

Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

Week - 10

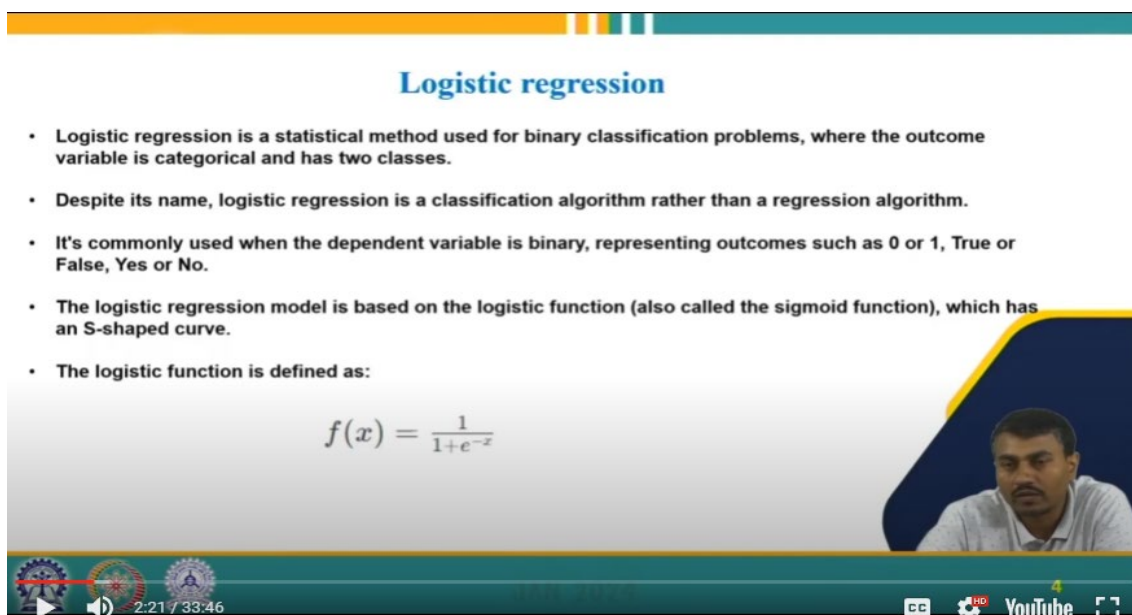
Lecture - 47

Logistic Regression

Welcome back to my course. Today, in this lesson, we are going to discuss logistic regression. It is a supervised machine learning method. So, let's deal with that. So, in this lecture, we are going to cover the following. Firstly, we will introduce you to logistic regression. Then, what are the assumptions required to use this machine learning method? And how to assess the performance of this model, which is an evaluation metric.

Then, we will discuss the advantages and disadvantages of this method. And finally, we will conclude this lesson by showing you some potential applications in the mining industry. So, what is logistic regression? Logistic regression is a statistical method used for binary classification problems. Whether you are going out to market or not, you are staying home.

Yes or no? So, in this kind of situation, daily, if you have some data to categorize based on the situation in the mining, yes or no? Or the probability of something occurring or not occurring? 0 and 1. Or whether it is good or bad. Only two distinct classes, two distinct categories. When the outcome is like this.



Logistic regression

- Logistic regression is a statistical method used for binary classification problems, where the outcome variable is categorical and has two classes.
- Despite its name, logistic regression is a classification algorithm rather than a regression algorithm.
- It's commonly used when the dependent variable is binary, representing outcomes such as 0 or 1, True or False, Yes or No.
- The logistic regression model is based on the logistic function (also called the sigmoid function), which has an S-shaped curve.
- The logistic function is defined as:

$$f(x) = \frac{1}{1+e^{-x}}$$

2:21 / 33:46

CC BY-NC-SA YouTube

So, logistic regression is a suitable method for this classification. Though the name suggests regression, logistic regression is the algorithm primarily used for the classification problem, particularly for the binary class, where the outcome is 0, true, false, yes, no, like that. That is done using the S-separate curve. The S-separate curve does that.

$$f(x) = \frac{1}{1+e^{-x}}$$

So, logistic regression is a suitable method for this classification. So some of the data points might be on this side, some of the data on this side, and some of the data might be on this side. So, this method aligns the data on two sides of this particular curve, the S-separate curve. This is mathematically done using this sigmoidal function one by e to the power minus x. I am assuming that this is for only a single variable, x.

