Worksheet No. 3 Machine Learning
By- Rakesh Shinde

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?
Ans: R-squared and Residual Sum of Squares (RSS) are both measures of the goodness of fit of a regression model, but they capture different aspects of the fit. R-squared, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, with a higher value indicating a better fit. R-squared is useful for comparing different models or for determining the proportion of the variability in the dependent variable that is explained by the model. On the other hand, Residual Sum of Squares (RSS) measures the total amount of unexplained variance in the dependent variable that remains after the model has been fit. It is the sum of the squared differences between the actual and predicted values of the dependent variable. A lower value of RSS indicates a better fit. RSS is useful for evaluating the accuracy of the predictions of the model.
In general, both measures are important and should be considered together when evaluating the goodness of fit of a model. However, R-squared is often considered to be a better measure of goodness of fit than RSS because it provides a single number that summarizes the proportion of variance in the dependent variable that is explained by the model, which is more interpretable and easier to compare across models.

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.
Ans:
The Total SS (TSS or SST) tells you how much variation there is in the dependent variable.
Total SS = Σ(Yi – mean of Y)^{2}.
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}.
The residual sum of squares tells you how much of the dependent variable's variation your model did not explain. It is the sum of the squared differences between the actual Y and the predicted Y:
Residual Sum of Squares = Σ e2
sum of squares in regression uses the equation:
$\Sigma(y - \overline{y})^2 = \Sigma(\hat{y} - \overline{y})^2 + \Sigma(y - \hat{y})^2$

3. What is the need of regularization in machine learning?
Ans: Regularization is an effective technique to prevent a model from overfitting. It allows us to reduce the variance in a model without a substantial increase in its bias. This method allows us to develop a more generalized model even if only a few data points are available in our dataset

4. What is Gini–impurity index?
Ans:

Gini-impurity tells us the what is the probability of misclassifying an observation, lower the probability purer the class.
Gini index varies between values 0 and 1, where 0 expresses the purity of classification, i.e. All the elements belong to a specified class or only one class exists there. And 1 indicates the random distribution of elements across various classes. The value of 0.5 of the Gini Index shows an equal distribution of elements over some classes.

5. Are unregularized decision-trees prone to overfitting? If yes, why?
Ans: Yes the unregularized decision trees are prone to overfitting, model overfits when it memorizes the noise of the training data and fails to capture important patterns.

6. What is an ensemble technique in machine learning?
Ans:
Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

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

Bagging	**Boosting**
Bagging involves fitting many decision trees on different samples of the dataset and averaging the predictions.	Boosting involves adding ensemble members sequentially to correct the predictions made by prior models and outputs a weighted average of the predictions.
The original dataset is divided into multiple subsets, selecting observations with replacement.	The new subset contains the components mistrained by the previous model.
This method combines predictions that belong to the same type.	This method combines predictions that belong to the different types.
Bagging decreases variance.	Boosting decreases bias.
Base classifiers are trained parallelly.	Base classifiers are trained sequentially.
The models are created independently.	The model creation is dependent on the previous ones.

8. What is out-of-bag error in random forests?
Ans:
OOB (out-of-bag) errors are an estimate of the performance of a random forest classifier or regressor on unseen data. In scikit-learn, the OOB error can be obtained using the oob_score_ attribute of the random forest classifier or regressor.

9. What is K-fold cross-validation?
Ans:

The data sample is split into ‘k’ number of smaller samples, hence the name: K-fold Cross Validation. E.g. 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.
If let us say that that we have 5 fold cross validation means data is divided into 5 equal parts and every time one part is test data and remaining 4 will go for training, this is done alternatively and the score is seen.

10. What is hyper parameter tuning in machine learning and why it is done?
Ans:
Hyperparameter tuning is the process of finding the optimal hyperparameters (a parameter of the model whose value influences the learning process and whose value cannot be estimated from the training data) for any given machine learning algorithm. We do it to increase the model predictivity by providing optimal set of parameters

11. What issues can occur if we have a large learning rate in Gradient Descent?
Ans:
Since gradient descent learning rate is large then we do not get the optimal or minimal solution it will converge to the suboptimal solution instead.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
Ans;
No. Logistic regression cannot be used for non linear data, because Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters. Or in other words, the output cannot depend on the product (or quotient, etc.) of its parameters!

13. Differentiate between Adaboost and Gradient Boosting.
Ans:

S.No	Adaboost	Gradient Boost
1	An additive model where shortcomings of previous models are identified by high-weight data points.	An additive model where shortcomings of previous models are identified by the gradient.
2	The trees are usually grown as decision stumps.	The trees are grown to a greater depth usually ranging from 8 to 32 terminal nodes.
3	Each classifier has different weights assigned to the final prediction based on its performance.	All classifiers are weighed equally and their predictive capacity is restricted with learning rate to increase accuracy.
4	It gives weights to both classifiers and observations thus capturing maximum variance within data.	It builds trees on previous classifier’s residuals thus capturing variance in data.

14. What is bias-variance trade off in machine learning?
Ans:
Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand, if our model has large number of parameters then it’s going to have high variance and low bias. So, we need to find the right/good balance without overfitting and underfitting the data.
To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.
Ans:
Linear Kernel
It is the most basic type of kernel, usually one dimensional in nature. It proves to be the best function when there are lots of features. The linear kernel is mostly preferred for text-classification problems as most of these kinds of classification problems can be linearly separated.
Linear kernel functions are faster than other functions.
Linear Kernel Formula
$F(x, xj) = sum(x.xj)$

Polynomial Kernel
It is a more generalized representation of the linear kernel. It is not as preferred as other kernel functions as it is less efficient and accurate.
Polynomial Kernel Formula
$F(x, xj) = (x.xj+1)^d$

Gaussian Radial Basis Function (RBF)
It is one of the most preferred and used kernel functions in svm. It is usually chosen for non-linear data. It helps to make proper separation when there is no prior knowledge of data.
Gaussian Radial Basis Formula
$F(x, xj) = exp(-gamma * ||x - xj||^2)$
The value of gamma varies from 0 to 1. You have to manually provide the value of gamma in the code. The most preferred value for gamma is 0.1.