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-1

Lecture-4

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my last classes, I discussed the concept of statistical pattern classification. In statistical pattern classification, I have to determine $P(\omega_i|X)$

. So, ω_i is the class. So, I have C number of classes, i is equal to 1 to C and X is the feature vector. So, I have to determine the probability of obtaining a particular class given the feature vector that is determine $P(X)$ and

that is the fundamental objective of statistical pattern classification.

In the regression, we have to find the relationship between two variables. One is the independent variable and another one is the dependent variable. Suppose X is the independent variable and Y is the dependent variable. So, I have to find the statistical relationship between X and Y and that is the main objective of regression.

So, X is the observed data point. Suppose I have some data points and that is represented by X, the variable is X and I have to fit a model so that it best fit the data points, the observed data points. So, that is the objective of the regression. And if I consider that relationship, that is actually the statistical relationship between X and Y and this is not the deterministic.

In deterministic by considering a mathematical equation, I can determine a particular value.

Suppose the temperature is given in the centigrade and I can determine the temperature in Fahrenheit by considering one mathematical equation. But in the regression, what we have to consider the statistical relationship between two variables, one is the dependent variable and another one is the independent variable. And suppose X is a vector. So, I have

some components X 1, X 2 like this X is a vector, then we have to consider multiple regression.

So, let us discuss about the concept of regression.

So, already I told you that in statistical machine learning, the objective is to determine $P(\omega_i|X)$. That is the objective of the statistical machine learning. So, we have to determine this.

So, I have C number of classes and X is the feature vector.

So, I have to determine the determine $P(X)$.

But in case of the linear regression, if I consider the linear regression, I have two variables, one is Y another one is X. So, suppose I have Y that is I can say it is the dependent variable and another one is the independent variable. So, I have to find a relationship between that is the statistical relationship between X and Y and that is the regression. So, if I consider X is a suppose vector, then I will be getting them the multiple regression. So, what is the linear regression?

So, linear regression is a statistical method that allows us to model the relationship between a scalar response that is the dependent variable and one or more independent variables.

So, this can be done by fitting a linear equation to the observed data. So, suppose I have the observed data observed data points are available. So, how to do the regression by fitting a linear equation to the observed data. So, this is the objective of the regression.

So, this dependent variable is called the response, sometimes this dependent variable is called response or also it is I can say outcome and this independent variable and this is called the predictor or I can say the regressor.

So, I have to predict Y from X that is the objective of the linear regression. So, if I consider only one independent variable, then it is a case of simple linear regression. But if I consider two or more independent variables, then it is called a multiple linear regression. So, we have to find the statistical relationship between the dependent variable and the independent variables. So, now let us consider the simple case simple regression.

So, in the simple regression, what we are considering only one dependent variable and one independent variable. So, in this example, I am considering only one dependent variable, dependent variable and only one independent variable. So, that means in the simple regression, the problem is very simple and this problem is actually the line fitting,

line fitting on a 2D XY plane. So, I have some data points, suppose these are main data points.

So, I have to find the best fit line.

So, suppose if I want to fit a line, so this is the best fit line, best fit line between the data points. So, these are the data points. So, the problem is mainly the line fitting on a 2D XY plane. So, we are considering a set of points in the XY plane. So, set of points in XY plane.

So, the points I can represent like this xi, yi, these are the points xi, yi and i is equal to 1, 2 up to n. So, I am considering n number of points in the XY plane. So, what is the objective of the linear regression? The linear regression attempts to find a line in 2D which best fits the points, the points are xi, yi. So, n number of points are available in the XY plane.

The most popular method of fitting of a line is the method of least square, least square method for fitting of a line.

So, we can consider this the least square method for fitting of a line. So, what we have to consider? We have to minimize, actually this method minimizes the sum of the squares of vertical distances from each data point to the line. So, I can show pictorially. So, suppose this is the 2D plane and I have the points, the points are like this.

These are the points in the XY plane.

