

Machine Learning (Sessional)

CSE - 452



Date: September 15, 2019

Remarks:

Name: Avishek Das

ID: 1504018

Section : A

Group: A1

Introduction to Apriori Algorithm:

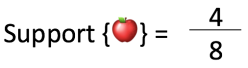
Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

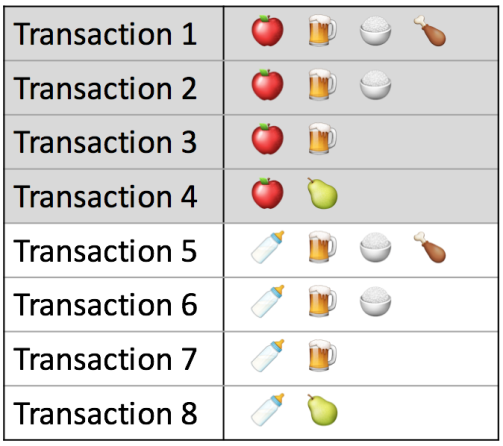
**Association rules**

Association rule learning is a prominent and a well-explored method for determining relations among variables in large databases.

Association rules analysis is a technique to uncover how items are associated to each other. There are three common ways to measure association.

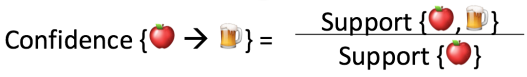
**Measure 1: Support**. This says how popular an itemset is, as measured by the proportion of transactions in which an itemset appears. In Table 1 below, the support of {apple} is 4 out of 8, or 50%. Itemsets can also contain multiple items. For instance, the support of {apple, beer, rice} is 2 out of 8, or 25%.



  
Table 1. Example Transactions

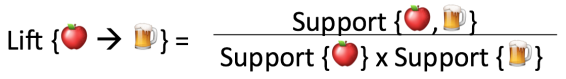
If you discover that sales of items beyond a certain proportion tend to have a significant impact on your profits, you might consider using that proportion as your *support threshold*. You may then identify itemsets with support values above this threshold as significant itemsets.

**Measure 2: Confidence**. This says how likely item Y is purchased when item X is purchased, expressed as {X -> Y}. This is measured by the proportion of transactions with item X, in which item Y also appears. In Table 1, the confidence of {apple -> beer} is 3 out of 4, or 75%.



One drawback of the confidence measure is that it might misrepresent the importance of an association. This is because it only accounts for how popular apples are, but not beers. If beers are also very popular in general, there will be a higher chance that a transaction containing apples will also contain beers, thus inflating the confidence measure. To account for the base popularity of both constituent items, we use a third measure called lift.

**Measure 3: Lift**. This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. In Table 1, the lift of {apple -> beer} is 1,which implies no association between items. A lift value greater than 1 means that item Y is *likely* to be bought if item X is bought, while a value less than 1 means that item Y is *unlikely* to be bought if item X is bought.



**Implementation of** Apriori Algorithm:

Dataset:

Groceries\_dataset.csv file which includes three attributes :  
• Member\_number: An ID that can help distinguish different purchases by different  
customers.  
• Date:The ate of transaction  
• ItemDescription: The description of the actual item that was bought

Goal:

Implement the Apriori algorithm to

 find all frequent item sets.

 generate association rules

Introduction to K-means clustering:

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity..

**How does the** K-means clustering **work?**

K-Means starts by randomly defining k centroids. From there, it works in iterative (repetitive) steps to perform two tasks:

* Assign each data point to the closest corresponding centroid, using the standard Euclidean distance. In layman’s terms: the straight-line distance between the data point and the centroid.
* For each centroid, calculate the mean of the values of all the points belonging to it. The mean value becomes the new value of the centroid.

Once step 2 is complete, all of the centroids have new values that correspond to the means of all of their corresponding points. These new points are put through steps one and two producing yet another set of centroid values. This process is repeated over and over until there is no change in the centroid values, meaning that they have been accurately grouped. Or, the process can be stopped when a previously determined maximum number of steps has been met.

