**1. What exactly is a feature? Give an example to illustrate your point.**

**A.** Features are the basic building blocks of datasets. The quality of the features in your dataset has a major impact on the quality of the insights you will gain when you use that dataset for machine learning.

Additionally, different business problems within the same industry do not necessarily require the same features, which is why it is important to have a strong understanding of the business goals of your data science project.

**2. What are the various circumstances in which feature construction is required?**

**A.** The features in your data will directly influence the predictive models you use and the results you can achieve. Your results are dependent on many inter-dependent properties. You need great features that describe the structures inherent in your data. Better features means flexibility.

**3. Describe how nominal variables are encoded.**

**A.** The features in your data will directly influence the predictive models you use and the results you can achieve. Your results are dependent on many inter-dependent properties. You need great features that describe the structures inherent in your data. Better features means flexibility.

**4. Describe how numeric features are converted to categorical features.**

**A.**  Converting categorical features into numeric features using domain knowledge. For example, we are given a list of countries and say we know the distance to these countries from India then we can replace it with distance from India. So, every country can be represented as its distance from India.

**5. Describe the feature selection wrapper approach. State the advantages and disadvantages of this approach?**

**A.** Wrapper methods measure the “usefulness” of features based on the classifier performance. In contrast, the filter methods pick up the intrinsic properties of the features (i.e., the “relevance” of the features) measured via univariate statistics instead of cross-validation performance.

The wrapper classification algorithms with joint dimensionality reduction and classification can also be used but these methods have high computation cost, lower discriminative power. Moreover, these methods depend on the efficient selection of classifiers for obtaining high accuracy.

**6. When is a feature considered irrelevant? What can be said to quantify it?**

**A.** Features are considered relevant if they are either strongly or weakly relevant, and are considered irrelevant otherwise.

Irrelevant features can never contribute to prediction accuracy, by definition. Also to quantify it we need to first check the list of features, There are three types of feature selection:

* **Wrapper methods** (forward, backward, and stepwise selection)
* **Filter methods** (ANOVA, Pearson correlation, variance thresholding)
* **Embedded methods** (Lasso, Ridge, Decision Tree).

**7. When is a function considered redundant? What criteria are used to identify features that could be redundant?**

**A.** If two features {X1, X2} are highly correlated, then the two features become redundant features since they have same information in terms of correlation measure. In other words, the correlation measure provides statistical association between any given a pair of features.

Minimum redundancy feature selection is an algorithm frequently used in a method to accurately identify characteristics of genes and phenotypes.

**8. What are the various distance measurements used to determine feature similarity?**

**A. :** Four of the most commonly used distance measures in machine learning are as follows:

* Hamming Distance.
* Euclidean Distance
* Manhattan Distance.

**9. State difference between Euclidean and Manhattan distances?**

**A.**

|  |  |
| --- | --- |
| **Euclidean Distance** | **Manhattan distances** |
| Euclidean Distance represents the shortest distance between two points | Manhattan Distance is the sum of absolute differences between points across all the dimensions**.** |
| The “Euclidean Distance” between two objects is the distance you would expect in “flat” or “Euclidean” space; it’s named after Euclid, who worked out the rules of geometry on a flat surface. | Manhattan distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. |
| The Euclidean is often the “default” distance used in e.g., K-nearest neighbors (classification) or K-means (clustering) to find the “k closest points” of a particular sample point | Applications of Manhattan includes Regression analysis, Compressed sensing, Frequency distribution. |
|  |  |

**10. Distinguish between feature transformation and feature selection.**

**A.**  Feature selection is for filtering irrelevant or redundant features from your dataset. The key difference between feature selection and extraction is that feature selection keeps a subset of the original features while feature extraction creates brand new ones.