WORKSHEET

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.
Answer - LINEAR KERNAL ----> k(x,y) = 1+xy +xy min(x,y) -((x+y)/2)min(x,y)^2 + (min(x,y)^3)/3
Linear kernal is used when the data is Linearly separable , that is , it can be separated using a single line. It is useful when dealing with large sparse data vectors. It is often used in text categorization. The splines kernel also performs well in regression problems.

POLYNOMIAL KERNAL --->K (x, y) = tanh (\gamma.{x^T y}+{r})^d
It represents the similarity of vectors in training set of data in a feature space over polynomials of the original variables used in kernel.

RBF KERNAL --->K (x, y) = e ^ - (\gamma{||x - y||^2})
When we have some large amount of data and we don't know much about how the data is link to each other than we use Radial Basic Functionand it is the most used type of kernel function . Because it has localized and finite response along the entire x-axis. It is used to perform transformation, when there is no prior knowledge about data.

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??
Answer --R-squared is a goodness-of-fit measure for regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. For instance, small R-squared values are not always a problem, and high R-squared values are not necessarily good!

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.
Answer-- TSS(TOTAL SUM OF SQUARES)--.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.

ESS(EXPLAINED SUM OF SQUARES)--> The explained sum of squares (ESS) is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model — for example, yi = a + b1x1i + b2x2i .In other words, ESS is how much the predicted values in the sample vary, while TSS is how much the actual values vary.

RSS(RESIDUAL SUM OF SQUARES)--> Residual sum of squares (RSS) is also known as the sum of squared residuals (SSR) or sum of squared errors (SSE) of prediction. It is an amount of the difference between data and an estimation model.
The residual vector
RSS=y^{T}y-y^{T}X(X^{T}X)^{{-1}}X^{T}y.
TSS==y^{T}y-2y^{T}{\bar y}+{\bar y}^{T}{\bar y}.
The explained sum of squares, defined as the sum of squared deviations of the predicted values from the observed mean of y, is

ESS=({\hat y}-{\bar y})^{T}({\hat y}-{\bar y})={\hat y}^{T}{\hat y}-2{\hat y}^{T}{\bar y}+{\bar y}^{T}{\bar y}.

TSS = ESS + RSS.#

4. What is Gini –impurity index?
Answer -- 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. Gini Impurity tells us what is the probability of misclassifying an observation. Note that the lower the Gini the better the split . it measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. A Gini Index of 0.5 denotes equally distributed elements into some classes.

5. Are unregularized decision-trees prone to overfitting? If yes, why?
Answer -- Unregularized 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?
Answwer -Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.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

7. What is the difference between Bagging and Boosting techniques?
Answer -- Bagging and Boosting are similar in that they are both ensemble techniques, where a set of weak learners are combined to create a strong learner that obtains better performance than a single one.

Bagging --->Bagging is used when the goal is to reduce the variance of a decision tree classifier. Here 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 their decision trees.

Boosting --->Boosting is used to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analysing data for errors. Consecutive trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree.

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 z= (x,y). The out-of-bag (OOB) error is the average error for each z calculated using predictions from the trees that do not contain z 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 -- 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.

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

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. What is bias-variance trade off in machine learning?
Answer -- 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?
Answer -- This is a form of 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. A simple relation for linear regression looks like this.

14. Differentiate between Adaboost and Gradient Boosting?
Answer -- Gradient boosting --> Gradient boosting generates learners during the learning process. It build first learner to predict the values/labels of samples, and calculate the loss (the difference between the outcome of the first learner and the real value). It will build a second learner to predict the loss after the first step. The step continues to learn the third, forth… until certain threshold.

Adaboost ---> Adaboost requires users specify a set of weak learners (alternatively, it will randomly generate a set of weak learner before the real learning process). It will learn the weights of how to add these learners to be a strong learner. The weight of each learner is learned by whether it predicts a sample correctly or not. If a learner is mispredict a sample, the weight of the learner is reduced a bit. It will repeat such process until converge.

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