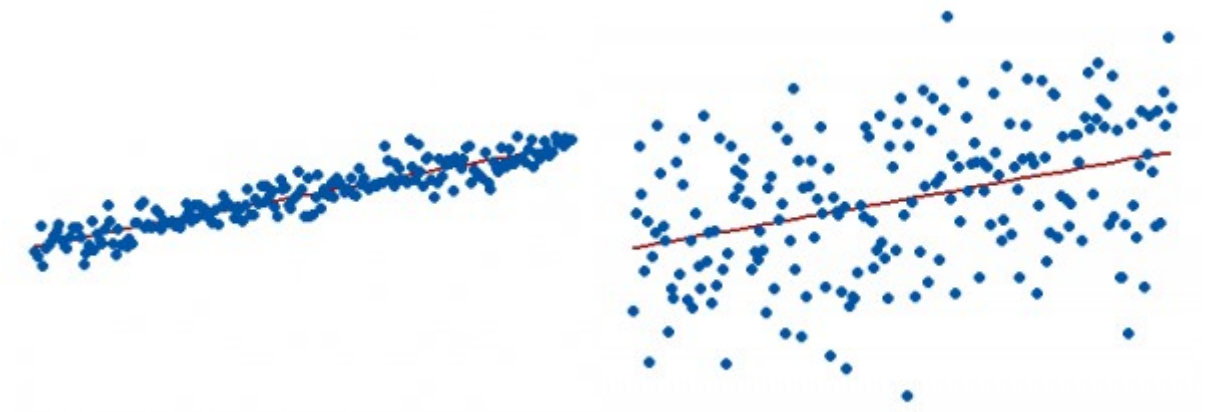
SOLUTIONS MACHINE LEARNING ASSIGNMENT – 5

Ans 1) The residual sum of squares (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

R-squared evaluates the scatter of the data points around the fitted regression line. It is also called the coefficient of determination, or the coefficient of multiple determination for multiple regression. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values.

R-squared is the percentage of the dependent variable variation that a linear model explains.



The R-squared for the regression model on the left is 15%, and for the model on the right it is 85%. When a regression model accounts for more of the variance, the data points are closer to the regression line. In practice, you'll never see a regression model with an R2 of 100%. In that case, the fitted values equal the data values and, consequently, all the observations fall exactly on the regression line.

R-squared cannot be used to determine whether the coefficient estimates and predictions are biased, which is why it is required to assess the residual plots.

R-squared does not indicate if a regression model provides an adequate fit to your data. A good model can have a low R2 value. On the other hand, a biased model can have a high R2 value!

ANS 2) The total sum of squares (TSS) measures how much variation there is in the observed data, while the residual sum of squares (RSS) measures the variation in the error between the observed data and predicted values.

The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable.

In some cases (if only we have an intercept (constant term) in our regression model),

Total sum of squares (TSS) = **Explained sum of squares (ESS)** + **Residual sum of squares (RSS)**

To start, let's break down the correlation between TSS, ESS, and RSS.

$$
\begin{aligned}
TSS &= \sum_i (y_i - \bar{y})^2 \\
&= \sum_i ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= ESS + RSS + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})
\end{aligned}
$$

We can see that there is a cross-term in the equation.

ANS 3) Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent over-fitting or under-fitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

ANS 4) The Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one. It favours mostly the larger partitions and is very simple to implement. In simple terms, it calculates the probability of a certain randomly selected feature that was classified incorrectly.

ANS 5) Decision trees are prone to over fitting, especially when a tree is particularly deep (unregularize). This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions. This small sample could lead to unsound conclusions. The depth parameter is one of the ways in which we can regularize the tree, or limit the way it grows to prevent over-fitting.

ANS 6) Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in machine learning.

ANS 7) Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data. Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification.

Bagging: It is a homogeneous weak learners' model that learns from each other independently in parallel and combines them for determining the model average.

Boosting: It is also a homogeneous weak learners' model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

ANS 8) The **out-of-bag (OOB) error** is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the Random Forest Classifier to be fit and validated whilst being trained.

