Machine learning

Q.1) R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

The residual sum of squares (RSS) is a statistical technique used to measure the amount of [variance](https://www.investopedia.com/terms/v/variance.asp) in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or [error term](https://www.investopedia.com/terms/e/errorterm.asp).

[Linear regression](https://www.investopedia.com/terms/r/regression.asp) is a measurement that helps determine the strength of the relationship between a dependent variable and one or more other factors, known as independent or explanatory variables.

Q.2) What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

What is the Total Sum of Squares?

The Total SS (TSS or SST) tells you how much variation there is in the [dependent variable](https://www.statisticshowto.com/dependent-variable-definition/).  
Total SS = Σ(Yi – mean of Y)2.  
Note: Sigma (Σ) is a mathematical term for [summation](http://www.columbia.edu/itc/sipa/math/summation.html) or “adding up.” It’s telling you to add up all the possible results from the rest of the equation.

Sum of squares is a measure of how a data set varies around a central number (like the [mean](https://www.statisticshowto.com/mean/)). You might realize by the phrase that you’re summing (adding up) squares—but squares of what? You’ll sometimes see this formula:  
[ss2](https://www.statisticshowto.com/wp-content/uploads/2015/04/ss2.jpg)

The residual sum of squares (RSS)

The residual sum of squares (RSS) is a statistical technique used to measure the amount of [variance](https://www.investopedia.com/terms/v/variance.asp) in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or [error term](https://www.investopedia.com/terms/e/errorterm.asp).

[Linear regression](https://www.investopedia.com/terms/r/regression.asp) is a measurement that helps determine the strength of the relationship between a dependent variable and one or more other factors, known as independent or explanatory variables.

Understanding the Residual Sum of Squares

In general terms, the [sum of squares](https://www.investopedia.com/terms/s/sum-of-squares.asp) is a statistical technique used in regression analysis to determine the dispersion of data points. In a regression analysis, the goal is to determine how well a data series can be fitted to a function that might help to explain how the data series was generated. The sum of squares is used as a mathematical way to find the function that [best fits](https://www.investopedia.com/terms/l/line-of-best-fit.asp) (varies least) from the data.

The RSS measures the amount of error remaining between the regression function and the data set after the model has been run. A smaller RSS figure represents a regression function that is well-fit to the data.

The RSS, also known as the sum of squared residuals, essentially determines how well a regression model explains or represents the data in the model.

How to Calculate the Residual Sum of Squares

*RSS = ∑ni=1(yi - f(xi))2*

*Where:*

*yi = the ith value of the variable to be predicted*

*f(xi) = predicted value of yi*

*n = upper limit of summation*

explained sum of squares (ESS)

In [statistics](https://en.wikipedia.org/wiki/Statistics), the explained sum of squares (ESS), alternatively known as the model sum of squares or sum of squares due to regression (SSR – not to be confused with the [residual sum of squares](https://en.wikipedia.org/wiki/Residual_sum_of_squares) (RSS) or sum of squares of errors), is a quantity used in describing how well a model, often a [regression model](https://en.wikipedia.org/wiki/Regression_analysis), represents the data being modelled. In particular, the explained sum of squares measures how much variation there is in the modelled values and this is compared to the [total sum of squares](https://en.wikipedia.org/wiki/Total_sum_of_squares) (TSS), which measures how much variation there is in the observed data, and to the [residual sum of squares](https://en.wikipedia.org/wiki/Residual_sum_of_squares), which measures the variation in the error between the observed data and modelled 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 a standard [regression model](https://en.wikipedia.org/wiki/Regression_model) — for example, *yi* = *a* + *b*1*x*1*i* + *b*2*x*2*i* + ... + *εi*, where *yi* is the *i* th observation of the [response variable](https://en.wikipedia.org/wiki/Response_variable), *xji* is the *i* th observation of the *j* th [explanatory variable](https://en.wikipedia.org/wiki/Explanatory_variable), *a* and *bj* are [coefficients](https://en.wikipedia.org/wiki/Coefficient), *i* indexes the observations from 1 to *n*, and *εi* is the *i* th value of the [error term](https://en.wikipedia.org/wiki/Error_term). In general, the greater the ESS, the better the estimated model performs.

If �^ and �^� are the estimated [coefficients](https://en.wikipedia.org/wiki/Coefficient), then

�^�=�^+�^1�1�+�^2�2�+⋯

is the *i*th predicted value of the response variable. The ESS is then:

ESS=∑�=1�(�^�−�¯)2.

where �^�the value estimated by the regression line .[[1]](https://en.wikipedia.org/wiki/Explained_sum_of_squares" \l "cite_note-1)

In some cases (see below): [total sum of squares](https://en.wikipedia.org/wiki/Total_sum_of_squares) (TSS) = explained sum of squares (ESS) + [residual sum of squares](https://en.wikipedia.org/wiki/Residual_sum_of_squares) (RSS).

Q.3) What is the need of regularization in machine learning?

