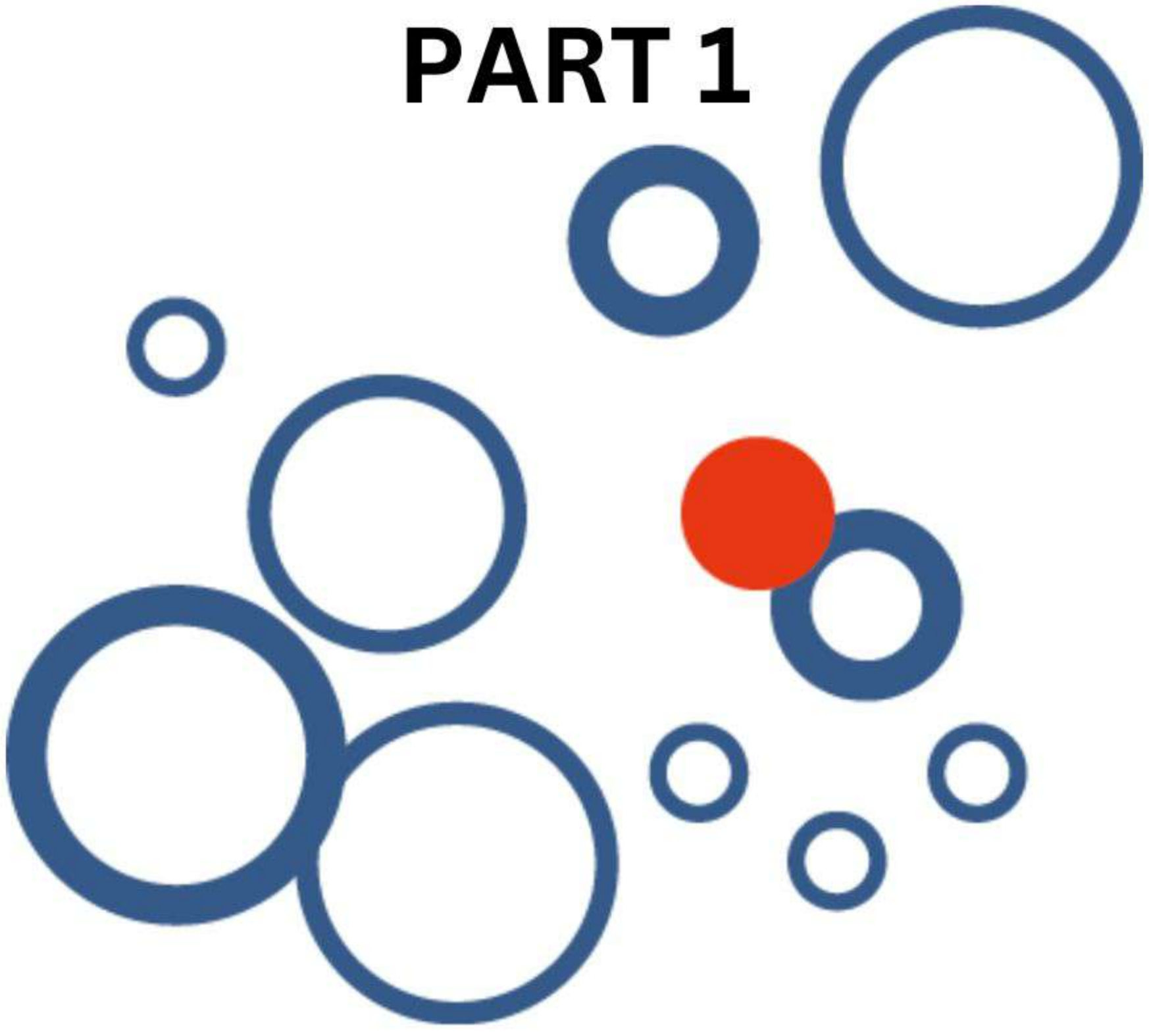
@AAKASH KANOJIYA

DATA SCIENCE INTERVIEW QUESTIONS

PART 1



What are the conditions for Overfitting and Underfitting?

