Introduction to Supervised Learning and Regression

Machine learning has found extensive use in solving real-life problems in today's world. For different kinds of problems, there are different techniques to solve them based on superiority of relevance in the corresponding concern. One such prior technique in machine learning is supervised learning. In supervised learning, the algorithm is provided with some **examples** or **sample records** to supervise the pattern existing among the data. This is why it is named supervised learning. Similar to machine learning, supervised learning further has multiple techniques to overcome a different range of problems. One such problem category is Regression.

A regression problem is comprised of the **below steps** -

1. **Formulation** - In this step, the identification of output and input features is done. It is all about settling the output and the input features.

2. **Solution** - Then the **relation** between the output and the input is developed using the given data for the corresponding problem. This is done by using a hypothesis function or a machine learning model.

3. **Interpretation** - Then the solution has to be interpreted to be deployed in business or any relevant place. The output of the model should be interpreted to deploy in real life and take relevant decisions.

4. **Performance assessment** - The performance of the model is assessed. This is required to see if the model can perform better than the current performance. In machine learning, there are multiple methods and parameters available that give clear indications of the performance of the model.

Understanding these steps will help us get an overview of the process. Before going ahead it seems necessary to grasp some basic definitions that will be useful in understanding the oncoming topics.

Record - A record is a single data point from the given dataset. It is also named an **instance**, **sample,** etc.

Feature - A feature is an attribute associated with the record that tells us about the record. For example, for a person height, weight, age etc are the features.

Independent features - These are the features that are independent of each other. A change in one can not assure a certain change in another. They associate the dependent variable with some relation to building the model.

Dependent features - These are the features that depend on the independent features under some relation. This leads to the formation of the model.

Model - A model is a **mathematical relation** between the output and the input features that are available. It is an anticipation of the actual relation existing between the set of features. But a model is never the truth. It is never a perfect relationship and always has a scope of improvement in itself. The utilization of the model is high because it is well designed and is problem-specific.

The big picture

The aim of the entire process is to make relevant predictions on the new or unseen data. To do so a trustworthy relationship between the independent and dependent variables is developed. Mathematically this relation is called the model in machine learning. Using this model predictions are done for the new data-point. So if X is the set of input/independent variables and Y is the output/dependent variable then for the new data-point, Y has to be predicted using X. In supervised learning, the outcome can be a continuous variable or a discrete variable. In the case of the output being a discrete variable, it is termed a classification problem. Whereas if the output variable is numeric, it is termed a regression problem. For example, the life expectancy of a person is a numeric feature. Hence it will be predicted using regression techniques. Whether it will rain or not today is a discrete feature, hence it will be solved by classification algorithms.

In general, there are two approaches to go over a data science problem -

a. **Data-ML-Prediction** - Here using the data a formula or a relation between the output and the input variable is developed. This formula is called the model. Using that model the final prediction is made on the seen/unseen data. This type of approach is called a machine learning approach.

b. **Data-Stats-Model-Prediction** - In the statistical approach, after getting the data the statistics play a vital role. Using some statistics an **empirical formula** or **statistical model** is prepared. Using that model, prediction over unseen data is done. Such an approach is called the **statistical** approach.

In today's world statistics and machine learning are inseparable fields. An expert in one has to be an expert in the other to function better while solving real-life problems. Nowadays it will not function well, being only a perfect statistician or a perfect machine learning expert. To understand this let us get into the differences between statistics and machine learning -

Statistics vs machine learning

- Statistics needs a deep understanding of simple models/methods while machine learning does not need to rely on theory. In machine learning, the theory does not always explain

the success of the model. A pivotal role is played by the algorithm that trains itself on the available training data and behaves accordingly on the unseen data.

Understanding the differences between statistics and machine learning lets us get into what is a basic statistical framework.

