Python Worksheet 1

1). Option C

2). Option B

3). Option C

4). Option A

5). Option D

6). Option C

7). Option A

8). Option C

9). Option A and C

10). Option A and B

Machine Learning Worksheet 1

1). Option D

2). Option C

3). Option B

4). Option C

5). Option D

6). Option B

7). Option C

8). Option C

9). Option A and B

10). Option A and D

11). Option C and D

Answer 12). If the training set has millions or more features then the Linear Regression Training algorithms that can be used are Stochastic gradient descent or Mini-batch gradient descent.

Stochastic Gradient Descent - This is a type of gradient descent processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is quite faster than batch gradient descent. When the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.

Mini Batch gradient descent: This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Even if the number of training examples is large, it is processed in batches of certain training examples but in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

If the number of training examples is large, then batch gradient descent is computationally very expensive. Hence if the number of training examples is large, then batch gradient descent is not preferred

Answer 13). If the features in your training set have very different scales, the cost function will have the shape of an elongated bowl, so the Gradient Descent suffers from features of different scales, because the model will take a longer time to reach the global maximum i.e. to converge. We can always scale the features to eliminate this problem

Statistics Worksheet 1

1). Option A

2). Option A

3). Option B

4). Option D

5). Option C

6). Option B

7). Option B

8). Option A

9). Option C

Answer 10). 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 graph form, normal distribution will appear as a bell curve. The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates a normal distribution.

Answer 11). Missing data is a huge problem for data analysis because it distorts findings. It's difficult to be fully confident in the insights when you know that some entries are missing values.

Best techniques to handle missing data: - The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical for large datasets. Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organisations can fill out the gaps manually. Pairwise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pairwise deletion saves more data compared to likewise deletion because the former only deletes entries where variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

Imputation Techniques: Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilising this method is the three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for non-numerical data only the median and mean are relevant).

Answer 12). A/B testing, at its most basic, is a way to compare two versions of something to figure out which performs better. A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal. Running an A/B test that directly compares a variation against a current experience lets you ask focused questions about changes to your website or app and then collect data about the impact of that change. Testing takes the guesswork out of website optimization and enables data-informed decisions that shift business conversations from "we think" to "we know." By measuring the impact that changes have on your metrics, you can ensure that every change produces positive results.

Answer 13). It is a non-standard, it uses Random Forest. It is use to predict the missing data. It also can be used for both i.e. continuous as well as categorical data and so it makes advantageous over other imputations. There are some limitations of it like Mean imputation does not preserve the relationship among variables. It preserves the mean of observed data. If data is missing completely at random, the estimate of the mean remains unbiased. Also, Mean Imputation leads to an underestimate of standard errors.

Answer 14). 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 regress and. 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: -

1. Simple linear regression
2. Multiple linear regressions
3. Logistic regression
4. Ordinal regression
5. Multinomial regression

Answer 15). The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of 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 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.