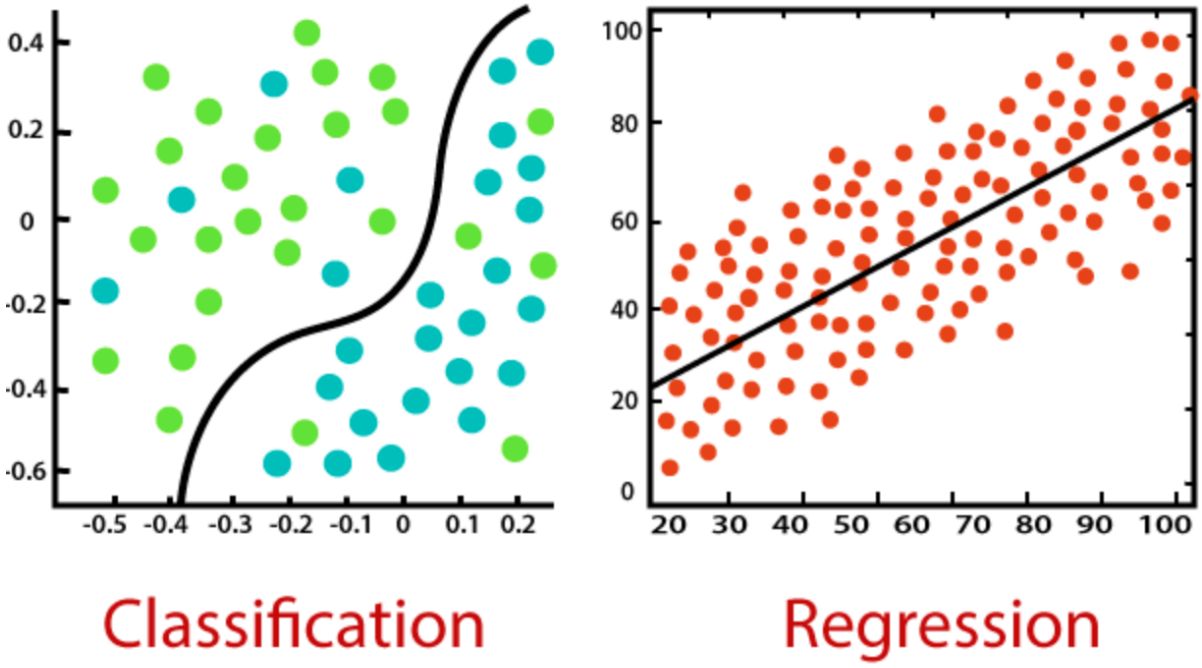
PRIYA MOHATA
PES UNIVERSITY

1. When should you use classification over regression?

- Classification is used when the output variable is a category such as “red” or “blue”, “spam” or “not spam”. It is used to draw a conclusion from observed values.
- The classification method is chosen over regression when the output of the model needs to yield the belongingness of data points in a dataset to a particular category.
- Regression is used when the output variable is a real or continuous value like “age”, “salary”, etc.



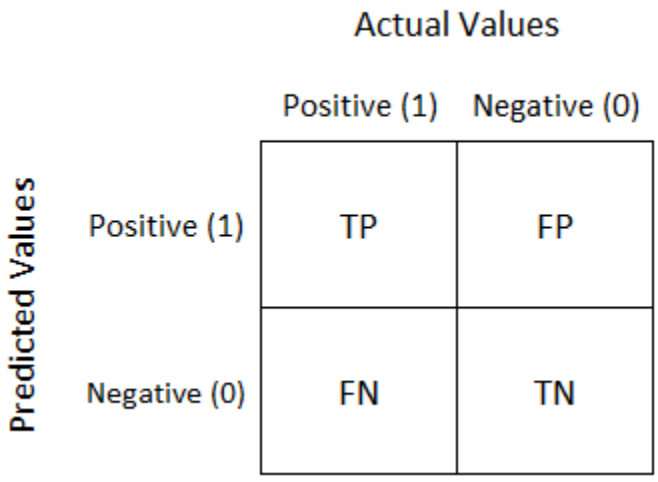
2. How do you deal with the class imbalance in a classification problem?

1. Use of right evaluation metrics (precision, sensitivity, F1 score)
2. Resampling the training set (random under-sampling, random over sampling)
3. Use of the k-cross-validation technique
4. Resampling of data with different ratios

5. Ensembling different resampled datasets(use of more data)
6. Clustering the abundant class
7. Design own, convenient models

3. What is a confusion matrix and why do you need it?

- 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.
- Confusion matrices are useful because they give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives.
- For eg- When the dataset of our model has more than 3 classes, we may get a classification accuracy of 80%, but we wouldn't know if that was because all classes are being predicted equally well or whether one or two classes are being neglected by the model. In such cases, the use of classification accuracy could be misleading and hence we make use of the confusion matrix.



4. What is the difference between sigmoid and softmax function?

The Sigmoid Activation Function is a mathematical function with a recognizable “S” shaped curve. It is used for the logistic regression and basic neural network implementation
The return value of Sigmoid Function is mostly in the range of values between 0 and 1 or -1 and 1.

$$f(x) = sigmoid(x) = \frac{1}{1 + e^{-x}}$$

The Softmax Activation Function is an activation function that takes vectors of real numbers as inputs, and normalizes them into a probability distribution proportional to the exponentials of the input numbers.
Each element will be in the range of 0 to 1.

$$\text{softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \text{ for j} = 1,\ldots,\text{K}$$

5. Why is logistic regression a type of classification technique and not a regression? Name the function it is derived from?

- The term “Logistic” is taken from the Logit function that is used in this method of classification.
- Logistic Regression is one of the most widely used algorithms to solve a classification problem. It is a supervised classification problem. In a classification problem, y(output or target variable) can only take discrete values for a given set of input (X).

- Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic regression is most commonly used when the data in question has binary output, when it is either a 0 or 1.