BMCS2114: Machine Learning

Lecture 01: Introduction to Machine Learning

1. Concepts of Machine Learning:

- Machine learning is a method of teaching computers to learn and make predictions based on data.
- It enables computers to learn without being explicitly programmed.
- Machine learning algorithms are trained on a dataset and then use that information to make predictions or decisions about new, unseen data.
- It opens up new possibilities for solving complex problems and making predictions in various fields, such as finance, healthcare, marketing, transportation and energy.

2. Types of Machine Learning:

- Supervised Learning: Training the model using labelled data to predict new and unseen data..
- Unsupervised Learning: Training the model using unlabeled data to discover hidden patterns or relationships in the data.
- Semi-Supervised Learning: A combination of supervised and unsupervised learning, using a small amount of labelled data and a large amount of unlabeled data.
- Reinforcement Learning: The model learns by taking actions in an environment and receiving feedback through rewards or penalties.

3. Terminology related to Machine Learning:

- Model: The mathematical representation of the relationship between the input and output variables in the data.
- Feature: An input variable that provides information to the model to make predictions.
- Label: The target variable to be predicted by the model.
- Training Data: Data used to train the model and create the relationship between the features and the label..
- Validation Data: Data used to evaluate the model's performance after training.
- Overfitting: When the model is too complex and performs well on the training data but poorly on new data.
- Underfitting: When the model is too simple and fails to capture the underlying relationships in the data.

4. SAS Viya for Machine Learning:

- SAS Viya is a cloud-based platform for advanced analytics and machine learning.
- It provides a comprehensive suite of data preparation, model building, deployment, and monitoring tools..
- With SAS Viya, organizations can accelerate their machine-learning initiatives and drive better business outcomes.

- It supports various ML algorithms, including decision trees, regression, clustering, and deep learning, and provides easy-to-use interfaces for technical and non-technical users.

Regression vs Classification:

- Regression and classification are two of the most common tasks in machine learning.
- Regression is used to predict a continuous outcome, such as a price, a score, or a probability.
- Regression aims to model the relationship between a set of input variables (also known as features) and a continuous output variable.
- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It help to predict real/continuous values.
- By performing the regression, we can confidently determine the most important factor, the least important factor, and how each factor is affecting the other factors.
- For example, using housing price data, a regression model could be trained to predict the price of a house based on its location, size, and other features.
- On the other hand, classification is used to predict a categorical outcome, such as a label, a class, or a binary outcome (yes/no).
- Classification aims to divide data into classes or categories based on their features.
- For example, using a dataset of handwritten digits, a classification model could be trained to recognize the digits 0-9 based on their image features.
- The output of a classification model is a categorical value, such as "digit 3".
- Another example is to predict crop yields based on weather data. A dataset of historical weather patterns can be used to train a machine learning model to predict crop yields for a given area based on features such as average temperature, rainfall, and sunshine hours. Farmers can use this information to plan planting and harvesting schedules, optimize irrigation systems, and decide which crops to grow.
- The main difference between regression and classification is that regression is used for predicting continuous outcomes, while classification is used for predicting categorical outcomes.

Exercise 1

1. With the aid of diagrams, discuss TWO (2) main purposes of regression analysis.

2. Based on the data set given in *Table 1* below, plot the diagram of house prices versus its size in square feet. Then draw a best-fit line to do our predictions of prices for the following houses with their sizes given in square feet:

   a) 2500

   b) 3000

   c) 3500

Table 1

Y-axis (Size in sqft)	X-Axis (Price in RM, in 1000's)
250	100
650	150

1150	210
1450	260
1250	300
2000	360
2100	400
2100	450
2100	500

Supervised Learning Overview:

- Supervised learning is a type of machine learning that involves using labelled data to make predictions or decisions.
- In supervised learning, the goal is to fit a model to the data that can then be used to make predictions.
- Fitting a model typically involves dividing the data into training and testing or validation sets.
- The training set is used to fit the model, and the testing set is used to evaluate its performance.
- The data in the training set is labelled and contains both the features and the target or output variables. It is also referred to as data with answers.
- The data in the testing set is unlabeled and is used to evaluate how well the model generalizes to new, unseen data. It is also referred to as data without answers.
- Once the model is trained, it can be applied to new data to make predictions or decisions.
- Fitting a model involves adjusting the model parameters to fit the training data well.
- Predicting using a model involves using the learned model to predict new, unseen data.
- In summary, supervised learning is a type of machine learning that involves training a model on labelled data and then using that model to make predictions on new, unseen data. The goal is to fit the best possible model by adjusting its parameters to fit the training data well and then using that model to make predictions for new data.

Top of Form

Bottom of Form

Top of Form

Regression with Movie Data:

- In this example, we will use supervised learning and regression to predict a movie's revenue based on its features.
- The data we will be using includes information about movies, such as budget, genre, and release date, along with the target variable of revenue.
- The goal is to fit a regression model to the data that can then be used to predict the revenue of a new movie based on its features.
- Fitting the model involves dividing the data into training and testing sets.
- The training set is used to fit the model, and the testing set is used to evaluate its performance.
- The data in the training set is labelled and contains both the features and the target variable of revenue.

- The data in the testing set is unlabeled and is used to evaluate how well the model generalizes to new, unseen data.
- Once the model is trained, it can then be applied to new data to make predictions for the revenue of a new movie.
- Fitting a model involves adjusting the model parameters to fit the training data well.
- Predicting using a model involves using the learned model to predict new, unseen data.
- In summary, this example demonstrates how supervised learning and regression can be used to predict a movie's revenue based on its features. The process involves dividing the data into a training set and a testing set, fitting a model to the training data, and then using that model to make predictions for new, unseen data.