In Overfitting the model performs well for the training data, but for any new data it fails to provide output. For Underfitting the model is very simple and not able to identify the correct relationship. Following are the bias and variance conditions.

1) Overfitting – Low bias and High Variance results in overfitted model. Decision tree is more prone to Overfitting.

2) Underfitting – High bias and Low Variance. Such model doesn't perform well on test data also. For example – Linear Regression is more prone to Underfitting.

Swipe

What are Type 1 and Type 2 errors? In which scenarios the Type 1 and Type 2 errors become significant?

Rejection of True Null Hypothesis is known as a Type 1 error. In simple terms, False Positive are known as a Type 1 Error.

Not rejecting the False Null Hypothesis is known as a Type 2 error. False Negatives are known as a Type 2 error.

1) Type 1 Error is significant where the importance of being negative becomes significant. For example – If a man is not suffering from a particular disease marked as positive for that infection. The medications given to him might damage his organs.

2) While Type 2 Error is significant in cases where the importance of being positive becomes important. For example – The alarm has to be raised in case of burglary in a bank. But a system identifies it as a False case that won't raise the alarm on time resulting in a heavy loss.

Swipe

What is the significance of Sampling? Name some techniques for Sampling?

For analyzing the data we cannot proceed with the whole volume at once for large datasets. We need to take some samples from the data which can represent the whole population. While making a sample out of complete data, we should take that data which can be a true representative of the whole data set.

There are mainly two types of Sampling techniques based on Statistics.

Probability Sampling and Non Probability Sampling

1) Probability Sampling – Simple Random, Clustered Sampling, Stratified Sampling.

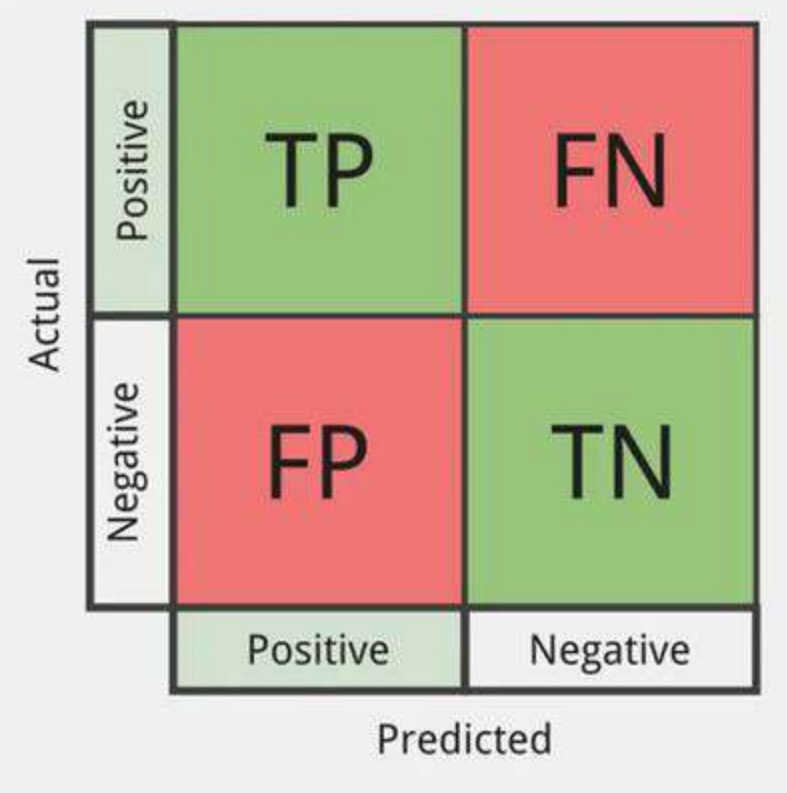
2) Non Probability Sampling – Convenience Sampling, Quota Sampling, Snowball Sampling.

Swipe

Why do we need Evaluation Metrics. What do you understand by Confusion Matrix ?

