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

Answer: 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 .As you may know, a “feature” is any measurable input that can be used in a predictive model — it could be the color of an object or the sound of someone's voice. Feature engineering, in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches.Feature can also mean to give special attention to something. The word feature has several other senses as a noun and a verb. A feature is a unique quality or characteristic that something has. Real-life examples: Elaborately colored tail feathers are peacocks' most well-known feature

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

Answer:Feature engineering can help data scientists by accelerating the time it takes to extract variables from data, allowing for the extraction of more variables. Automating feature engineering will help organizations and data scientists create models with better accuracy.Feature selection techniques can be used if the requirement is to maintain the original features, unlike the feature extraction techniques which derive useful information from data to construct a new feature subspace.Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. Step 1: Preprocessing of Data.

Step 2: Feature Engineering.

Step 3: Diverse Algorithms.

Step 4: Algorithm Selection.

Step 5: Training and Tuning.

Step 6: Ensembling.

Step 7: Head-to-Head Model Competitions.

Step 8: Human-Friendly Insights.

3. Describe how nominal variables are encoded.

Answer:Nominal Encoding 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.Nominal data are used to label variables without any quantitative value. Common examples include male/female (albeit somewhat outdated), hair color, nationalities, names of people, and so on. In plain English: basically, they're labels (and nominal comes from "name" to help you remember).Label encoding is mostly suitable for ordinal data. Because we give numbers to each unique value in the data. If we use label encoding in nominal data, we give the model incorrect information about our data. The model algorithm can act as if there is a hierarchy among the data.Encoding or continuization is the transformation of categorical variables to binary or numerical counterparts. An example is to treat male or female for gender as 1 or 0. Categorical variables must be encoded in many modeling methods (e.g., linear regression, SVM, neural networks).You can use zero-one encoding: male is encoded as 0 and female is encoded as 1. A dependent variable that has three or more possible values is usually encoded using one-hot encoding or ordinal encoding. If you are trying to predict color, which can be red, blue or green, ordinal encoding is red = 0, blue = 1, green =2.

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

Answer:Generally, for numerical features, we use mean and median values to replace them, in case of categorical features we tend to use mode. Let's understand this with the help of an example. In the above example, the value for height and hair color is missing.Meaning, if I consider the above example of hair color then assigning black hair=1, grey hair=2, brown hair=3, and so on. Though this approach is easy to handle categorical variables, there is a problem with it. The problem is, by assigning a number to each of the types of hair we are creating an artificial order. In terms of numbers we can say that 2>1 or 3>2 but when we consider the actual meaning of the number that is the color of the hair, we are actually trying to say that grey hair > black hair and brown hair > grey hair. This is an absurd statement to make. So, by converting it to simple numbers, we are creating an artificial order which is not genuine, and to solve this problem we use another method called one-hot encoding.

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

Answer: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.Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. Decreases over-fitting. Fewer redundant data means fewer chances of making decisions based on noise. Improves Accuracy. Less misleading data means better modeling accuracy. Reduces Training Time. Less data means quicker algorithms. There are three categories of feature selection methods, depending on how they interact with the classifier, namely, filter, wrapper, and embedded methods

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

Answer:A feature may be deemed irrelevant if it has poor data quality. We should be concerned whenever we expect recorded values in our dataset are different from true values. For example, data could be missing or entered incorrectly. Often poor data quality means a feature would not be predictive and excluded anyways.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.Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.The concept is really straightforward: We measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature. A feature is “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction.

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

Answer:Redundant Features Slow Down the Training Process

The more features you have, the slower the calculations are. However, there is another hidden factor that slows down training significantly. Having correlated features in the training set makes the loss landscape ill-conditioned (definition comes later).An attribute (column or feature of data set) is called redundant if it can be derived from any other attribute or set of attributes. Inconsistencies in attribute or dimension naming can also lead to the redundancies in data set.A redundant system consists of at least two systems that are interconnected and designed for the same purpose. There are many different types of redundant system configurations available, and different implementations of the system provide unique approaches to how to keep a system up at all times. It can be used for feature selection by evaluating the Information gain of each variable in the context of the target variable. Chi-square Test.

Fisher's Score.

Correlation Coefficient.

Dispersion ratio.

Backward Feature Elimination.

Recursive Feature Elimination.

Random Forest Importance.

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

Answer:

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. Perhaps four of the most commonly used distance measures in machine learning are as follows: Hamming Distance. Euclidean Distance. Manhattan Distance Similarity. 1)Cosine Similarity: 2) Manhattan distance: 3) Euclidean distance: 4) Minkowski distance. 5) Jaccard similarity:

9. State difference between Euclidean and Manhattan distances?

Answer:2) 1) Euclidean distance: The Euclidean distance between two points in either the plane or 3-dimensional space measures the length of a segment connecting the two points. It is the most obvious way of representing distance between two points.Euclidean distance calculates the distance between two real-valued vectors.

You are most likely to use Euclidean distance when calculating the distance between two rows of data that have numerical values, such a floating point or integer values.

If columns have values with differing scales, it is common to normalize or standardize the numerical values across all columns prior to calculating the Euclidean distance. Otherwise, columns that have large values will dominate the distance measure.

Manhattan distance: Manhattan distance is a metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. In a simple way of saying it is the total sum of the difference between the x-coordinates and y-coordinates.The Manhattan distance, also called the Taxicab distance or the City Block distance, calculates the distance between two real-valued vectors.

It is perhaps more useful to vectors that describe objects on a uniform grid, like a chessboard or city blocks. The taxicab name for the measure refers to the intuition for what the measure calculates: the shortest path that a taxicab would take between city blocks (coordinates on the grid).

It might make sense to calculate Manhattan distance instead of Euclidean distance for two vectors in an integer feature space.

Manhattan distance is calculated as the sum of the absolute differences between the two vectors.

ManhattanDistance = sum for i to N sum |v1[i] – v2[i]|

Formula: In a plane with p1 at (x1, y1) and p2 at (x2, y2)

10. Distinguish between feature transformation and feature selection.

Answer:The main difference:- Feature Extraction transforms an arbitrary data, such as text or images, into numerical features that is understood by machine learning algorithms. Feature Selection on the other hand is a machine learning technique applied on these (numerical) features.While both methods are used for reducing the number of features in a dataset, there is an important difference. Feature selection is simply selecting and excluding given features without changing them. Dimensionality reduction transforms features into a lower dimensionThus, feature selection and feature importance sometimes share the same technique but feature selection is mostly applied before or during model training to select the principal features of the final input data, while feature importance measures are used during or after training to explain the learned model.

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

1.SVD (Standard Variable Diameter Diameter)

2. Collection of features using a hybrid approach

3. The width of the silhouette

4. Receiver operating characteristic curve

Answer: 1.SVD (Standard Variable Diameter Diameter)SVD,--->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.The Singular Value Decomposition (SVD) of a matrix is a factorization of that matrix into three matrices. It has some interesting algebraic properties and conveys important geometrical and theoretical insights about linear transformations. It also has some important applications in data science

2. Collection of features using a hybrid approach---->.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--->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.The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

    4. Receiver operating characteristic curve---->A Receiver Operating Characteristic Curve (ROC) is a standard technique for summarizing classifier performance over a range of trade-offs between true positive (TP) and false positive (FP) error rates (Sweets, 1988).The ROC curve is used to assess the overall diagnostic performance of a test and to compare the performance of two or more diagnostic tests. It is also used to select an optimal cut-off value for determining the presence or absence of a disease.