PRIOR KNOWLEDGE

Date	19 November 2022
Team ID	PNT2022TMID32768
Project Name	Web Phishing Detection
Maximum Marks	2 Marks

To understand and work out the project, we must have prior knowledge on the following concepts:

- Supervised Learning
- Unsupervised Learning
- Regression Classification and Clustering
- Logistic Regression
- Flask

Supervised Learning:

Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

Supervised learning can be separated into two types of problems when data mining—classification and regression:

Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the data set and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, which are described in more detail below.

Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

Unsupervised Learning:

Unsupervised learning, also known as **unsupervised machine learning**, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

MARKET BASKET ANALYSIS

Nowadays Machine Learning is helping the Retail Industry in many different ways. You can imagine that from forecasting the performance of sales to identify the buyers, there are many applications of machine learning(ML) in the retail industry. “Market Basket Analysis” is one of the best applications of machine learning in the retail industry. By analyzing the past buying behavior of customers, we can find out which are the products that are bought frequently together by the customers.

Regression Classification and Clustering:

Regression and Classification are types of supervised learning algorithms while Clustering is a type of unsupervised algorithm. When the output variable is continuous, then it is a regression problem whereas when it contains discrete values, it is a classification problem.

The first one is clustering. Clustering is an unsupervised technique. With clustering, the algorithm tries to find a pattern in data sets without labels associated with it. This could be a clustering of buying behaviour of customers. Features for this would be the household income, age, … and clusters of different consumers could then be built.

The next one is classification. In contrast to clustering, classification is a supervised technique. Classification algorithms look at existing data and predicts what a new data belongs to. Classification is used for spam for years now and these algorithms are more or less mature in classifying something as spam or not. With machine data, it could be used to predict a material quality by several known parameters (e.g. humidity, strength, colour,etc).The output of the material prediction would then be the quality type (either “good” or “bad” or a number in a defined space like 1-10).

The last technique for this post is regression. Regression is often confused with clustering, but it is still different from it. With a regression, no classified labels (such as good or bad, spam or not spam) are predicted. Instead, regression outputs continuous, often unbound, numbers. This makes it useful for financial predictions and alike. A common known sample is the prediction of housing prices, where several values (features) are known, such as distance to specific landmarks, plot size. The algorithms could then predict a price for your house and the amount you can sell it for.

SEMANTIC CLUSTERING:

We describe a semantic clustering method designed to address shortcomings in the common bag-of-words document representation for functional semantic classification tasks. The method uses Word Net-based distance metrics to construct a similarity matrix, and expectation maximization to find and represent clusters of semantically-related terms. Using these clusters as features for machine learning helps maintain performance across distinct, domain-specific vocabulary while reducing the size of the document representation. We present promising results.

Logistic Regression:

In statistics, the logistic model is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression is estimating the parameters of a logistic model.

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd using logistic regression.Many other medical scales used to assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, gender, body mass index, results of various blood tests, etc.).The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc.In economics it can be used to predict the likelihood of a person ending up in the labor force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

Flask:

Flask is a micro web framework written in Python. It is classified as a micro framework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

A Flask extension typically has flask in its name as a prefix or suffix. If it wraps another library, it should include the library name as well. This makes it easy to search for extensions, and makes their purpose clearer. A general Python packaging recommendation is that the install name from the package index and the name used in import statements should be related. The import name is lowercase, with words separated by underscores (_). The install name is either lower case or title case, with words separated by dashes (-). If it wraps another library, prefuse the same case as that library’s name.

Here are some example install and import names:

- Flask-Name imported as flask name
- flask-name-lower imported as flask name lower
- Flask-Combo Name imported as
- Name-Flask imported as Configuration Techniques
- Configuration per application instance, through app configuration values. This is configuration that could reasonably change for each deployment of an application. A common example is a URL to an external resource, such as a database. Configuration keys should start with the extension’s name so that they don’t interfere with other extensions.
- Configuration per extension instance, through init arguments. This configuration usually affects how the extension is used, such that it wouldn’t make sense to change it per deployment.