# Worksheet 5 Solutions

## Statistic Solutions:

1. d) Expected
2. d) All of these
3. c) 6
4. c) Chisqared distribution
5. c) F- Distribution
6. b) Hypothesis
7. a) Null Hypothesis
8. a) Two-tailed test
9. b) Research Hypothesis
10. a) np

## Machine Learning Solutions:

1. RSS gives us the total square of the distance of actual points from the regression line. But if we focus on a single residual, we can say that it is the distance that is not captured by the regression line. Therefore, RSS as a whole gives us the variation in the target variable that is not explained by our model.R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.
2. Total variation in target variable is the sum of squares of the difference between the actual values and their mean.

   $$TSS = \sum (y_i - \bar{y})^2$$

   Residual Sum of Squares

   As we discussed before, RSS gives us the total square of the distance of actual points from the regression line. But if we focus on a single residual, we can say that it is the distance that is not captured by the regression line. Therefore, RSS as a whole gives us the variation in the target variable that is not explained by our model. If TSS gives us the total variation in Y, and RSS gives us the variation in Y not explained by X, then TSS-RSS gives us the variation in Y that is explained by our model! We can simply divide this value by TSS to get the proportion of variation in Y that is explained by the model. And this our R-squared statistic!

   R-squared = (TSS-RSS)/TSS

   = Explained variation/ Total variation

   = 1 – Unexplained variation/ Total variation

Sum of squares is a measure of how a data set varies around a central number (like the mean). The Explained SS tells you how much of the variation in the dependent variable your model explained :

Explained SS = Σ(Y-Hat – mean of Y)2

3. Overfitting is a phenomenon that occurs when a Machine Learning model is constraint to training set and not able to perform well on unseen data.Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. The commonly used regularisation techniques are :
    - L1 regularisation
    - L2 regularisation
    - Dropout regularisation
4. 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 if there exists only one class, and 1 denotes that the elements are randomly distributed across various classes. A Gini Index of 0.5 denotes equally distributed elements into some classes.
5. 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.Ideally, we would like to minimize both error due to bias and error due to variance.Overfitting can be one problem that describes if your model no longer generalizes well. Overfitting happens when any learning processing overly optimizes training set error at the cost test error. While it's possible for training and testing to perform equality well in cross validation, it could be as the result of the data being very close in characteristics, which may not be a huge problem. In the case of decision tree's they can learn a training set to a point of high granularity that makes them easily overfit. Allowing a decision tree to split to a granular degree, is the behavior of this model that makes it prone to learning every point extremely well — to the point of perfect classification — ie: overfitting.
6. Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. a machine learning ensemble consists of only a concrete finite set of alternative models, but typically allows for much more flexible structure to exist among those alternatives.
7. Bagging is a method of merging the same type of predictions. Boosting is a method of merging different types of predictions.Bagging decreases variance, not bias, and solves over-fitting issues in a model. Boosting decreases bias, not variance. In Bagging, each model receives an equal weight. In Boosting, models are weighed based on their performance.Models are built independently in Bagging. New models are affected by a previously built model's performance in Boosting. In Bagging, training data subsets are drawn randomly with a replacement for the training dataset. In Boosting, every new subset comprises the elements that were misclassified by previous models. Bagging is usually applied where the classifier is unstable and has a high variance. Boosting is usually applied where the classifier is stable and simple and has high bias.
8. Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging). Bagging uses subsampling with replacement to create training samples for the model to learn from. OOB error is the mean

prediction error on each training sample $x_i$, using only the trees that did not have $x_i$ in their bootstrap sample

9. Cross-validation,sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set). The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).
10. In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned. The same kind of machine learning model can require different constraints, weights or learning rates to generalize different data patterns. These measures are called hyperparameters, and have to be tuned so that the model can optimally solve the machine learning problem. Hyperparameter optimization finds a tuple of hyperparameters that yields an optimal model which minimizes a predefined loss function on given independent data.The objective function takes a tuple of hyperparameters and returns the associated loss.Cross-validation is often used to estimate this generalization performance.
11. When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training errorFor gradient descent to reach the local minimum we must set the learning rate to an appropriate value, which is neither too low nor too high. This is important because if the steps it takes are too big, it may not reach the local minimum because it bounces back and forth between the convex function of gradient descent (see left image below). If we set the learning rate to a very small value, gradient descent will eventually reach the local minimum but that may take a while.
12. Logistic regression is known and used as a linear classifier. It is used to come up with a hyperplane in feature space to separate observations that belong to a class from all the other observations that do not belong to that class. The decision boundary is thus linear. Robust and efficient implementations are readily available (e.g. scikit-learn) to use logistic regression as a linear classifier.
13. 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.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. In the case of AdaBoost, the shifting is done by up-weighting observations that were misclassified before, while Gradient Boosting identifies the difficult observations by large residuals computed in the previous iterations.

14. In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters. The bias–variance dilemma or bias–variance problem is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set:[1][2]

- The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).
- The bias–variance decomposition is a way of analyzing a learning algorithm's expected generalization error with respect to a particular problem as a sum of three terms, the bias, variance, and a quantity called the irreducible error, resulting from noise in the problem itself.

15. Polynomial Kernel : 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 : 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. Training a SVM with a Linear Kernel is Faster than with any other Kernel.Radial Bias Function(RBF) : 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.Since the value of the RBF kernel decreases with distance and ranges between zero (in the limit) and one (when x = x'), it has a ready interpretation as a similarity measure.[2] The feature space of the kernel has an infinite number of dimensions.

## SQL Worksheet solution :

1. SELECT * FROM movie;
2. SELECT `title`, MAX(`runtime `) AS longest_runtime FROM movie ;
3. SELECT `title`, MAX(`revenue `) AS highest_revenue FROM movie limit 1;
4. SELECT `title`, MAX(`runtime `) AS longest_runtime FROM movie limit 1;
5. SELECT `m.title`, `p.person_name`,`g.gender`,`c.character_name`, `c.cast_order` from movie AS m INNER JOIN movie_cast AS c ON m.movie_id = c.movie_id INNER JOIN person AS p ON c.person_id = p.person_id INNER JOIN gender AS g ON c.gender_id = g.gender_id;
6. Select c.country_name, count(m.title) as movie_counts from movie as m inner join production_country as pc on pc.movie_id = m.movie_id inner join country c as c.country_id = pc.country_id;
7. SELECT * FROM genre;
8. SELECT l.language_name,COUNT( m.title) FROM movie AS m INNER JOIN movie_languages AS ml ON ml.movie_id = m.movie_id INNER JOIN language AS l ON l.language_id = ml.language_id;
9. SELECT m.title AS movie_name, COUNT(cr.person_id) AS no_of_crews, COUNT(ca.person_id) as no_of_cast FROM movie as m INNER JOIN movie_crew as cr ON cr.movie_id = m.movie_id INNER JOIN movie_cast ca ON ca.person_id = cr.person_id;
10. SELECT `title`FROM movie order by popularity desc limit 10;
11. SELECT `title`, revenue from movie where revenue in ( select revenue from movie order by revenue desc limit 3) order by revenue limit 1;
12. SELECT * FROM MOVIE where movie_status = ‘rumoured’;
13. Select m.title, max(m.revenue) from movie as m inner join production_country as pc on pc.movie_id = m.movie_id inner join country as c on c.country_id = pc.country_id AND c.country_name = ‘United States of America’;
14. Select m.title, pn.company_name from movie as m inner join movie_company as mc on mc.movie_id = m.movie_id inner join production_company pc on pc.company_id = mc.company_id;
15. Select title from movie order by budget desc limit 20;