SET 1 Machine Learning Assignment 8

Q 1 B

Q2 A

Q3 B

Q4 A

Q5 D

Q6 C

Q7 C

Q8 A and B

Q9 C and D

Q10 A and C

Q11 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.

Q12 Two approaches to make a balanced dataset out of an imbalanced one are **under-sampling and over-sampling**

Q13 The key difference between ADASYN and SMOTE is that **the former uses a density distribution**, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed ...

Q14 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.

Q15 There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: **Mean Squared Error (MSE). Root Mean Squared Error (RMSE). Mean Absolute Error (MAE)**

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

To better understand, let’s take an example you have input data and output data and use Linear Regression, which draws a best-fit line.

- The MAE you get is in the same unit as the output variable.
- It is most Robust to outliers.

- MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.

The graph of MSE is differentiable, so you can easily use it as a loss function.

RMSE is clear by the name itself, that it is a simple square root of mean squared error.

- The output value you get is in the same unit as the required output variable which makes interpretation of loss easy.