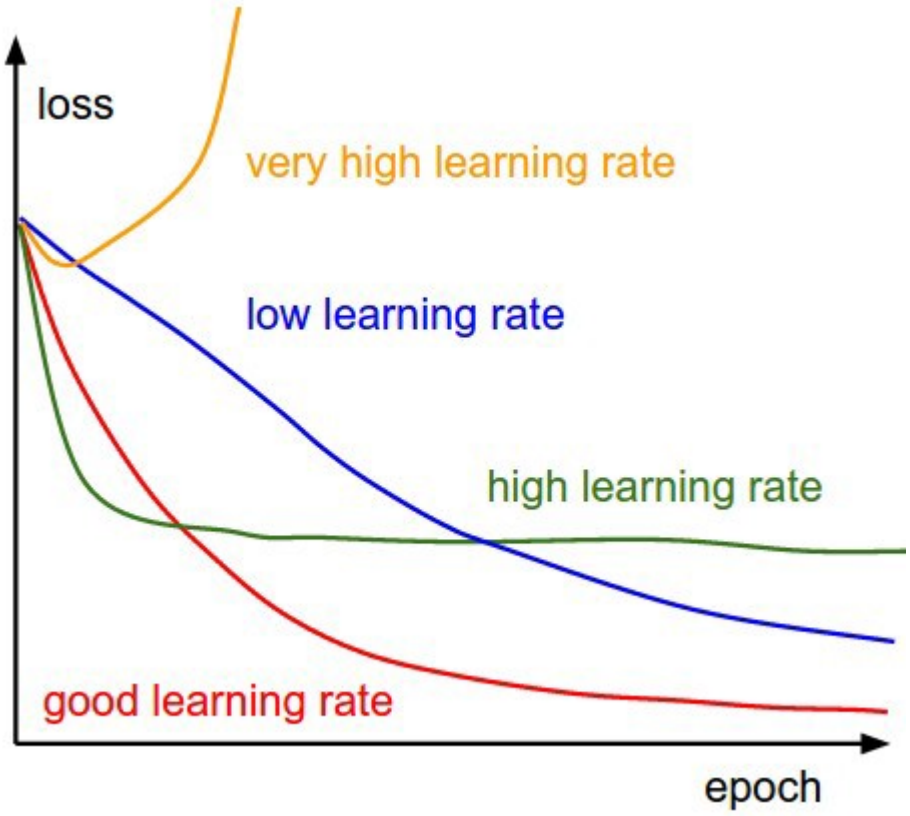
ANS 9) Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

ANS 10) Hyper parameter tuning consists of finding a set of optimal hyper parameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyper parameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

ANS 11) In order for Gradient Descent to work, we must set the learning rate to an appropriate value. This parameter determines how fast or slow we will move towards the optimal weights. If the learning rate is very large we will skip the optimal solution.



ANS 12) Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios. Logistic regression is indeed non linear in terms of Odds and Probability, however it is linear in terms of Log Odds.

ANS 13) Following are the differences based on some parameters:

Loss Function:

The technique of Boosting uses various loss functions. In case of Adaptive Boosting or AdaBoost, it minimises the exponential loss function that can make the algorithm sensitive to the outliers. With Gradient Boosting, any differentiable loss function can be utilised. Gradient Boosting algorithm is more robust to outliers than AdaBoost.

Flexibility

AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

Benefits

AdaBoost minimises loss function related to any classification error and is best used with weak learners. The method was mainly designed for binary classification problems and can

be utilised to boost the performance of decision trees. Gradient Boosting is used to solve the differentiable loss function problem. The technique can be used for both classification and regression problems.