A basic statistical framework -

There are mainly two types of features, discrete and continuous in machine learning. A discrete feature is one that takes a **finite** number of values within a certain range. While a **continuous** feature is one that is free to take an **infinite** number of values within a certain range. In real scenarios, the dependent and independent variables might be of any type among the above-mentioned ones. To be able to make the predictions there has to be a mapping between the independent variable to the dependent variable. This mapping is done using a function called an **estimator**. An **estimator** is a function that is developed by using the available data. It gets trained on the available data and makes predictions on the unseen data. In general, this estimator is also called a model or a mathematical relation or rule-based relation between the output and the input variables. Let x_1, x_2..... x_n be the n independent variables while y is the dependent one. Let g be the estimator function. Then it can be shown as follows -

$$y = g(x_1, x_2, x_3, \ldots\ldots x_n)$$

In statistics, there are many methods to come to the estimator. A few of them are listed below -

1. **Plugin** - It is related to directly using the function that depicts the relation between the output and the input features. The **plug-in principle** says that a feature of a given distribution can be approximated by the same feature of the empirical distribution of a sample of observations drawn from the given distribution. The feature of the **empirical distribution** is called a **plug-in** estimate of the feature of the given distribution. For example, a quantile of a given distribution can be approximated by the analogous quantile of the empirical distribution of a sample of draws from the given distribution.

2. **Maximum likelihood** - It is one of the important ways of getting the correct hypothesis between the output and the input features. Maximum Likelihood Estimation is a probabilistic framework for solving the problem of density estimation. It is used to construct the **loss** or **cost function** of an algorithm. In this function, the variables are the parameters of the model that is set under a hypothesis on the available data. The likelihood function is then maximized using calculus to get the values of the parameters. These parameters are considered **best** for the construction of the actual model/hypothesis that depicts the relation between the output and the input feature. So on maximization, it leads to giving an empirical relation between the output and the input feature.

One of the important aspects of regression analysis is to select the features suitable for the model. To do so one can do the correlation analysis.

Correlation

Correlation gives an idea about the strength and type of relationship existing between two features, specially between the output and the input feature. Mathematically it gives the correlation coefficient between the two features. It is denoted by r. it ranges from -1 to +1. A positive value of r signifies that two features are positively correlated i.e. if one feature is increased the other will also increase. The negative value of r shows the inverse relation. If one feature increases the other will decrease. Now in terms of magnitude higher the value (either positive or negative) stronger is the relation(Closer the points are to the regression line).
Apart from correlation test, to understand the significance of the features for the model **Wald's test** is used in machine learning. Higher the significance of the model more important it is to the model and should be included in it.

To understand the entire regression process let us take an example.

An example - advertising and sales

Data across **200** markets about spending for TV, Radio, and Newspapers is collected. Companies in this field are spending a lot of money in advertising these products in the market. They are trying to enhance sales by doing advertisements for the corresponding products in the market. We are trying to see whether the budget for advertisement is affecting the number of sales or not. As the output feature is the number of sales, which is a **continuous** feature, it is a supervised learning regression problem. If the relationship exists then how can we predict sales given the channel budget on advertisement?

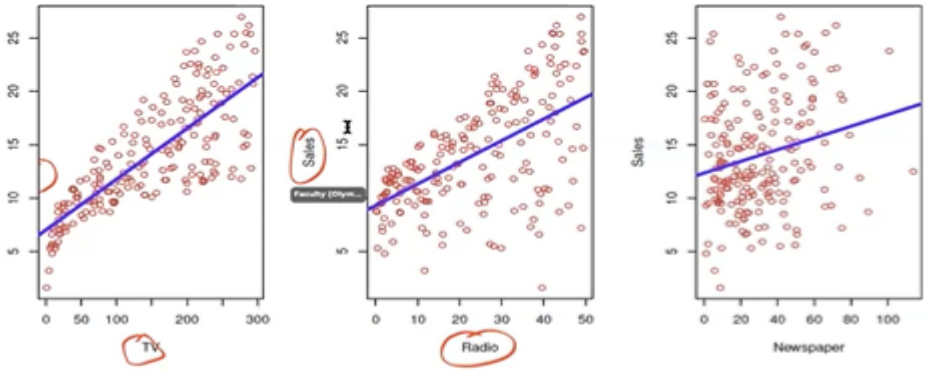
In the process, the first suitable action is to do the visualization. The below figure depicts the **scatter plot** of sales for **three** different products namely **TV, Radio and Newspaper**. The blue line is considered to be the **best fit** line for the available data that can represent the relation between the sales and the advertising budget of the mentioned products. The plots are depicting whether the relation between the sales and advertisement budget is existing or not.

