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

Ans: Features are nothing but the independent variables in ML models. In any ML problem is a set of these features (independent variables), coefficients of these features, and parameters for coming up with appropriate functions or models (also termed hyperparameters) is required to be learned. represents a few examples of features of ML models:

* A model for predicting the risk of cardiac disease may have features such as the following:
  + Age
  + Gender
  + Weight
  + Whether the person smokes
  + Whether the person is suffering from diabetic disease, etc.
* A model for predicting whether the person is suitable for a job may have features such as the educational qualification, number of years of experience, experience working in the field etc
* A model for predicting the size of a shirt for a person may have features such as age, gender, height, weight, etc.

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

Ans: Feature creation of new input or target features from existing features always aims to create ones that do a better job of representing a machine learning problem to the model.

If existing dataset has entries of city,date and adding the *district* to make inferences at the district level. The date is broken down into *year* and *month* components to look for patterns across time as well as seasonal ones. By spending time performing feature engineering, focus on high-quality features, rather than throwing raw data at a model can be given

For eg. using data from the World Happiness Report, we create a new feature, happiness\_band, by binning the happiness feature into low, medium, and high bands or apply binning to categorical features where the countries binned into their global region.

Splitting can be used to split up an existing feature into multiple new features.

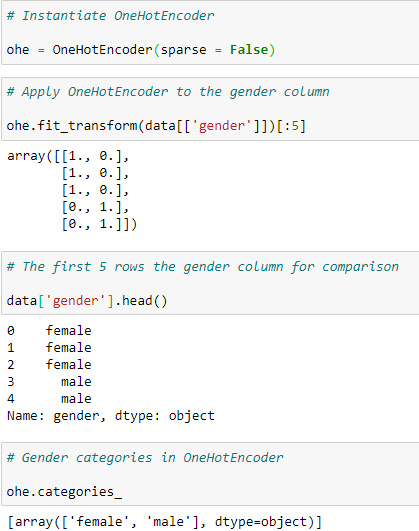
1. Describe how nominal variables are encoded.

Ans: Nominal variables are encoded using OneHotEncoder or [get\_dummies](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html).

Consider for e.g following are nominal variables in the dataset :

* Gender
* Race/ethnicity
* Lunch
* Test preparation course

## Scikit-learn OneHotEncoder



OneHotEncoder has created two columns to represent the two categories in the gender column, one for male and one for female. Female students will receive a value of 1 in the female column and a value of 0 in the male column.

## Pandas get\_dummies

Graphical user interface, text, application

Description automatically generated

Similarly, get\_dummies has encoded the gender feature by creating two columns to represent the two categories in the gender feature, male and female.

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

Ans: numeric features are converted to categorical features using Bucketing /Binning techniques

1. Equal width binning: It is also known as “Uniform Binning” since the width of all the intervals is the same. The algorithm divides the data into N intervals of equal size. The width of intervals is:
   * + 1. w=(max-min)/N
2. Equal frequency binning or “Quantile Binning”: The algorithm divides the data into N groups where each group contains approximately the same number of values. Consider, we want 10 bins, that is each interval contains 10% of the total observations. Here the width of the interval need not necessarily be equal. Handles outliers better than the previous method and makes the value spread approximately uniform(each interval contains almost the same number of values).
3. K-means binning: This technique uses the clustering algorithm namely ” K-Means Algorithm”.This technique is mostly used when our data is in the form of clusters.

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

Ans: Wrapper methods: it is a greedy algorithms train the algorithm by using a subset of features in an iterative manner. Based on the conclusions made from training in prior to the model, addition and removal of features takes place. Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved.

Graphical user interface, text

Description automatically generated with medium confidence

Wrapper Methods Implementation

Some techniques used are:

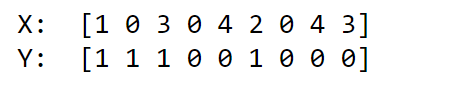
* Forward selection – This method is an iterative approach where we initially start with an empty set of features and keep adding a feature which best improves our model after each iteration. The stopping criterion is till the addition of a new variable does not improve the performance of the model.
* Backward elimination – This method is also an iterative approach where we initially start with all features and after each iteration, we remove the least significant feature. The stopping criterion is till no improvement in the performance of the model is observed after the feature is removed.
* Bi-directional elimination – This method uses both forward selection and backward elimination technique simultaneously to reach one unique solution.
* Exhaustive selection – This technique is considered as the brute force approach for the evaluation of feature subsets. It creates all possible subsets and builds a learning algorithm for each subset and selects the subset whose model’s performance is best.
* Recursive elimination – This greedy optimization method selects features by recursively considering the smaller and smaller set of features. The estimator is trained on an initial set of features and their importance is obtained using feature\_importance\_attribute. The least important features are then removed from the current set of features till we are left with the required number of features.

The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive which is drawback.

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

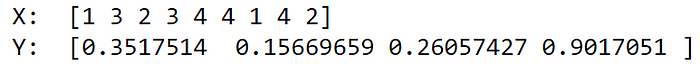
Ans: When a feature does not affect output,it is said to be irrelevant feature.The diifernt test are used to quantify like:

* Chi-squre test for categorical feature *Categoric*al  response:



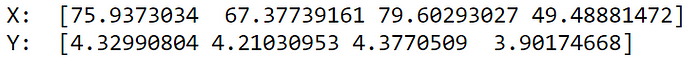
*The* essence of the Chi-square test is assuming no relationship between the input and output and checking how valid that assumption is. If the feature truly affects the response, P value must be low else p value is high for irrelevant feature.