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

Q.4) What is Gini–impurity index?

Gini Index or Gini impurity measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.

But what is actually meant by ‘impurity’?

If all the elements belong to a single class, then it can be called pure. The degree of Gini Index varies between 0 and 1,

where,  
'0' denotes that all elements belong to a certain class or there exists only one class (pure), and  
'1' denotes that the elements are randomly distributed across various classes (impure).

A Gini Index of '0.5 'denotes equally distributed elements into some classes.

Q.5) Are unregularized decision-trees prone to overfitting? If yes, why?

Decision trees are prone to overfitting, especially when a tree is particularly deep. 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.

Q.6) What is an ensemble technique in machine learning?

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.

Q.7) What is the difference between Bagging and Boosting techniques?

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. It attempts to increase the weight of an observation if it was erroneously categorized. Boosting creates good predictive models in general.

## What are the similarities between bagging and boosting?

Bagging and boosting are ensemble strategies that aim to produce N learners from a single learner. They sample at random and create many training data sets. They arrive at their final decision by averaging N learners' votes or selecting the voting rank of the majority of them. They reduce variance and increase stability while reducing errors.

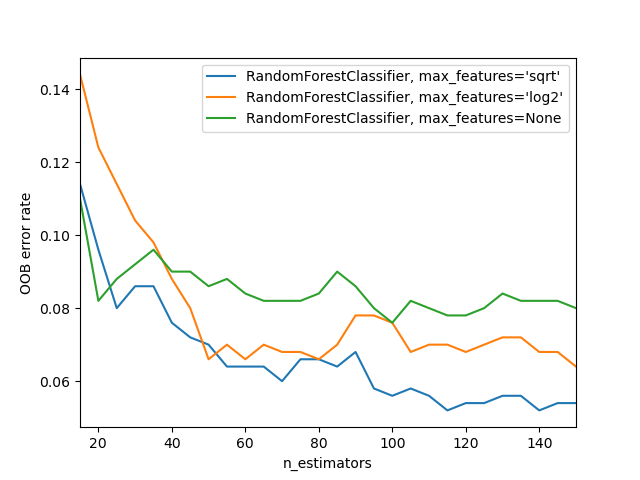
Q.8) What is out-of-bag error in random forests?

The RandomForestClassifier is trained using bootstrap aggregation, where each new tree is fit from a bootstrap sample of the training observations ��=(��,��). 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 RandomForestClassifier to be fit and validated whilst being trained [[1]](https://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html#id2).

The example below demonstrates how the OOB error can be measured at the addition of each new tree during training. The resulting plot allows a practitioner to approximate a suitable value of n\_estimators at which the error stabilizes.

[[1](https://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html#id1)]

T. Hastie, R. Tibshirani and J. Friedman, “Elements of Statistical Learning Ed. 2”, p592-593, Springer, 2009.



Q.9) What is K-fold cross-validation?

Cross validation is an evaluation method used in machine learning to find out how well your machine learning model can predict the outcome of unseen data. It is a method that is easy to comprehend, works well for a limited data sample and also offers an evaluation that is less biased, making it a popular choice.

The data sample is split into ‘k’ number of smaller samples, hence the name: K-fold Cross Validation. You may also hear terms like four fold cross validation, or ten fold cross validation, which essentially means that the sample data is being split into four or ten smaller samples respectively.

# How is k-fold cross validation performed?

The general stratergy is quite straight forward and the following steps can be used:

1. First, shuffle the dataset and split into k number of subsamples. (It is important to try to make the subsamples equal in size and ensure k is less than or equal to the number of elements in the dataset).
2. In the first iteration, the first subset is used as the test data while all the other subsets are considered as the training data.
3. Train the model with the training data and evaluate it using the test subset. Keep the evaluation score or error rate, and get rid of the model.
4. Now, in the next iteration, select a different subset as the test data set, and make everything else (including the test set we used in the previous iteration) part of the training data.
5. Re-train the model with the training data and test it using the new test data set, keep the evaluation score and discard the model.
6. Continue iterating the above k times. Each data subsamples will be used in each iteration until all data is considered. You will end up with a k number of evaluation scores.
7. The total error rate is the average of all these individual evaluation scores.

Q.10) What is hyper parameter tuning in machine learning and why it is done?

Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

Q.11) What issues can occur if we have a large learning rate in Gradient Descent?

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.

Q.12) Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

Non-linear problems can't be solved with logistic regression because it has a linear decision surface.

Q.13) Differentiate between Adaboost and Gradient Boosting

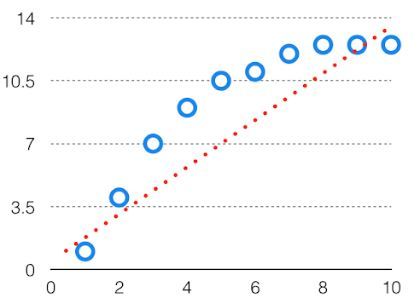
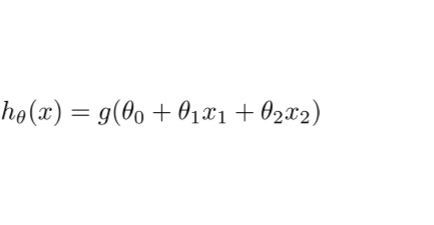
AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

Q.14) What is bias-variance trade off in machine learning?

