Course Name: Machine Learning and Deep learning - Fundamentals and Applications

Professor Name: Prof. M. K. Bhuyan

Department Name: Electronics and Electrical Engineering

Institute Name: Indian Institute of Technology, Guwahati

Week-5

Lecture-21

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my earlier classes, I discussed the concept of linear regression. Linear regression is nothing but fitting of a line or fitting of a curve or polynomial between the sample points. Today I am going to discuss the concept of logistic regression. Linear regression handles the problem of regression. On the other hand, logistic regression handles the problem of classification.

So, it is a classification problem and logistic regression is a statistical model which can analyze and predict the outcomes of a binary event and that is the fundamental concept of the logistic regression. So, it is a statistical model, which can analyze and predict the outcomes of a binary event. So, that is the fundamental of logistic regression. So, let us start this class,

we will mainly discuss the concept of the classification and the concept of the logistic regression.

So, what is actually the classification? So, classification problem actually already I have discussed, I can define like this goal is to learn a mapping from input is x from input x to outputs y. So, input is x and output is y and where y I am considering c number of classes 1 to c. So, you can see the goal is to learn a mapping from input x to output y and y I am considering c number of classes. So, where c is the number of classes. So, that is the definition of a classification and suppose if I consider c is equal to 2 that means only two classes are considered that means it is a binary classification problem.

So, binary classification. So, corresponding to c is equal to 2 it is a binary classification that means y has only two outcomes or values 0 and 1. So, only two classes and if I consider c is greater than 2 and if I consider c is greater than 2 that is actually the multiclass

classification, multiclass classification. So, the logistic regression considers this classification problem linear regression considers the fitting of a line or the fitting of a curve or fitting of a polynomial between the sample points. So, how to consider this classification problem by logistic regression let us discuss about this.

So, move to the next slide. So, in this case I am showing one example the example is given body mass index this is also called BMI. Does a patient have type 2 diabetes. So, type 2 diabetes T to D suppose. So, this type 2 diabetes depends on the body mass index.

So, what is actually the body mass index BMI is actually body mass divided by square of the body height. So, that means body mass divided by square of the body height and the unit will be kg per meter square. So, this is the definition of the BMI. Now, the condition is our input is the BMI and we have to determine whether the patient have type 2 diabetes or not that I have to determine. So, for this problem can we use the linear regression concept.

So, in this figure you can see I am showing BMI the body mass index and also the T to D that is the type 2 diabetes and you can see I am showing the sample points. So, corresponding to body mass index these corresponding to the body mass index these you can see the type 2 diabetes is 0 and corresponding to the body mass index if you see here this body mass index corresponding to this the type 2 diabetes is 1. That means it is a binary classification problem. So, T to D it may be 0 or it may be 1 based on the BMI the body mass index.

So, whether this problem can be considered by linear regression formulation that we have to see.

So, suppose I have some readings like this the BMI versus type 2 diabetes. So, I have some data suppose BMI is 23 and the type 2 diabetes is 0 that means and there is no type 2 diabetes. Suppose body mass index is 26 then type 2 diabetes is 1 that means the patient has diabetes and similarly if I consider another input 16 type 2 diabetes is 0 and like this I have this data suppose 31 that means corresponding to the body mass index 30 the type 2 diabetes is 1. So, you can see I am plotting these points suppose and you can see it is a binary classification problem.

So, you can see here we have 2 variables one is the X and another one is Y.

X is the independent variable and Y is the dependent variable. So, here X is the independent variable and Y is the dependent variable. So, if I consider BMI I can consider it as X that is the independent variable and if I consider the type 2 diabetes then this I can consider as output and this output is Y that is the dependent variable. So, whether this

problem can be considered by considering the linear regression that is the classification as a regression. So, let us consider this problem.

So, here you can see I am considering the same problem that is I am showing the BMI versus type 2 diabetes. Now we are considering it as a linear regression problem. So, for this we can find or we can fit a line between these data points. So, you can see what is the importance of the logistic regression you can understand now.

The classification is considered as a regression.