Evaluation Metrics are statistical measures of model performance. They are very important because to determine the performance of any model it is very significant to use various Evaluation Metrics. Few of the evaluation Metrics are, Accuracy, Log Loss, Confusion Matrix.

Confusion Matrix is a matrix to find the performance of a Classification model. It is in general a 2×2 matrix with one side as prediction and the other side as actual values.



Swipe

What is Natural Language Processing? State some real life example of NLP.

Natural Language Processing is a branch of Artificial Intelligence that deals with the conversation of Human Language to Machine Understandable language so that it can be processed by ML models.

Examples – NLP has so many practical applications including chatbots, google translate, and many other real time applications like Alexa.

Some of the other applications of NLP are in text completion, text suggestions, and sentence correction.

Swipe

What is Linear Regression. What are the Assumptions involved in it?

Linear Regression is a mathematical relationship between an independent and dependent variable. The relationship is a direct proportion, relation making it the most simple relationship between the variables.

$$Y = mX+c$$

- Y – Dependent Variable
- X – Independent Variable
- m and c are constants

Assumptions of Linear Regression :

- ✓ The relationship between Y and X must be Linear.
- ✓ The features must be independent of each other.
- ✓ Homoscedasticity – The variation between the output must be constant for different input data.
- ✓ The distribution of Y along X should be the Normal Distribution.

Swipe

How does Confusion Matrix help in evaluating model performance?

We can find different accuracy measures using a confusion matrix. These parameters are Accuracy, Recall, Precision, F1 Score, and Specificity.

Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall performance of model
Precision	$\frac{TP}{TP + FP}$	How accurate the positive predictions are
Recall Sensitivity	$\frac{TP}{TP + FN}$	Coverage of actual positive sample
Specificity	$\frac{TN}{TN + FP}$	Coverage of actual negative sample
F1 score	$\frac{2TP}{2TP + FP + FN}$	Hybrid metric useful for unbalanced classes

Difference between Regression and Classification?

The major difference between Regression and Classification is that Regression results in a continuous quantitative value while Classification is predicting the discrete labels.

Regression

- ✓ Regression predicts the quantity.
- ✓ We can have discrete as well as continuous values as input for regression.
- ✓ If input data are ordered with respect to the time it becomes time series forecasting.

Classification

- ✓ The Classification problem for two classes is known as Binary Classification.
- ✓ Classification can be split into Multi- Class Classification or Multi-Label Classification.
- ✓ We focus more on accuracy in Classification while we focus more on the error term in Regression.

Swipe ➤

What is Logistic Regression? What is the loss function in LR?

Logistic Regression is a Binary Classification function. It is a statistical model that uses the logit function on the top of the probability to give 0 or 1 as a result.

The loss function in LR is known as the Log Loss function. The equation for which is given as :

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

Swipe

What do you mean by Normalization? Difference between Normalization and Standardization?

Normalization is a process of bringing the features in a simple range, so that model can perform well and do not get inclined towards any particular feature. For example – If we have a dataset with multiple features and one feature is the Age data which is in the range 18-60 , Another feature is the salary feature ranging from 20000 – 2000000. In such a case, the values have a very much difference in them. Age ranges in two digits integer while salary is in range significantly higher than the age. So to bring the features in comparable range we need Normalization.

Both Normalization and Standardization are methods of Features Conversion. However, the methods are different in terms of the conversions.The data after Normalization scales in the range of 0-1. While in case of Standardization the data is scaled such that it means comes out to be 0.

Standardisation

	Age	Salary
0	0.758874	7.494733e-01
1	-1.711504	-1.438178e+00
2	-1.275555	-8.912655e-01
3	-0.113024	-2.532004e-01
4	0.177609	6.632192e-16
5	-0.548973	-5.266569e-01

Max-Min Normalization

	Age	Salary
0	0.739130	0.685714
1	0.000000	0.000000
2	0.130435	0.171429
3	0.478261	0.371429
4	0.565217	0.450794
5	0.347826	0.285714