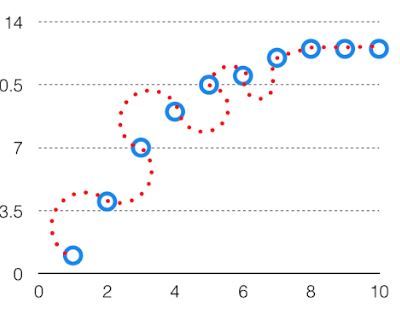
In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

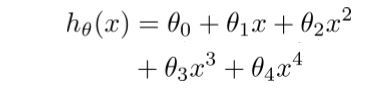
It is important to understand prediction errors (bias and variance) when it comes to accuracy in any machine learning algorithm. There is a tradeoff between a model’s ability to minimize bias and variance which is referred to as the best solution for selecting a value of **Regularization** constant. Proper understanding of these errors would help to avoid the overfitting and underfitting of a data set while training the algorithm.

**Bias**  
The bias is known as the difference between the prediction of the values by the ML model and the correct value. Being high in biasing gives a large error in training as well as testing data. Its recommended that an algorithm should always be low biased to avoid the problem of underfitting.  
By high bias, the data predicted is in a straight line format, thus not fitting accurately in the data in the data set. Such fitting is known as **Underfitting of Data**. This happens when the hypothesis is too simple or linear in nature. Refer to the graph given below for an example of such a situation.

 In such a problem, a hypothesis looks like follows.  
  
Variance  
The variability of model prediction for a given data point which tells us spread of our data is called the variance of the model. The model with high variance has a very complex fit to the training data and thus is not able to fit accurately on the data which it hasn’t seen before. As a result, such models perform very well on training data but has high error rates on test data.  
When a model is high on variance, it is then said to as **Overfitting of Data**. Overfitting is fitting the training set accurately via complex curve and high order hypothesis but is not the solution as the error with unseen data is high.  
While training a data model variance should be kept low.

The high variance data looks like follows.



In such a problem, a hypothesis looks like follows.  
  
Bias Variance Tradeoff

If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex ( hypothesis with high degree eq.) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Trade-off or Bias Variance Trade-off.

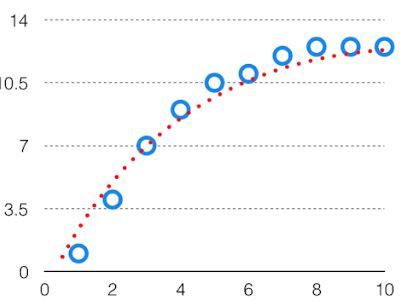
This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time. For the graph, the perfect tradeoff will be like.

Q.15) Give short description each of Linear, RBF, Polynomial kernels used in SVM.

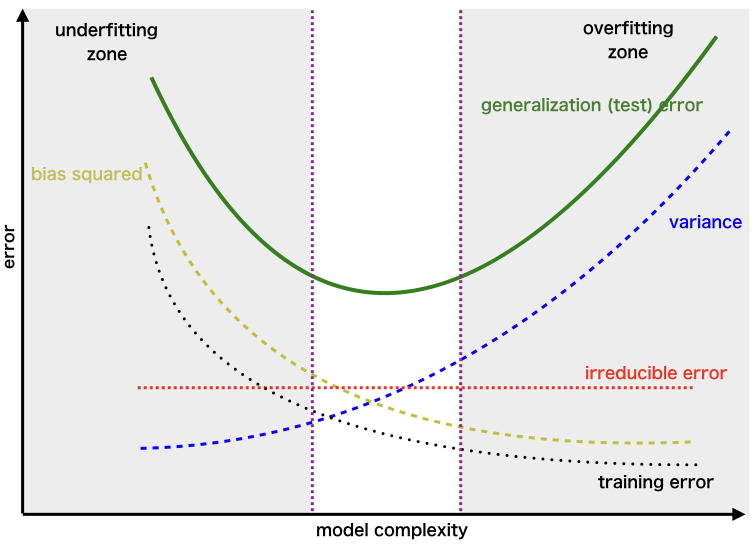
In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

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.



The best fit will be given by hypothesis on the tradeoff point.

The error to complexity graph to show trade-off is given as –  
  
This is referred to as the best point chosen for the training of the algorithm which gives low error in training as well as testing data.

Statistics

Q.1) d

Q.2) c

Q.3) c

Q.4) b

Q.5) d

Q.6) b

Q.7) a

Q.8) a

Q.9) b

Q.10) a