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

A:

Each feature, or column, represents a measurable piece of data that can be used for analysis: Name, Age, Sex, Fare, and so on. Features are also sometimes referred to as “variables” or “attributes.” Depending on what you're trying to analyze, the features you include in your dataset can vary widely.

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

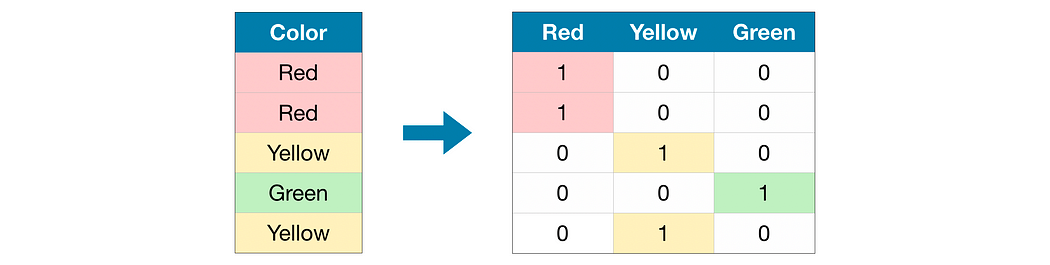
A:

Feature extraction/construction is a process through which a set of new features is created. They are used either in isolation or in combination. All attempt to improve performance such as estimated ac- curacy, visualization, and comprehensibility of learned knowledge.

3. Describe how nominal variables are encoded.

A:

When we have a feature where variables are just names and there is no order or rank to this variable's feature.



For example: City of person lives in, Gender of person, Marital Status, etc.In the above example, We do not have any order or rank, or sequence.

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

A:

Discretization: It is the process of transforming continuous variables into categorical variables by creating a set of intervals, which are contiguous, that span over the range of the variable's values. It is also known as “Binning”, where the bin is an analogous name for an interval.

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

approach?

A

In wrapper methods, the feature selection process is based on a specific machine learning algorithm that we are trying to fit on a given dataset. It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion.

The wrapper method has the advantages of better generalization and robust interaction with the classifier used for feature selection.

There are several disadvantages with WFS such as computationally intensive, discriminative power, lower shorter training times, classifier dependent selection, and higher risk of over-fitting than deterministic algorithms.

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

A:

One general definition for relevance is that a feature can be regarded as irrelevant if it is conditionally independent of the class labels or it does not influence the class labels; in these cases, it can be discarded.

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

be redundant?

A:

For example, 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.

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

A:

Euclidean Contours. The most common distance function used for numeric attributes or features is the Euclidean distance which is defined in the following formula: Euclidean distance between two points in n-dimensional space.

9. State difference between Euclidean and Manhattan distances?

A:

Euclidean Distance represents the shortest distance between two points.

Most machine learning algorithms including K-Means use this distance metric to measure the similarity between observations.

Manhattan Distance is the sum of absolute differences between points across all the dimensions.

10. Distinguish between feature transformation and feature selection.

A:

feature transformation: transformation of data to improve the accuracy of the algorithm; feature selection: removing unnecessary features.

11. Make brief notes on any two of the following:

1.SVD (Standard Variable Diameter Diameter)

SVD, or Singular Value Decomposition, is one of several techniques that can be used to reduce the dimensionality, i.e., the number of columns, of a data set. Why would we want to reduce the number of dimensions? In predictive analytics, more columns normally means more time required to build models and score data.

2. Collection of features using a hybrid approach

A: A hybrid feature selection method is proposed for classification in small sample size data sets. The filter step is based on instance learning taking advantage of the small sample size of data. A few candidate feature subsets are generated since their number corresponds to the number of instances.

3. The width of the silhouette

A:

The Average Silhouette Width (ASW) of a clustering is. a ( i ) is the average distance of to points in the cluster to which it was assigned, and is the average distance of to the points in the nearest cluster to which it was not assigned.

1. Receiver operating characteristic curve

A:

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.