So, I have to represent this point that means, based on the least square method, I have to determine the line. So, this line I am determining like this between the sample points between the data points. Now, what we have to consider I have to minimize the sum of the square of vertical distances from each data point to the line. So, this is the best fit line suppose, best fit line I am considering suppose between the data points and now we are considering the distance between the vertical distance between the data points and the line. So, this is the distance between the data points and the line.

So, these distances we are considering and we have to minimize the distances all the distances we have to minimize and based on this minimization condition, I can get the best fit line. So, you can see I have considered n number of points n data points we are considering and this in this n data points that is the observed data points, I am fitting a line and that is the best fit line I want to determine and the condition is the least square method. So, it has to minimize the sum of the squares of the vertical distances from each data point to the line. So, the question is how to find the best line how to find the best line. So, we have to consider this least square method and based on this method we have to minimize the sum of the squares of vertical distances from each data point to the line.

Now, let us consider the equation of a line. So, this line is represented by two parameters one is the slope another one is the intercept. The slope is suppose a the slope is suppose a and another parameter is intercept. So, for representing a straight line I need two parameters one is the slope another one is the intercept.

So, these two parameters we are considering to determine the best fit line.

So, corresponding to this you can see and this equation of a line that means I am estimating the line. So, yi I am estimating that is the predicted value I can say $\hat{y} = a.x_i + b$. So, this is a equation of a straight line. Now, you can see I have to consider the least square method that means I have to minimize the vertical distance. So, this is my vertical distance.

So, this is the vertical distance. So, I have to consider all the vertical distances. So this is the equation of a line. So, this is suppose the predicted line $a.x_i + b$ and as per the condition that is the least square method, I have to find the error between the axial point and the line because already I told you that I have to find the vertical distance and I have to minimize the distance the vertical distance.

The error between the axial point and the line that error I can write like this ei is the error this yi is the axial and this y hat is the predicted value of y.

So, now I have to compute the error, error I have to determine error is the e.

$$E=\frac{1}{n}\sum_{i=1}^{n} ((y_i - \hat{y}_i)^2)$$

So, this is the nothing but the mean square error we are determining. So, this can be represented like this 1 by n summation from i is equal to 1 to n yi and in place of that $\hat{y}_i$.I am writing $a.x_i + b$ and whole square.

So, I am writing this. Now the objective is to find the slope a and the a and the intercept b which gives minimum error e that is the objective of the linear regression. So, we have to find a and b that is the objective which gives minimum error e. So, that is the objective. So, that is why what we are considering we are differentiating the error with respect to a this partial derivative of e with respect to a and the partial derivative with respect to b we have to determine and equating it to 0 and equating it to 0.

So, that is we are considering. So, suppose this is equation number 1 and this is equation number 2.

$$\frac{\partial e}{\partial a} = 0$$

$$\frac{\partial e}{\partial b} = 0$$

So, after putting the value of e in the above equation equation is 1. So, from equation 1 what I will be getting.

$$\frac{2}{n}\sum_{i=1}^{n} (y_i - ax_i - b)x_i = 0$$

So, this is with respect to a I am differentiating. So, that I can write like this summation it is actually summation for the variable i, i is from 1 to n xi yi minus a summation for i, i is equal to from 1 to n. So, it is xi square minus b summation i xi is equal to 0. So, corresponding to this a i xi square plus b summation i is equal to 1 to n.

So, b xi is equal to i xi yi I can write like this.

So, this is suppose equation number 3. And similarly, we can apply the same procedure for the equation number 2. So, that means the differentiation of e with respect to b and I will be getting a xi plus n b is equal to yi. So, I will be getting this one from the equation number 2. So, I am getting these two equations equation number 3 and equation number 4.

This is these are linear equations in two variables. So, from these two equations, I can directly determine the parameters a and b and this a and b I can determine and based on this I can determine the line. So, that means I can determine the base fit line I can determine. So, the base fit line between the observed data points.

So, that is the objective of the linear regression.

Now let us discuss in detail because I am considering now only the linear line fitting only we are considering a straight line, but maybe we can consider the polynomial curve fitting we can consider. So, in my next slide, I will be explaining the concept of the polynomial curve fitting. The case is more complicated than this case, because in this case only we are considering the problem of linear line fitting. So, move to the next slide. So, what is the problem statement of regression? So, I can say this is a problem statement.

The problem statement is suppose I have n observations of x. So, n number of observations. So, I can say x is equal to x 1 x 2 up to x n. So, n number of observations we are considering and suppose t is the target value all the target values we are considering corresponding to these observations t 1 t 2 t n. So, what is the goal? The goal is to exploit

training set to predict the value of t from x that is the goal.

So, I can write the goal is goal is to exploit training set to predict the value of t. So, that is I can say t hat that is the predicted value from the observations from x. So, this is the goal of this regression. So, this is a very difficult problem and this probability theory allows us to make a prediction.

So, in this case in this figure you can see we are considering 10 data points.

So, n is equal to 10 we are considering 10 data points and how to generate data this data data generation. So, this is this data space uniformly in the range of 0 and 1. So, if you see in the x axis. So, you can see x is the observation and t is the target value.

So, I can write it again. So, x and t in the x axis it is x and this observations 10 observations and t is the target value. So, these data generations what we are considering space uniformly in the range of 0 to 1 in the range 0 and 1. So, we have all these points n is equal to 10 points for this case this example and this is generated by. So, this is generated by generated by a function the function is we are considering sine suppose twice pi x. So, we are considering this sinusoidal function sine twice pi x and with the help of this we are generating this data and

maybe we can consider maybe addition of some noise maybe we can consider Gaussian noise.

So, sine twice pi x considered and maybe we can consider some noises we can consider the Gaussian noise and these noises are typically unobserved variables. So, these noises are typically unobserved variables. So, you can see this approximation if I consider these data points that can be generated by the function the sine twice pi x plus some noises we can consider. So, now how to predict these values and these are the observation points and we have to predict the value of t that is the t hat.

So, for this we can consider the case of polynomial fitting.

So, maybe we can consider a polynomial function. So, we can explain this concept in the next slide. So, what is the polynomial fitting? For this polynomial fitting we are considering a polynomial function the polynomial function may be like this. The polynomial function I can consider suppose x w, w is the weight vector. So, w naught these are the coefficients of this polynomial function plus w 2 x square plus w m x to the power m. So, which can be written like this summation j is equal to 0 to m j is equal to 0 to m w j x j.

So, I can write like this. So, in this case m is the order of the polynomial order of the

polynomial. So, these coefficients we are considering the coefficients are w naught w 1 these are the coefficients which can be represented by the weight vector which can be represented by the vector w. These are the coefficients a coefficients w naught w 1 w m that can be represented by the weight vector. So, this model we can consider for polynomial fitting. So, this is a nonlinear function of x linear function of the coefficients w.

So, maybe it can be called as a linear model and with the help of this model we can consider the polynomial fitting. So, now let us discuss about this concept what is the polynomial fitting because for polynomial fitting we have to minimize an error and that is nothing but the sum of square of the error between the predicted value for each of the data points and the target value. So, that error we have to minimize.

So, let us discuss this concept.

So, what is this error here you can see we are considering the red line is the best polynomial fit.

So, if you see this red line is the best polynomial fit and we have shown the data points and you can see the target value the target value is the t, t is the target value and x is the observation. So, in case of the polynomial fitting we are considering this is the error E w is the error that is the mean square error 1 by 2 summation n is equal to 1 to n because we are considering capital n number of data points y x n w minus t n, t n is the target value. So, what actually what is the meaning of this expression the meaning of this expression is sum of squares of the errors between the prediction of the error for each data point x n and the target value t n target value is t n. So, that means you can see this is the error suppose this is the vertical distance between the predicted value for each data points and the target value. So, this predicted value is nothing but the red line that is the best polynomial fit and you can see the distance between this one is the target value this is the target value the points in the red line that is nothing but the predicted value.

So, that distance should be minimum this vertical distance should be minimum. So, this is nothing but the error. So, for each and every data points I have to determine the error and I have to minimize. So, this is the sum of squares of the errors between the prediction for each data point x n and the target value t n.