There might be a possibility that x has several x, x1, x2, x3, and up to xn. So, it is a simplistic representation of the logistic regression curve. This is called the logistic regression curve. So, this is the function we are estimating. Then what will we do? We will calculate the probability of this particular, assuming a new value and data.

This is a new data. Now, whether this new data belongs to this class or belongs to this class depends on what? Depending on this particular value line, if it is above 0.5, then it is class 1. This is class 1, class 0, so if it is above 1.

5, so it will be classified as class 1. So, this logistic regression method convincingly classifies the data points into two distinct classes. So, this is a handy tool. It is not computationally expensive, is straightforward, and can efficiently classify a linear data set using this method without requiring much learning. So, this is the significant advantage of why the machine learning community still relies on these logistic regression methods. Now, let us discuss the workflow of this particular method because this is ultimately a model, a process, and a technique that you have to follow.

So, to apply that method, you need to follow the rules and the methods of this particular algorithm to be used for a specific data set. So first of all, we have to prepare the data. As I mentioned in the last slide, the data is a binary class: yes, no, true, false, or 0, 1. So, data should be represented in a single observation, with each column of a different variable, and the target variables should be binary: 0, 1, true, false, or yes, no. So based on that, we are trying to train the model because ultimately, you have to train the model on the data, and based on the training, we will try to minimize the error in the training data example because training data is class is already known.

So, we will fine-tune the model during the model's training process. Now, the third stage of this workflow is to evaluate the model. We have to consider the model based on the unseen data, a new data set, or a new data point, as I mentioned in the last slide. Whether this particular data point is correctly classified to its suitable class or not, we will evaluate the model, and we can use it for the prediction.

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

So, based on the new data, it will correctly classify the particular data into its actual class. These are the four stages we follow for the logistic regression methods. This is a broader representation of the logistic regression equation where the probability of Y is equal to 1 is one by one plus e to the power minus beta 0 plus beta 1 X 1 beta 2 X 2 up to beta n and X n. So here, the probability Y is equal to 1, which is the probability of the dependent variable being in class 1. As I represented in the sigmoidal curve, the above one is class 1, and this is class 0.

So if it is above 50 percent 0.5, it will be classified as class 1. Epsilon e is the natural logarithm e, beta 0 is the intercept, and beta one beta two up to beta n is the coefficient associated with the independent variable of X 1 X 2 up to X n. So this logistic regression model predicts the log odd or logit of the probability of the event Y is equal to 1. So, these log odds are then transformed into probability using the logistic function, and the coefficients are estimated using methods such as maximum likelihood estimation.

Our goal is to find this value and maximize the likelihood of the observed data. So, this logistic regression method has a wide application in social science and engineering problems, particularly in classifying emails, whether they are spam emails or if they are to be sent to the inbox. So, this kind of classification can quickly be done using this method. Algorithms and codes are available in Python and Scikit-learn, as well as on our platform. So, let us see the difference between linear regression and logistic regression.

Suppose there is a data point like this on the left side. So these are the data points. These are the data points. Linear regression will try to fit the maximum between these points. However, this is not a correct way of classifying or segregating these two classes of data points into two classes.

Here, it is a correct representation. This S separates the curve significantly and categorically subdivides the data set point into class 0 and class 1. So, this is a better way of classifying the data when the data set is like this. So, let us see one example. Here, we have data of 0, 1, 2, 3, 4, and 5 in the feature space.

