17 March Assignmnet

May 22, 2023

[ ]:
```
Q1: What are missing values in a dataset? Why is it essential to handle missing␣
 ↪values? Name some
algorithms that are not affected by missing values.
```

ANS -

[ ]:
```
      Missing values in a dataset are the values that are not present for some␣
 ↪variable/s in the given dataset. Many machine learning
algorithms fail if the dataset contains missing values. However, algorithms␣
 ↪like K-nearest and Naive Bayes support data with missing values.
Missing data can lead to a lack of precision in the statistical analysis1. It␣
 ↪is common to identify missing values in a dataset and replace
them with a numeric value. This is called data imputing or missing data␣
 ↪imputation.

  It is essential to handle missing values because it can lead to biased␣
 ↪machine learning models, leading to incorrect results if the missing
values are not handled properly1. Missing values provide a wrong idea about the␣
 ↪data itself, causing ambiguity. For example, calculating an
average for a column with half of the information unavailable or set to zero␣
 ↪gives the wrong metric.

Some algorithms that are not affected by missing values include K-nearest and␣
 ↪Naive Bayes.
```

[ ]:
```

```

[ ]:
```

```

[ ]:
```
Q2: List down techniques used to handle missing data. Give an example of each␣
 ↪with python code.
```

ANS -

[ ]:
```
1. MCAR - Missing completely at random  :

    This happens if all the variables and observations have the same␣
 ↪probability of being missing. Imagine providing a child with Lego of
```

```
different colors to build a house. Each Lego represents a piece of information,␣
 ↪like shape and color. The child might lose some Legos during
the game. These lost legos represent missing information, just like when they␣
 ↪can't remember the shape or the color of the Lego they had.
That information was lost randomly, but they do not change the information the␣
 ↪child has on the other Legos.
```

[ ]:
```
2. MAR - Missing at random:

     For MAR, the probability of the value being missing is related to the␣
 ↪value of the variable or other variables in the dataset. This means
that not all the observations and variables have the same chance of being␣
 ↪missing. An example of MAR is a survey in the Data community where
data scientists who do not frequently upgrade their skills are more likely not␣
 ↪to be aware of new state-of-the-art algorithms or technologies,
hence skipping certain questions. The missing data, in this case, is related to␣
 ↪how frequently the data scientist upskills.
```

[ ]:
```
3. MNAR- Missing not at random :

    MNAR is considered to be the most difficult scenario among the three types␣
 ↪of missing data. It is applied when neither MAR nor MCAR apply.
In this situation, the probability of being missing is completely different for␣
 ↪different values of the same variable, and these reasons can
be unknown to us. An example of MNAR is a survey about married couples. Couples␣
 ↪with a bad relationship might not want to answer certain
questions as they might feel embarrassed to do so.
```

[ ]:
```

```

[ ]:
```

```

[ ]:
```
Q3: Explain the imbalanced data. What will happen if imbalanced data is not␣
 ↪handled?
```

ANS -

[ ]:
```
  Imbalanced data can cause issues when training a model, especially when the␣
 ↪dataset is small. A model needs many observations of each
target class to be able to generalize adequately, so for small datasets there␣
 ↪can simply not be enough minority class observations for the
model to learn from. Algorithms may get biased towards the majority class and␣
 ↪thus tend to predict output as the majority class. Minority
class observations look like noise to the model and are ignored by the model.␣
 ↪Imbalanced dataset gives misleading accuracy score.
```

[ ]:

[ ]:

[ ]:
```
Q4: What are Up-sampling and Down-sampling? Explain with an example when␣
 ↪up-sampling and down-
sampling are required.
```