So, this error we have to minimize. So, we have to solve this one solve by selecting or maybe I can write solve by choosing value of w w is the weight vector for which the error should be minimum E w the error is as small as possible. So, that means this regressor minimizes the error what is the error? Error already I have defined. So, it is 1 by 2 summation n is equal to 1 to n y x n that is the predicted value y is the predicted value and

t is the target value and we are considering a mean square error. So, that means the meaning is the minimizing the error on the training samples. So, I can write simply I am writing the minimizing minimizing the error on the training samples.

So, how to learn the weights the learning of the weights? So, move to this next slide learning the weights. So, how to learn the weights? So, already we have defined the error function E w is nothing but 1 by 2 summation n is equal to 1 to n y x n comma w that is the predicted value and the target value is t n square mean square we are considering. So, error function is a quadratic in coefficients w. So, this error function is a quadratic in coefficients w.

So, we have to minimize this error. So, that is why we have to take the derivative with respect to the coefficients. So, that means the derivative with respect to coefficients will be linear elements of the weight vector w. So, we have to determine the minimum error. So, this unique minimum unique minimum is denoted by suppose w star. So, this is after derivative and equating it to 0 I am getting the minimum and corresponding to this the resulting polynomial will be just y x w star.

So, this is the resulting polynomial. So, we have to learn the weights and objective is to minimize the error. So, first I am taking the derivative and equating it to 0 and we are getting the unique minimum vector that is the w star and corresponding to this I can determine the resulting polynomial that is y x comma w star. Now, how to select the order of this polynomial fit. So, move to the next slide.

Here I am showing that predictions for different model values. So, here I have shown that model order is 0 1 3 9 and we have 10 data points n is equal to 10. So, how to select the particular model. So, in the first figure if you see m is equal to 0 we are considering this green line corresponds to our observed data and you can see the predicted line the predicted line is the red line. So, that means this predicted line it cannot perfectly represent the input data the observed data. Similarly, if I consider m is equal to 1 corresponding to m is equal to 1 I am getting this red line and also in this case it is nothing but the poor representation of the input function the function is sine twice pi x because we have generated data by considering the function sine twice pi x.

But if I consider order of the polynomial is 1 m is equal to 1 then if I do the prediction based on this then it is nothing but the poor representation of the input function the function is sine twice pi x. So, it cannot perfectly represent the observed data points. Similarly, if I consider m is equal to 3 in the third figure so corresponding to m is equal to 3 I am getting the best fit corresponding to the function sine twice pi x. So, you can see the red line the red curve and the green curve. So, they are very close to each other because we are

considering the complex model m is equal to 3 is a complex model and with this model I can do the prediction perfectly almost perfectly not exactly because this red and green are exactly not exactly equal but almost overlapping.

So, that is why I can say it is a best fit corresponding to the input function the function is sine twice pi x. Suppose in the fourth case I am considering m is equal to 9 that is the more complex model we are considering. So, if I consider in the more complex model then you can see overfitting take place that is nothing but the poor representation of sine twice pi x. Here the main problem is we have limited number of training data samples that is only 10 data points we are considering and we are considering a very complex model m is equal to 9 and that corresponds to the curse of dimensionality that we have limited number of training samples but we are considering the complex model and corresponding to this you can see the overfitting take place that is nothing but the poor representation of the function sine twice pi x.

So, you can see the oscillations. So, red line you can see the red line is nothing but the red curve is nothing but the predicted one and you can see the oscillations in the prediction. So, you can see how to select the model order. So, because we have the limited number of training samples. So, we have to consider a simple model rather than a very complex model the complex model is m is equal to 9.

So, let us discuss the performance of the regressor. So, how to determine the performance of regression. The performance of regression. So, suppose during the testing we are considering 100 test point. So, test set we are considering test set of 100 points we are considering.

So for each value of m we have to evaluate the error function. So, for each value of m m is the order of the polynomial or I can say it is order of the model evaluate what I have to evaluate the error I have to determine. So, error is E star is equal to 1 by 2 n is equal to 1 to n y x n w star w star is the optimum weight that we have obtained during the training whole square and in this case what is y x w star that is nothing but I can write like this j is equal to 1 to m w j star that is nothing but the polynomial I am writing the polynomial equation I am writing x to the power j. So, for each value of m we have to evaluate this E w star we have to evaluate for training data for both training data and the test data. So, we have to evaluate this function E w we have to determine and corresponding to this we can determine the RMS error we can determine RMS error we can determine the root mean square error we can determine that is nothing but E RMS is equal to root over 2 E w star divided by n. So, in this case we are dividing by n so that means the meaning is the division by n allows different sizes of n to be compared on equal footing.