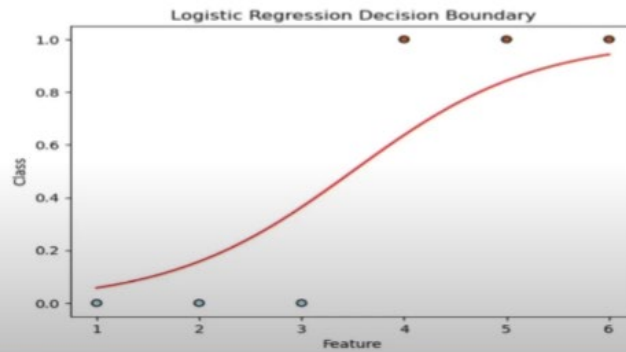
Simple linear regression

Feature: $X = 1, 2, 3, 4, 5, 6$

Class : $Y = 0, 0, 0, 1, 1, 1$

Hence, we have two classes : 0 and 1

After applying logistic regression:



The class is for 1, 2, 3, and 0, and the 4, 5, and 6 are for class 1. Because starting from 4, if we go there, it is above 0.5. So, it is classified as class 1. So here we have only two classes, 0 and 1.

And we are applying the regression model. This regression curve can substantially classify the data into two classes, 0 and 1. So, this is another better representation when the number of data points is much more. These are classified into classes A and B, the logistic regression curve. This is the sigmoidal function used to get the class probability between 0 and 1.

So, similar to the other machine learning model, there are some assumptions in this logistic regression model. So, the first assumption is the binary outcome. Because the logistic regression is specifically designed for binary classification tasks. However, the dependent or target variable has only two possible outcomes or classes. So, alternative approaches such as multinomial logistic regression or one versus-rest classification may be more appropriate if there are more than two classes.

The second is the linearity of log odds. The logistic regression assumes that the relationship between the featured independent variable and the dependent variable's log odds is linear. This is a fundamental assumption, and if this assumption is violated, then there might be a problem in significantly segregating these data into two classes using these methods. So, these assumptions imply that the log odds of the outcome variable are a linear combination of the predictor variable.

Independence of observation. Logistic regression assumes that the observation data points are independent of each other. This is a critical observation and assumption we must follow and obey. In other words, one observation does not affect the occurrence of different observations. This is crucial for estimating the validity of the statistical inference that we are following. And the following assumption is no multicollinearity.

Logistic regression assumes little or no multicollinearity among the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated. High multicollinearity can lead to unstable parameter estimates and inflate the standard error. Another critical assumption is the large sample size, though there are no strict guidelines for this method's sample size requirement. However, it is always better when the data set is significant; it is always beneficial for getting a reliable solution, estimation, and accurate predictions.

Though smaller, the size can often lead to overfitting problems or unstable estimates. Linearity in logit. Logistic regression assumes that the relationship between the independent variable and the log odds of the dependent variable is linear. So these imply that the effect of changing variables, changing one predictor variable, is constant across all levels of other predictor variables. The logistic regression method is susceptible to outliers.

So, there should be a manageable number of outliers, especially if there is a significant influence on the model parameters. We must check for outliers and consider strategies such as robust regression techniques or data transformation to mitigate their impact because logistic regression is very sensitive to outliers.

Absence of perfect separation. Perfect separation occurs when the values of one or more independent variables perfectly predict the outcome variable, resulting in infinite parameter estimates. So, the logistic regression may fail to converge or produce unreliable estimates in cases of perfect separation. So techniques such as fifth penalized likelihood or separation diagnostic can be used to address this issue. So, it is essential to note that violating this assumption can lead to biased estimates, poor model performance, or erroneous conclusions. So, we have to follow and ensure that when applying this logistic regression for a data set, we are looking to classify the data into two classes.