1. In the first plot for **TV**, it can be seen that there is a strong **dependency** between the output and the input features. For TV the sales are increasing steeply with an increase in the budget for advertisement. Here the figure indicates that they are dependent on each other.

2. In the second plot for **Radio**, the relation is still there but it is a bit weaker than the first one. The steepness of increase in the number of sales is less than that in **TV**.

3. In the third plot, the steepness of the line is flat. That means it depicts the weakest relation among all the three products

This way it can be concluded that the advertisement budget has an impact on the sales of the product. In every case adding an extra budget to the advertisement is increasing the sales of the product. This is how visualization helps in selecting features and having an idea about the existing patterns and relationships between features. But it is not saying anything about the expression between the output and input.



After getting this useful visual insight from the above plots let us get into finding the expression of the actual relation between the output and the input.

Regression -

While we look into the dataset, each record has a dimension m, that is the number of features available. Using the set of input features X we need to come up with a relation g between the output y and the input set X. Once this is done, to make the prediction it takes a new X and gives y as the output. To find g, it requires capitalization of a certain function called a **loss function** or **objective function**. For regression, the loss function is the **sum of squares** of the **residual** terms. A residual can be defined as the **difference** between the **actual value** and the **predicted target value** by the model. This function has to be capitalized to gain the **weights** of the hypothesis function.

Now let us have a look at the mathematical expression of the linear regression model. It is given as follows -

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \ldots\ldots + a_nx_n$$

Here a_0, a_1, a_2, ... a_n are the **coefficients** of the model. They are also called as **weights** of the model.

$x_1, x_2, x_3, \ldots\ldots, x_n$ is the input features available in the data.

y is the output variable. In this case, it is the sales amount of the product.

For linear regression, the loss function is the sum of squared residual terms. A residual can be expressed mathematically as follows -

$$Residual = (y_i - (a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots\ldots + a_n x_n))$$

The loss function is formed by squaring the sum of such residuals. It can be given as follows -

$$Loss\ Function = \sum_{i=1}^{m} (y_i - (a_0 + a_1 x_1^i + a_2 x_2^i + a_3 x_3^i + \ldots\ldots + a_n x_n^i))^2$$

This loss function represents the **cost** of making wrong predictions. It adds the residual corresponding to each and every training example. **Minimizing** such terms will lead to a better model. Here **squares** of each residual term in the training example are taken to remove the impact of the sign in the algebraic expression. To minimize it is differentiated to converge to a certain point. Doing so will fetch to n different linear equations in the parameters. Solving them we will get the corresponding values of $a_1, a_2, a_3, a_4, \ldots\ldots a_n$, once finding them the final equation in the input features can be given as follows -

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots\ldots + a_n x_n$$

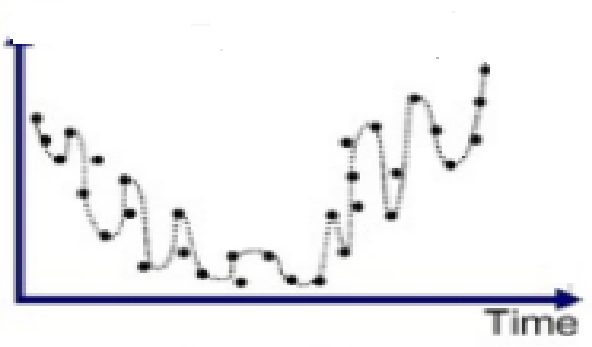
Now the coefficients are known and hence the relationship is established to make predictions on the unseen data. The above equation is suitable for making predictions for one output feature. If we have to predict multiple dependent variables then we need to train multiple models corresponding to each output feature.

A better model is one that is **generalizing** on unseen data at the **most** accurate level. It happens in machine learning that the model is performing very well on the training data but not on the testing data and also the model performs badly on both training and the test set data. These two cases can be explained below -

1. **Overfitting of the model** - When a machine learning model is very accurate on the training set but unable to generalize over the unseen data then it is called an overfitting model. It is a high bias model. Learning each and every nuance of the data is not healthy from a prediction perspective. Such a model might be very accurate but it is too complex to generalize the new data. They have high bias and high variance. Bias can be interpreted as the training error and variance of the test set error.