So, I am repeating this sentence because it is nothing but the normalization. So, that is the division by n allows different sizes of n to be compared on equal footing and the square root ensures that E RMS is measured in the same units as t. So, that is why we are taking the square root. So, this is nothing but the RMS error we are determining. So, this performance of the regression I can determine based on this RMS error and pictorially I can show you in the next slide here you can see the blue line represents the training and the red represents the testing. So, if I consider model order is suppose 0 or maybe 1 or 2 if you see this is the model order.

So, this the error E RMS error during the testing it is high because poor due to inflexible polynomials. So, if I consider a very very simple model then I cannot perfectly represent the observed data. So, that is why I am getting the error during the training and the error during the testing. But if I increase the model order suppose if I consider this one model order maybe 4 m is equal to 4 m is equal to 3 small error during the training and during the testing because if I consider this model order m is equal to 3 or 4 it can more or less represent the observed data. But if I increase the model order suppose m is equal to 9 that is the order is 9 then it means 10 degree of freedom.

So, if I consider m is equal to 9 that means the 10 degrees of freedom. So, what will happen during the training I can consider a very complex model and during the training you can see that is the blue line the error is minimum. But during the testing that is the red one the error is maximum because we are considering limited number of training samples. But we are considering a very complex model m is equal to 9. So, in this case this complex model cannot perfectly represent the observed data. So, that is why we are getting the high error during the testing

but in the training it is minimum the error is minimum

but during the testing I am getting the significant error.

So, you can see you can see that exactly 10 training data points and this is because of oscillations in the polynomials. So, oscillations in the polynomial means sometimes that is polynomial value will be very high again it will be very low again it will be very high like this oscillation take place. So, we have to consider a simple model but very very simple model will not be also good. So, maybe we can consider m is equal to 3 4 like this we can consider and we cannot also consider a very complex model

because we have limited number of training samples.

If I had more number of training samples then we can consider a complex model. But in this case we are considering n is equal to 10. So, corresponding to this we cannot afford to consider a very complex model. So, in summary I can write like this the complex model

models have large oscillations in the learned weight. So, this is nothing but the case of overfitting. That means overfitting is taking place because of we are fitting a higher order model with limited training samples.

So, in this example we have considered only 10 data points training samples. That means the meaning is we are making the model complex that means a complex model. So, overfitting take place corresponding to m is equal to 9 and corresponding to m is equal to 9 already I told you the oscillations in polynomial it take place. So, that can be explained in my next slide. Here you see I am considering these are the coefficients all these are coefficients in the polynomial fitting and for different polynomial orders we are considering m is equal to 0 m is equal to 1 m is equal to 6 m is equal to 9. So, corresponding to m is equal to 9 already I told you that is the overfitting take place and oscillation take place in the polynomial coefficients.

So, you can see sometimes it is a very low value again high value again very low value again high value again low value again high value like this we have these oscillations. But corresponding to the simple model like this you can see m is equal to 6 we have this coefficients, coefficients value are 0.31 7.99 like this. But if I consider a very complex model like m is equal to 9 you can see overfitting take place and you can see I have the oscillations in the values of the coefficients.

So, that is the observation. So, as m increases the magnitude of the coefficients increases at m is equal to 9 finally tune to a random noise in target values. Now how to control this overfitting? So, in my next slide I will explain how to control the overfitting. So, that is the techniques I can write the techniques techniques for controlling controlling the overfitting. So, how to control the overfitting? So, we may consider more number of data points. So, suppose if I consider n is equal to 15 I can consider or maybe n is equal to 100 that means more number of data points we can consider.

