WORKSHEET-1

MACHINE LEARNING

1. 0(n)
2. Logistic Regression
3. Gradient descent
4. Lasso
5. All of the above
6. False
7. scaling cost function by half makes gradient descent converge faster.
8. Correlation
9. We don't have to choose the learning rate, It becomes slow when number of features are very large.
10. Linear Regression will have high bias and low variance, Polynomial with degree 5 will have low bias and high variance
11. It discovers causal relationship, No inference can be made from regression line.
12. We can use batch gradient descent, stochastic gradient descent or mini batch gradient descent. The best method would be SGD and MBGD because they need not have to load the entire dataset into memory in order to take 1 step of gradient descent.
13. The Gradient Descent suffers from features of different scales, because the model will take a longer time to reach the global maximum. We can always scale the features to eliminate this problem.

PYTHON

1. %
2. 0
3. 24
4. 2
5. 6
6. The finally block will be executed no matter if the try block raises an error or not.

7. It is used to raise an exception.
8. in defining a generator.
9. _abc and abc2
10. Yield,raise

STATISTICS

1. True
2. Central limit theorem
3. Modeling bounded count data
4. All of the mentioned
5. Poisson
6. False
7. Hypothesis
8. 0
9. Outliers cannot conform to the regression relationship
10. Normal distribution tells us that the probability distribution is symmetric about the mean i.e. the majority of the data is accumulated near the mean. It appears as a bell curve.
11. Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. There are some common methods in order to replace the missing data:-
    i. Mean or median imputation-When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data, we use different techniques to impute data that have been described below.
    ii. Multivariate imputation by chained equation(MICE)-MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables,

logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

iii. Random forest- Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB (out of bag) imputation error estimates.

One caveat is that random forest works best with large datasets and using random forest on small datasets runs the risk of overfitting. The extent of overfitting leading to inaccurate imputations will depend upon how closely the distribution for predictor variables for non-missing data resembles the distribution of predictor variables for missing data.

12. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

13. 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.

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 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.

15. There are two branches of statistics- Descriptive and Inferential.

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 experiments.

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.