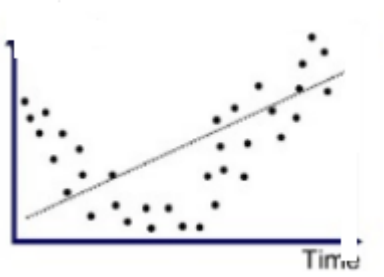
   The below figure depicts an overfitting model. It can be seen easily that the model line is

following each and every training example. It is such a complex model that it is very hard to even interpret.



2. **Underfitting of the model** - If a model is performing badly on training and test set both then it is called underfitting of the model. Such models are less accurate as well as unable to generalize the unseen data. They have low bias but high variance.

   The below figure shows an example of the underfitting model. It can be seen that the best fit line is far away from the actual data points. It creates a high bias in the model as the training error is going to be high.



Overfitting and underfitting both the cases are not considered suitable for a machine learning model. In an ideal case, the model should have low bias and low variance both. Such a model is supposed to generalize better over the unseen data.

The machine learning model can have different perspectives to look upon. If we look at the model equation of the linear regression algorithm it looks as follows -

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + ... + a_nx_n$$

If we consider weights as one **vector** say A and the input features X as another vector. Then the vector product of these two vectors will lead to the above equation. The output y can be represented as follows -

$$y = A \cdot X$$

Where X is $1, x_1, x_2, x_3 .... x_n$ is the vector of input features.

A is 1, a_1, a_2, a_3.... a_n are the weight vectors.

Weights are estimated in such a way that the loss function is minimum or the sum squared or error is as small as possible.

Now it is important to understand how the equation is established between the output and the input variable. The process begins with minimizing the loss function.

The solution to the regression problem -

Once the loss function is established it is crucial to find the set of weights/parameters that are best for the prediction on unseen data. Let us look at the loss function once -

$$Loss\ Function = \sum_{i=1}^{m} (y_i - (a_0 + a_1 x^i_1 + a_2 x^i_2 + a_3 x^i_3 + ... + a_n x^i_n))^2$$

Where symbols have their usual meaning.

Differentiating the loss function with respect to the weights because the input features are known in terms of the data available.

$$\frac{d(Loss\ function)}{d(a_j)} = 2 \times \sum_{i=1}^{m} (y_i - (a_0 + a_1 x^i_1 + a_2 x^i_2 + a_3 x^i_3 + ... + a_n x^i_n)) \times x^i_j = 0$$

The above expression will lead to equations of the following format.

For each J in 1 to n

$$k^j_0 a_0 + k^j_1 a_1 + k^j_2 a_2 + ... + k^j_n a_n = 0$$

Solving these n equations will lead to the values of the weights for the model. Using these weights the final equation between the output and the input features will be set.

Nowadays solving such linear equations is a very easy task for computers. It is much faster than manual computations.

Let us have a look at the outcomes received in the case of the example we began with.

Results for our example -

Doing all the requisite processings we found the weights belonging to the final model. These weights are associated with the corresponding features to make the final model. It is shown below -

$$Sales = 2.94 + 0.046 \times tv + 0.01 \times news$$

Where sales is the number of sales done, tv is the budget spent on the advertisement of tv and news is the budget spent on advertisement through the news.

A simple linear regression model is one where only one **independent** variable is used to make a prediction of the output feature. In the present case, it can be shown as follows -

$$Sales = 12.35 + 0.065 \times News$$

Such a model is called a simple linear regression model.

Interpretation and justification: empirical risk minimization

There are a lot of factors that influence a machine learning model. One such factor is the quality of the sample taken from the population. When we consider the entire population as our training data then the training becomes very complex due to the processing of huge data. Doing so is not easily possible in every case. So in general we take **finite samples** from the population itself. A finite sample is a sample of finite size drawn from the population that should represent the population to the best.

It happens that using different finite samples for the same set of variables will fetch different linear regression models. As the size of the finite sample increases, it tends to become the population itself. The predictions of the model trained on a finite sample that is a good representative of the population are trustworthy than a bad finite sample model.

Now after training and testing the model it is required to do the **performance assessment** of the model. Doing so ensures a better performance by the model.

Performance assessment

Performance assessment is the process of assessing the performance of a machine learning model. It is required because it tells us about the possible scope of changes and improvements to be done to make the model perform better. In machine learning, there are a certain number of methods to do the performance assessment. Such methods ensure the best possible fit of the model.