So, that means for a given model complexity overfitting problem is less severe as the size of data set increases. So, that means if I increase the size of the data set then we can afford a complex model. That means the larger the data set the more complex we can afford to fit their data. So, the meaning is I can write in a summary the larger the data set the more we can afford to fit the data. So, that means we can consider more number of training samples that means data points and corresponding to this we can consider a complex model relatively complex model.

And also you can see here in this figure we are considering n is equal to 15 and corresponding to n is equal to 15 you can see the one is the green is the that is the observed data points and we are considering by sine twice pi x and what is the predicted one the

predicted one is the red one the red curve. So, you can see almost perfectly represent the observed data points. But if I increase these data points n is equal to 100 then we can consider a complex model very complex model and you can see this red curve and the green curve they are almost coincide. So that means it is a perfect representation of the input data. So, that means we can consider a very complex model if I have the large number of training samples that is the meaning of this.

So, this one technique is we may consider the large number of training data samples. So, another technique is by considering the regularizer. So, what is the regularizer we can explain in my next slide. So, regularization of least squares. So, using relatively complex model with data sets of limited size. You can see in the previous slide the oscillation take place because we are considering a very complex model.

Now we can add a penalty term to error function to discourage coefficients from reaching the large values. So that means what we can consider add a penalty term to error function to discourage from reaching large values. So, this is the another technique we are considering a regularizer that the penalty term we are considering and corresponding to this my error function will be E w will be already I have explained this that one that is this is my original error function. So, this is my original error function and we are considering a penalty term lambda by 2 w square norm we are considering.

So, a penalty term we are considering here. So, this is the penalty term. So, this lambda what is the meaning of this lambda? Lambda determines the relative importance of the regularization term to error term. So, in brief I can write it determines the relative importance relative importance of regularization term to error term. There is the meaning of this lambda is a parameter. So, based on this we are considering this one. So, this error function the modified error function I can write like this and this w square norm is nothing but already you know this is w transpose w that is equal to w square plus w 1 square w m square.

So, we can determine like this. So, what is the effect of this regularizer? So, we can see in the next slide. So, m is equal to 9 we are considering that is a polynomial using the regularized error function the order is m is equal to 9. So, here we are considering log lambda we are considering. So, you can see ln lambda minus 18 we are considering ln lambda is equal to 0 that we are considering and in this case the lambda is equal to 0. So, that means lambda is equal to 0 means there is no regularization and if there is no regularization then oscillation will be there, but if I consider this optimum ln lambda is equal to minus 18 you can see the effect of the regularizer there is no oscillations. So, if I consider the large regularizer also then in this case also the problem is there that is not the exact representation if I see the predicted predicted one the predicted one is the red one.

So, if I consider this one that is not the exact representation of the input function the input function is sine twice pi x. So, if I consider this optimal value the optimal log lambda is equal to minus 18 then what will happen. So, it is almost exact representation of the input data and lambda is equal to 0 means there is no penalty term that is effect of regularizer is not there. So, you can see here the values of the coefficients corresponding to no regularization. So, ln lambda is equal to minus infinity that means no regularization then you can see the oscillation take place and corresponding to lambda is equal to minus 18 then in this case you have this values of the coefficients and that is the most optimal representation.

So, that is the effect of the regularizer to minimize the overfitting because the problem is because of the overfitting. So, move to the next slide. So, in this case you can see that lambda controls the complexity of the model. So, what is the impact of regularization on error because we are considering the regularizer.

So, what is the impact of the regularization on error. So, this lambda parameter we have considered actually this lambda controls the complexity of the model complexity of the model. So, it is very similar to the model parameter M the model order is M. So, what we have to consider what is the approach for this. So, for the training set we are considering a training set. So, with the help of this training set first we have to determine coefficients W coefficients of W for different values of M or lambda and after this we have to consider the validation set.

So, during the validation to optimize model complexity model complexity M or lambda. So, like this we have to consider this approach. So, for the training what we have to consider we have to determine the coefficients of W and for different values of M or lambda and for the validation what we have to consider we have to optimize the model complexity by considering M or lambda. So, this is the approach. So, these are the techniques for controlling the overfitting. So, that means I can say that partitioning I can say the partitioning data into training set to determine coefficients of this weight vector W and separate validation set to optimize model complexity M or lambda.