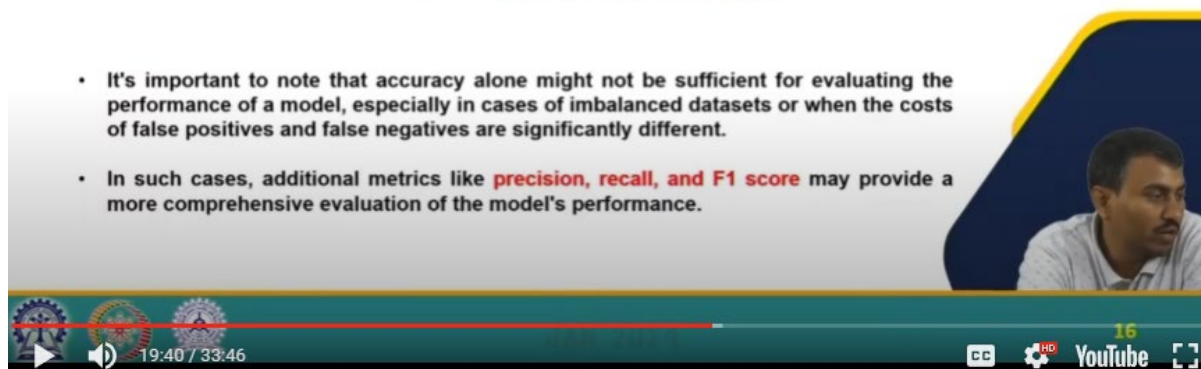
We should also consider religiously these assumptions. How can we assess the performance of the method, how efficient is this method, how is this method performing, whether this performance is up to the level, and what is the accuracy? So these are critical parameters, and in logistic regression, we have several metrics we will now discuss. So, one of the significant metrics is accuracy. So, the accuracy of the logistic regression model is a measure of its performance in correctly predicting the class levels of the data set.

Evaluation Metrics

- The **accuracy** of a logistic regression model is a measure of its performance in correctly predicting the class labels of the dataset.
- It's calculated as the ratio of the number of correctly predicted instances to the total number of instances in the dataset. Mathematically, accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- It's important to note that accuracy alone might not be sufficient for evaluating the performance of a model, especially in cases of imbalanced datasets or when the costs of false positives and false negatives are significantly different.
- In such cases, additional metrics like **precision, recall, and F1 score** may provide a more comprehensive evaluation of the model's performance.



It calculates the ratio of correctly predicted instances to the model number of the cases in the data set. So, the number of correct predictions is the accuracy divided by the total number of predictions. And it is a reasonable estimate that if it is nearly 80%, 90%, or 95%, it is good. However, with this data, some data between percentages between 75%, 70%, and 80% is not sufficient to say about the performance of this particular method. It will not alone justify, or it will explain whether, yes, the method performance is up to the mark or not.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

So, we need further studies because false positives and negatives significantly affect predictions. So, in those cases, we have to rely on new parameters called precision, recall, and F1 score. So, let us see what all these recall precision and F1 scores are. So precision is another metric that we often use for the performance analysis of the logistic regression model for the binary classification task. It basically measures the correctly predicted positive instances, true positives, and positives, as well as whether they are true.

So, out of all instances, it was predicted as positive, including both true and false positives. So, the mathematical representation of the precision is true positive divided by true positive plus false positive. So, these false positives mean they are misclassifying a negative instance as positive. So, it is an error. So, this ratio precision is an excellent estimate, and this particular value ranges from 0 to 1.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

So, the higher the value, the better the precision. It is essential to interpret precision, other metrics, such as recall and accuracy, and the F1 score. So, what is a recall? So recall is very sensitive to actual positive rate, and it is a metric used to evaluate the performance of the classification model, particularly in the binary classification task. And it measures the proportion of correctly predicted valid positive instances out of all actual positive instances. So, a true positive is divided by a true positive plus a false negative. So, a false negative misclassifies a positive example as unfavorable.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

