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.
Kernelization with svm implies internal feature transformation. With appropriate kernel function we can even seperate non linearly seperable datasets which could not be solved using simple logistic regression

Rbf kernel (radial basis kernel most popular or general based kernel)

Kernel(x1,x2) = exp(-||x1-x2||/2*(sigma^2)) —— Mathematical formula

To be more intutive ,we can think of kernel function as a measure of simmilarity and distance as a measure of dissimilarity

RBF kernel is simmilar to a gauusian PDF.RBF kernel is a nice approximation of K nearest neighbours Sigma used in K-NN is similalar to K . as K increases we are overfitting and trying to minimise the loss function similar is case with sigma. Always remember if we are not sure about the best kernel function we can use RBF kernel
The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized.

$$k(x, y) = (\alpha x^T y + c)^d$$

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-square can take on any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model. For example, an R-square value of 0.8234 means that the fit explains 82.34% of the total variation in the data about the average. If we increase the number of fitted coefficients in your model, R-square will increase although the fit may not improve in a practical sense. To avoid this situation, you should use the degrees of freedom adjusted R-square statistic described below. the residual sum of squares, also known as the sum of squared residuals or the sum of squared estimate of errors, is the sum of the squares of residuals. It is a measure of the discrepancy between the data and an estimation model.

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 coefficient of determination, R^2, is a statistical measure that shows the proportion of *variation* explained by the estimated regression line. Variation refers to the sum of the squared differences between the values of *Y* and the mean value of *Y*, expressed mathematically as

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$

R^2 always takes on a value between 0 and 1. The closer R^2 is to 1, the better the estimated regression equation fits or explains the relationship between *X* and *Y*.

The expression

- $$ESS=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$$
- **Residual sum of squares (RSS):** This expression is also known as *unexplained variation* and is the portion of total variation that measures discrepancies (errors) between the actual values of *Y* and those estimated by the regression equation.

  You compute the RSS with the formula

  $$RSS=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$$

The smaller the value of RSS relative to ESS, the better the regression line fits or explains the relationship between the dependent and independent variable.

The sum of RSS and ESS equals TSS.

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$

R^2 is the ratio of explained sum of squares (ESS) to total sum of squares (TSS):

$$R^2=\frac{ESS}{TSS}$$

4. What is Gini –impurity index?

Gini index and entropy are the criteria for calculating information gain. Decision tree algorithms use information gain to split a node.

Both gini and entropy are measures of impurity of a node. A node having multiple classes is impure whereas a node having only one class is pure. Entropy in statistics is analogous to entropy in thermodynamics where it signifies disorder. If there are multiple classes in a node, there is disorder in that node.

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

Over-fitting is the phenomenon in which the learning system tightly fits the given training data so much that it would be inaccurate in predicting the outcomes of the untrained data.

In decision trees, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus it ends up with branches with strict rules of sparse data. Thus this effects the accuracy when predicting samples that are not part of the training set.

One of the methods used to address over-fitting in decision tree is called pruning which is done after the initial training is complete. In pruning, you trim off the branches of the tree, i.e., remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed. This is done by segregating the actual training set into two sets: training data set, D and validation data set, V. Prepare the decision tree using the segregated training data set, D. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set, V.

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. Its a powerful collaborative filtering algorithm. Ensemble is a Machine Learning concept in which the idea is to train multiple models using the same learning algorithm. The ensembles take part in a bigger group of methods, called multiclassifiers, where a set of hundreds or thousands of learners with a common objective are fused together to solve the problem.

The second group of multiclassifiers contain the hybrid methods. They use a set of learners too, but they can be trained using different learning techniques.

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

Bagging and Boosting get N learners by generating additional data in the training stage. N new training data sets are produced by **random sampling with replacement** from the original set. By sampling with replacement some observations may be repeated in each new training data set.In the case of Bagging, any element has the same probability to appear in a new data set. However, for Boosting the observations are weighted and therefore some of them will take part in the new sets more often

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

Random forests technique involves sampling of the input data with replacement (bootstrap sampling). In this sampling, about one thrird of the data is not used for training and can be used to testing.These are called the out of bag samples. Error estimated on these out of bag samples is the out of bag error.

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.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

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?

Parameters which define the model architecture are referred to as hyperparameters and thus this process of searching for the ideal model architecture is referred to as *hyperparameter tuning*.

hyperparameters are not model parameters and they cannot be directly trained from the data. *Model parameters* are learned during training when we optimize a loss function using something like gradient descent.

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

A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution, whereas a learning rate that is too small can cause the process to get stuck.The challenge of training deep learning neural networks involves carefully selecting the learning rate. It may be the most important hyperparameter for the model.

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

In statistics and machine learning, 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.

- The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

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

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.*Here Y represents the learned relation and β represents the coefficient estimates for different variables or predictors(X).*

$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + ...+ \beta pXp$The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Now, this will adjust the coefficients based on your training data. If there is noise in the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.

14. Differentiate between Adaboost and Gradient Boosting

Both are boosting algorithms which means that they convert a set of weak learners into a single strong learner.At each iteration, adaptive boosting 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. The weak learner thus focuses more on the difficult instances. After being trained, the weak learner is added to the strong one according to his performance (so-called alpha weight). The higher it performs, the more it contributes to the strong learner.

On the other hand, gradient boosting doesn't modify the sample distribution. Instead of training on a newly sample distribution, the weak learner trains on the remaining errors (so-called pseudo-residuals) of the strong learner. It is another way to give more importance to the difficult instances. At each iteration, the pseudo-residuals are computed and a weak learner is fitted to these pseudo-residuals. Then, the contribution of the weak learner (so-called multiplier) to the strong one isn't computed according to his performance on the newly distribution

sample but using a gradient descent optimization process. The computed contribution is the one minimizing the overall error of the strong learner.

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

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. Solving for the model parameters can be more challenging but the optimization modules in scipy can help.