So, this we have to follow. So, this is the fundamental concept of the regression. So, now I will explain the mathematics behind the regression. So, how to learn the weight vector. So, what is the mathematics behind regression and actually how to learn the weight vector. So, we are given n samples of d dimension with scalar target values.

So, that means what we are given n number of training samples of d dimensions with scalar target values. So, this is given. So, I can represent like this x1 that is the

corresponding to input the target value is t1 corresponding to the second input x2 the target value is t2 like this we have given this training samples. So, xn is given and the target value is tn it is given. So, now let us define M plus 1 non-linear basis functions. So, maybe we can consider M plus 1 number of non-linear basis function phi x that is nothing but suppose phi naught x phi 1 x phi 2 x phi m x this is the non-linear basis function maybe we can consider exponential function as a non-linear basis function.

So, for a polynomial for a polynomial feed we can consider this function phi x maybe like this 1 x x to the power m. So, we can consider this polynomial feed case that is we are considering the non-linear basis functions and based on this we can consider the polynomial feed. Now, let us move to the next slide. So, we have to define the weight vector and we have to define the vector corresponding to the predicted value and also the another vector corresponding to the targets the vector of the targets. So, this is the weight vector w is the weight vector and you can see these are the components of the weight vector w naught w 1 like this up to w m this is the weight vector we are considering this is the vector of weights.

We can consider another vector that is y. So, this is y 1 y 2 y n. So, this is a vector of predicted values vector of the predicted value. So, in the first weight vector dimension is m plus 1 cross 1 the second one is the dimension is n cross 1 and the target vector is t t 1 t 2 t n n cross 1. So, this is nothing but the vector of targets. So, we are defining these vectors.

So, what is this predicted value the predicted value we can define like this the predicted values. The predicted values this is nothing but y 1 is equal to phi transpose x 1 w y 2 is equal to these are the predicted values phi transpose x 2 w like this all these we have the predicted value y n is equal to phi t x n w. So, you can see this linear combination of these basis functions gives the predicted value for each of the samples. So, we have this predicted value and how to get this predicted value the linear combination of these basis functions gives the predicted value for each of the samples.

And in this case what we have to do we have to learn m plus 1 number of weights. So, meaning is from this you can see. So, what is the objective the objective is to learn the m plus 1 number of weights that is the objective. So, you can see how to get the predicted value in terms of the basis functions. So, the linear combination of these basis functions give the predicted value for each of the samples.

Now we have to define the error function. So, in my next slide I can explain what is the error function in this case. So, move to the next slide. So, what is the error function? This error function I can write like this. So, this is my weight vector w 1 w 2 w m this basis

function it is the order is n cross m this weight vector the order is the size is m cross 1 and I am getting the predicted vector the predicted vector is y 1 y 2 up to y n.

So, what is the size? The size is size is n cross 1 because it is n up to n. So, size will be n cross 1. So, this phi I can write the phi if I write like the phi like this phi is equal to suppose phi transpose x 1 phi transpose x 2 phi transpose x n that is represented by this matrix phi n into m plus 1 the size is n into m plus 1.

So, size is n cross m plus 1. So, you can represent like this. So, we can write in this form. So, y is equal to phi w. So, from this I can write y is equal to nothing but phi into w and based on this we can write the error function. So, what is the error function? So, error function I can write the error function is E w is equal to 1 by 2 summation i is equal to 1 to n t i minus y i whole square and we can write in this form 1 by 2 t minus phi w.

