Machine Learning Session 2: Key Questions and Insights by Dr. Monali Mavani

In the second session of the **Machine Learning (S1-24_AIMLCZG565)** course, Dr. Monali Mavani engaged with students by addressing several important questions that arise when applying machine learning algorithms. These questions are fundamental in helping students understand the nuances of choosing the right algorithm and its application based on the type of problem, target variables, and domain knowledge.

This article provides a detailed explanation of the key questions discussed during the session, along with the answers and reasoning behind them.

---

1. How Do We Determine Which Algorithm to Use: Regression or Classification?

The Question:

Dr. Monali Mavani asked how to determine whether a machine learning problem should be treated as a **regression problem** or a **classification problem**. This distinction is critical because it dictates the choice of algorithms to use.

Answer:

The choice between **regression** and **classification** depends on the type of target variable (also called the dependent variable):

- **Regression Problem**: If the target variable is **continuous**, such as predicting prices, temperatures, or sales figures, the problem is categorized as **regression**. Examples include **linear regression** and **polynomial regression**.
- **Classification Problem**: If the target variable is **categorical**, such as predicting whether an email is spam or non-spam, or whether a customer will churn, the problem is classified as **classification**. Examples include **logistic regression**, **decision trees**, and **support vector machines (SVM)**.

It is essential to first identify whether the output is a continuous value or a category, which determines whether regression or classification should be used.

---

2. When the Output is Between 0 and 1, Should I Use Linear or Logistic Regression?

The Question:

A follow-up question was raised regarding the scenario when the output values are between 0 and 1. The question asked whether **linear regression** or **logistic regression** should be used in this case.

Answer:

Dr. Monali Mavani explained that the range of the output (between 0 and 1) alone is not sufficient to decide between **linear regression** and **logistic regression**. Instead, the key factor is **what the output represents**:

- **Logistic Regression**: If the output represents a **probability** (such as the probability of a customer making a purchase, or the probability of an event occurring), **logistic regression** is the appropriate choice. Logistic regression uses the **sigmoid function** to map the output to a value between 0 and 1, which is interpreted as a probability.
- **Linear Regression**: If the output is **continuous** and represents a physical quantity (e.g., the speed of an object, price of a product, or temperature), **linear regression** should be used, even if the output happens to fall between 0 and 1. Linear regression will give a continuous output that might fall within any range, including between 0 and 1.

Key Point: The distinction lies in the **interpretation of the output**. If it's a probability, use logistic regression; if it's a continuous value, use linear regression.

---

3. How Do You Interpret the Output in Logistic Regression?

The Question:

A question about **logistic regression** asked how to interpret the output, especially when it is a value between 0 and 1. Students were curious about whether this output should be interpreted as a continuous quantity or as a probability.

Answer:

Dr. Monali emphasized that in **logistic regression**, the output should always be interpreted as a **probability**. The logistic regression model uses the **sigmoid function** to output a value between 0 and 1, which indicates the probability of an event occurring.

- **Interpretation**: If the output is 0.8, for example, it means that the model predicts an 80% probability of the event occurring (e.g., the email being spam). The threshold for classification (whether it is classified as "spam" or "not spam") is usually set at 0.5, but this can vary depending on the specific problem and context.

In contrast, if the target variable is continuous and the task is to predict a value (e.g., the price of a product), **linear regression** would be used.

---

4. When Should You Use Logistic Regression vs. Linear Regression?

The Question:

Another key question was raised about when to use **logistic regression** versus **linear regression**, particularly when the output is between 0 and 1.

Answer:

The decision to use **logistic regression** or **linear regression** is not determined by the range of the output but rather by the nature of the target variable:

- **Logistic Regression**: Used when the target variable represents **categorical data** or **probability**. For example, in binary classification problems like determining whether a customer will buy a product (yes/no), logistic regression is used to predict the probability of the positive class.
- **Linear Regression**: Used when the target variable is **continuous**. Even if the predicted value falls between 0 and 1, if it represents a continuous quantity (like the price of a product or a temperature), linear regression should be used. The goal of linear regression is to fit a line (or hyperplane in multiple dimensions) that minimizes the error between predicted and actual values.

---

5. How Do You Choose Between Regression and Classification Based on Domain Knowledge?

The Question:

One student asked how domain knowledge could influence the choice between regression and classification.

Answer:

Dr. Monali explained that domain knowledge plays a **critical role** in determining the appropriate approach:

- **Understanding the Target Variable**: If the target variable is something that naturally falls into categories (e.g., customer churn, spam detection), then **classification** algorithms should be used. On the other hand, if the target is continuous, like predicting prices, temperatures, or quantities, then **regression** should be applied.
- **Nature of the Problem**: Sometimes, the nature of the problem requires the model to predict a **probability** rather than a continuous quantity. For example, in medical diagnosis, predicting the probability that a patient has a disease requires **logistic regression**. Conversely, predicting the actual cost of a treatment would be a regression problem.

In short, domain expertise allows you to understand the problem better and select the model based on whether the target variable is categorical or continuous, as well as whether you need a probability or an exact value.

---

6. How Does Generalization Fit into the Discussion of Regression and Classification?

The Question:

Another question discussed the concept of **generalization** and how it fits into the decision between regression and classification algorithms.

Answer:

Dr. Monali highlighted that **generalization** is one of the most important aspects of machine learning:

- **Overfitting vs. Underfitting**: The key challenge in machine learning is to avoid **overfitting** (where the model memorizes the training data) and **underfitting** (where the model fails to capture the underlying patterns). A model that **generalizes well** is able to make accurate predictions on new, unseen data.
- **In Regression**: Overfitting can occur if the model is too complex or has too many features. In classification, this could lead to an excessively detailed decision boundary that doesn't generalize well.

- **Regularization**: Techniques like **L1 (Lasso)** and **L2 (Ridge)** regularization are used to ensure that the model is simple and generalizes well. These techniques penalize large weights and encourage simpler models, helping to prevent overfitting.

Conclusion: The Key Takeaways from the Session

In this session, Dr. Monali Mavani provided clarity on several important concepts in machine learning:

- **Regression vs. Classification**: Choose based on the type of target variable (continuous vs. categorical).
- **Linear vs. Logistic Regression**: The output's interpretation determines which model to use. Logistic regression is for probabilities, while linear regression is for continuous values.
- **Domain Knowledge**: Understanding the problem and target variable is key to choosing the appropriate machine learning algorithm.
- **Generalization**: The goal is to build models that generalize well to unseen data, preventing overfitting and underfitting.

By asking these important questions, students can refine their understanding of machine learning algorithms and apply them more effectively to real-world problems. As the course progresses, these foundational concepts will be crucial for tackling more advanced topics and algorithms.