6COSC020W Applied AI

Tutorial Week 07

Machine Learning

In this tutorial, we will review some definitions of machine learning, discussed in the lecture, and the implementation of important machine learning algorithms. In the first part of the implementation, you will learn how to work with data as one of the essential steps that every machine learning engineer must know. Then you will be given four use cases (challenges), and you need to do one of them in the tutorial session either independently or by your classmate (max two people including yourself).

1. Definitions and Concepts:

2. What is a supervised learning algorithm? What are the types of supervised learning?

   Supervised learning is a type of machine learning method in which we provide sample **labeled data** to the machine learning system to train it, and on that basis, it predicts the output. The goal of supervised learning is to map input data with output data. Supervised learning is based on supervision, and it is the same as when a student learns things under the supervision of the teacher.

   There are two types of supervised learning algorithms: regression and classification.

3. What does it mean by feature(s)? Please indicate different types of features.

   Features are individual independent variables that act as input in your system. While making the predictions, models use such features to make the predictions.

   Numerical and categorical features are the most common types of features.

4. You have been asked to predict the sales of an online store for a period between 2023 to 2025. Which machine learning algorithm you would use and why?

   For continuous data, linear regression could be a good solution.

5. What is an unsupervised learning algorithm?

   Unsupervised machine learning is the training of models on raw and unlabelled training data. It is often used to identify patterns and trends in raw datasets or to cluster similar data into a specific number of groups.

6. Please explain the theory behind the K-means and SVM algorithms.

   K-Means clustering is an unsupervised learning algorithm that learns properties of a set of data points and forms partitions called clusters, that represent data with similar properties. For continuous data, each cluster is represented by the centroid which is the mean of cluster members. K-Means uses squared Euclidean distance as the similarity measure for cluster membership.

   Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

2. Machine Learning Algorithms Implementation

In this part of the tutorial, you will learn how to implement machine learning algorithms. You will also learn almost the whole process of the machine learning pipeline from data preprocessing to model building. To do the exercises, you need to log in to the blackboard and download a ZIP file called Tutorial_Week07_ML under the Week 07 folder. After extracting the ZIP file, I recommend you import it to your JupyterLab environment so that you have access to all files. Correspondingly, you see four Jupyter Source files (01-Data Exploration, 02-Regression, 03-Classification, and 04-Clustering).

These exercises are a good starting point to understand almost all machine learning basics that we discussed during the lecture. Please try to understand each line of codes that you are working with and if you find them difficult to understand, please ask your instructor.

3. Challenge

In this part of the tutorial, you will need to do one of four challenges during the tutorial session, either independently or with one of your classmates. The data for all challenges have been already provided. You can find those challenges under the challenges folder (01-Flights Challenge, 02-Real Estate Regression Challenge, 03-Wine Classification Challenge, and 04-Clustering Challenge).