ANS -

[ ]:
```
   Up-sampling and down-sampling are techniques used in signal processing and␣
 ↪image processing to change the resolution of a signal or image.

Down-sampling is the process of reducing the number of samples in a signal or␣
 ↪image. It is used to reduce the size of an image or signal while
retaining its essential features. For example, when you take a high-resolution␣
 ↪image and want to display it on a low-resolution screen, you
can down-sample the image to fit the screen.

Up-sampling is the process of increasing the number of samples in a signal or␣
 ↪image. It is used to increase the resolution of an image or
signal while retaining its essential features. For example, when you have a␣
 ↪low-resolution image and want to print it at high resolution, you
can up-sample the image.

Both up-sampling and down-sampling are required in many applications such as␣
 ↪audio processing, image processing, and machine learning.
For example, in machine learning, up-sampling is used to balance imbalanced␣
 ↪datasets by duplicating observations from the minority class.
Down-sampling is used to reduce computation time by reducing the dimensionality␣
 ↪of features while losing some information.
```

[ ]:

[ ]:

[ ]:
```
Q5: What is data Augmentation? Explain SMOTE.
```

ANS -

[ ]:
```
   Data augmentation is a technique used to increase the size of a dataset by␣
 ↪adding slightly modified copies of already existing data.
It is used to improve the performance of machine learning models by increasing␣
 ↪the amount and diversity of data available for training.
SMOTE (Synthetic Minority Over-sampling Technique) is one such data␣
 ↪augmentation technique that is used to address the problem of class
```

```
imbalance in machine learning datasets. SMOTE creates synthetic examples of the␣
 ↪minority class by interpolating between existing examples.
This helps to balance the class distribution and improve the performance of␣
 ↪machine learning models on imbalanced datasets.
```

[ ]:

[ ]:

[ ]:
```
Q6: What are outliers in a dataset? Why is it essential to handle outliers?
```

ANS -

[ ]:
```
   An outlier in a dataset is a value that is much larger or smaller than the␣
 ↪others or that lies an abnormal distance from the
rest of the observations. Outliers can affect the mean of the data and may␣
 ↪indicate errors, variabilities, or novelties.

It is essential to handle outliers because they can have a significant impact␣
 ↪on statistical analyses and skew the results of any hypothesis
tests. Outliers can give helpful insights into the data you're studying, and␣
 ↪they can potentially help you discover inconsistencies and
detect any errors in your statistical processes.
```

[ ]:

[ ]:

[ ]:
```
Q7: You are working on a project that requires analyzing customer data.␣
 ↪However, you notice that some of
the data is missing. What are some techniques you can use to handle the missing␣
 ↪data in your analysis?
```

ANS -

[ ]:
```
There are several ways to handle missing data in analysis. One common method is␣
 ↪imputation, which involves filling in missing values with
estimated values. There are many different methods of imputation, including␣
 ↪multiple imputations. Another method is weighting, which adjusts
the value of each observation in the data set according to its importance.␣
 ↪Missing data can also be handled by excluding observations with
missing values. Other methods include deleting rows with missing values,␣
 ↪imputing missing values for continuous or categorical variables,
using algorithms that support missing values, and imputing using deep learning␣
 ↪libraries. Another approach is to substitute the variable with
a similar indicator. The choice of method depends on the type of data and the␣
 ↪purpose of the analysis.
```

[ ]:

[ ]:

[ ]:
```
Q8: You are working with a large dataset and find that a small percentage of␣
 ↪the data is missing. What are
some strategies you can use to determine if the missing data is missing at␣
 ↪random or if there is a pattern
to the missing data?
```

ANS -

[ ]:
```
There are several methods to determine if the missing data is missing at random␣
 ↪or if there is a pattern to the missing data. One common
method is mean or median imputation. When data is missing at random, we can use␣
 ↪list-wise or pair-wise deletion of the missing observations.
However, there can be multiple reasons why this may not be the most feasible␣
 ↪option1. Another method is Multivariate Imputation by Chained
Equations (MICE) which assumes that the missing data are Missing at Random␣
 ↪(MAR)1. A third method is Random Forest.

Where data is identified as Missing Not at Random (MNAR), we can consider using␣
 ↪a model which handles missing values well - such as a Decision
Tree or Naïve Bayes model. These models can consider missingness as explanatory.
```

[ ]:

[ ]:

[ ]:
```
Q9: Suppose you are working on a medical diagnosis project and find that the␣
 ↪majority of patients in the
dataset do not have the condition of interest, while a small percentage do.␣
 ↪What are some strategies you
can use to evaluate the performance of your machine learning model on this␣
 ↪imbalanced dataset?
```

