# Machine Learning
# Answer sheet

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?**

   R-squared is a better measure of goodness of fit model in regression.

   R-squared value always lies between 0 and 1. A higher R-squared value indicates a higher amount of variability being explained by our model and vice-versa.
   If we had a really low RSS value, it would mean that the regression line was very close to the actual points. This means the independent variables explain the majority of variation in the target variable. In such a case, we would have a really high R-squared value.

   $$\uparrow \text{R-squared} = 1 - \frac{\text{RSS}\downarrow}{\text{TSS}}$$

   On the contrary, if we had a really high RSS value, it would mean that the regression line was far away from the actual points. Thus, independent variables fail to explain the majority of variation in the target variable. This would give us a really low R-squared value.

   $$\downarrow \text{R-squared} = 1 - \frac{\text{RSS}}{\text{TSS}} \uparrow$$

   So, this explains why the R-squared value gives us the variation in the target variable given by the variation in independent variables. Therefore R-squared is a better measure.

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.**

   Total Sum of Squares:- TSS is the sum of square of difference of each data point from the mean value of all the values of target variable (y).

   Explained Sum od Squares:- the explained sum of squares (ESS), alternatively known as the model sum of squares or sum of squares due to regression, is a quantity used in describing how well a model, often a regression model, represents the data being modelled.

   Residual Sum of Squares:- The residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regression model. The smaller the residual sum of squares, the better your model fits your data; the greater the residual sum of squares, the poorer your model fits your data.

Below mentioned is the equation relating these three metrics with each other

TSS = ESS + RSS

where TSS is Total Sum of Squares, ESS is Explained Sum of Squares and RSS is Residual Sum of Squares. The aim of Regression Analysis is explain the variation of dependent variable Y.

**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.

**4. What is Gini–impurity index?**

Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

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

Yes, unregularized decision-trees 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. they are very data intensive - that is, they examine the data in a lot of ways. At each node, they look at every possible split of every independent variable.

**6. What is an ensemble technique in machine learning?**

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in several machine learning competitions, where the winning solutions used ensemble methods.

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

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.

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.

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

   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. The out-of-bag error is an error estimation technique often used to evaluate the accuracy of a random forest and to select appropriate values for tuning parameters, such as the number of candidate predictors that are randomly drawn for a split.

9. **What is K-fold cross-validation?**

   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.

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

    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.

    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.

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

    The learning rate can see as step size, η. As such, gradient descent is taking successive steps in the direction of the minimum. If the step size η is too large, it can (plausibly) "jump over" the minima we are trying to reach, i.e.. we overshoot.

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

    No, we cannot use the Logistic Regression for classification.

    It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set. It is very fast at classifying unknown records. Non-linear problems can't be solved with logistic regression because it has a linear decision surface.

**13. Differentiate between Ada boost 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.

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.

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

Bias-variance trade-off is tension between the error introduced by the bias and the error produced by the variance. To understand how to make the most of this trade-off and avoid underfit or overfit our model, lets first learn that Bias an Variance.

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

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 default kernel used within the sklearn SVM classification algorithm and can be described with the following formula: where gamma can be set manually and has to be >0.

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.

# SQL
# Answer Sheet

1. **Write SQL query to show all the data in the Movie table.**
   ```
   Select * from movie;
   ```
2. **Write SQL query to show the title of the longest runtime movie.**
   ```
   Select title from movie order by runtime desc limit 1;
   ```
3. **Write SQL query to show the highest revenue generating movie title**.
   ```
   Select title from movie order by revenue desc limit 1;
   ```
4. **Write SQL query to show the movie title with maximum value of revenue/budget.**
   ```
   Select title from movie order by budget desc limit 1;
   ```
5. **Write a SQL query to show the movie title and its cast details like name of the person, gender, character name, cast order.**
   ```
   Select title, gender, character_name, cast_order, person_name from movie
   a inner join movie_cast b on a.movie_id=b.movie_id inner join gender
   c on c.gender_id=b.gender_id inner join person d on d.person_id= b.person_id;
   ```
6. **Write a SQL query to show the country name where maximum number of movies has been produced, along with the number of movies produced.**
   ```
   Select country_name, count(country_name) as count from country as
   a inner join production_country as b on b.country_id=a.country_id
   group by country_name order by count desc limit 1;
   ```
7. **Write a SQL query to show all the genre_id in one column and genre_name in second column.**
   ```
   Select * from genre;
   ```
8. **Write a SQL query to show name of all the languages in one column and number of movies in that particular column in another column.**
   ```
   Select language_name,movie_id,count(language_name) from movie_languages as
   a join language as b on a.language_id=b.language_id
   group by language_name
   order by count(language_name) desc;
   ```
9. **Write a SQL query to show movie name in first column, no. of crew members in second column and number of cast members in third column.**
   ```
   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. **Write a SQL query to list top 10 movies title according to popularity column in decreasing order.**
    ```
    Select title from movie order by popularity desc limit 10;
    ```
11. **Write a SQL query to show the name of the 3rd most revenue generating movie and its revenue.**
    ```
    Select title from movie order by revenue desc offset 3 limit 1;
    ```
12. **Write a SQL query to show the names of all the movies which have “rumoured” movie status.**
    ```
    Select title from movie where movie_status like ‘rumored’;
    ```

13. **Write a SQL query to show the name of the “United States of America” produced movie which generated maximum revenue.**
    Select title, revenue from movie a inner join production_country b on b.movie_id = a.movie_id inner join country c on c.country_id = b. country_id where country_name= ‘United State of America’;
14. **Write a SQL query to print the movie_id in one column and name of the production company in the second column for all the movies.**
    Select m.movie_id, pc.company_name from movie m inner join movie_company mc on mc.movie_id = m.movie_id inner join production_company pc on pc.company_id =mc.company_id;
15. **Write a SQL query to show the title of top 20 movies arranged in decreasing order of their budget.**
    Select title from movie order by budget desc limit 20;

# Statistics
# Answers Sheet

1. D
2. C
3. C
4. B
5. C
6. B
7. A
8. A
9. B
10. A