UNIT-4

CLASSIFICATION

Probabilistic Classification:

Probabilistic classification in machine learning is a method of classifying data points based on the probability that they belong to a certain class. This is done by assigning a probability value to each class for a given data point, and then choosing the class with the highest probability as the predicted class for that data point. This approach is often used in situations where the data is noisy or uncertain, or when there are multiple possible classes for a given data point.

One can have a classification model that is probabilistic in nature, in particular, these models can give probability of an instance belonging to positive or negative class. Then it is up to us to decide whether the instance is positive or negative based on the probabilities given by the model. The two commonly used forms of probabilistic models are:

1. **Generative Models:** A generative model includes the distribution of the data itself, and tells you how likely a given example is. For example, models that predict the next word in a sequence are typically generative models An example of such a classification model is Naive Bayes.
2. **Discriminative Models:** The discriminative model refers to a class of models used in Statistical Classification, mainly used for supervised machine learning. These types of models are also known as conditional models since they learn the boundaries between classes or labels in a dataset. These models are not capable of generating new data points like Generative Models. An example of such a classification model is Logistic Regression.

LOGISTIC REGRESSION:

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not. It is a kind of statistical algorithm, which analyze the relationship between a set of independent variables and the dependent binary variables. It is a powerful tool for decision-making. For example email spam or not.

The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

Terminologies involved in Logistic Regression:

- **Independent variables**: The input characteristics or predictor factors applied to the dependent variable's predictions.
- **Dependent variable:** The target variable in a logistic regression model, which we are trying to predict.
- **Logistic function**: The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.
- **Odds:** It is the ratio of something occurring to something not occurring. it is different from probability as probability is the ratio of something occurring to everything that could possibly occur.
- **Log-odds**: The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of the dependent variable are modeled as a linear combination of the independent variables and the intercept.
- **Coefficient:** The logistic regression model's estimated parameters, show how the independent and dependent variables relate to one another.
- **Intercept:** A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
- **Maximum likelihood estimation:** The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

WORKING:

Probability always ranges between 0 (does not happen) and 1 (happens). Using our Covid-19 example, in the case of binary classification, the probability of testing positive and not testing positive will sum up to 1. We use logistic function or sigmoid function to calculate probability in logistic regression. The logistic function is a simple S-shaped curve used to convert data into a value between 0 and 1.

$$h\Theta(x) = 1/1 + e - (\beta o + \beta 1X)$$

'hΘ(x)' is output of logistic function , where 0 ≤ hΘ(x) ≥ 1
'β1' is the slope
'βo' is the y-intercept
'X' is the independent variable

*(βo + β1*x) - derived from equation of a line Y(predicted) = (βo + β1*x) + Error value*

ASSUMPTIONS

Before diving into the implementation of logistic regression, we must be aware of the following assumptions about the same −

- In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1.
- There should not be any multi-collinearity in the model, which means the independent variables must be independent of each other .
- We must include meaningful variables in our model.
- We should choose a large sample size for logistic regression.

TYPES OF LOGISTIC REGRESSION

The three types of logistic regression

- **Binary logistic regression** - When we have two possible outcomes, like our original example of whether a person is likely to be infected with COVID-19 or not.
- **Multinomial logistic regression** - When we have multiple outcomes, say if we build out our original example to predict whether someone may have the flu, an allergy, a cold, or COVID-19.
- **Ordinal logistic regression** - When the outcome is ordered, like if we build out our original example to also help determine the severity of a COVID-19 infection, sorting it into mild, moderate, and severe cases.

STEPS FOR LOGISTIC REGRESSION:

- Define the problem: Identify the dependent variable and independent variables and determine if the problem is a binary classification problem.
- Data preparation: Clean and preprocess the data, and make sure the data is suitable for logistic regression modeling.
- Exploratory Data Analysis (EDA): Visualize the relationships between the dependent and independent variables, and identify any outliers or anomalies in the data.
- Feature selection: Choose the independent variables that have a significant relationship with the dependent variable, and remove any redundant or irrelevant features.
- Model building: Train the logistic regression model on the selected independent variables and estimate the coefficients of the model.
- Model evaluation: Evaluate the performance of the logistic regression model using appropriate metrics such as accuracy, precision, recall, F1-score, or AUC-ROC.
- Model improvement: Based on the results of the evaluation, fine-tune the model by adjusting the independent variables, adding new features, or using regularization techniques to reduce overfitting.
- Model deployment: Deploy the logistic regression model in a real-world scenario and make predictions on new data.