Machine Learning Worksheet 3

1. The linear, polynomial and RBF kernel are simply different in case of making the hyperplane

decision boundary between the classes. The kernel functions are used to map the dataset

(linear/nonlinear) into a higher dimensional space with view to making it linear dataset.

Usually linear and polynomial kernels are less time consuming and provides less accuracy than

the RBF kernel.

2. R-Squared gives a better measure of goodness of fit of model in regression because R-square

represents the proportion of variance in our data, the closer to one, the better the fit. But RSS is

the sum of the squared distances between the actual values and the predicted values.

3. RSS is a statistical technique used to measure the amount of variance in a dataset.

ESS tells how much of the variation between observed data and predicted data is being explained

by the model proposed. Mathematically, it is the sum of squares of difference between predicted

data and mean data.

TSS is defined as the sum of squares of difference between actual data and mean of data.

TSS = RSS + ESS

4. Gini impurity index measures the probability of a particular variable being wrongly classified

When it is randomly chosen. A Gini impurity index of 0.5 denotes equally distributed elements in

same classes. The lower the Gini impurity index, the better the split.

5. Yes, un-regularized decision trees prone to overfitting, because in un-regularized decision tree,

the tree can continue to fit till each data point to a different leaf in a tree. This results to

overfitting of the model.

6. Ensemble methods is a machine learning technique used that combines several base models

in order to produce one optimal predictive model. Ensemble techniques is usually used to

average the predictions of different models to get a better prediction. Using of ensemble

techniques will decrease the variance and bias and improves predictions.

7. Bagging:

Bagging technique is used to reduce the variance of a model. The objective is to create

several subsets of data from training sample chosen randomly with replacement.

Each collection of subset data is used to train the models. As a result, we get an

ensemble of different models. Average of all the predictions from the different models

are used which is more robust than single model.

Boosting:

Boosting is used to create a collection of predictors. Boosting grants machine learning models

to improve their accuracy of predictions. Boosting algorithms converts weak learners to strong

learners. Firstly, a model is built from the training data. Then the second model is built which

tries to correct the errors present in the first model. This procedure is continued and models are

added until the complete training dataset is predicted correctly.

8. Out-of-bag error is a method of measuring the prediction error of random forests, boosted

Decision trees and other machine learning models utilizing bootstrap aggregating to sub-sample

data samples used for training.

9. K-Fold CV is where a given dataset is split into a K number of folds/sections where each

fold is used as testing set at some point. Let’s take the scenario of 5-Fold CV(K=5).

Here, the dataset is split into 5 folds. In the first iteration, the first fold is used to test the model

and the remaining folds are used to train the model. In the second iteration, the second fold is

used to test the model and the remaining folds are used to train the model. This process is

repeated until each fold of the 5 folds are used as testing set.

10. Hyper-parameters cannot be directly learned from the regular training process. They are

Usually fixed before the actual training process begins. These parameters express important

properties of the model such as complexity or how fast it should learn. Models can have many

hyper-parameters and finding the best combination of parameters is done by hyper-parameter

tuning. It includes two strategies i.e., GridSearchCV and RandomizedSearchCV.

11. When the learning rate is too large, gradient descent can increase rather than decrease

decrease the training error. When using high learning rates, it is possible to encounter a

positive feedback loop in which large weights induce large gradients which then induce

a large update to the weights. If these updates consistently increase the size of the weights,

the weights rapidly move’s away from the origin until numerical overflow occurs.

12. 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. Bias-Variance trade-off is the

property of a set of predictive models whereby models with a lower bias in parameter

estimation have a higher variance of the parameter estimates across samples and vice versa.

13. Regularizations are the techniques used to reduce the error by fitting a function appropriately

on the given training set and avoid overfitting.

14. Ada-boost classifier fits a sequence of weak learners on different weighted training data.

It starts by predicting original data set and gives equal weight to each observation. If

prediction is incorrect using the first learner then it gives higher weight to observation which

have been predicted incorrectly. Being an iterative process, it continues to add learners

until a limit is reached in the number of models or accuracy.

In Gradient boosting, it trains many models sequentially. Each new model gradually minimizes

the loss function of the whole system using gradient descent method. The principle idea

behind this algorithm is to construct new base learners which can be maximally correlated

with negative gradient of the loss function, associated with the whole ensemble.

15. No, logistic regression cannot be used for classification of non-linear data. Logistic Regression

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.