So, in this case if I consider it as a regression problem I have to fit a line I have to fit a line or maybe hyper plane to this data points to data. So, corresponding to this we have to find the weights. So, if I see this figure you can see this line I am getting that is the from linear regression I can get this line between the sample points between the data points and you can see I have to find the weights. So, W is the weight, weight is already in my linear regression class I have shown. It can be obtained like this $(\phi^T\phi)^{-1}\phi^T t$, t is the target vector.

So, this expression already I have shown in my lecture in linear regression. So, that is also it is equivalent to actually $(X^TX)^{-1}X^TY$. So, I have to determine the weights. The weight is W and you can see I am determining the best fit line this is the best fit line between the data points. Now, if I consider this line for classification.

So, what will happen I have to estimate Y corresponding to X, X is my input and Y I have to estimate. So, I have to estimate Y that is the output and it has 2 outcomes either 1 or 0. So, 1 means the patient has diabetes and 0 means the patient has no diabetes. So, that is the interpretation of this.

So, Y_{test} we can determine from the weight vector W_t and input is X_{test}.

So, I can determine the Y_{test} and how to do the prediction. So, prediction we can do like this the prediction will be predict Yes that means the patient has diabetes if Y_{test} is greater than suppose 0.5. So, in the figure I have shown it is the 0.

5 is here. So, I can say the prediction is Yes if Y_{test} is greater than 0.5 and also you can see the predict no if no means the patient has no diabetes. So, if no that means Y_{test} is less than Y_{test} is less than 0.5. So, this is the decision criteria.

So, predict Yes if Y_{test} is greater than equal to 0.5 and predict No if Y_{test} is less than 0.5. So, if I consider this one here you can see this sample number 1 sample number 2 okay sample number 3 sample number 4 sample number 5 and sample number 6.

So, this sample should be correctly classified.

So, this sample should be correctly classified and so we have to see which are the samples which are not correctly classified. So, here you can see this sample and also this sample these are not correctly classified by this principle. So, that means I can say it is misclassification. So, you can see the sample number 1 2 3 4 5 6 they are correctly classified this regression line. So, line is defined by the weight vector,

but the sample points I can show the sample points 7 and 8 these are not correctly classified by this principle that is the misclassification.

Corresponding to the point 1 the prediction is yes corresponding to the point 2 the prediction is yes like this corresponding to the point 3 the prediction is no corresponding to the point 4 the prediction is no, but if you see corresponding to the point 7 it should be actually yes, but now the predicted is as a no. So, it is predicted as no and corresponding to the sample point 8 and there is also misclassification. So, in this case you can see for these 2 sample point 7 and 8 we have misclassifications. So, you can see you can see that linear regression is not possible to solve this problem.

So, that is why we have to consider logistic regression.

So, what is the fundamental principle of the, So, let us move to the next slide. So, logistic regression because you have seen that there is a misclassification. So, now we are we are considering the logistic regression for this problem and this is the probabilistic approach to classification. So, I can say the probabilistic approach to classification. So, the probability of y is equal to 1 given x.

So, we are determining like this and corresponding to this what is $f(x)$. So, what $f(x)$ I should consider. So, for this I am considering one logistic function that is $f(s) = \frac{1}{1+e^{-s}}$. So, this function and this is very similar to the sigmoid function actually it is a sigmoid function we are considering.

So, based on this logistic function we can consider the logistic regression model.

So, what is the logistic regression model logistic regression model. So, my model is $P(y = 1|x)$. So, it is $\frac{1}{1+e^{-w^T x}}$ we are considering this model that is that this model is called the logistic regression model. So, in this model you can see I am considering a linear term inside the sigmoid function. So, this function this function is nothing, but a sigmoid function and in this formulation of the logistic regression model we are considering a linear term $w^T x$ that is a linear term inside the sigmoid function.

So, this is in the figure I have shown one example of the sigmoid function that is the $f(s)$. So, in the x direction I am plotting s and in y direction I am plotting $f(s)$ that is the logistic function that is the sigmoid function. And in the expression of the logistic regression model you can see we are considering a linear term inside the sigmoid function based on this logistic model can I do the classification. So, the previous problem again we are considering and with the help of this logistic regression model can I classify because the problem is the binary classification problem.

