Pattern Recognition
Task:2

1-**What is overfitting and underfitting:**

The solution

Overfitting: refers to a model that models the training data too well.

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

For example, decision trees are a nonparametric machine learning algorithm that is very flexible and is subject to overfitting training data. This problem can be addressed by pruning a tree after it has learned in order to remove some of the detail it has picked up.

Underfitting: Underfitting refers to a model that can neither model the training data nor generalize to new data.

An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

Underfitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithms. Nevertheless, it does provide a good contrast to the problem of overfitting.

2- Why we test the model on both train set and test set

training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model.

After a model has been processed by using the training set, you test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct.

3- What are the common techniques of regularization?

The solution

L2 and L1 are the most common types of regularization. Regularization works on the premise that smaller weights lead to simpler models which in results helps in avoiding overfitting. So to obtain a smaller weight matrix, these techniques add a ‘regularization term’ along with the loss to obtain the cost function.

Cost function = Loss + Regularization term

The difference between L1 and L2 regularization techniques lies in the nature of this regularization term. In general, the addition of this regularization term causes the values of the weight matrices to reduce, leading simpler models.

In L2, we depict cost function as

$$Cost\,function = Loss + \frac{\lambda}{2m} * \sum \|w\|^2$$

Here, lambda is the regularization parameter which is the sum of squares of all feature weights. L2 technique forces the weight to reduce but never makes them zero. Also referred to as ridge regularization, this technique performs best when all the input features influence the output, and all the weights are of almost equal size.

In the L1 regularization technique,

$$Cost\,function = Loss + \frac{\lambda}{2m} * \sum \|w\|$$

Unlike in the case of L2 regularization, where weights are never reduced to zero, in L1 the absolute value of the weights are penalised. This technique is useful when the aim is to compress the model. Also called Lasso regularization, in this technique, insignificant input features are assigned zero weight and useful features with non-zero.

Dropout

Another most frequently used regularization technique is dropout. It essentially means that during the training, randomly selected neurons are turned off or 'dropped' out. It means that they are temporarily obstructed from influencing or activating the downward neuron in a forward pass, and none of the weights updates is applied on the backward pass.

So if neurons are randomly dropped out of the network during training, the other neurons step in and make the predictions for the missing neurons. This results in independent internal representations being learned by the network, making the network less sensitive to the specific weight of the neurons. Such a network is better generalized and has fewer chances of producing overfitting.

Early Stopping

It is a kind of cross-validation strategy where one part of the training set is used as a validation set, and the performance of the model is gauged against this set. So if the performance on this validation set gets worse, the training on the model is immediately stopped.

The main idea behind this technique is that while fitting a neural network on training data, consecutively, the model is evaluated on the unseen data or the validation set after each iteration. So if the performance on this validation set is decreasing or remaining the same for the certain iterations, then the process of model training is stopped.

Data Augmentation

The simplest way to reduce overfitting is to increase the data, and this technique helps in doing so.

Data augmentation is a regularization technique, which is used generally when we have images as data sets. It generates additional data artificially from the existing training data by making minor changes such as rotation, flipping, cropping, or blurring a few pixels in the image, and this process generates more and more data. Through this regularization technique, the model variance is reduced, which in turn decreases the regularization error.

4- **What are the advantages of using Bayesian linear regression? (i.e., what Bayesian regression is good for?)**

The solution

doing Bayesian regression is not an algorithm but a different approach to statistical inference. The major advantage is that, by this Bayesian processing, you recover the whole range of inferential solutions, rather than a point estimate and a confidence interval as in classical regression.