So, in the vector form I can write t minus phi w. So, I can write in this form. So, that is the error function. So, this can be simplified or it can be expanded this error function now I can write like this error function E w is already I have defined it is 1 by 2 t minus phi w t t minus phi w. So, this can be expanded. So, it is 1 by 2 t transpose t minus w transpose phi t t minus t transpose phi w plus w transpose phi transpose phi w. So, I can write after expanding and also we have this expression we know this w transpose phi transpose t is equal to t transpose phi w we can write this is equal to this w transpose phi transpose t is equal to t transpose phi w.

$$E(w) = \frac{1}{2}(tTt - 2wT\phi Tt + wT\phi T\phi w)$$

So, I can write like this. So, after this I have to find the optimal weight vector. So, that is why what I have to consider I have to do the differentiation and after this it is equating to 0 then I will be getting the optimal weight vector. So, move to the next slide. So, what is the optimal weight vector? So, just I have to do the differentiation. So, if I take the differentiation I am taking the partial differentiation because I have to consider differentiation with respect to all the components of w. So, the optimal weight vector that is the optimal weight vector is I can write d E w star w star is the optimal weight vector divided by del w is equal to 0 I am equating to 0.

So, corresponding to this I will be getting phi transpose t minus phi transpose phi w star is equal to 0 and corresponding to this we can determine w star w star is equal to phi transpose phi to the power of minus 1 phi transpose t. So, this is the expression for the weight vector that is the optimal weight vector. So, with regularization we can write this

error function the E w is equal to 1 by 2 summation i is equal to 1 to n t i minus w transpose phi x i square plus I am considering the penalty term that is the regularization term lambda by 2 w transpose w we are considering.

So, this is with the regularization. So, corresponding to this this optimal weight vector will be it can be determined. So, phi transpose phi plus lambda i, i is the identity matrix to the power minus 1 phi transpose t that can be determined like this. So, corresponding to this case we can determine the optimal weight vector that the phi transpose phi plus lambda i, i is the identity matrix to the power minus 1 phi transpose t. So, this expression I can determine.

So, now let us consider this case the extension to high dimensional input feature. So, how to extend to high dimensional input feature. So, the same principle is applicable. So, move to the next slide. This is the extension to high dimension input feature.

So, extension to high dimension input feature. So, corresponding to this I can write y x w that is the predicted value. So, w plus w 1 x 1 plus w m this is nothing but the polynomial fit that already I have explained the polynomial fit and corresponding to this that weight vector is corresponding to w naught w 1 up to w m. So, size is m plus 1 into 1 and what is the basis function phi x. So, basis function phi x is equal to 1 x 1 x 2 up to x m m plus 1 into 1. So, corresponding to this I can write the predicted value y x w is equal to w transpose phi x.

So, what is x here x is the input vector and that is also m dimensional training feature vector. So, this is nothing but the m dimensional training feature vector. So, this is nothing but the extension of the previous equations. And based on this how I can write the predicted value. So, move to the next slide. What are the predicted values? So, how to write the predicted values y 1 I can write y 1 is nothing but phi transpose x 1 w y 2 is nothing but phi transpose x 2 w like this y up to y n I can predict phi transpose x n w.

So, you can see the linear combinations of basis functions phi x gives the predicted value for each of the samples. And already I told you the idea is to learn m number of weights the m number of weights we have to consider. So, ultimately what we have to consider the idea is to idea is to learn the m number of weights from the training samples. So, this expressions is the predicted values. What is the meaning of this that is the linear combinations of basis functions phi x gives the predicted value for each of the samples that is the meaning of this.

So, we can derive a similar expression like this. So, what is the optimum value of w? So, optimum value of w is nothing but phi transpose phi plus lambda i to the power minus 1

phi transpose t. So, we can write like this what is phi that is the matrix corresponding to this basis function. So, it is phi transpose x 1 phi transpose x 2 like this phi transpose x n.

So, this is n into m plus 1. So, we can determine the predicted value like this and this is the optimal weight. So, this is nothing but the optimal weight vector. So, we can explain this concept of regression like this. So, in this class I discussed the concept of regression. I explained how to do the polynomial fit and you can see the concept of the overfitting.

The overfitting take place when I am considering limited number of training samples and if I consider very complex model. So, if I consider a very complex model with limited number of training samples, then what will happen the overfitting take place and during the overfitting you can see the oscillations of the coefficients the polynomial coefficients. So, how to avoid this overfitting? If I consider more number of training samples, then I can minimize this overfitting problem. So, also I can consider a regularizer that is the penalty term in the error function I can consider and with the help of this regularizer I can minimize the overfitting. So, that is the concept of the regression. So, let me stop here today. Thank you. .