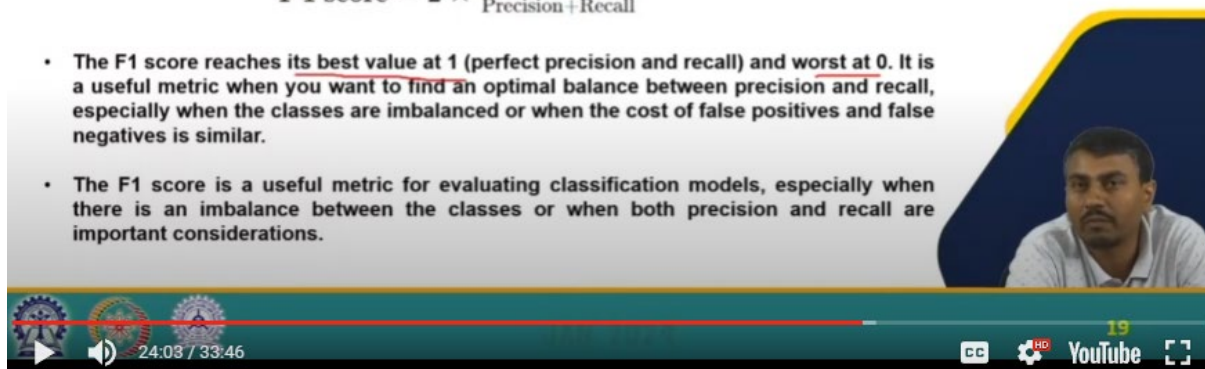
It is also an error. So recall is useful when the cost of a false negative is high. We also want to minimize the number of false negatives in our prediction. So, the value lies between 0 and 1; a higher value indicates better recall. So, it is essential to interpret recall along with other metrics such as precision, accuracy, and F1 score to analyze the model's performance comprehensively. So, what is the F1 score? F1 score is a combined metric representing both the precision and recall into a single value and provides a balanced evaluation of a classification model performance, particularly in binary classification tasks.

Evaluation Metrics

- The **F1 score** is a metric that combines both precision and recall into a single value, providing a balanced evaluation of a classification model's performance, particularly in binary classification tasks.
- It is the harmonic mean of precision and recall, calculated as:
- Mathematically, recall is defined as:

$$F1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- The F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. It is a useful metric when you want to find an optimal balance between precision and recall, especially when the classes are imbalanced or when the cost of false positives and false negatives is similar.
- The F1 score is a useful metric for evaluating classification models, especially when there is an imbalance between the classes or when both precision and recall are important considerations.



$$F1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

So, it is a harmonic mean of precision and recall. So, the F1 score is nothing but two into precision into recall divided by precision plus recall. So, the F1 score reaches its best value of 1, with perfect precision and recall, and worst at 0. So, finding an optimum balance between precision and recall is a handy metric, especially when the classes are imbalanced or when the cost of false positives and false negatives is similar. So, the F1 score is a valuable metric for evaluating the classification model, especially when there is an imbalance between the classes or when both precision and recall are important considerations. So, let us see what are the advantages of logistic regression.

So, the significant advantage of logistic regression is that it is straightforward and interpretable. It is straightforward and interpretable and relatively easy to use and interpret in terms of feature extraction. It is also beneficial. It is computationally less expensive and very efficient in training in a large data set. However, we might have to go for another classification model for a very complex model and data set. However, a simple model will perform better than a neural network and this algorithm.

Probabilistic prediction, logistic regression output, is the probability that an instance belongs to a particular class rather than just binary prediction. So, this can be particularly useful when assessing the confidence of your model prediction. And it is significantly less prone to overfitting. So, with the help of proper regularization techniques like L1 or L2 regularization, logistic regression can handle overfitting reasonably well, especially in high dimensional space. Regarding the importance of features, logistic regression provides a clear indication of the relative importance of each feature in predicting the outcome.

