Machine Learning Assignment

1) A) Least Square Error
2) A) Linear regression is sensitive to outliers
3) B) Negative
4) C) Both of them
5) C) Low bias and high variance
6) B) Predictive modal
7) D) Regularization
8) D) SMOTE
9) A) TPR and FPR
10) A) True
11) B) Apply PCA to project high dimensional data
12) A) We don't have to choose the learning rate. B) It becomes slow when number of features is very large.
13) Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.
14) The commonly used regularization techniques are:
    (A) L1 regularization - A regression model which uses **L1 Regularization** technique is called **LASSO (Least Absolute Shrinkage and Select**ion Operator) regression. **Lasso Regression** adds "absolute value of magnitude" of coefficient as penalty term to the loss function(L).
    (B) L2 regularization - **A regression model that uses** L2 regularization **technique is called** Ridge regression. Ridge regression adds" square magnitude" off coefficient as penalty term to the loss function (L).
    (C) Dropout regularization - Dropout is a regularization technique for reducing overfitting in neural networks by preventing complex co-adaptations on training data. It is a very efficient way of performing model averaging with neural networks. The term "dropout" refers to dropping out units (both hidden and visible) in a neural network.
15) Within a linear regression model tracking a stock's price over time, the error term is the difference between the expected price at a particular time and the price that was actually observed. In instances where the price is exactly what was anticipated at a particular time, the price will fall on the trend line and the error term will be zero.

Machine Learning Assignment

1) C) %
2) B) 0
3) C) 24
4) A) 2
5) D) 6
6) B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block.
7) A) It is used to raise an exception.
8) C) in defining a generator
9) C) abc2 & C) abc2
10) A) yield & B) raise

11)
```
num = (5)

 factorial = 1

 if num < 0:

 print("Sorry, factorial does not exist for negative numbers")

 elif num == 0:

print("The factorial of 0 is 1")

else:

for i in range(1,num + 1):

factorial=factorial*i

print("The factorial of",num,"is",factorial)
```

12)
```
n= int(200)

 if(n ==0 or n == 1):

printf(n,"Number is neither prime nor composite")
```

elif n>1 :

for i in range(2,n):

if(n%i == 0):

print(n,"is not prime but composite number")

break

else:

print(n,"number is prime but not composite number")

else :

print("Please enter positive number only ")

```
n= int(200)
if(n ==0 or n == 1):
    printf(n,"Number is neither prime nor composite")
elif n>1 :
    for i in range(2,n):
        if(n%i == 0):
            print(n,"is not prime but composite number")
            break
    else:
        print(n,"number is prime but not composite number")
else :
    print("Please enter positive number only ")
```

```
200 is not prime but composite number
```

13) num=int(250)

temp=num

rev=0

while(num>0):

dig=num%10

rev=rev*10+dig

num=num//10

if(temp==rev):

print("The number is palindrome!")

else:

print("Not a palindrome!")

```
num=int(250)
temp=num
rev=0
while(num>0):
    dig=num%10
    rev=rev*10+dig
    num=num//10
if(temp==rev):
    print("The number is palindrome!")
else:
    print("Not a palindrome!")
```

14) def pythagoras(opposite_side,adjacent_side,hypotenuse):

if opposite_side == str("x"):

return ("Opposite = " + str(((hypotenuse**2) - (adjacent_side**2))**0.5))

elif adjacent_side == str("x"):

return ("Adjacent = " + str(((hypotenuse**2) - (opposite_side**2))**0.5))

elif hypotenuse == str("x"):

return ("Hypotenuse = " + str(((opposite_side**2) + (adjacent_side**2))**0.5))

else:

return "You know the answer!"

print(pythagoras(3,4,'x'))

print(pythagoras(3,'x',5))

print(pythagoras('x',4,5))

print(pythagoras(3,4,5))

```
def pythagoras(opposite_side,adjacent_side,hypotenuse):
        if opposite_side == str("x"):
            return ("Opposite = " + str(((hypotenuse**2) - (adjacent_side**2))**0.5))
        elif adjacent_side == str("x"):
            return ("Adjacent = " + str(((hypotenuse**2) - (opposite_side**2))**0.5))
        elif hypotenuse == str("x"):
            return ("Hypotenuse = " + str(((opposite_side**2) + (adjacent_side**2))**0.5))
        else:
            return "You know the answer!"

print(pythagoras(3,4,'x'))
print(pythagoras(3,'x',5))
print(pythagoras('x',4,5))
print(pythagoras(3,4,5))
```

```
Hypotenuse = 5.0
Adjacent = 4.0
Opposite = 3.0
You know the answer!
```

