MACHINE LEARNING - Worksheet Set 1

1. What is the advantage of hierarchical clustering over K-means clustering?
A) Hierarchical clustering is computationally less expensive
B) In hierarchical clustering you don’t need to assign number of clusters in beginning
C) Both are equally proficient
D) None of these

Ans: B) In hierarchical clustering you don’t need to assign number of clusters in beginning

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?
A) max_depth
B) n_estimators
C) min_samples_leaf
D) min_samples_splits

Ans: A) max_depth

3. Which of the following is the least preferable resampling method in handling imbalance datasets?
A) SMOTE
B) RandomOverSampler
C) RandomUnderSampler
D) ADASYN

Ans: B) RandomOverSampler

4. Which of the following statements is/are true about “Type-1” and “Type-2” errors?
1. Type1 is known as false positive and Type2 is known as false negative.
2. Type1 is known as false negative and Type2 is known as false positive.
3. Type1 error occurs when we reject a null hypothesis when it is actually true.

A) 1 and 2
B) 1 only
C) 1 and 3
D) 2 and 3

Ans: C) 1 and 3

5. Arrange the steps of k-means algorithm in the order in which they occur

   1. Randomly selecting the cluster centroids
   2. Updating the cluster centroids iteratively
   3.Assigning the cluster points to their nearest center

   A) 3-1-2
   B) 2-1-3
   C) 3-2-1
   D) 1-3-2

   Ans: D) 1-3-2

6. Which of the following algorithms is not advisable to use when you have limited CPU resources and time, and when the data set is relatively large?

   A) Decision Trees
   B) Support Vector Machines
   C) K-Nearest Neighbors
   D) Logistic Regression

   Ans: C) K-Nearest Neighbors

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?

   A) CART is used for classification, and CHAID is used for regression.
   B) CART can create multiway trees (more than two children for a node), and CHAID can only create binary trees (a maximum of two children for a node).
   C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)
   D) None of the above

   Ans: C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)

8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?

   A) Ridge will lead to some of the coefficients to be very close to 0
   B) Lasso will lead to some of the coefficients to be very close to 0
   C) Ridge will cause some of the coefficients to become 0
   D) Lasso will cause some of the coefficients to become 0

   Ans: A) Ridge will lead to some of the coefficients to be very close to 0
   D) Lasso will cause some of the coefficients to become 0

9. Which of the following methods can be used to treat two multi-collinear features?

A) remove both features from the dataset
B) remove only one of the features
C) Use ridge regularization
D) use Lasso regularization

Ans: B) remove only one of the features
C) Use ridge regularization
D) use Lasso regularization

10. After using linear regression, we find that the bias is very low, while the variance is very high. What are the possible reasons for this?

A) Overfitting
B) Multicollinearity
C) Underfitting
D) Outliers

Ans: A) Overfitting
D) Outliers

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Below are the situations in which One-hot encoding must be avoided.

- ➔ When the categorical features present in the dataset are ordinal
- ➔ Dataset is large
- ➔ Due to High cardinality in categorical variables, when we encode the categorical data to numerical

Alternative Encoding Technique

- ➔ **Label encoding:** This method assigns a unique integer value to each category based on the natural ordering of the categories. It avoids the curse of dimensionality and allows capturing the order of the categories
- ➔ **Binary encoding:** This method represents a categorical variable as a binary code, where each digit corresponds to a category. It allows for the capture of the order of the categories
- ➔ **Count encoding:** This method replaces a categorical variable with the number of times each category shows up in the dataset
- ➔ **Target encoding:** This method replaces a categorical variable with the average value of the target variable for each category.
- ➔ **Helmert encoding:** This method creates a new feature for each category: the difference between the mean of the target variable for that category and the mean for the last category.

- **Frequency encoding:** Replaces a categorical variable with the number of times each category appears in the dataset.

- **Backward difference encoding:** This method creates a new feature for each category: the difference between the mean of the target variable for that category and the mean for all categories below it.

- **Leave-one-out encoding:** This method replaces all but the current instance of a categorical variable with the mean value of the target variable.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly

Resampling is a widely-adopted technique for dealing with imbalanced datasets. Resampling changes the dataset into a more balanced one by adding instances to the minority class or deleting ones from the majority class.
Class imbalance problem can occur in binary classification problems as well as multi-class classification problems, but most techniques can be used on either.

Oversampling Techniques

Oversampling methods duplicate examples in the minority class to obtain a balanced dataset, below are the widely used oversampling methods

- **Random Oversampling:** Random oversampling method randomly duplicate records from the minority class. This technique got high chance of over-fitting as it simply replicates random minority class.

