**Company Name - Siemens**
**Role - Data Scientist**

### Q. How to join tables in python?

A. The concat() function pf pandas can be used to concatenate two Data frames table by adding the rows of one to the other.

### Q. What is the benefit of shuffling a training dataset when using a batch gradient descent algorithm for optimizing a neural network?

A. it helps the training converge fast. it prevents any bias during the training. it prevents the model from learning the order of the training.

### Q. Why it is not advisable to use a softmax output activation function in a multi-label classification problem for a one-hot encoded target?

A. The final score for each class should be independent of each other. Thus, we cannot apply softmax activation, because softmax converts the score into probabilities taking other scores into consideration.

### Q. What does the cost parameter in SVM stand for?

A. The cost parameter decides how much an SVM should be allowed to “bend” with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

### Q. What is Softmax Function? What is the formula of Softmax Normalization?

A. The softmax function is an activation function that turns numbers into probabilities which sum to one. The softmax function outputs a vector that represents the probability distributions of a list of outcomes. It is also a core element used in deep learning classification tasks. Formula for softmax normalization can be given as:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

### Q. How to add a border that is filled with 0's around an existing array?

A. For doing this the pad function of numpy can be used by passing a pad width of 1 and constant values as 0.

**Q. An array has shape (7,4,2) what is the index (x,y,z) of the 20th element?**

A. The answer is (2, 2, 0). This can be found out using the unravel_index function of numpy. Use "print (np.unravel_index(20, (7,4,2)))"

**Q. For a model an accuracy of 100% is obtained but with the validation set, the accuracy score is 75%. What should be looked out for?**

A. There could be an issue with overfitting. When analysing machine learning algorithms, there are a few key strategies to avoid overfitting: To estimate model accuracy, use a resampling strategy, hold back a validation dataset, use extra training data, or pick relevant features.

**Q. How is skewness different from kurtosis?**

A. Skewness is a metric for symmetry, or more specifically, the lack of it. If a distribution, or data collection, looks the same to the left and right of the centre point, it is said to be symmetric. Kurtosis is a measure of how heavy-tailed or light-tailed the data are in comparison to a normal distribution. Data sets having a high kurtosis are more likely to contain heavy tails, or outliers. Light tails or a lack of outliers are common in data sets with low kurtosis. The main distinction between skewness and kurtosis is that the former refers to the degree of symmetry in the frequency distribution, whilst the latter refers to the degree of peakedness.

**Q. Explain Data cleaning steps?**

A. Step 1: Remove duplicate or irrelevant observations. Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations.

Step 2: Fix structural errors.

Step 3: Filter unwanted outliers.

Step 4: Handle missing data.

Step 5: Validate your data if it's appropriate according to problem statement .

## Q. Upsampling and downsampling methods?

A. In a classification task, there is a high chance for the algorithm to be biased if the dataset is imbalanced. An imbalanced dataset is one in which the number of samples in one class is very higher or lesser than the number of samples in the other class.

To counter such imbalanced datasets, we use a technique called up-sampling and down-sampling.

In up-sampling, we randomly duplicate the observations from the minority class in order to reinforce its signal. The most common way is to resample with replacement.

In down-sampling, we randomly remove the observations from the majority class. Thus, after up-sampling or down-sampling, the dataset becomes balanced with same number of observations in each class.

## Q. Imputation methods?

**A.** They are:

- List wise or case deletion
- Pairwise deletion
- Mean substitution
- Regression imputation
- Maximum likelihood.

## Q. Inter quartile ranges?

A. The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR. Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q 1 , Q 2 and Q 3 , respectively. Q 2 is nothing but the median.

## Q. Probabilistic and statistical modelling?

A. Probabilistic modelling is a statistical technique used to take into account the impact of random events or actions in predicting the potential occurrence of future outcomes. Statistical modelling is the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.

---

**NOTE:**
If you want to learn Data Science from scratch and become interview ready, check out our popular Data science course.

https://ai.cloudyml.com/Learn-Data-Science-from-Scratch