So, let us move to the next slide. So, you can see the problem is the problem $P(y = 1|x) = \frac{1}{1+e^{-w^T x}}$. So, in my in my previous slide I have shown this one this is the logistic regression model. So, here you can see if I consider this a sigmoid function this is the sigmoid function. So, with the help of this sigmoid function I can perfectly classify all the samples. So, you can see the sample number 1, sample number 2, sample number 3, sample number 4, sample number 5, sample number 6, sample number 7 and sample number 8.

So, all these sample points are correctly classified. So, you can see if it is greater than 0.5 the prediction is yes if it is less than 0.5 then the prediction is no.

So, that means we can perfectly classify all the sample points.

And in this case one important point is we are not using the measure the sum of squared error for fitting in case of the linear regression we consider the sum of squared error condition for fitting of a line or fitting of a polynomial. But in this case we are not considering this measure the measure is the sum of squared errors. So, that is we are not considering. So, you can see perfectly we can do the classification with this logistic regression model. So, that is the fundamental concept of the logistic regression.

So, now let us discuss about the formulation of this problem. So, linear regression assumes a normal distribution with a fixed variance and mean given by the linear model. So, in case of the linear regression that already we have discussed we considered normal distribution with a fixed variance and mean given by a linear model.

So, actually this mathematically I can represent like this $P(y|x)$.

So, it is actually a normal distribution. So, mean is given by a linear model $w^T x$ and variance is σ^2. The concept is like this we have this sample points in case of the linear regression and the problem is fitting of a line or fitting of a curve between the sample points. So, if you consider this distribution of this data so that we can consider as normal distribution. And after this we considered the measure the sum of squared error we

considered for determining the best fit line or for determining the best fit curve or the polynomial we considered that measure that is the sum of squared error. But in case of the logistic regression we have to consider the Bernoulli distribution.

So, let us move to the next slide in case of the logistic regression in case of the logistic regression we considered the Bernoulli distribution. So, this Bernoulli distribution we are considering with parameters given by logistic transform of linear model. In case of the logistic regression we consider a Bernoulli distribution with parameter given by logistic transform of a linear model. So, mathematically I can represent like this $P(y|x)$. So, it is the $B(y|\theta(x) = \frac{1}{1+e^{-w^T x}}$ is nothing but the that logistic regression model.

So, you can see it is a Bernoulli distribution and we are considering $\theta(x) = \frac{1}{1+e^{-w^T x}}$ that is nothing but the logistic regression model. So, this Bernoulli distribution is defined like this $P(1) = \theta$ and $P(0) = 1 - \theta$ that is the Bernoulli distribution and for θ, I have 2 values θ is either 0 or 1. So, equivalently I can write $P(y) = \theta^y(1-\theta)^{(1-y)}$ for Y again output has 2 values 0 and 1 because I have only the binary outcomes either 0 or 1. So, that is the concept of the logistic regression and we are considering the Bernoulli distribution.

Now, let us consider about the training. So, how to do the training in case of the logistic regression. So, logistic regression is a supervised technique and we have to do the training and training we can consider as maximizing likelihood estimation. So, maximizing so that I can write the training I can consider as maximizing likelihood estimation.

So, that is the maximum likelihood estimation.

Assuming independence of probability of data. So, based on this I can write $P(y_1, \ldots, y_n|x_1, x_2, \ldots, x_n)$ is equal to because we are considering the independence condition. So, it is $\Pi_{i=1}^{n} P(y_i|x_i)$ okay, and if I consider the Bernoulli distribution. So, in case of the Bernoulli distribution $P(y_i|x_i)$ will follow the Bernoulli distribution.

So, $\theta(x_i)^{y_i}(1-\theta(x_i))^{1-y_i}$.

So, where $\theta(x)$ is nothing, but the the regression model the logistic regression model $\frac{1}{1+e^{-w^T x}}$.