- **SMOTE - SMOTE stands for Synthetic Minority Oversampling Technique** it consists of creating or synthesizing elements or samples from the minority class rather than creating copies based on those that exist already. This is used to avoid model over-fitting. To create a synthetic instance, SMOTE finds the K-nearest neighbours of each minority instance, randomly selects one of them and then calculates linear interpolations to produce a new minority instance in the neighborhood. Changing this instance features one at a time by a random amount the new points are added between the neighbors.

- **ADASYN - ADASYN stands for Adaptive Synthetic sampling**, and as SMOTE does, ADASYN measures the K-nearest neighbours for all minority instances, then calculates the class ratio of the minority and majority instances to create new samples, this technique differs from SMOTE based on density distributions.

Undersampling Techniques

Undersampling methods delete or select a subset of examples from the majority class, below are the widely used oversampling methods.

- **Random Undersampling:** Random undersampling randomly deletes records from the majority class.

- **NearMiss Undersampling:** It’s based on the nearest neighbour algorithm and has a lot of variations. NearMiss-1 selects examples from the majority class that have the smallest average distance to the three closest examples from the minority class. NearMiss-2 selects examples from the majority class that have the smallest average distance to the three furthest examples from the minority class. NearMiss-3 involves selecting a given number of majority class examples for each example in the minority class that are closest.

- ➔ **Tomek Links Undersampling:** Tomek links are pairs of very close instances that belong to different classes. Samples are near the borderline between classes, by removing the examples of the majority class of each pair, it increase the space between the two classes and move toward balancing the dataset by deleting those points.
- ➔ **Edited Nearest Neighbor Rule (ENN):** This technique removes any instance from the majority class whose class label is different from the class label of at least two of its three nearest neighbor.
- ➔ **Neighborhood Cleaning Rule (NCR):** Neighborhood Cleaning Rule (or NCR) deals with the majority and minority samples separately when sampling the datasets.
  If the instance belongs to the majority class and the classification given by its three nearest neighbors is the opposite of the class of the chosen instance - then the chosen instance is removed
  If the instance belongs to the minority class and it’s misclassified by its three nearest neighbors - then the nearest neighbor that belong to the majority class are removed

13. What is the difference between SMOTE and ADASYN sampling techniques?

SMOTE finds the K-nearest neighbors of each minority instance, it randomly selects one of them and then calculates linear interpolations to produce a new minority instance in the neighborhood. Changing this instance features one at a time by a random amount the new points are added between the neighbors.

ADASYN Its purpose is to generate data for minority class samples that are harder to learn, as compared to those minority samples that are easier to learn measures the K-nearest neighbors for all minority instances, then calculates the class ratio of the minority and majority instances to create new samples.

Both the methods are oversampling techniques but ADASYN takes into account the density of distribution to distribute the data points evenly.

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

GridSearchCV is a method to search the best parameters from the grid of given parameters and is useful when we are looking for the best parameter for the target model and dataset. Grid Search uses a different combination of all the specified parameters and their values and calculates the performance for each combination and the best parameters is extracted to apply for a predictive model.

We can use GridSearchCV on large datasets as we will apply it on the train dataset (20% to 30% sample), applying for the entire dataset is not advisable.
Moreover GridSearchCV makes the process time-consuming and expensive based on the number of parameters involved.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief

Evaluation metric is an integral part of regression models. Loss functions take the model’s predicted values and compare them against the actual values, by knowing the amount of deviation between the predicted value and the actual value, we can train our model accordingly. This difference between the actual value and the predicted value is called the loss. A high loss value means the model has poor performance.
The Sklearn.metrics module implements several loss, score, and utility functions to measure regression performance. Below are the few evaluation metrics for regression model.

- ➔ **Mean Squared Error (MSE):** MSE is one of the most common regression loss functions. In MSE we calculate the error by squaring the difference between the predicted value and actual value and averaging it across the dataset. It is always non negative and values close to zero are better.

- ➔ **Root Mean Squared Error (RMSE):** RMSE is computed by taking the square root of MSE It measures the average magnitude of the errors and is concerned with the deviations from the actual value. RMSE value with zero indicates that the model has a perfect fit. The lower the RMSE, the better the model and its predictions. A higher RMSE indicates that there is a large deviation from the residual to the ground truth.

- ➔ **Mean Absolute Error (MAE):** It is calculated by taking the absolute difference between the predicted values and the actual values and averaging it across the dataset. MAE is the arithmetic average of absolute errors. The lower the MAE, the higher the accuracy of a model.

- ➔ **R2_Score**: It represents the proportion of variance that has been explained by the independent variables in the model. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model, depending on the ratio of total deviation of results described by the model. A higher value of R2_score indicates better results.