning is the science of programming computers.
    - Machine learning is the field of allowing robots to act intelligently.
    - ***Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.***
    - Machine learning learns from labeled data.

    *This definition depicts the essence of machine learning - the ability of systems to improve through experience rather than through explicit programming of rules.*