So, you can see so training means we have to maximize this expression the expression number 1 suppose if I consider expression and this expression is suppose 1. So, training means I have to maximize the expression number 1 with respect to the weight vector W. So, to consider this so instead of maximizing the likelihood we may consider minimizing

negative log likelihood. So, what is the negative log likelihood I will explain in my next slide. So, what is the negative log likelihood this is actually $-log\Pi_{i=1}^{n} P(y_i|x_i)$ this multiplication is converted into addition because we are considering the $-\sum_{i=1}^{n} logP(y_i|x_i)$ and that is also equal to $-\sum_{i=1}^{n} log(\theta(x_i)^{y_i}(1-\theta(x_i))^{1-y_i})$, just I am putting the values that is from the Bernoulli distribution I am just putting the value.

So, that is equal to $-\sum_{i=1}^{n}(y_i log(\theta(x_i)) + (1-y_i)log(1-\theta(x_i))$. So, you can see this mathematics is a very simple mathematics. So, what is the negative log likelihood it is nothing but $-\sum_{i=1}^{n}((y_i-1)w^T x_i - log(1+e^{-w^T x_i}))$. So, this is the negative log likelihood NLL.

So, this is the training principle we have to consider minimizing the negative log likelihood.

So, this minimization problem we can consider from the cross entropy point of view. So, let us see what is the cross entropy. So, we can also determine the cross entropy what is the cross entropy. So, that is minimizing from cross entropy point of view.

So, from this point of view I can also consider this problem. So, cross entropy is a method for comparing two distributions and the cross entropy is a measure of a divergence between a reference distribution and the estimated distribution. So, I can write here cross entropy is a measure of a divergence between a reference distribution and an estimated distribution. So, cross entropy already I told you. So, it is a method for comparing two distributions.

So, mathematically how to write this one H that is the cross entropy one distribution is g and another one is $\hat{g}$.

So, g is the reference distribution here g is the reference distribution and $\hat{g}$ is the estimated distribution. So, this is equal to $\sum_{a\in A} g(a) log(\hat{g}(a))$ and A I have only two outcomes 0 or 1. So, I can write like this. So, for logistic regression minimizing negative block likelihood can be considered as minimizing cross entropy. So, I can write like this for logistic regression what is the case minimizing NLL that is the negative block likelihood can be considered as minimizing cross entropy.

So, mathematically I can show like this $g(1) = y_i$ and $g(0) = 1 - y_i$ and $\hat{g}(1) = \theta(x_i)$ and $\hat{g}(0) = 1 - \theta(x_i)$. So, I can consider like this. So, that means we can also consider the cross entropy measure.

So, we have to minimize the cross entropy and based on this we can do the training.

So, this is about the training of the logistic regression model. So, the problem is there is no close form solution. So, that is why we have to consider iterative methods. So, maybe we can consider stochastic gradient descent algorithm and that already I have discussed. So, these type of algorithms we can consider for getting the minimization or for solving the minimization problem because there is no close from solution and the problem is strictly convex that is the good news.

The good news is the problem is it is convex. So, convex it is like that convex. So, we can apply some iterative techniques like the stochastic gradient descent algorithms to get the solution. So, it is a convex problem. So, how to do the predictions finally, I can show how to do the prediction in case of the logistic regression. So, how to do the prediction with the help of this logistic regression model.

So, prediction we can do. So, if we threshold the output probability at 0.5, we can consider a decision rule, the decision rule will be something like this $\hat{y}(x) = 1$. So, my rule is like this $P(y = 1|x) > 0.5$. So, this is the my decision rule and based on this I can take a classification decision.

So, corresponding to this figure you can see we have considered the sigmoid function and with the help of this function you can see all the samples I can correctly classify and in this case the threshold is considered as 0.5. So, the output will be 1 if $P(y = 1|x) > 0.5$. So, that is the fundamental principle of the logistic regression, but the limitation is it cannot consider nonlinear data.

So, the problem is cannot consider cannot deal I can write cannot deal with nonlinear data. So, that is the one problem and also it is not very robust to outliers. So, another problem I can mention not very robust to outliers and also the problem of the overfitting.

So, sometimes overfitting may take place. So, I can write prone to overfitting. So, these are the main limitations of the logistic regression model. In this class I explained the concept of the logistic regression. So, it is a statistical method for analyzing and predicting the outcomes of a binary event. And I have explained how the sigmoid function can be considered for this classification and you can see the difference between the logistic regression and the linear regression models. So, let me stop here today. Thank you.