**MACHINE LEARNING ASSIGNMENT 8**

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

In machine learning, features are individual independent variables that act like a input in your system. Actually, while making the predictions, models use such features to make the predictions. And using the feature engineering process, new features can also be obtained from old features in machine learning.

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](https://www.datarobot.com/wiki/insights/) you will gain when you use that dataset for [machine learning](https://www.datarobot.com/wiki/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](https://www.datarobot.com/wiki/data-science/) project.

You can improve the quality of your dataset’s features with processes like [feature selection](https://www.datarobot.com/wiki/feature-selection/) and [feature engineering](https://www.datarobot.com/wiki/feature-engineering/)**,**which are notoriously difficult and tedious. If these techniques are done well, the resulting optimal dataset will contain all of the essential features that might have bearing on your specific business problem, leading to the best possible model outcomes and the most beneficial insights.

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

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. We do this by including or excluding important features without changing them. It helps in cutting down the noise in our data and reducing the size of our input data.

Feature Selection Models

Feature selection models are of two types:

1. Supervised Models: Supervised feature selection refers to the method which uses the output label class for feature selection. They use the target variables to identify the variables which can increase the efficiency of the model
2. Unsupervised Models: Unsupervised feature selection refers to the method which does not need the output label class for feature selection. We use them for unlabelled data.

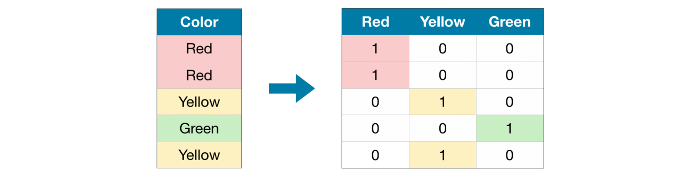
**3. Describe how nominal variables are encoded.**

When we are working on some of the datasets, we found that some of the features are categorical. We all know that machines can’t understand categorical data. Models only work with numerical values. For this reason, it is necessary to convert the categorical values of the features into numerical ones, So the machine can learn from those data and gives the right model. This process of converting categorical data into numerical data is called Encoding.

There are two most popular types of encoding,

1. Nominal Encoding

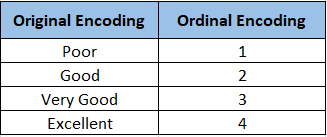
When we have a feature where variables are just names and there is no order or rank to this variable's feature.

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. All the variables in the respective feature are equal. We can't give them any orders or ranks. Those features are called Nominal features.

2. Ordinal Encoding

When we have a feature where variables have some order/rank.



For example: Student’s performance, Customer’s review, Education of person, etc…

In the above example, we have orders/ranks/sequences. We can assign ranks based on student’s performance, based on feedback given by customers, based on the highest education of the person. Those features are called Ordinal features.

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

A fundamental task in the field of machine learning is converting numeric data to categorical data. For example, if you have a data set of people’s heights in inches, such as 59.5, 64.0 and 75.5, you might want to convert this numeric data into categorical data, for example 0, 1, and 2, to represent short, medium, and tall. Informally, this process is sometimes called binning data. In machine-learning literature, the process is usually called discretization of continuous data.

There are several scenarios where you might want to discretize data. Many machine-learning algorithms, such as naive Bayes classification and prediction, work only with categorical data. So if your raw data is numeric and you want to apply naive Bayes, you have to discretize the data. You might also have mixed numeric and categorical data, such as the data often found in an Excel spreadsheet. Very few machine-learning algorithms work with mixed data, so you can convert the numeric data to categorical data and then use a machine-learning algorithm that works with categorical data. Data clustering using category utility is an example.

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

## Wrapper methods

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 evaluation criterion is simply the performance measure which depends on the type of problem, for e.g. For regression evaluation criterion can be p-values, R-squared, Adjusted R-squared, similarly for classification the evaluation criterion can be accuracy, precision, recall, f1-score, etc. Finally, it selects the combination of features that gives the optimal results for the specified machine learning algorithm.

