Machine Learning Worksheet 8

Q1- What is the advantage of hierarchical clustering over K-means clustering?
Answer- B)-In hierarchical clustering you don't need to assign number of clusters in beginning

Q2- Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?
Answer-A) max_depth

Q3- Which of the following is the least preferable resampling method in handling imbalance datasets?
Answer- B) RandomOverSampler

Q4- Which of the following statements is/are true about "Type-1" and "Type-2" errors?
Answer- C) 1 and 3

Q5- Arrange the steps of k-means algorithm in the order in which they occur:
Answer- D) 1-3-2

Q6- 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?
Answer- B) Support Vector Machines

Q7- What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?
Answer- 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)

Q8- In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?
Answer-A) Ridge will lead to some of the coefficients to be very close to &
D) Lasso will cause some of the coefficients to become 0.

Q9- Which of the following methods can be used to treat two multi-collinear features
Answer- B) remove only one of the features &
D) use Lasso regularization

Q10- 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?
Answer- A) Overfitting & D) Outliers

Q11- In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?
Answer- We should not use the One Hot Encoding method when :

- When the categorical features present in the dataset are ordinal i.e for the data being like Junior, Senior, Executive, Owner.
- When the number of categories in the dataset is quite large. One Hot Encoding should be avoided in this case as it can lead to high memory consumption.

To fight the curse of dimensionality, binary encoding might be a good alternative to one-hot encoding because it creates fewer columns when encoding categorical variables. Ordinal encoding is a good choice if the order of the categorical variables matters.

Q12- In case of data imbalance problem in classification, what techniques can be used to balance the dataset? ? Explain them briefly.
Answer- 5 Techniques to Handle Imbalanced Data For a Classification Problem

1. **Choose Proper Evaluation Metric:-** For an imbalanced class dataset F1 score is a more appropriate metric. It is the harmonic mean of precision and recall and the expression is –

   F1=2*P*R/(P+R)

   Where,
   P=Precision
   R=Recall
2. **Resampling** (Oversampling and Undersampling)- This technique is used to upsample or downsample the minority or majority class. When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called undersampling.
3. **SMOTE** - Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class .SMOTE looks into minority class instances and use k nearest neighbor to select a random nearest neighbor, and a synthetic instance is created randomly in feature space.
4. **K-fold Cross Validation**- This technique involves cross validating the dataset after it is generated by the process of oversampling since it makes predicting the minority class easier.
   - Exclude some amount of data for validation that will not be used for oversampling, feature selection, and model building;
   - Follow up by oversampling the minority class without the excluded data in the training set;
   - Depending on the number of folds, i.e., ‘K’—Repeat it ‘K’ times

5. **Ensembling resampled datasets**- The most obvious—but not an all round way—to handle imbalanced data is to use more data. Therefore, ensembling different resampled datasets is another technique that can overcome problems while generalising using random forest or logistic regression. This comes along with identifying the rare class that was discarded during generalising the training dataset.

Q13- What is the difference between SMOTE and ADASYN sampling techniques?
Answer- **SMOTE(Synthetic Minority Oversampling Technique)**
SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.
ADASYN(Adaptive Synthetic Sampling Approach)
ADASYN is a generalized form of the SMOTE algorithm. This algorithm also aims to oversample the minority class by generating synthetic instances for it. But the difference here is it considers the density distribution, ri which decides the no. of synthetic instances generated for samples which difficult to learn. Due to this, it helps in adaptively changing the decision boundaries based on the samples difficult to learn. This is the major difference compared to SMOTE.

Q14- . What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?
Answer- **GridSearchCV is also known as GridSearch cross-validation**: an internal cross-validation technique is used to calculate the score for each combination of parameters on the grid.
GridSearchCV is a technique for finding the optimal parameter values from a given set of parameters in a grid. It’s essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made.
For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible. We can shift to Random Search CV where the algorithm will randomly choose the combination of parameters.

Q15- List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief?
Answer- **There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model**

- **Mean absolute error (MAE) :** Represents average error
- **Mean squared error (MSE) :** Similar to MAE but noise is exaggerated and larger errors are “punished”. It is harder to interpret than MAE as it’s not in base units, however, it is generally more popular.

- **Root mean squared error (RMSE):** Most popular metric, similar to MSE, however, the result is square rooted to make it more interpretable as it’s in base units. It is recommended that RMSE be used as the primary metric to interpret your model