MACHINE LEARNING WORKSHEET

Q1- Least square error

Q2-linear regression is sensitive to out layers

Q3-negative

Q4-regression

Q5-low bias and high variance

Q6-Predictive model

Q7-regularisation

Q8-SMOTE

Q9-sensitivity and specificity

Q10-false

Q11-appling PCA to project higher dimensional data

Q12-We need to iterate

SUBJECTIVE TYPE

Q13- REGULARIZATION-Every machine learning algorithm comes with built-in assumptions about the data. In some cases, these assumptions are reasonable and ensure good performance, but often they can be relaxed to produce a more general learner that might perform better on new examples. This process of relaxing an assumption is known as regularization .

Let me introduce three popular types of regularization: L1 (or Lasso) Penalization , L2 (Ridge) Penalization , and Elastic Net Regression. All codes are part of scikit-learn.

Q14-There are three main regularization techniques, namely:

1. Ridge Regression (L2 Norm)
2. Lasso (L1 Norm)
3. Dropout

Q15-Linear regression most often uses mean square error(MSE) to calculate the error of the model. MSE is calculated by: measuring the distance of the observed y-values from the predicted y-values at each value of x; squaring each of these distances; calculating the mean of each of the squared distances.

Python worksheet

Q1-#

Q2-0

Q3-24

Q4-true

Q5-0

Q6-It encloses the line of code which will be executed if any error occurs while executing the lines of code in the try block.

Q7-It is used to raise exception or errors.

Q8-defining a generator

Q9-_abc

Q10-all of the above

STATISTICS WORKSHEET

Q1-True

Q2-Central limit theorem

Q3-Modeling bounded count data

Q4-All of the mentiomed

Q5-Poisson

Q6-False

Q7-Hypothesis

Q8-0

Q9-Outliers cannot conform to the regression relstion ship

Q10- NORMAL DISTRIBUTION

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

In graphical form, the normal distribution appears as a "bell curve".

KEY TAKEAWAYS

- The normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- Many naturally-occurring phenomena tend to approximate the normal distribution.
- In finance, most pricing distributions are not, however, perfectly normal.

Q11-**Techniques for Handling the Missing Data**

1. Listwise or case deletion. ...
2. Pairwise deletion. ...
3. Mean substitution. ...
4. Regression imputation. ...
5. Last observation carried forward. ...
6. Maximum likelihood. ...

7. Expectation-Maximization. ...
8. Multiple imputation.

Imputation Techniques

- Complete Case Analysis(CCA):- This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing. ...
- Arbitrary Value Imputation. ...
- Frequent Category Imputation.

Q12- A/B testing

A/B testing is a type of split testing and is commonly used to drive improvements to specific variables or elements by measuring user or audience engagement. The approach is commonly used to optimise marketing campaigns or digital assets like websites. In A/B testing a specific variable is altered such as a title, image, or element layout. A sample of the audience is shown the control version and the altered version in a 50/50 split. Half traffic will interact with the original version, the other half will interact with the newer version. Engagement or the completion of a defined goal is the metric that is compared between the versions after a set period of time.

A/B testing can be used to:

- Refine marketing campaign messaging and design.
- Improve conversion rates through enhancements to user experience.
- Continuously optimise assets like web pages by considering user engagement.

Q13- Mean imputation of missing data is a bad practice in general

- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by "pulling" estimates of the correlation toward zero

Q14- LINEAR REGRESSION-

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest

form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

TYPES OF LINEAR REGRESSION-

Simple linear regression
1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

Multiple linear regression
1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

Logistic regression
1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

Ordinal regression
1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

Multinomial regression
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

Discriminant analysis
1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

Q15-Statistics is generally of two types .

1-Descriptive statistics

2-inferential statistics

DESCRIPTIVE STATISTICS-

Descriptive statistics 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.

INFERENTIAL STATISTICS-

Inferential statistics as the name suggests, 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.