* *ANOVA test for categorical feature continuous response:*

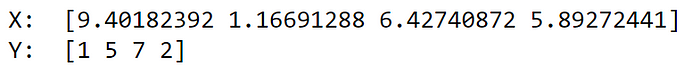


*The means of different groups (mean of responses for each categorical feature data) and tests whether the intergroup difference is statistically significant. If a feature is relevant, we expect to see significant differences between the means of different group*

* Correlation or Chi-square/ANOVA:Continuous Feature Continuous Response *quare/ANOVA*



*The most common correlation coefficient is called*[*Pearson’s coefficient*](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)*, which tests the linear relationship between the feature and the responsePearson’s correlation returns low values when the relationship is non-linear. A better choice is called*[*Spearman’s correlation*](https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php)*, which tests if a monotonic relationship exists.*

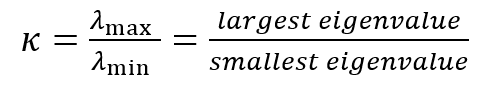
* *Continuous Feature Categorical Response — Chi-Square Test continuous feature categorical response*

*For this problem, we can simply discretize the feature by splitting data into different ‘groups’ After discretization, we can apply the Chi-Square Test again.*

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

Ans:. The more features slow down the training time of models and if highly correlated features requires much more iterations to train the models.Thus thses redundant features can be identified. .Criteria to Identify redundant features :

to detect multicollinearity in data is called Eigensystem Analysis, which uses the concept of Condition Number. The definition of Condition Number(K) is:

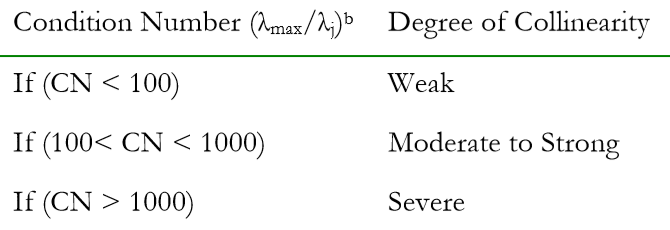


Given a matrix(input data X), find the correlation matrix of X Corr(X):

Table

Description automatically generated

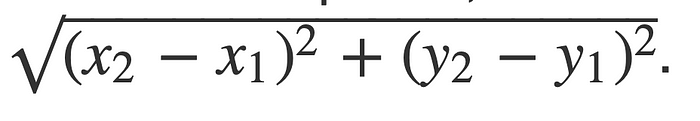
If the correlation matrix of X has a large condition number, it indicates serious collinearity.



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

# Ans: **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.  Euclidean distance between them is

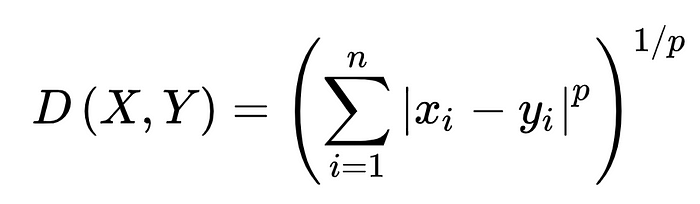


# **Minkowski distance**

Minkowski distance is a generalisation of the Euclidean and Manhattan distances.

**Formula:** The Minkowski distance of order p between two points is defined as



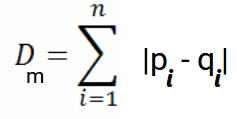


1. Manhattan Distance

Manhattan Distance is the sum of absolute differences between points across all the dimensions. the Manhattan distance in a 2-dimensional space is given as:

manhattan distance formula | distance metrics

And the generalized formula for an n-dimensional space is given as:



Where,

* n = number of dimensions
* pi, qi = data points

Hamming Distance measures the similarity between two strings of the same length. The Hamming Distance between two strings of the same length is the number of positions at which the corresponding characters are different.e.g even characters are different, whereas two characters (the last two characters) are similar:

euclidean and manhattan distances

Hence, the Hamming Distance here will be 7.

9. State difference between Euclidean and Manhattan distances?

Ans: While Euclidean distance gives the shortest or minimum distance between two points, Manhattan Distance is the sum of absolute differences between points across all the dimensions.For example, if we were to use a Chess dataset, the use of Manhattan distance is more appropriate than Euclidean distance.

10. Distinguish between feature transformation and feature selection.

Ans:It is a technique by which we can boost our model performance. Feature transformation is a mathematical transformation in which we apply a mathematical formula to a particular column(feature) and transform the values which are useful for our further analysis. transformation of data to improve the accuracy of the algorithm while feature selection  removes unnecessary features. Feature transformations(transforming them to make sense)Feature transformation deals with:

* Normalization and changing distribution(Scaling)
* Interactions
* Filling in the missing values(median filling etc)

Feature selection(building your model on these selected features)

* Statistical approaches
* Selection by modeling
* Grid search
* Cross Validation etc

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

1.SVD (Standard Variable Diameter Diameter)

2. Collection of features using a hybrid approach

1. The width of the silhouette:

Ans:The width of the silhouette:  The Average Silhouette Width (ASW) is a popular cluster validation index to estimate the number of clusters. The ASW is an intuitive and simple measurement of cluster quality that does not rely on statistical model assumptions. Given that it is widely used and trusted for comparing the qualities of clusterings produced by various clustering methods over different numbers of clusters, it seems natural to investigate optimal ASW quality clustering not only over k but also for fixed k in order to integrate the problem for fixed k and the problem of finding the best k.

1. Receiver operating characteristic curve

Ans: 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

True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

TPR=TP/TP+FN

False Positive Rate (FPR) is defined as follows:

FPR=FN/TP+FN

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