This can be valuable for feature selection and understanding the underlying relationship between features and the target variable. So, let us see what are the disadvantages of this method. Limited expressiveness, so logistic regression assumes a linear relationship between the feature and the log odds of the response. So this means it may not capture more complex relationships in the data. So, if the relationship is highly non-linear, logistic regression might underperform compared to more flexible models like decision trees or neural networks.

Assumption of linearity, so logistic regression is based on the belief that the independent variable and the logit transform mission of the dependent variable is linear. If this assumption is violated, model prediction will be inaccurate. Binary classification only, logistic regression is inherently designed for binary classification problems. Sensitive to outliers: I have already said that logistic regression can be very sensitive to outliers, especially if the outliers are present in the independent variable. So, outliers can disproportionately influence the parametric estimation process, leading to biased results.

Feature engineering dependency: the performance of logistic regression heavily relies on feature engineering. The model's predictive performance may suffer if informative features are

not selected or irrelevant features are included. So overall, logistic regression is a valuable tool in the machine learning practitioners' toolbox, particularly for binary classification tasks where interpretability and efficiency are essential. So, let us see some of the applications and potential applications of logistic regression in the mining industry. Exploration is an integral part of the mining and exploration targeting because the mining starts based on this exploration.

Whether mining will be done, whether mining will be economical, and whether the resource has a sufficient reserve. So, to estimate that kind of situation, we cannot drill many holes because that is very costly. So, based on the data pattern, I can substantially predict the presence of the ore body using the data. So, these data and their uses and classification are helpful examples in the mining industry to apply logistic regression. So, the historical exploration data, including the drilling results and known mineral occurrences, can serve as level data for training the logistic regression model.

The model then predicts the mineralization probability in unexplored regions and guides exploration efforts toward high-potential areas. Risk assessment for mine safety is critical because ensuring the miners' safety and the mining industry's different activities is very important. So we can apply this method as well. So, logistic regression can be employed to assess the risk of accidents or incidents occurring during mining operations. Here, the logistic regression model can identify patterns and risk factors associated with accidents by analyzing the historical and incident data.

So, the model can then predict the likelihood of an accident occurring under specific conditions, allowing for proactive measures to mitigate risk and improve safety protocols—mineral deposit classification. Similar to the exploration, different types of mineral deposits require different mining technology and processing methods. So, classifying mineral deposits accurately is crucial for efficient resource extraction and processing. So, by training the logistic regression model with on-level data representing different mineral deposit types such as porphyry copper, epithermal gold, or massive sulfide deposit, the model can learn to distinguish between them.

This classification aids in targeting exploration efforts and planning operations accordingly—mineral prospecting mapping. So, mining companies need to prioritize areas for exploration based on their potential for hosting economically viable mineral deposits. So here, the logistic regression model can predict the probability of similar occurrences in unexplored areas by analyzing the occurrences of mineral deposits along with the geological features associated with mineralization. So, this information helps identify prospective regions for further exploration and investment.

Environmental impact assessment is a critical point. We all know that mining operations can have significant ecological impacts, particularly habitat destruction, water pollution, air

pollution, and so many other environmental impacts. So, by analyzing the project locations, terrain characteristics, and proximity to sensitive ecosystems, proposed mitigation measures and how they affect the logistic regression model can predict the probability of various environmental impacts occurring. So, this is a precious tool in the mining industry to assess the conditions and how these operations will impact the mining environment. By applying this method, the mining industry can plan well ahead of time for the reclamation works and the mitigation works to be done on the environment so that the climate is less affected by the mining industry.

So these are the references. So, let me summarize in a few sentences what we have covered in this lesson. So, we have discussed and introduced a new classification method: logistic regression. It is a supervised method, and we have discussed its assumptions and the different evaluation metrics used for measuring the performance of this model. We have also discussed the advantages and disadvantages of this model and a few real-world examples or applications of these logistic regression methods in the mining industry. Thank you.