Shaina Mehta
7CSE 4Y
A2305219268

Assignment 2.2

Q1) Explain the difference between Data Selection and Data Extraction.

Ans) Feature Selection: The feature selection 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. Examples: Variance Thresholds, Correlation Thresholds, Genetic Algorithms (GA), etc.

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. Examples: PCA Analysis, LDA Analysis, Autoencoders, etc.

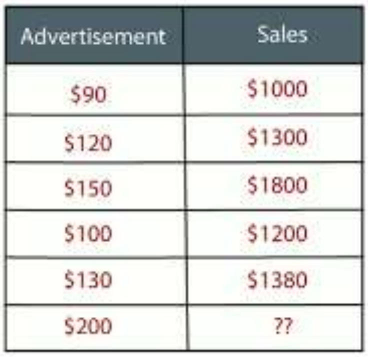
The main difference is: Feature Extraction transforms arbitrary data, such as text or images, into numerical features that are understood by machine learning algorithms. Feature Selection on the other hand is a machine learning technique applied to these (numerical) features.

Q2) What is Regression analysis?

Ans) Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as temperature, age, salary, price, etc.

We can understand the concept of regression analysis using the below example:

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:



Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Now, the company wants to do the advertisement of $200 in the year 2019 and wants to know the prediction about the sales for this year. So, to solve such type of prediction problems in machine learning, we need regression analysis. Regression is a supervised learning technique that helps in finding the correlation between variables and enables us to predict the continuous output variable based on one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

In Regression, we plot a graph between the variables which best fit the given data points, using this plot, the machine learning model can make predictions about the data. In simple words, *"Regression shows a line or curve that passes through all the datapoints on the target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum."* The distance between data points and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

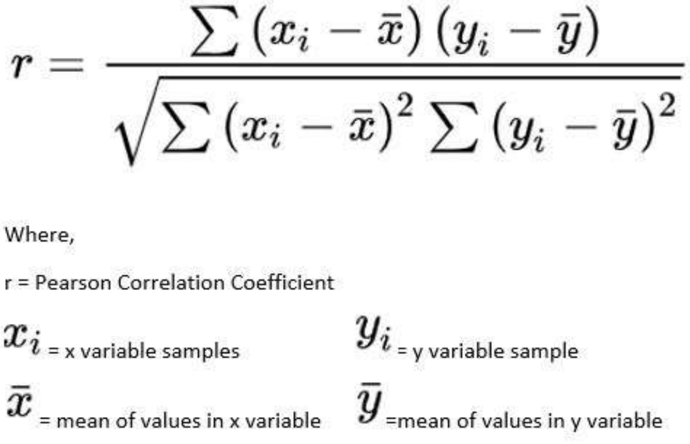
- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables. Here we are discussing some important types of regression which are given below:

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- Ridge Regression
- Lasso Regression

Q3) What is Pearson correlation coefficient?

Ans) Pearson's Correlation coefficient is represented as 'r', it measures how strong is the linear association between two continuous variables using the formula:



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

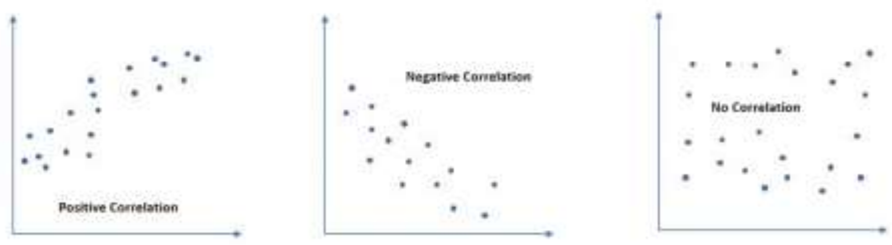
Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

Value of ‘r’ ranges from ‘-1’ to ‘+1’. Value ‘0’ specifies that there is no relation between the two variables. A value greater than ‘0’ indicates a positive relationship between two variables where an increase in the value of one variable increases the value of another variable. Value less than ‘0’ indicates a negative relationship between two variables where an increase in the value of one decreases the value of another variable.



Pearson correlation attempts to draw a line of best fit through the spread of two variables. Hence, it specifies how far away all these data points are from the line of best fit. Value of ‘r’ equal to near to +1 or -1 that means all the data points are included on or near to the line of best fit respectively. Value of ‘r’ closer to the ‘0’ data points is around the line of best fit.

Assumptions for a Pearson Correlation:

1. Data should be derived from random or least representative samples, draw a meaningful statistical inference.
2. Both variables should be continuous and normally distributed.
3. There should be Homoscedasticity, which means the variance around the line of best fit should be similar.
4. Extreme outliers influence the Pearson Correlation Coefficient. You need to consider outliers that are unusual only on one variable, called as ‘univariate variable’ or for both of the variables known as ‘multivariate outliers’. 2 variables are measured independently from each other pairs. e.g. If we plot age vs amount then, we can certainly, see that there is a correlation between the age of a person and loan the amount is given to that person, as age increases the loan amount given to the person decreases and vice versa. But if we plot the loan amount vs age, it is not possible to draw any conclusion from it. It would violate the assumption.

Q4) Describe Factor Analysis in detail. (5-8 lines)

Ans) Factor Analytics is a special technique reducing the huge number of variables into a few numbers of factors is known as factoring of the data, and managing which data is to be present in sheet comes under factor analysis. It is completely a statistical approach that is also used to describe fluctuations among the observed and correlated variables in terms of a potentially lower number of unobserved variables called factors.

The factor analysis technique extracts the maximum common variance from all the variables and puts them into a common score. It is a theory that is used in training the machine learning model and so it is quite related to data mining. The belief behind factor analytic techniques is that the information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset.

Factor analysis is a very effective tool for inspecting changeable relationships for complex concepts such as social status, economic status, dietary patterns, psychological scales, biology, psychometrics, personality theories, marketing, product management, operations research, finance, etc. It can help a researcher to investigate the concepts that are not easily measured in a much easier and quicker way directly by the cave in a large number of variables into a few easily interpretable fundamental factors.

Types of factor analysis:

1. Exploratory factor analysis (EFA)
2. Confirmatory factor analysis (CFA)
3. Multiple Factor Analysis
4. Generalized Procrustes Analysis (GPA)

Factor Loadings

In addition, factors are created with equality; some factors have more weights some have low. In a simple example, imagine your car company says Maruti Suzuki is conducting a survey includes, using – telephonic survey, physical survey, google forms, etc. for customer satisfaction and the results show the following factor loadings:

VARIABLE	F1	F2	F3
Problem 1	0.985	0.111	-0.032
Problem 2	0.724	0.008	0.167
Problem 3	0.798	0.180	0.345

Here –

F1 – Factor 1

F2 – Factor 2

F3 – Factor 3

The factors that affect the question the most (and therefore have the highest factor loadings) are bolded. Factor loadings are similar to correlation coefficients in that they can vary from -1 to 1. The closer factors are to -1 or 1, the more they affect the variable.

Note: A factor loading of 0 indicates no effect.

Q5) Write the various applications of Proximity measures.

Ans) The applications of proximity measures are given as:

1. It is used by K-Nearest Neighbour Algorithm for classification purposes.
2. It is used by clustering algorithms such as K-Means Clustering, Agglomerative Hierarchical Clustering Algorithms, etc. to find the groups of similar characteristics in a given dataset.

3. It is used by Anomaly detection algorithms to find out the most dissimilar objects in a given dataset.