Experiment No. 5

Aim: To develop and evaluate logistic regression models for multi-class classification tasks using machine learning.

Platform used: Google Colab

Theory:

Logistic Regression

Logistic regression is a statistical method used for binary classification problems, where the goal is to predict the probability of a data point belonging to a specific category.

Multinomial Logistic Regression

Multinomial logistic regression is a statistical method used to predict the probability of a categorical outcome with more than two categories, where the categories have no inherent order. It generalizes binary logistic regression to handle multiple classes, making it suitable for situations with more than two possible outcomes.

Key Concepts:

- **Multiclass Classification:**

Multinomial logistic regression is designed for situations where the dependent variable (the outcome you're trying to predict) has more than two categories.

- **Nominal Variables:**

The categories in the outcome variable are typically nominal, meaning there's no inherent order or ranking among them. For example, types of fruit (apple, banana, orange) are nominal, while educational levels (high school, bachelor's, master's) are ordinal.

- **Log-Odds:**

Like logistic regression, multinomial logistic regression models the log-odds of belonging to a particular category, using a log-linear relationship with the predictor variables.

- **Maximum Likelihood Estimation:**

The model parameters (coefficients) are estimated using maximum likelihood estimation, which finds the values that maximize the probability of observing the given data.

How it Works:

- **Reference Category:** One category is chosen as the reference category.
- **Pairwise Comparisons:** The model calculates the log-odds of belonging to each other category relative to the reference category.
- **Multiple Equations:** This results in multiple logistic regression equations, one for each non-reference category.
- **Simultaneous Estimation:** These equations are estimated simultaneously, not separately.
- **Interpretation:** The coefficients in each equation indicate how changes in the predictor variables affect the odds of belonging to that specific category compared to the reference category.

Example:

Imagine predicting a student's choice of college major (e.g., Arts, Sciences, Engineering) based on their high school GPA and test scores. If you choose "Arts" as the reference category, the model would calculate:

The log-odds of choosing "Sciences" versus "Arts".

The log-odds of choosing "Engineering" versus "Arts".

Advantages:

- Handles multiple classes effectively.
- Provides insights into the relationships between predictors and different outcome categories.
- Can be used with various types of predictor variables (continuous, categorical).

Disadvantages:

- Requires careful consideration of the reference category.
- Can become complex with many categories and predictors.

Applications:

- Predicting customer behavior (e.g., product choice, brand preference).
- Classifying documents into categories.
- Analyzing survey data with multiple response options.
- Modeling transportation mode choice (car, train, bus, etc.).
- Medical diagnosis.

Multinomial Logistic Regression function:

The Softmax Function

In the case of Logistic Regression, we use a sigmoid function for predicting the results. The generalized version of the sigmoid is known as the softmax function. The softmax function accepts an input vector z = [z1, z2, ..., zk] and produces another output vector of probability distributions. It can be expressed as :

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where:

σ = Softmax

vector z = Input vector of the given data.

e^(zi) = Exponential function for input vector.

e^(zj) = Exponential function for output vector.

K = Number of classes.

Conclusion: Thus, we successfully developed and evaluated logistic regression models for multi-class classification tasks using machine learning.