For example, let us understand this with the R^2 value, (read as R squared value).

It is a performance assessment parameter that tells us about the quality of fit. Mathematically it is given as follows -

$$R^2 = 1 - (\frac{RSS}{TSS})$$

Here RSS is the residual sum of squares of the actual model. TSS is the total sum of squares of the model. It is calculated by taking the mean of the dependent variable as the prediction and then calculating the residual sum of squares. This is the maximum possible value of error a model can depict. So the term RSS/TSS is always less than or equal to 1. A lesser RSS means a better model and that will lead to R^2 tending to 1. While a higher RSS means a bad model and that will lead to R^2 tending to 0. So R^2 tending to 1 will assure the best fit of the model and the value 0 will assure that to be the worst model.

$R^2 = 1,$

It means that the $RSS = 0$ and the model is perfectly fitting to the training data. Such models are fitting to the best of the data but are not useful in real-life scenarios due to overfitting. It explains a lot about the dependency of the variables.

$R^2 = 0,$

It means that the RSS and TSS are equal. It means the model is as bad as a **naive** model where all the predictions are guessed to be the mean of the existing output values. It explains very little about the dependency of variables.

Take an example of the model

$$Sales = 2.94 + 0.046 \times tv + 0.19 \times radio - 0.001 \times news$$

Here $R^2 = 0.897$ a high value. Hence all the budgets together explain a lot about the sales.

Considering the simple linear regression models -

$$Sales = 12.35 + 0.055 \times news$$

$R^2 = 0.05$ Newspaper budget explains a little

For tv alone $R^2 = 0.61$

For radio alone $R^2 = 0.33$

How noisy/reliable are my estimates of weights

Estimating weights is one of the aspects of the model preparation process. Along with this we also need to ensure how reliable it is. There is randomness in the weights also because it is dependent on the **input features** where noise is already there. Due to this, the weights are random variables that are normally distributed. Weights also have their ground truth values.

Let the ground truth values be a^* while the estimated values are a. Then the expected value is given as $E(a - a^*)^2$. It can be broken into two components namely **bias** and **variance**.

$$E[(a - a^*)^2] = (E[a] - a^*)^2 + var(a)$$

In the above equation, the first term is the **bias** term while the second term is the **variance**. For the linear regression model, the bias term is 0 so the only **variance** is taken into consideration in terms of error.

Let us know about the distribution of the weights of the model.

The distribution of weights

Weights are random variables but they are approximately normally distributed. The lesser the standard error of these random variables, the more trustworthy the model is. If it is high the model is less trustworthy because it is supposed to give erroneous outcomes when deployed to unseen data.

Confidence interval

A confidence interval is an interval in which if the accuracy of a model falls, it is considered to be a trustworthy model. In general, let us take an example of a 95% confidence interval. It means that out of 100 trials 95 times the model will make the correct prediction while 5 times it is allowed to fail or make wrong predictions. While reporting the outcome of a model it is a good practice to add the confidence interval. It becomes more impactful then.

To test the compatibility of the data with the weights let us test the null hypothesis that $a^* = 0$.

Testing the hypothesis

If the estimated value of weights deviates a lot from 0 then the null hypothesis will be rejected. It would be considered that enough evidence is found against the null hypothesis.

If the estimated value is close to 0 it is considered that enough evidence could not be found against the null hypothesis. In such a case the null hypothesis is not rejected and is considered to be the true statement.

In general, while reporting we use the p-value corresponding to the hypothesis test. If it is greater than the **critical value** then the **null hypothesis** is rejected and if it is less than the critical value then the **null hypothesis** is accepted.

Let us understand the p-value first. The p-value is the probability of seeing something **at least** as extreme as the observed value. Let us consider 5% to be the critical value. In concern of the hypothesis testing the null hypothesis is rejected if the p-value is less than 5% and it is acceptable if it is greater than 5%.

Sources of error

In the end, let us understand the different sources of errors while building the regression model. For example, suppose the following model is built -

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + ... + a_nx_n$$

Corresponding to such models there are two sources of errors -

1. **The noise of the data** - It is something that can not be treated because it is inherent to the data.

2. **Variance or error** - It is the error associated with the model. It can be reduced and modified. It is related to inaccuracies of the weights.