Most commonly used techniques under wrapper methods are:

1. Forward selection
2. Backward elimination
3. Bi-directional elimination(Stepwise Selection)

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

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

Identifying the redundant features in a high-dimensional space is quite complex since the feature selection method needs a special mechanism to identify the relevancy among the features. Therefore, redundancy analysis takes more computations than the relevancy analysis and it needs a suitable relevancy measure. In order to overcome these problems in the redundancy analysis phase of the feature selection process, the clustering techniques can be employed.

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

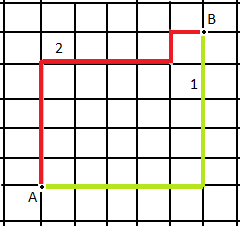
Many Supervised and Unsupervised machine learning models such as K-NN and K-Means depend upon the distance between two data points to predict the output. Therefore, the metric we use to compute distances plays an important role in these models.

In this blog post, we are going to learn about some distance metrics used in machine learning models. They are:-

* Minkowski distance
* Manhattan distance
* Euclidean distance
* Hamming distance
* Cosine distance

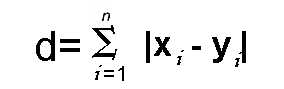
**9. State difference between Euclidean and Manhattan distances?**

**Manhattan Distance:**We use Manhattan distance, also known as **city block distance**, or **taxicab geometry**if we need to calculate the distance between two data points in a grid-like path. Manhattan distance metric can be understood with the help of a simple example.



In the above picture, imagine each cell to be a building, and the grid lines to be roads. Now if I want to travel from Point A to Point B marked in the image and follow the red or the yellow path. We see that the path is not straight and there are turns. In this case, we use the Manhattan distance metric to calculate the distance walked.

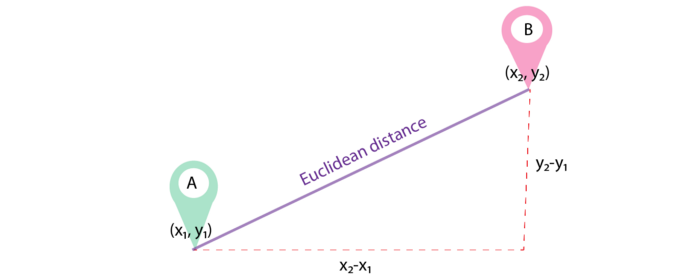
We can get the equation for Manhattan distance by substituting p = 1 in the Minkowski distance formula. The formula is:-



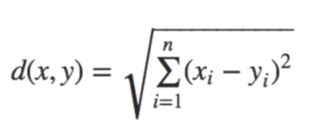
**When is Manhattan distance metric preferred in ML?**Quoting from the paper, [“On the Surprising Behavior of Distance Metrics in High Dimensional Space”](https://bib.dbvis.de/uploadedFiles/155.pdf), by Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Kiem. “ for a given problem with a fixed (high) value of the dimensionality d, it may be preferable to use lower values of p. This means that the L1 distance metric (Manhattan Distance metric) is the most preferable for high dimensional applications.”

Thus, Manhattan Distance is preferred over the Euclidean distance metric as the dimension of the data increases. This occurs due to something known as the ‘curse of dimensionality’. For further details, please visit this [link](https://bib.dbvis.de/uploadedFiles/155.pdf).

## Euclidean Distance: Euclidean distance is the straight line distance between 2 data points in a plane.



It is calculated using the Minkowski Distance formula by setting ‘p’ value to 2, thus, also known as the L2 norm distance metric. The formula is:-



This formula is similar to the Pythagorean theorem formula, Thus it is also known as **the Pythagorean Theorem.**

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

**Feature Selection**:- This module is used for feature selection/dimensionality reduction on given datasets. This is done either to improve estimators’ accuracy scores or to boost their performance on very high-dimensional datasets.

**Feature Extraction**:- This module is used to extract features in a format supported by machine learning algorithms from the given datasets consisting of formats such as text and image.