ANS -

[ ]:
```
When working with an imbalanced dataset, it is important to use evaluation␣
 ↪metrics that are appropriate for the problem. Some strategies that
can be used to evaluate the performance of a machine learning model on an␣
 ↪imbalanced dataset are:

Confusion matrix: A confusion matrix is a table that is used to evaluate the␣
 ↪performance of a classification model. It shows the number of
true positives, false positives, true negatives, and false negatives. This can␣
 ↪help you understand how well your model is performing on the
```

```
imbalanced dataset.

Precision and recall: Precision and recall are two metrics that are commonly␣
 ↪used to evaluate the performance of a classification model on an
imbalanced dataset. Precision is the number of true positives divided by the␣
 ↪number of true positives plus false positives. Recall is the
number of true positives divided by the number of true positives plus false␣
 ↪negatives.

F1 score: The F1 score is a metric that combines precision and recall into a␣
 ↪single score. It is calculated as 2 * (precision * recall) /
(precision + recall).

ROC curve: The ROC curve is a plot of the true positive rate (TPR) against the␣
 ↪false positive rate (FPR) at various threshold settings.
It can help you understand how well your model is performing on the imbalanced␣
 ↪dataset.

Class weights: You can assign class weights to your model to give more␣
 ↪importance to the minority class. This can help improve the performance
of your model on the imbalanced dataset.
```

[ ]:

[ ]:

[ ]:
```
Q10: When attempting to estimate customer satisfaction for a project, you␣
 ↪discover that the dataset is
unbalanced, with the bulk of customers reporting being satisfied. What methods␣
 ↪can you employ to
balance the dataset and down-sample the majority class?
```

ANS -

[ ]:
```
There are several ways to deal with an unbalanced dataset. One option is to␣
 ↪collect more data to try to balance the dataset, but this is often
not possible1. Another option is to downsample the majority class or upsample␣
 ↪the minority class12. Decomposing the larger class into smaller
classes or using a one-class classifier are also options3. We can oversample␣
 ↪the minority class using replacement or randomly delete rows from
the majority class to match them with the minority class2. We can also use a␣
 ↪weighted training criterion or resample the dataset to have a
greater proportion of the minority class45. The presence of an imbalance is␣
 ↪very rarely a justification in itself for resampling the data,
but rather because tasks with an imbalance often have unequal misclassification␣
 ↪costs.
```

[ ]:

[ ]:

[ ]:
```
Q11: You discover that the dataset is unbalanced with a low percentage of␣
 ↪occurrences while working on a
project that requires you to estimate the occurrence of a rare event. What␣
 ↪methods can you employ to
balance the dataset and up-sample the minority class?
```

ANS -

[ ]:
```
There are several methods that can be employed to balance an unbalanced dataset␣
 ↪and up-sample the minority class. Some of these methods
include:

1.Up-sampling the minority class: This involves adding more examples from the␣
 ↪minority class to the dataset. This can be done by randomly
                                    duplicating observations from the minority␣
 ↪class in order to reinforce its signal.

2.Down-sampling the majority class: This involves removing samples from the␣
 ↪majority class in order to balance the dataset.

3.Synthetic Minority Over-sampling Technique (SMOTE): This technique involves␣
 ↪creating synthetic samples from the minority class instead of
duplicating existing ones. SMOTE works by selecting examples that are close in␣
 ↪the feature space, drawing a line between the examples in the
feature space and drawing a new sample at a point along that line.

4.Random Under-sampling: This involves randomly removing examples from the␣
 ↪majority class until the dataset is balanced.

5.Cluster-based Over-sampling: This technique involves clustering the minority␣
 ↪class instances and generating new instances in each cluster.

6.Adaptive Synthetic Sampling (ADASYN): This technique is similar to SMOTE but␣
 ↪generates more synthetic samples for minority instances that
are harder to learn.

7.Cost-sensitive learning: This involves assigning misclassification costs␣
 ↪differently for different classes.
```