Answer -1

Reason for Gap in Training accuracy and test accuracy-

The reason this happens is because we use different criteria to train the model and then test its efficiency. As we know, a model is trained by maximizing its accuracy on the training dataset. But its performance is determined on its ability to perform well on unknown data. In this situation, overfitting occurs when our model tries to memorize the training data as opposed to try to generalize from patterns observed in the training data.

overfitting occurs when a learning model customizes itself too much to describe the relationship between training data and the labels. Overfitting tends to make the model very complex by having too many parameters. By doing this, it loses its generalization power, which leads to poor performance on new data.

For example, you may want to stop training your model once the accuracy stops improving. In this situation, there will be a point where the accuracy on the training set continues to improve but the accuracy on unseen test data starts to degrade.

How to Deal with it-

we can solve the problem of overfitting using

- Regularization technique
- Drop out
- Cross Validation

Answer-2

Difference between L1 and L2 regularization

Lasso

L1 regularization does feature selection. It does this by assigning insignificant input features with zero weight and useful features with a non zero weight.

In L1 regularization we penalize the absolute value of the weights.

L1 or Lasso produces a model that is simple, interpretable and contains a subset of input features

Ridge

In L2 regularization, regularization term is the sum of square of all feature weights .

L2 regularization forces the weights to be small but does not make them zero and does non sparse solution.

L2 is not robust to outliers as square terms blows up the error differences of the outliers and the regularization term tries to fix it by penalizing the weights

L2 or Ridge regression performs better when all the input features influence the output and all with weights are of roughly equal size

L1 Regularization

- It penalizes sum of absolute value of weights.
- It has a sparse solution
- It has multiple solutions
- It has built in feature selection
- It is robust to outliers
- It generates model that are simple and interpretable but cannot learn complex patterns

L2 Regularization

- It penalizes sum of square weights.
- It has a non sparse solution
- It has one solution
- It has no feature selection
- It is not robust to outliers
- It gives better prediction when output variable is a function of all input features
- It regularization is able to learn complex data patterns

Answer -3

Given two linear models:

L1: y = 39.76x + 32.648628

And

L2: y = 43.2x + 19.8

We will choose L2 Model because it is simple model as Occams Razor principle says choose simplest model first. Once the set of candidate models has been chosen, the statistical analysis allows us to select the best of these models. What is meant by best is controversial. A good model selection technique will balance goodness of fit with simplicity. More complex models will be better able to adapt their shape to fit the data (for example, a fifth-order polynomial can exactly fit six points), but the additional parameters may not represent anything useful. (Perhaps those six points are really just randomly distributed about a straight line.) Goodness of fit is generally determined using a likelihood ratio approach, or an approximation of this, leading to a chi-squared test. The complexity is generally measured by counting the number of parameters in the model.

Model selection techniques can be considered as estimators of some physical quantity, such as the probability of the model producing the given data. The bias and variance are both important measures of the quality of this estimator; efficiency is also often considered.

Answer - 4

Robust Generalised Model

A Robust model is resistant to errors in the results. Robust statistics are statistics with good performance for data drawn from a wide range of probability distributions, especially for distributions that are not normal.The generalised linear model is the general tool that can be used with all such types of response variables. This generalised linear model allows the experimenter to model the response variables by any distribution within a large family of distributions, namely the exponential family, and the expected response by any (suitably smooth) function of the explanatory variables, with the only restriction that this function should depend on the explanatory variables linearly. As a special case, it also includes the ordinary linear regression problem; the study of the generalised linear model helps us to deal with a very large super family of parametric regression problems.The classical procedure to estimate the parameters of the generalised linear regression model is the maximum likelihood estimation method generating most efficient estimators. This theoretical advantage of the maximum likelihood estimator is, however, tempered by its known lack of robustness to outliers and model misspecification.

A method for robustness in linear models is to assume that there is a mixture of standard and outlier observations with a different error variance for each class. For generalised linear models (GLMs) the mixture model approach is more difficult as the error variance for many distributions has a fixed relationship to the mean. This model is extended to GLMs by changing the classes to one where the standard class is a standard GLM and the outlier class which is an over dispersed GLM achieved by including a random effect term in the linear predictor. The advantages of this method are it can be extended to any model with a linear predictor, and outlier observations can be easily identified. Using simulation the model is compared to an *M*-estimator, and found to have improved bias and coverage.

Answer – 5

optimal value of lambda for ridge- 50

optimal value of lambda for Lasso- 500

We will prefer Lasso over Ridge because Lasso minimizes complexity of computation and provide sparse output. Advantage in using **LASSO** method, first of all it can provide very good prediction accuracy, because shrinking and removing the coefficients can reduce variance. Lasso tries to determine the optimum coefficient that minimize a cost function, such as the Mean Squared Error (MSE) and a new term is added to the MSE that penalizes large absolute values for the coefficients. Its goal is to avoid overfitting your data.