15) string = "Yolo Life"

for i in string:

frequency = string.count(i)

print(str(i) + ": " + str(frequency), end=", ")

```
string = "Yolo Life"

for i in string:
    frequency = string.count(i)
    print(str(i) + ": " + str(frequency), end=", ")
```

```
Y: 1, o: 2, l: 1, o: 2,  : 1, L: 1, i: 1, f: 1, e: 1,
```

STATISTICS WORKSHEET

1) a) True
2) c) Centroid Limit Theorem
3) b) Modeling bounded count data
4) d) All of the mentioned
5) c) Poisson
6) b) False
7) b) Hypothesis
8) (a) 0
9) c) Outliers cannot conform to the regression relationship
10) In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center. Normal distributions are also called Gaussian distributions or bell curves because of their shape. The probability density function of normal or gaussian distribution is given by;

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

11) Missing data can be dealt with in a variety of ways Here are the most common ways of handling missing data

- Zero Replacement: Here, you replace the missing value with zero irrespective of everything.

Min or Max Replacement: Replace the missing value with the minimum or maximum value of a feature.

- Mean/ Median/ Mode Replacement: Replace missing value with mean or median or most frequent feature value.

- Also, one can replace the value of the missing cell with the previous cell's value. This kind of technique is popular while inputting time series data. For example, if the price of an instrument is missing on the i-th day, it makes sense to replace it with the (i-1)-th day's price.

    I. Mean imputation

Calculate the mean of the observed values for that variable for all non-missing people.It has the advantage of maintaining the same mean and sample size, but it also has a slew of drawbacks. Almost all of the methods described below are superior to mean imputation.

II. Substitution

Assume the value from a new person who was not included in the sample.To put it another way, pick a new subject and employ their worth instead.

III. Hot deck imputation

A value picked at random from a sample member who has comparable values on other variables.To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

One benefit is that you are limited to just feasible values. In other words, if age is only allowed to be between 5 and 10 in your research, you will always obtain a value between 5 and 10.Another factor is the random element, which introduces some variation. For exact standard errors, this is crucial.

IV. Cold deck imputation

A value picked deliberately from an individual with similar values on other variables.In most aspects, this is comparable to Hot Deck, but without the random variance. As an example, under the same experimental condition and block, you can always select the third individual.

V. Regression imputation

The result of regressing the missing variable on other factors to get a predicted value.As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

VI. Stochastic regression imputation

The predicted value of a regression plus a random residual value.This has all of the benefits of regression imputation plus the random component's benefits.The majority of multiple imputation is based on stochastic regression imputation.

VII. Interpolation and extrapolation

An estimate based on other observations made by the same person. It generally only works with data that is collected over time.Proceed with caution, though. For a variable like height in children–one that cannot be reduced through time–interpolation would make more sense. Extrapolation entails estimating beyond the data's true range, which necessitates making more assumptions than is necessary.

VIII. Single or Multiple Imputation

- Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.
- The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.
- It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.
- When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.
- Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.
- The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set.

12) A/B testing is a type of experiment in which you split your web traffic or user base into two groups, and show two different versions of a web page, app, email, and so on, with the goal of comparing the results to find the more successful version. With an A/B test, one element is changed between the original (a.k.a, “the control”) and the test version to see if this modification has any impact on user behavior or conversion rates.

13) Mean imputation is typically considered terrible practice since it ignores feature correlation. The disadvantages of mean imputations are:

    1. **Using the mean.** You can fill in missing values with the mean of the variable over the time period of observation. You if your data has a trend (if the rolling-mean is increasing over time) your added values may make your charting look odd. Also, this is not acceptable if your variables have an odd distribution that makes the mean value meaningless.
    2. **Rolling forward the last value**. This is generally only acceptable if you have a time series, such that you have sequential observations. Here, you "roll forward" the last value of the variable into the missing space. You may roll forward an anomaly-type value (outlier) if you are unlucky.
    3. **Exclude the observation**. You can choose to simply exclude any observation with missing variables. This reduces the total number of observations, obviously. You lose some information due to the exclusion, which can be harmful to your analysis. Not useful for datasets with few observations (this is completely subjective, generally you would like n>50n>50).
    4. **Simulate the value using distribution of the variable**. You can use the values you *do* have for the variable in question to determine an approximate distribution of the variable. Then, use a random generator to simulate a variable out of the distribution you

determined. Alternatively, you could add all variable values into a bag and randomly pull one out, which is less accurate but ensures you are sampling values directly from an observable set of values. Computationally expensive, requires knowledge of statistics and/or computing, more difficult to explain, assumes your observations are i.i.d.i.i.d..

14) Linear regression models the relationships between at least one explanatory variable and an outcome variable. These variables are known as the independent and dependent variables, respectively. When there is one independent variable (IV), the procedure is known as simple linear regression. When there are more IVs, statisticians refer to it as multiple regression.

15) There are 2 branches of statistics

(A) Descriptive statistics – it deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.
Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

(B) Inferential statistics – it involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.
Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.
While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.