# Assignment 08

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

**Answer:**

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 ?

**Answer:**

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 ?

**Answer:**

Nominal data is made of discrete values with no numerical relationship between the different categories — mean and median are meaningless. Animal species is one example. For example, pig is not higher than bird and lower than fish.

##### 4. Describe how numeric features are converted to categorical features ?

**Answer:**

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 ?

**Answer:**

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 ?

**Answer:**

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 ?

**Answer:**

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 ?

**Answer:**

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 ?

**Answer:**

Euclidean & Hamming distances are used to measure similarity or dissimilarity between two sequences. Euclidean distance is extensively applied in analysis of convolutional codes and Trellis codes.

Hamming distance is frequently encountered in the analysis of block codes.

##### 10. Distinguish between feature transformation and feature selection ?

**Answer:**

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.

##### 11. Make brief notes on the following:

##### 1. Collection of features using a hybrid approach

##### 2. The width of the silhouette

##### 3. Receiver operating characteristic curve

**Answer:**

**1. Collection of features using a hybrid approach**: In the context of data analysis or machine learning, the collection of features refers to the process of selecting and extracting relevant characteristics or attributes from a dataset. A hybrid approach typically implies combining multiple methods or techniques to accomplish this task. For example, it could involve a combination of manual feature engineering (where domain knowledge is used to select relevant features) and automatic feature extraction algorithms (such as deep learning models). This hybrid approach aims to leverage the strengths of different methods to obtain a comprehensive set of informative features.

**2. The width of the silhouette**: The silhouette width is a measure used in cluster analysis to assess the quality of clustering results. It quantifies how well each data point fits within its assigned cluster compared to other clusters. The silhouette width is calculated for each data point and ranges from -1 to 1. A value close to 1 indicates that the data point is well-matched to its cluster and is far from neighboring clusters. A value close to -1 suggests that the data point may have been assigned to the wrong cluster. A silhouette width around 0 indicates overlapping or ambiguous clusters. In general, a higher average silhouette width indicates better-defined and more separated clusters.

**3. Receiver operating characteristic curve**: The Receiver Operating Characteristic (ROC) curve is a graphical representation used to assess the performance of a binary classification model. It plots the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds. The TPR, also known as sensitivity or recall, measures the proportion of positive instances correctly classified as positive. The FPR represents the proportion of negative instances incorrectly classified as positive. By varying the classification threshold, the ROC curve provides a comprehensive view of the model's trade-off between true positives and false positives. A model with a higher area under the ROC curve (AUC) generally indicates better discrimination and performance.