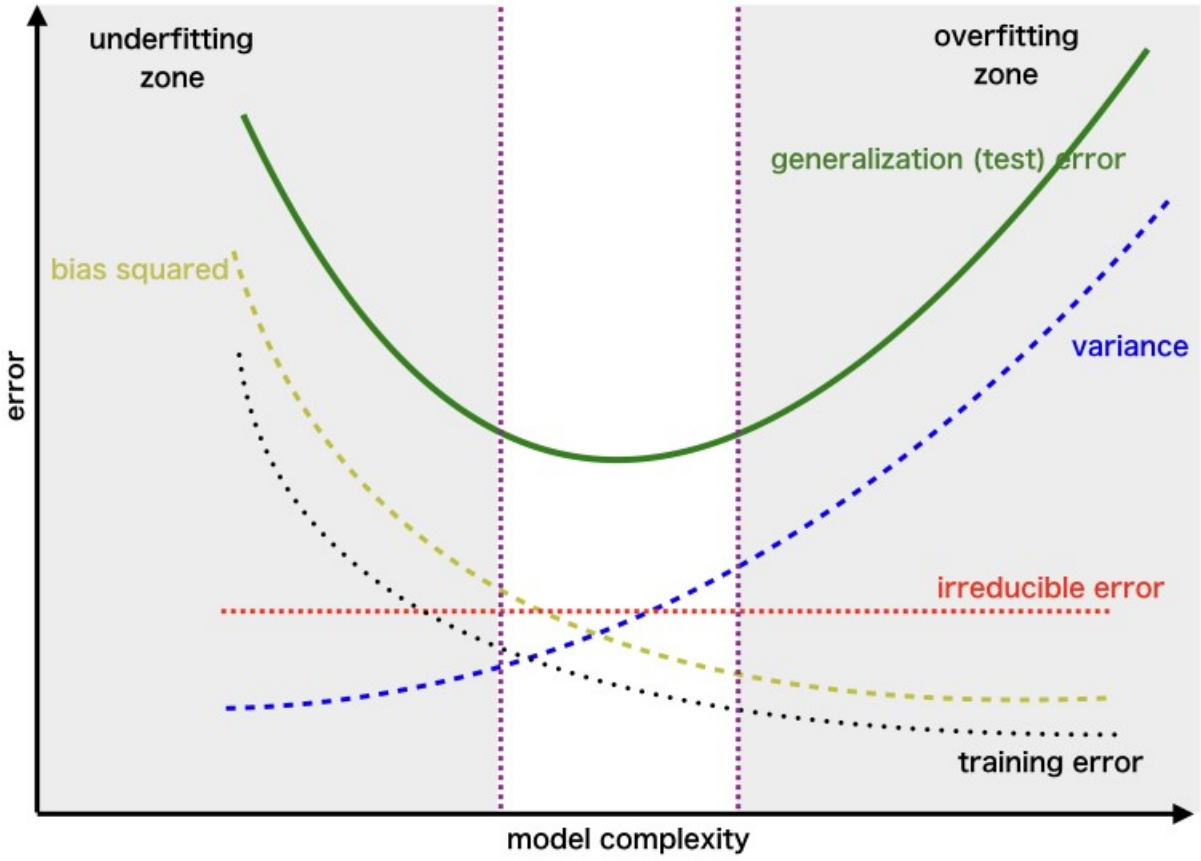
Shortcomings

In the case of Gradient Boosting, the shortcomings of the existing weak learners can be identified by gradients and with AdaBoost, it can be identified by high-weight data points.

Wrapping Up

Though there are several differences between the two boosting methods, both the algorithms follow the same path and share similar historic roots. Both the algorithms work for boosting the performance of a simple base-learner by iteratively shifting the focus towards problematic observations that are challenging to predict.

ANS 14) There is a trade-off between a model's ability to minimize bias and variance which is referred to as the best solution for selecting a value of Regularization constant. Proper understanding of these errors would help to avoid the over-fitting and under-fitting of a data set while training the algorithm.



ANS 15) SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form.

Different SVM algorithms use different types of kernel functions. These functions can be different types for example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid.

Introduce Kernel functions for sequence data, graphs, text, images, as well as vectors. The most used type of kernel function is RBF because it has localized and finite response along the entire x-axis.

Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set.

RBF is the most popular support vector machine kernel choice, and the default one used in sklearn . RBF is short for "radial basis function", a type of function that is used to approximate other functions in the literature.

Polynomial kernel represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.