Classification with Email Data:

- In this example, we will use supervised learning and classification to predict whether an email is a spam.
- The data we will be using includes information about emails, such as the sender, subject line, and content, along with the target variable of whether the email is spam.
- The goal is to fit a classification model to the data that can then be used to predict a new email's class (spam or not spam).
- Fitting the model involves dividing the data into training and testing sets.
- The training set is used to fit the model, and the testing set is used to evaluate its performance.
- The training dataset is labelled and contains the features and the target variable of whether the email is spam.
- The data in the testing set is unlabeled and is used to evaluate how well the model generalizes to new, unseen data.
- Once the model is trained, it can then be applied to new data to make predictions for the class of a new email.
- Fitting a model involves adjusting the model parameters to fit the training data well.
- Predicting using a model involves using the learned model to predict new, unseen data.
- In summary, this example demonstrates how supervised learning and classification can be used to predict whether an email is a spam or not. The process involves dividing the data into a training set and a testing set, fitting a model to the training data, and then using that model to make predictions for new, unseen data.

Top of Form

Bottom of Form

Examples of machine learning in our daily lives:

1. **Web search**: Machine learning algorithms play a significant role in web search. The algorithms analyze and rank the relevance of search results based on factors such as the user's search history, location, and previous queries. The algorithms also learn from user behaviour and adjust the results accordingly to provide a better search experience.

2. **Postal mail routing**: Machine learning algorithms are used in postal mail routing to optimize the delivery process. The algorithms analyze the characteristics of the mail, such as size, weight, and destination, to determine the most efficient routing and delivery process.

3. **Fraud detection**: Machine learning algorithms detect fraud in various industries, including finance, insurance, and healthcare. The algorithms analyze behaviour patterns and transactions to identify potential fraud, reducing the risk of financial losses.

4. **Vehicle driver assistance**: Machine learning algorithms are used in vehicle drivers assistance systems, such as adaptive cruise control, lane departure warning systems, and automatic emergency braking. The algorithms analyze sensor data and make real-time decisions to enhance driver safety and improve the driving experience.

5. **Web advertisements**: Machine learning algorithms are used to target the right audience. The algorithms analyze user behaviour and demographics to determine the most relevant advertisements. As a result, it will increase the chances of the user clicking on the advertisement.

6. **Social networks**: Machine learning algorithms are used to personalize the user experience. The algorithms analyze the user's activity, interests, and connections to suggest new friends, groups, and content to the user.

7. **Speech recognition**: Machine learning algorithms, such as Siri and Google Assistant, are used in speech recognition systems. The algorithms analyze speech patterns and language to transcribe speech into text and understand the intent behind the speech, allowing users to interact with devices and applications hands-free.

In conclusion, machine learning algorithms are used in various applications in our daily lives to make our lives easier, safer, and more efficient. For example, machine learning significantly shapes our world, from optimizing postal mail delivery to improving driver safety.

Revision

Q1) Define and discuss the differences among artificial intelligence, machine learning and deep learning.

Answer

Deep learning (DL) is a subset of machine learning (ML), and machine learning is a subset of artificial intelligence (AI).

AI: A branch of computer science dealing with the simulation of intelligent behavior in computers. It is a program that can sense, reason, act and adapt.

ML: The algorithms can learn from past data or experiences without explicitly programming.

DL: Structures algorithms in multilayered neural networks to learn from vast amounts of data.

Q2) Discuss the machine learning life cycle in **FIVE (5)** phases.

Answer

i. Data gathering

• Identify various data sources (files, database, mobile)
• Collect Data
• Integrate the data collected from different sources (dataset)

ii. Data preparation and Data Wrangling

• Randomize the ordering of data
• Data exploration (is to understand the quality of data better). In this, we find correlations, general trends, and outliers.
• Cleaning and converting raw data into a usable format to make it more suitable for analysis.
• Missing values, duplicate data, invalid data and noise

iii. Data Analysis

• Build a machine learning model to analyze the data using various analytical techniques.
• First, we need to determine the type of problems and select the machine learning techniques such as Classification, Regression, Cluster analysis, and Association.
• Last, we build the model.

iv. Train and Test Models

• We use our dataset to train our model to improve its performance for better outcomes and also make our model understand the various rules, patterns, and features

• After we trained our model, we tested the model. In this step, we checked the accuracy of our model by providing a test dataset.

v. Deployment

• If our model produces an accurate result, then we deploy the model in a real system.
• Before deploying our model, we will check whether it is improving its performance using available data or not.

Q3) Provide examples of application using regression technique and explain the purpose of using regression analysis.

Answer

- Prediction of rain using temperature and other factors.
- Determining market trends
- Prediction of road accidents due to rash driving.

Q4) What are residuals in a regression task?

Answer

The difference between the predicted value and the actual value

Q5) Discuss **FIVE (5)** terminologies related to regression analysis.

Answer

- **Dependent variable**

  ➢ The main element in regression analysis we need to forecast or comprehend is the dependent variable.

  ➢ It is likewise called the target variable.

- **Independent variable**

  ➢ The elements that influence the dependent variable or are utilized to forecast the value of the dependent variable are called independent variables.

  ➢ This is also known as predictors.

- **Outliers**

  ➢ Outliers are observations which involve very low or very high values when compared to other observations.

  ➢ Outliers may affect the results and this should be avoided.

- **Multicollinearity**

  ➢ When the correlation between independent variables is higher than other variables.

  ➢ It should not be in the dataset as it will cause problems when locating the most affected variables.