Worksheet 1

Machine Learning

Q10 to Q15 are subjective answer type questions, Answer them briefly.

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

One hot encoding is process of converting the categorical data variables to be provided to machine and deep learning algorithms which in turn improve predictions as well as classification accuracy of a model. One Hot Encoding is a common way of preprocessing categorical features for machine learning models. This type of encoding creates a new binary feature for each possible category and assigns a value of 1 to the feature of each sample that corresponds to its original category.

One Hot Encoding is to be avoided 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.
- 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

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

Imbalanced datasets are a special case for classification problem where the class distribution is not uniform among the classes. Typically, they are composed by two classes: The majority (negative) class and the minority (positive) class

1.Use the right evaluation metrics

Choosing the right evaluation metric is pretty essential whenever we work with imbalanced datasets. Generally, in such cases, the F1 Score is important. The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall.

2. Resample the training set

The common method for dealing with highly imbalanced datasets is resampling. It consists of removing samples from the majority class (under-sampling) and/or adding more examples from the minority class (over-sampling).

3.Ensemble Different Resampled Datasets

The easiest way to successfully generalize a model is by using more data.

4. Resample with Different Ratios

The previous approach can be fine-tuned by playing with the ratio between the rare and the abundant class.

5. Cluster the abundant class

Instead of relying on random samples to cover the variety of the training samples, clustering the abundant class in r groups, with r being the number of cases in r.

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

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 distributions. The latter generates the same number of synthetic samples for each original minority sample.

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

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. The performance of a model significantly depends on the value of hyperparameters.

There is no way to know in advance the best values for hyperparameters so, we need to try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources and thus we use GridSearchCV to automate the tuning of hyperparameters.

We pass predefined values for hyperparameters to the GridSearchCV function. This is done by defining a dictionary in which we mention a particular hyperparameter along with the values it can take.

For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible.

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

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)

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.

RMSE is a simple square root of mean squared error.

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

Worksheet-1

Statistics

Q13 to Q15 are subjective answers type questions.

Answers them in their own words briefly.

13. What is Anova in SPSS?

Analysis of variance i.e. ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables.

It is used as the test of means for two or more populations.

ANOVA in SPSS must have a dependent variable which should be metric (measured using an interval or ratio scale).

ANOVA in SPSS must also have one or more independent variables, which should be categorical in nature.

14. What are the assumptions of Anova?

There are three primary assumptions in ANOVA:

The responses for each factor level have a normal population distribution.

These distributions have the same variance.

The data are independent.

15. What is the difference between one way Anova and two way Anova?

The only difference between one-way and two-way ANOVA is the number of independent variables.

A one-way ANOVA has one independent variable, while a two-way ANOVA has two.

Worksheet-2

Statistics

Q13 to Q15 are subjective answers type questions,

Answers them in their own words briefly.

13.What is T distribution and Z distribution?

The T-Distribution is a measure of probability (p-value). It is used to find the statistical significance when the sample size is small, i.e., less than 30, with an obscure standard deviation. The mean of a T-Distribution is evaluated as zero, and the variance is derived as v/(v-2), where v is the degree of freedom.

The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1. Any normal distribution can be standardized by converting its values into z scores. Z scores tell you how many standard deviations from the mean each value lies.

Both the z-distribution (normal distribution) and the t-distribution are very similar. They both are bell-shaped and symmetrical, and they also can be used to model the distribution of a sample mean.

14.Is the T distribution normal?

The t-distribution is a type of normal distribution that is used for smaller sample sizes. Normally-distributed data form a bell shape when plotted on a graph, with more observations near the mean and fewer observations in the tails.

15.What does the T distribution tell us?

The t-distribution describes the standardized distances of sample means to the population mean when the population standard deviation is not known, and the observations come from a normally distributed population.

The *t*-distribution forms a bell curve when plotted on a graph. It can be described mathematically using the mean and the standard deviation.

Worksheet-2

Python

Q11 to Q13 are subjective questions, answer them briefly

11. Differentiate between a list, tuple, set and dictionary.

A list is a collection of *ordered* data. Lists are *mutable.* Lists are declared with square braces. The append() method adds a single item at the end of the list without modifying the original list. The pop() method removes the item at the given index from the list and returns it. The sort() method sorts the elements of a given list in a specific ascending or descending order. index() searches for a given element from the start of the list and returns the lowest index where the element appears. The count() method returns the number of times the specified element appears in the list. The reverse() method reverses the elements of the list.

A tuple is an *ordered* collection of data. Tuples are *immutable*. Tuples are enclosed within parenthesis. An element cannot be added to the tuple as it is immutable. Though tuples are ordered, the elements cannot be sorted. Searches the tuple for a specified value and returns the position of where it was found. The count() method returns the number of times a specified value occurs in a tuple. The reverse() method is not defined in tuples, as they are unchangeable

A set is an *unordered* collection. Sets are *mutable* and have *no duplicate elements*. Sets are represented in curly brackets. The set add() method adds a given element to a set. The pop() method removes a random item from the set. Elements in the set cannot be sorted as they are unordered. The index of a particular element is not retrieved as they are unordered. There are no count() methods in sets as they do not allow any duplicates. The sets are unordered, which refrains from applying the reverse() method.

12. Are strings mutable in python? Suppose you have a string “I+Love+Python”, write a small code to replace ‘+’ with space in python.

Strings are not mutable in Python. Strings are immutable data types which means that its value cannot be updated.

```
txt = " I+Love+Python "

x = txt.replace("+", " ")

print(x)
```

13. What does the function ord() do in python? Explain with an example. Also, write down the function for getting the data type of a variable in python.

The ord() function returns the number representing the unicode code of a specified character.

```
value1 = ord('A')

# prints the unicode value
print (value1)
```

To determine the type of a variable in Python, use the built-in type() function.

```
age = 50
print("The type is : ",type(age))
```

Worksheet-3

Machine Learning

Q14 and Q15 are subjective answer type questions, Answer them briefly.

14. Explain Linear Regression?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable.
Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.
There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

15. What is difference between simple linear and multiple linear regression?

Simple linear regression has only one x and one y variable. Multiple linear regression has one y and two or more x variables. For instance, when we predict rent based on square feet alone that is simple linear regression.

A linear regression model extended to include more than one independent variable is called a multiple regression model. It is more accurate than to the simple regression. The purpose of multiple regressions are: i) planning and control ii) prediction or forecasting.

Q13 to Q15 are subjective answers type questions. Answers them in their own words briefly.

13. How do you find the test statistic for two samples?

The test statistic for a two-sample independent t-test is calculated by taking the difference in the two sample means and dividing by either the pooled or unpooled estimated standard error. The estimated standard error is an aggregate measure of the amount of variation in both groups.

Generally, the test statistic is calculated as the pattern in your data (i.e. the correlation between variables or difference between groups) divided by the variance in the data (i.e. the standard deviation).

14. How do you find the sample mean difference?

The mean difference, or difference in means, measures the absolute difference between the mean value in two different groups. In clinical trials, it gives you an idea of how much difference there is between the averages of the experimental group and control groups.

The sample mean is a statistic obtained by calculating the arithmetic average of the values of a variable in a sample. If the sample is drawn from probability distributions having a common expected value, then the sample mean is an estimator of that expected value.

15. What is a two sample t test example?

The two-sample *t*-test (also known as the independent samples *t*-test) is a method used to test whether the unknown population means of two groups are equal or not.

What is an example of a 2 sample t test?

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The following are a few real-life examples where two-sample T-test for independent samples can be used:
Comparing the average test scores of two classes from two different schools.

Comparing the average weights of two different groups of people.

Measuring the difference in height between men and women.