Prior Knowledge

Date	04 September 2022
Project Name	Project- WEB PHISHING DETECTION
Team ID	PNT2022TMID43816

Prior Knowledge:

One should have knowledge on the following Concepts

1. Supervised and unsupervised learning

Supervised learning

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

Unsupervised learning

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

2. Regression Classification and 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 behavior of customers. Features for this would be the household income, age,… and clusters of different consumers could then be built.

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, color,…). 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). Another well known sample is if someone would survive the titanic – classification is done by “true” or “false” and input parameters are “age”, “sex”, “class”. If you are 55, male and in 3rd class, chances are low. But if you are 12, female and in first class, chances are rather high.

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 prediciton 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.

Logistic Regression:

This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi))

ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k

Flask:

Flask is a web framework, it’s a Python module that lets you develop web applications easily. It’s has a small and easy-to-extend core: it’s a microframework that doesn’t include an ORM (Object Relational Manager) or such features.

It does have many cool features like url routing, template engine. It is a WSGI web app framework.