**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.

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

**1.SVD (Standard Variable Diameter Diameter)**

The SVD is one of the most well used and general purpose tools from linear algebra for data processing!

Methodologically

* Dimension reduction (e.g. images, gene expression data, movie preferences)
* Used as a first step in many data reduction and machine learning approaches
* Taylor a coordinate system based on the data we have
* Solve system of linear equations for non-square matrices: e.g. linear regression
* Basis for principal component analysis (PCA) and multidimensional scaling (MDS).
  + PCA is one of the most widely used methods to study high dimensional data and to understand them in terms of their dominant patterns and correlations

Applications:

* At the heart of search engines: Google
* Basis of many facial recognition methods: e.g. Facebook
* Recommender systems such as Amazon and Netflix
* A standard tool for data exploration and dimension reduction in Genomics

2. Collection of features using a hybrid approach

Most learning algorithms used in ML are really good at completing one task or working with one dataset. While helpful and infinitely better than doing it manually, these algorithms won’t help you realize the full potential of AI across all of your data.

That’s where hybrid machine learning (HML) comes in. Multiple simple algorithms work together to complement and augment each other. Together they can solve problems that alone they were not designed to solve.

Within HML there are various types of techniques that interact with the data in different ways. Which technique you use depends on the problems you’re trying to solve, the technical expertise available, and the tools you’re using.

Here are some types of hybrid machine learning.

#### Semi-supervised learning

In semi-supervised learning, you provide the algorithm with a small set of labelled data. Then, you give it a much larger set of unlabeled data and put it to work. This type of algorithm is helpful when you need (or have) to start with a smaller batch of data upfront. It learns from all the data, not just the labelled data, and helps you organize it.

#### Self-supervised learning

A self-supervised learning model combines unsupervised and supervised learning problems, then applies a supervised learning algorithm. You can create the model for the algorithm to follow, and it begins applying that to unlabeled data.

This type of learning is commonly used on unlabeled images and defines actions that can be taken on those images—like rotating them, identifying color or grayscale, or distinguishing between real and fake photos.

#### Multi-instance learning

Multi-instance learning is a method where you are labeling groups or collections of data, rather than the individual members of the group. This is a helpful method when you’re working with large sets of similar data and have a lot of duplicates.

This method uses supervised learning models to identify labels for groups of data. You train the models to recognize attributes of a few pieces of data within a group, and then it predicts labels for future groups based on attributes of some of the data within the new groups.

### Tools to support hybrid machine learning

The point of incorporating ML into your data science and analytics processes is to allow you to begin looking forward with your data. Rather than relying on datasets that have been cleaned and organized, you’ll be able to quickly group and label data in real-time for the most accurate analysis and forecasts.

When considering how to manage this data, you’ll need a few key features in your [business intelligence tools](https://www.domo.com/why-domo) to support this type of advanced analysis. Your tool will need to support:

**Integration from all your data sources**  
You’ll need one place to manage your data and train your ML models. Find a tool that will allow for[easy integration of all your data sources](https://www.domo.com/data-integration).

**Real-time analysis**  
Many of the HML models mentioned here function best as they’re learning from new data. Find tools that will support [real-time ingestion and analysis](https://www.domo.com/platform/leverage-the-cloud), and then will push that data out to workers who can use it to improve performance right then.

**Automatic decisions**  
Find a tool that will support [automatic decisions](https://www.domo.com/glossary/what-is-automated-machine-learning/) for your team, with [alerts and notifications](https://www.domo.com/business-intelligence/features/alerts) for when your data passes specific thresholds.

.

**3. The width of the silhouette**

Silhouette width is a widely used index for assessing the fit of individual objects in the classification, as well as the quality of clusters and the entire classification. Silhouette combines two clustering criteria, compactness and separation, which imply that spherical cluster shapes are preferred over others—a property that can be seen as a disadvantage in the presence of complex, nonspherical clusters, which is common in real situations. We suggest a generalization of the silhouette width using the generalized mean.

**4.Receiver operating characteristic curve**

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.