MACHINE LEARNING ASSIGNMENT - 5

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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?
Answer...
R-squared: R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

The Formula for R-Squared Is

R2=1− (Unexplained Variation/Total Variation)

Residual Sum of Squares(RSS): The residual sum of squares measures the amount of error remaining between the regression function and the data set. A smaller residual sum of squares figure represents a regression function. Residual sum of squares–also known as the sum of squared residuals–essentially determines how well a regression model explains or represents the data in the model.

Before assessing numeric measure of the goodness of fit, like r-squared, you should evaluate the residual plot. Residual plot can expose a bias model for more effectively than the numeric output by displaying problemetic patterns in the residuals. If your model is biased, you can not trust the result. If your residual plots looks good, go ahead and access your r-squared and other statistics.

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.
Answer...
Total Sum of Squares-->The Total SS (TSS or SST) tells you how much variation there is in the dependent variable.

$TSS = \Sigma(Yi - \text{mean of } Y)^2$.

Explained Sum of Squares--> The Explained SS tells you how much of the variation in the dependent variable your model explained.

$ESS = \Sigma(Y\text{-Hat} - \text{mean of } Y)^2$.

Residual Sum of Squares---> 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:

$\text{Residual Sum of Squares} = \Sigma\, e2$

The relationship between the three types of sum of squares can be summarized by the following equation:

Relationship Formula $TSS=SSR+SSE$

3. What is the need of regularization in machine learning?
Answer...
Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. It is a technique which makes slight modification to the learning algorithm such that model generalizes better.

The commonly used regularisation techniques are :
a- L1 regularisation also Lasso
b- L2 regularisation also known as Ridge
c- ElasticNet

4. What is Gini–impurity index?
Answer...
The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits.

Def: Gini Impurity tells us what is the probability of misclassifying an observation. Note that the lower the Gini the better the split.

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

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

technique multiple weak learners are ensemble into one unit make them boosted and amplify their learning and compute the results.
There are two ensemble technique: a)-->Bagging also called parallel technique.
b)-->Boosting also called sequential technique.

We use these technique when large number of rows and columns are present.

7. What is the difference between Bagging and Boosting techniques?
Answer...
Differences Between Bagging and Boosting :-

Sr.No	BAGGING	BOOSTING
1	Simplest way of combining predictions that belong to the same type.	A way of combining predictions that belong to the different types.
2	Aim to decrease variance, not bias.	Aim to decrease bias, not variance.
3	Each model receives equal weight.	Models are weighted according to their performance.
4	Each model is built independently.	New models are influenced by performance of previously built models.
5	Different training data subsets are randomly drawn with replacement from the entire training dataset.	Every new subsets contains the elements that were misclassified by previous models.

6	Bagging tries to solve over-fitting problem.	Boosting tries to reduce bias.
7	If the classifier is unstable (high variance), then apply bagging.	If the classifier is stable and simple (high bias) the apply boosting.
8	Random forest	Gradient boosting.

8. What is out-of-bag error in random forests?
Answer...
The RandomForestClassifier is trained using bootstrap aggregation, where each new tree is fit from a bootstrap sample of the training observations z1=(x1,y1)
The out-of-bag (OOB) error is the average error for each z1 calculated using predictions from the trees that do not contain z1 in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained.

9. What is K-fold cross-validation?
Answer...
K-Fold CV is where a given data set is split into a K number of sections/folds where each fold is used as a testing set at some point. Lets take the scenario of 5-Fold cross validation(K=5). Here, the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

10. What is hyper parameter tuning in machine learning and why it is done?
Answer....
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.
Hyperparameters are important because they directly control the behaviour of the training algorithm and have a significant impact on the performance of the model is being trained. “A good choice of hyperparameters can really make an algorithm shine” .....Easy to manage a large set of experiments for hyperparameter tuning.

11. What issues can occur if we have a large learning rate in Gradient Descent?
Answer...
When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error. When the learning rate is too small, training is not only slower, but may become permanently stuck with a high training error.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
Answer...
No 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.
Answer...

Adaboost is more about ‘voting weights’ and Gradient boosting is more about ‘adding gradient optimization’. Adaboost increases the accuracy by giving more weightage to the target which is misclassified by the model. At each iteration, Adaptive boosting algorithm changes the sample distribution by modifying the weights attached to each of the instances. It increases the weights of the wrongly predicted instances and decreases the ones of the correctly predicted instances.

Gradient boosting calculates the gradient (derivative) of the Loss Function with respect to the prediction (instead of the features). Gradient boosting increases the accuracy by minimizing the Loss Function (error which is difference of actual and predicted value) and having this loss as target for the next iteration.
Gradient boosting algorithm builds first weak learner and calculates the Loss Function. It then builds a second learner to predict the loss after the first step. The step continues for third learner and then for fourth learner and so on until a certain threshold is reached.

14. What is bias-variance trade off in machine learning?
Answer...
Bias: Amount of error introduced by approximating real world phenomena with a simplified model.

Variance: It shows how much your model's test error changes based on variation in the training data.

Trade-off: It is tension between the error introduced by the bias and variance.

Bias-Variance Trade off-> It is the property of a set of predictive models where by models with a lower bias in parameter estimation have higher variance of the parameter estimates across samples and vice versa.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.
Answer...
Linear kernel: A linear kernel is used as normal dot product of any two given observation. The product between two vectors is the sum of the multiplication
of each pair of input values.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.

k(x,x1) = sum(x*x1)

RBF(Radial Basis Function): In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning
algorithms.RBF can map an input space in infinite dimensional space. In particular, it is commonly used in support vector machine
classification.

k(x,x1) = exp(-gamma*sum(x-x1^2))
Here gamma is a parameter which ranges from 0 to 1.

Polynomial: A polynomial kernel is a more generalized form of the linear kernel. The polynomial kernel can distinguish curved or non linear input space.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.

$k(x,x1) = 1+sum(x*x1)^d$ where d = degree of polynomial