Cognitive Application

Samarthya Gupta , IIT Delhi , ce1200290@iitd.ac.in (email changed due to branch change **new email : am1200290@iitd.ac.in**)

Quest1 - What is overfitting and how to avoid it?

Ans1 - Overfitting in linear regression happens when the model tries to fit the curve through every training data set. In this way the model picks up each small disturbance and random fluctuations in data set and is learned as concepts by the model. This negatively impacts the performance of the ml because the model is so precisely fit to the curve that for any new data set the results produced are prone to large error and impacts the model's ability to generalize.

Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning from a dataset. Ways to avoid overfitting.

1. Splitting : the data set into two parts first testing and another training data set usually in (20 : 80 ratio).This will not only let the model fit into a smaller data moreover will give a good generalization capability since the testing set represents unseen data that were not used for training. A better approach will be two divides into n groups and then iterate until each group is used as a testing data set.
2. Increasing the size of data gathered.
3. If there is a small data set with large number of features the selecting only those features which are important so that our model doesn't need to learn for so many features and eventually overfit. We can simply test out different features, train individual models for these features, and evaluate generalization capabilities, or use one of the various widely used feature selection methods.
4. Regularization is a technique to constrain our network from learning a model that is too complex, which may therefore overfit. In L1 or L2 regularization, we can add a penalty term on the cost function to push the

estimated coefficients towards zero (and not take more extreme values). L2 regularization allows weights to decay towards zero but not to zero, while L1 regularization allows weights to decay to zero.

5. Stopping the training when finished with a smaller data set and saving the model after analyzing the loss function.

Quest 2 - What is line of best fit?

Ans2 - A line of best fit is a straight line that is the best approximation of the given set of data. It is used to study the nature of the relation between two variables. In a linear regression for a two-dimension model the line of best fit would be a straight line passing close through most of the points of the data set such the cost function (least square) is minimized. Error is also less.

Quest3 - Explain multivariant in linear regression with a real-life example.

Ans3 - Multivariate Regression is a supervised machine learning algorithm involving multiple data variables for analysis and training the model. A Multivariate regression is an extension of multiple regression with one dependent variable and multiple independent variables. Based on the number of independent variables, we try to predict the output. Multivariate regression tries to find out a formula that explains how factors in variables respond simultaneously to changes in others. It is basically predicting the output using large number of features for a key value moreover these features will be used to learn patterns and predict important results.

Example : If we want to estimate the price of a house, we can collect details such as the location of the house, number of bedrooms, size in square feet, amenities available, or not.

These details price of the house can be predicted and how each variable are interrelated. Many features are used to predict the prize of house and interrelate various features.

Quest4 - How can we improve the accuracy of a linear regression model?

Ans4 - The ways with which we can increase the accuracy of a regression models are as follows:

1. Large data : Presence of more data results in better and accurate models giving a larger data set for training and testing the model. Data could be increase using data argumentation that is producing data for the given data set
2. Treat missing and Outlier values : The unwanted presence of missing and outlier values in the training data often reduces the accuracy of a model or leads to a biased model. It leads to inaccurate predictions. This is because we don't analyze the behavior and relationship with other variables correctly. So, it is important to treat missing and outlier values well.
3. Feature Selection : Feature Selection is a process of finding out the best subset of attributes which better explains the relationship of independent variables with target variable.
4. Multiple algorithms and refining them : Hitting at the right machine learning algorithm is the ideal approach to achieve higher accuracy. Large number of algorithms produces more and more accuracy.

Quest5 - What is RMSE and MSE.

Ans5 - The Mean Squared Error (MSE) is perhaps the simplest and most common loss function, often taught in introductory Machine Learning courses. To calculate the MSE, you take the difference between your model's predictions and the ground truth, square it, and average it out across the whole dataset. The MSE will never be negative since we are always squaring the errors.

Root mean square error (RMSE) or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance. To compute RMSE, calculate the residual (difference between prediction and truth) for each data point, compute the norm of residual for each data point, compute the mean of residuals and take the square root of that mean.

Both these methods are used to get the loss function of a