**MACHINE LEARNING INTERVIEW QUESTIONS**

**What is Machine Learning?**

**Machine learning** is a data analytics technique, getting computers to learn and also act like a human. **Machine learning** algorithms use computational methods to “learn” by feeding data and information. And it is also known as a field of data analytics to make predictions depends on trends and insights of the data.

Machine learning algorithms use statistical tools to find meaningful connections in large amounts of data. The data here are; covers many things like numbers, words, images, clicks. If this data can be stored digitally, it can be fed into a machine learning algorithm.

Briefly, the main basis of machine learning is to create algorithms that can process its data and to predict the result using statistical analysis. The new data obtained during this process continues to be learned and used to update and improve the results.

**Q: What Is Linear Regression ?**  
**A:**In simple terms, [linear regression](https://lms.clarusway.com/mod/lesson/view.php?id=12837" \o "Linear Regression) is a method of finding the best straight line fitting to the given data and also finding the best linear relationship between the independent and dependent variables. We use Least Squares Method  to decide which line is the best fit the model.

**Q: What’s the trade-off between bias and variance?**  
 **A:** The bias-variance trade-off is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set. Our goal is to find the optimal middle ground where the errors from both over-fitting and under-fitting are minimal.  We want a model that is highly enough to capture the signals in our data, but not too complex that it can’t be applied to new data.

Q: Why can’t [linear regression](https://lms.clarusway.com/mod/lesson/view.php?id=9409" \o "Linear Regression) be used in place of [logistic regression](https://lms.clarusway.com/mod/lesson/view.php?id=12841" \o "Logistic Regression) for binary classification?  
A: In [linear regression](https://lms.clarusway.com/mod/lesson/view.php?id=9409" \o "Linear Regression), the output is continuous. In case of binary classification, an output of a continuous value does not make sense. For binary classification problems, [linear regression](https://lms.clarusway.com/mod/lesson/view.php?id=9409" \o "Linear Regression) may predict values that can go beyond 0 and 1. If we want the output in the form of probabilities, which can be mapped to two different classes, then its range should be restricted to 0 and 1. As the [logistic regression](https://lms.clarusway.com/mod/lesson/view.php?id=12841" \o "Logistic Regression) model can output probabilities with logistic/sigmoid function, it is preferred over [linear regression](https://lms.clarusway.com/mod/lesson/view.php?id=9409" \o "Linear Regression).

Q: Why is accuracy not a good measure for classification problems?  
A: Accuracy is not a good measure for classification problems because it gives equal importance to both false positives and false negatives. However, this may not be the case in most business problems. For example, in case of cancer prediction, declaring cancer as benign is more serious than wrongly informing the patient that he is suffering from cancer. Accuracy gives equal importance to both cases and cannot differentiate between them.

## What are precision and recall?

Precision is the proportion of true positives out of predicted positives. To put it in another way, it is the accuracy of the prediction. It is also known as the ‘positive predictive value’.  
**Precision = TP/TP+FP**

Recall refers to the ratio of positives correctly predicted from all the true labels. In simple words, it is the frequency of correctly predicted true labels.  
**Recall = TP/TP+FN**

**Q:** **Why should we not use KNN algorithm for large datasets?**  
**A:** KNN works well with smaller dataset because it is a lazy learner. It needs to store all the data and then makes decision only at run time. It needs to calculate the distance of a given point with all other points. So if dataset is large, there will be a lot of processing which may adversely impact the performance of the algorithm. KNN is also very sensitive to noise in the dataset. If the dataset is large, there are chances of noise in the dataset which adversely affect the performance of KNN algorithm.

**Q: What is the SVM machine learning algorithm?**

**A:** SVM stands for support vector machine, it is a supervised machine learning algorithm which can be used for both Regression and Classification. If you have n features in your training data set, SVM tries to plot it in n-dimensional space with the value of each feature being the value of a particular coordinate. SVM uses hyper planes to separate out different classes based on the provided kernel function.

**Q: What is the difference between Entropy and Information Gain?**

**A:** Entropy is an indicator of how messy your data is. It decreases as you reach closer to the leaf node. The Information Gain is based on the decrease in entropy after a dataset is split on an attribute. It keeps on increasing as you reach closer to the leaf node.

**Q: What are the pros and cons of the K-Means algorithm?**

**Pros:**

* Simple to understand,
* Easy to implement

**Cons:**

* Need to pick number of clusters
* Sensitive to outliers
* Needs standardization

**Q: Explain Principle Component Analysis (PCA)**

**A:** PCA is a method for transforming features by combining them into uncorrelated linear combinations. These new principal components, sequentially maximize the variance represented (i.e. the first principal component has the most variance, the second principal component has the second most, and so on).As a result, PCA is useful for dimensionality reduction and visulization of the data.