MACHINE LEARNING – WORKSHEET 3

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

1. Give short description for each Linear, RBF, Polynomial kernels used in SVM.

The linear, RBF or Gaussian, and Polynomial kernels are used differently for making the hyperplane decision boundary between the classes.

The kernel functions are used to map the original dataset (linear/nonlinear) into a higher dimensional space with a view to make it a linear dataset.

Usually linear and polynomial kernels are less time consuming and provide less accuracy than the rbf or Gaussian kernels.

The k cross validation is used to divide the training set into k distinct subsets. Then every subset is used for training and others k-1 are used for validation in the entire training phase. This is done for the better training of the classification task.

Usually it should be done but not mandatory.

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 is a better measure of goodness of fit model in regression because it has a range (0 to 1) while RSS does not have an upper limit (more the value is RSS, poorer the model fitting is, whereas more the value in R squared, better the model fitting).

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 (TSS or SST) is a quantity that appears as part of a standard way of presenting results of such analyses. For a set of observations $y_i, i \leq n$, it is defined as the sum over all squared differences between the observations and their overall mean.

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The explained sum of squares (ESS), alternatively known as the model sum of squares or sum of squares due to regression, is a quantity used in describing how well a model, often a regression model, represents the data being modelled. In particular, the explained sum of squares measures how much variation there is in the modelled values.

If $\hat{a}$ and $\hat{b}_i$ are the estimated coefficients, then

$$\hat{y}_i = \hat{a} + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \cdots$$

is the i^{th} predicted value of the response variable. The ESS is then:

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

A residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model. Regression is a measurement that helps determine the strength of the relationship between a dependent variable and a series of other changing variables or independent variables.

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.

Equation relating these three metrics with each other:

TSS = ESS + RSS

4. What is Gini –impurity index?

Gini - Impurity index is the probability of *incorrectly* classifying a randomly chosen element in the dataset if it were randomly labeled *according to the class distribution* in the dataset. It's calculated as

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

where C is the number of classes and $p(i)$ is the probability of randomly picking an element of class i.

When training a decision tree, the best split is chosen by **maximizing the Gini Gain**, which is calculated by subtracting the weighted impurities of the branches from the original impurity.

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

Yes. Unregularized decision trees will have low bias and high variance.

6. What is an ensemble technique in machine learning?

Ensemble method is a technique that creates multiple models and then combines them to produce improved results. Ensemble methods usually produce more accurate solutions than a single model would.

It is a supervised learning technique for combining multiple weak learners/ models to produce a strong learner. Ensemble models work better when we ensemble them with low correlation.

They are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking).

They can be classified as:

- Voting and Averaging Based Ensemble Methods (Majority Voting, Weighted Voting, Simple Averaging, Weighted Averaging)
- Stacking Multiple Machine Learning Models
- Bootstrap Aggregating (Bagging)
- Boosting - Converting weak models to strong ones

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, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging) to sub-sample data samples used for training. OOB is the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample.

Subsampling allows one to define an out-of-bag estimate of the prediction performance improvement by evaluating predictions on those observations which were not used in the building of the next base learner.

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. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups
3. For each unique group:
    1. Take the group as a hold out or test data set
    2. Take the remaining groups as a training data set
    3. Fit a model on the training set and evaluate it on the test set
    4. Retain the evaluation score and discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

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

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 derived via training.

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. ... Efficiently search the space of possible hyperparameters. 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?

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, then [the weights] rapidly moves away from the origin until numerical overflow occurs increasing the

training error.

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

The goal of any supervised machine learning algorithm is to achieve low bias and low variance. The parameterization of machine learning algorithms is often a battle to balance out bias and variance.

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.

- Increasing the bias will decrease the variance.
- Increasing the variance will decrease the bias.

High bias and low variance leads to underfitting whereas low bias and high variance leads to overfitting.

The bias–variance tradeoff 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. What is the need of regularization in machine learning?

Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting by adding information

The commonly used regularisation techniques are :

1. L1 regularisation
2. L2 regularisation
3. Dropout regularisation

14. Differentiate between Adaboost and Gradient Boosting

Adaboost	Gradient Boosting
AdaBoost stands for Adaptive Boosting. In Adaboost, ‘shortcomings’ are identified by high-weight data points.	In Gradient Boosting, ‘shortcomings’ (of existing weak learners) are identified by gradients.

Adaboost is more about ‘voting weights’.	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. It first builds 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?

No. Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters.

Logistic regression has traditionally been used to come up with a hyperplane that separates the feature space into classes. But if we suspect that the decision boundary is nonlinear we may get better results by attempting some nonlinear functional forms for the logit function.