**MACHINE LEARNING – WORKSHEET 3**

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

1.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. One of the examples where there are a lot of features, is **Text Classification**, as each alphabet is a new feature. So we mostly use Linear Kernel in Text Classification.

The Radial basis function kernel, also called the RBF kernel, or Gaussian kernel, is a kernel that is in the form of a radial basis function (more speciﬁcally, a Gaussian function)

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), the polynomial kernel is a [kernel function](https://en.wikipedia.org/wiki/Kernel_function) commonly used with [support vector machines](https://en.wikipedia.org/wiki/Support_vector_machine) (SVMs) and other [kernelized](https://en.wikipedia.org/wiki/Kernel_trick" \o "Kernel trick) 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.

2. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit of model in regression and why?

R-squared because the RSS depends on the scale of the dependent variable whereas the R2 does not.

3. 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.

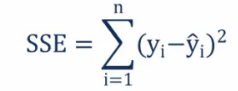
The total sum of squares is a variation of the values of a [dependent variable](https://corporatefinanceinstitute.com/resources/knowledge/terms/dependent-variable/) from the sample mean of the dependent variable. Essentially, the total sum of squares quantifies the total variation in a [sample](http://www.webmath.com/sampledata.html). It can be determined using the following formula:



The Explained sum of squares describes how well a regression model represents the modeled data. A higher regression sum of squares indicates that the model does not fit the data well.



The residual sum of squares essentially measures the variation of modeling errors. In other words, it depicts how the variation in the dependent variable in a regression model cannot be explained by the model. Generally, a lower residual sum of squares indicates that the regression model can better explain the data while a higher residual sum of squares indicates that the model poorly explains the data.



Total sum of squares ( TSS ) = explained sum of squares (ESS)+ residual sum of squares (RSS)

4. What is Gini –impurity index?

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

Yes, because unregularized tress grow to a depth until they don’t get pure leaf nodes. Sometimes, it may happen that there is only one point in each leaf nodes. Now, changing a single point in dataset will cause a change in the entire model, which is nothing but over-fitting.

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 a number of machine learning competitions, where the winning solutions used ensemble methods.

Bagging, boosting and stacking are examples of ensmebling techniques used widely.

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

Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification.

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

Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of [random forests](https://en.wikipedia.org/wiki/Random_forest), [boosted decision trees](https://en.wikipedia.org/wiki/Gradient_boosting), and other [machine learning](https://en.wikipedia.org/wiki/Machine_learning) models utilizing [bootstrap aggregating](https://en.wikipedia.org/wiki/Bootstrap_aggregating) (bagging) to sub-sample data samples used for training.

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.

In the k-fold cross validation method, all the entries in the original training data set are used for both training as well as validation. Also, each entry is used for validation just once.

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

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), hyperparameter optimization or tuning is the problem of choosing a set of optimal [hyperparameters](https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)" \o "Hyperparameter (machine learning)) for a learning algorithm. A hyperparameter is a [parameter](https://en.wikipedia.org/wiki/Parameter) whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

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

When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error. We might overshoot the minima, and keep bouncing along the ridges of the "valley" without ever reaching the minima.

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

Bias is the simplifying assumptions made by the model to make the target function easier to approximate. Variance is the amount that the estimate of the target function will change given different training data. Trade-off is tension between the error introduced by the bias and the variance.

13. What is the need of regularization in machine learning?

Regularization is a regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

Regularization adds a penalty on the different parameters of the model to reduce the freedom of the model. Hence, the model will be less likely to fit the noise of the training data and will improve the generalization abilities of the model.

14. Differentiate between Adaboost and Gradient Boosting

The basic idea of boosting (an ensemble learning technique) is to combine several weak learners into a stronger one. The general idea of boosting algorithms is to try predictors sequentially, where each subsequent model attempts to fix the errors of its predecessor.  
  
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.

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

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. Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries