# Non-Graded Assignment

Instructions:

- This is a non- graded assignment. However: **Students who submit the assignment will get additional 2 marks towards the final score**
- You will find 14 questions below that are commonly asked in Interviews.
- Kindly answer all the questions. We request you to submit a correct and wrong answer for each question.
- Each question is followed by a hint, which is an acceptable correct answer.
- You will have to look at the hint and come up with your own answers for each question.
- Please do not copy and paste the answers we have given while submitting.
- You can make your submission as either doc file or pdf file
- This exercise will help you prepare for interviews and your responses will be used for our own research on how to assess descriptive types of questions.

# Questions

1. How would you define Machine Learning?

   Hint:

   Machine learning: It is an application of artificial intelligence (AI) that provides systems the ability to learn automatically and to improve from experiences without being programmed. It focuses on the development of computer applications that can access the data and used it to learn for themselves. The process of learning starts with the observations or data, such as examples, direct experience, or instruction, to look for the patterns in data and to make better decisions in the future based on examples that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.

Wrong Answer:

Machine Learning is the process of understanding how a computer system works and learning its impact on our daily lives. It is also about processes related to other technologies that connect and exchange data with other devices and systems over the Internet or other communications networks.

2. **What is Overfitting, and how can you avoid it?**

Hint:

Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data. When a model is given the training data, it shows 100 percent accuracy—technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting. There are multiple ways of avoiding overfitting, such as:

- Regularization. It involves a cost term for the features involved with the objective function
- Making a simple model. With lesser variables and parameters, the variance can be reduced
- Cross-validation methods like k-folds can also be used
- If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

Wrong Answer:

Overfitting is the condition in which the designed model accurately fits for any set of parameters apart from its training dataset. In other words, Overfitting is when a model accurately predicts for any dataset which is similar to its training dataset.

3. Difference between logistic and linear regression?

   Hint:

   Linear and Logistic regression are the most basic form of regression which are commonly used. The essential difference between these two is that Logistic regression is used when the dependent variable is binary. In contrast, Linear regression is used when the dependent variable is continuous, and the nature of the regression line is linear.

   Key Differences between Linear and Logistic Regression

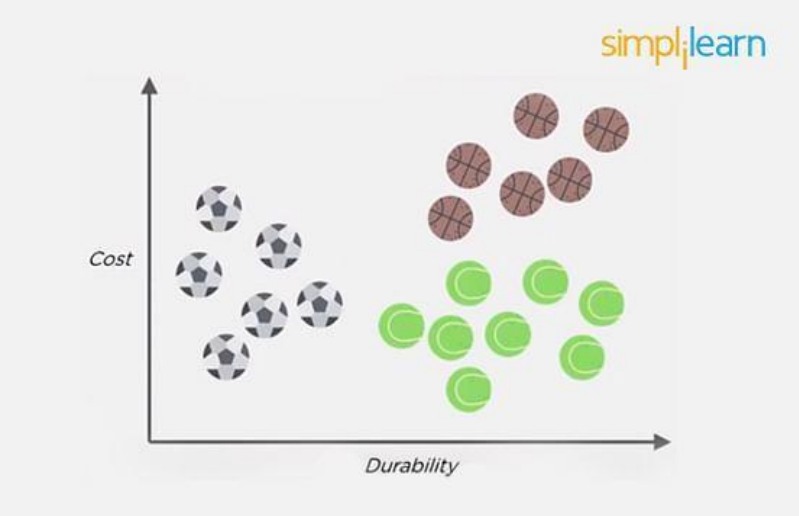
   Linear regression models data using continuous numeric values. As against, logistic regression models the data in the binary values. Linear regression requires to establish the linear relationship among dependent and independent variables, whereas it is not necessary for logistic regression. In linear regression, the independent variables can be correlated with each other. On the contrary, in logistic regression, the variables must not be correlated with each other.

4. Explain the K Nearest Neighbor Algorithm.

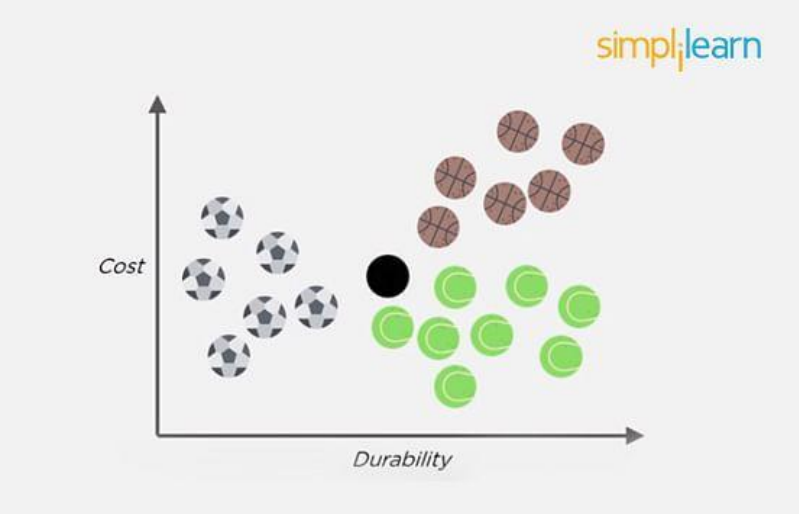
   Hint:

   K nearest neighbor algorithm is a classification algorithm that works in a way that a new data point is assigned to a neighboring group to which it is most similar. In K nearest neighbors, K can be an integer greater than 1. So, for every new data point, we want to classify, we compute which neighboring group it is closest to. Let us classify an object using the following example. Consider there are three clusters:

   - Football
   - Basketball
   - Tennis ball




Let the new data point to be classified is a black ball. We use KNN to classify it. Assume K = 5 (initially). Next, we find the K (five) nearest data points, as shown.




Observe that all five selected points do not belong to the same cluster. There are three tennis balls and one each of basketball and football. When multiple classes are involved, we prefer the majority. Here the majority is with the tennis ball, so the new data point is assigned to this cluster.

5. What is the Confusion Matrix?

Hint:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

This is the key to the confusion matrix.

It gives us insight not only into the errors being made by a classifier but, more importantly, the types of errors that are being made.

| | Class 1 (Predicted) | Class 2(Predicted) |
|---|---|---|
| Class 1 (Actual) | TP | FN |
| Class2 (Actual) | FP | TN |

Here,

Class 1: Positive

Class 2: Negative

Definition of the Terms:

- Positive (P): Observation is positive (for example: is an apple).
- Negative (N): Observation is not positive (for example: is not an apple).
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

6. What is multicollinearity, and how do you treat it?

Hint:

Multicollinearity means independent variables are highly correlated to each other. In regression analysis, it's an important assumption that the regression model should not be faced with a problem of multicollinearity. If two explanatory variables are highly correlated, it's hard to tell, which affects the dependent variable.

Let's say Y is regressed against X1 and X2 and where X1 and X2 are highly correlated. Then the effect of X1 on Y is hard to distinguish from the effect of X2 on Y because any increase in X1 tends to be associated with an increase in X2. Another way to look at the multicollinearity problem is: Individual t-test P values can be misleading. It means a P-value can be high, which means the variable is not important, even though the variable is important.

Correcting Multicollinearity:

- Remove one of the highly correlated independent variables from the model. If you have two or more factors with a high VIF, remove one from the model.
- Principle Component Analysis (PCA) - It cut the number of interdependent variables to a smaller set of uncorrelated components. Instead of using highly

correlated variables, use components in the model that have eigenvalue greater than 1.

- Run PROC VARCLUS and choose the variable that has a minimum (1-R2) ratio within a cluster.
- Ridge Regression - It is a technique for analyzing multiple regression data that suffer from multicollinearity.
- If you include an interaction term (the product of two independent variables), you can also reduce multicollinearity by "centering" the variables. By "centering," it means subtracting the mean from the values of the independent variable before creating the products.

7. What is the Central limit theorem?

   Hint:

   The theorem states that as the size of the sample increases, the distribution of the mean across multiple samples will approximate a Gaussian distribution (Normal). Generally, sample sizes equal to or greater than 30 are considered sufficient for the CLT to hold. It means that the distribution of the sample means is normally distributed. The average of the sample means will be equal to the population mean. This is the key aspect of the theorem.

   Assumptions:

   - The data must follow the randomization condition. It must be sampled randomly
   - Samples should be independent of each other. One sample should not influence the other samples
   - Sample size should be no more than 10% of the population when sampling is done without replacement
   - The sample size should be sufficiently large.

8. What is Bayes' Theorem? How is it useful in a machine learning context?

   Hint:

   Bayes' Theorem gives you the posterior probability of an event given what is known as prior knowledge.

   Mathematically, it's expressed as the true positive rate of a condition sample divided by the sum of the false positive rate of the population and the true positive rate of a condition. Say you had a 60% chance of actually having the flu after a flu test, but out of people who had the flu, the test will be false 50% of the time, and the overall population only has a 5% chance of having the flu. Would you actually have a 60% chance of having the flu after having a positive test?

   Bayes' Theorem says no. It says that you have a (.6 * 0.05) (True Positive Rate of a Condition Sample) / (.6*0.05)(True Positive Rate of a Condition Sample) + (.5*0.95) (False Positive Rate of a Population) = 0.0594 or 5.94% chance of getting a flu.

   Bayes' Theorem is the basis behind a branch of machine learning that most notably includes the Naive Bayes classifier. That's something important to consider when you're faced with machine learning interview questions.

9. How is a decision tree pruned?

   Hint:

   Pruning is what happens in decision trees when branches that have weak predictive power are removed in order to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can happen bottom-up and top-down, with approaches such as reduced error pruning and cost complexity pruning.

Reduced error pruning is perhaps the simplest version: replace each node. If it doesn't decrease predictive accuracy, keep it pruned. While simple, this heuristic actually comes pretty close to an approach that would optimize for maximum accuracy.

10. What is the correlation and coefficient?

    Hint:

    The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. We use it to measure both the strength and direction of a linear relationship between two variables the values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation.

    Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where: 1 indicates a strong positive relationship. -1 indicates a strong negative relationship. A result of zero indicates no relationship at all.

11. Describe Hypothesis Testing. How is the statistical significance of an insight assessed?

    Hint:

    Hypothesis Testing in statistics is used to see if a certain experiment yields meaningful results. It essentially helps to assess the statistical significance of insight by determining the odds of the results occurring by chance. The first thing is to know the null hypothesis and then state it. Then the p-value is calculated, and if the null hypothesis is true, other values are also determined. The alpha value denotes the significance and is adjusted accordingly.

If the p-value is less than alpha, the null hypothesis is rejected, but if it is greater than alpha, the null hypothesis is accepted. The rejection of the null hypothesis indicates that the results obtained are statistically significant.

12. What are the assumptions required for linear regression?

Hint:

Four major assumptions for linear regression are as under –

- There's a linear relationship between the predictor (independent) variables and the outcome (dependent) variable. It means that the relationship between X and the mean of Y is linear.
- The errors are normally distributed with no correlation between them. This process is known as Autocorrelation.
- There is an absence of correlation between predictor variables. This phenomenon is called multicollinearity.
- The variation in the outcome or response variable is the same for all values of independent or predictor variables. This phenomenon of assumption of equal variance is known as homoscedasticity.

13. Explain dimensionality reduction in Machine Learning?

Hint:

- Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset.
- More input features often make a predictive modeling task more challenging to model, more generally referred to as the curse of dimensionality.
- High-dimensionality statistics and dimensionality reduction techniques are often used for data visualization. Nevertheless these techniques can be used in applied

machine learning to simplify a classification or regression dataset in order to better fit a predictive model.

14. Define Principal component analysis:

    Hint:

    Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

    Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.