MACHINE LEARNING ASSIGNMENT – 5

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

   Answer:-

   R-squared is a relative measure, and it does not provide information about the absolute goodness of fit or the quality of the model's predictions.

   A smaller RSS indicates a better fit because it means that the model's predictions are closer to the actual observed values.

   RSS is useful when you want to evaluate the absolute goodness of fit, focusing on the magnitude of the prediction errors. If minimizing prediction errors is a primary concern, RSS is a more appropriate choice.

2. **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 = ESS + RSS, where TSS is Total Sum of Squares, ESS is Explained Sum of Squares and RSS is Residual Sum of Suqares. The aim of Regression Analysis is explain the variation of dependent variable Y

3. **What is the need of regularization in machine learning.**

Answer

The main aim of regularization is to reduce the over-complexity of the machine learning models and help the model learn a simpler function to promote generalization.

4. What is Gini–impurity index?

Answer:

Gini impurity has a maximum value of 0.5, which is the worst we can get, and a minimum value of 0 means the best we can get.

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

Answer:

Decision trees are a popular and powerful method for data mining, as they can handle both numerical and categorical data, and can easily interpret the results. However, decision trees can also suffer from overfitting, which means that they learn too much from the training data and fail to generalize well to new data.

6. **What is an ensemble technique in machine learning?**

Answer:

Ensemble learning is a machine learning technique that enhances accuracy and resilience in forecasting by merging predictions from multiple models. It aims to mitigate errors or biases that may exist in individual models by leveraging the collective intelligence of the ensemble.

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

Answer:

In bagging, models are trained independently in parallel on different random subsets of the data. Whereas in boosting, models are trained sequentially, with each model learning from the errors of the previous one.

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

Answer:

in machine learning and data science, it is crucial to create a trustful system that will work well with the new, unseen data. Overall, there are a lot of different approaches and methods to achieve this generalization. Out-of-bag error is one of these methods for validating the machine learning model.

9. What is K-fold cross-validation?

Answer:

K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold are averaged to estimate the model's generalization performance.

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

Answer:

When you're training machine learning models, each dataset and model needs a different set of hyperparameters, which are a kind of variable. The only way to determine these is through multiple experiments, where you pick a set of hyperparameters and run them through your model. This is called hyperparameter tuning.

11. 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 suffer from divergence. This means that weights increase exponentially, resulting in exploding gradients which can cause problems such as instabilities and overly high loss values.

12. Differentiate between Adaboost and Gradient Boosting.

Answer:

The main difference between these two algorithms is that Gradient boosting has a fixed base estimator i.e., Decision Trees whereas in AdaBoost we can change the base estimator according to our needs.

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

Answer:

Bias in ML is an sort of mistake in which some aspects of a dataset are given more weight and/or representation than others. A skewed outcome, low accuracy levels, and analytical errors result from a dataset that is biased that does not represent a model's use case accurately.

14. 15. Give short description each of Linear, RBF, Polynomial kernels used in SVM

Answer:

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

What is polynomial kernel in SVM?

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized 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.

What is a linear kernel in SVM?

A linear kernel is a type of kernel function used in machine learning, including in SVMs (Support Vector Machines). It is the simplest and most commonly used kernel function, and it defines the dot product between the input vectors in the original feature space.