**Implementation of** K-means clustering **With Tensorflow:**

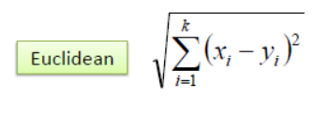
Dataset : The data set (wine.xlsx) that we are going to analyze in this lab is a result of a chemical analysis of wines grown in a particular region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. The attributes are: Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, and Proline. The data set has 178 observations and no missing values

Goal : To group similar observations together and determine the number of possible clusters (region).

Introduction to KNN (K-Nearest Neighbours) :

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. It is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbours are classified.

**How does the KNN algorithm work?**

In the classification setting, the K-nearest neighbor algorithm essentially boils down to forming a majority vote between the K most similar instances to a given “unseen” observation. Similarity is defined according to a distance metric between two data points. A popular one is the Euclidean distance method.

Other methods are Manhattan, Minkowski, and Hamming distance methods. For categorical variables, the hamming distance must be used.

**Implementation of KNN With Sci-kit Learn:**

Dataset : The dataset (train.csv) used to implement KNN contains information about the passengers in TITANIC. The features of this dataset are PassengerId, Survived, Pclass, Name, Sex, Age, Sibsp, Parch, Ticket, Fare, Cabin, Embarked.

Goal : Given the (test.csv) dataset we have to find out for the information of a passenger whether the passenger Survived or Not Survived.

Introduction to Decision Tree:

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any boolean function on discrete attributes using the decision tree.

**How does the Decision Tree algorithm work?**

* At the beginning, we consider the whole training set as the root.
* Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
* On the basis of attribute values records are distributed recursively.
* We use statistical methods for ordering attributes as root or the internal node.

In Decision Tree the major challenge is to identification of the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

1. Information Gain
2. Gini Index

**Implementation of** **Decision Tree With Sci-kit Learn:**

Dataset : The dataset (train.csv) used to implement Decision Tree contains information about the passengers in TITANIC. The features of this dataset are PassengerId, Survived, Pclass, Name, Sex, Age, Sibsp, Parch, Ticket, Fare, Cabin, Embarked.

Goal : Given the (test.csv) dataset we have to find out for the information of a passenger whether the passenger Survived or Not Survived.

Introduction to An artificial neuron network (ANN):

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output.ANNs are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found.

**How does the An artificial neuron network (ANN) work?**

Artificial Neural Networks can be best viewed as weighted directed graphs, where the nodes are formed by the artificial neurons and the connection between the neuron outputs and neuron inputs can be represented by the directed edges with weights. The Artificial Neural Network receives the input signal from the external world in the form of a pattern and image in the form of a vector. These inputs are then mathematically designated by the notations x(n) for every n number of inputs.Each of the input is then multiplied by its corresponding weights (these weights are the details used by the artificial neural networks to solve a certain problem). In general terms, these weights typically represent the strength of the interconnection amongst neurons inside the artificial neural network. All the weighted inputs are summed up inside the computing unit (yet another artificial neuron).If the weighted sum equates to zero, a bias is added to make the output non-zero or else to scale up to the system’s response. Bias has the weight and the input to it is always equal to 1. Here the sum of weighted inputs can be in the range of 0 to positive infinity. To keep the response in the limits of the desired value, a certain threshold value is benchmarked. And then the sum of weighted inputs is passed through the activation function.

The activation function, in general, is the set of transfer functions used to get the desired output of it. There are various flavors of the activation function, but mainly either linear or non-linear sets of functions. Some of the most commonly used set of activation functions are the Binary, Sigmoidal (linear) and Tan hyperbolic sigmoidal (non-linear) activation functions. Now let us take a look at each of them, to certain detail:

**Implementation of An artificial neuron network (ANN) With Tensorflow:**

Dataset : The dataset to be audited was provided which consists of a wide variety of intrusions simulated in a military network environment. It created an environment to acquire raw TCP/IP dump data for a network by simulating a typical US Air Force LAN. The LAN was focused like a real environment and blasted with multiple attacks. A connection is a sequence of TCP packets starting and ending at some time duration between which data flows to and from a source IP address to a target IP address under some well-defined protocol. Also, each connection is labelled as either normal or as an attack with exactly one specific attack type. Each connection record consists of about 100 bytes. For each TCP/IP connection, 41 quantitative and qualitative features are obtained from normal and attack data (3 qualitative and 38 quantitative features) .The class variable has two categories: • Normal • Anomalous.

Goal : Given the (train.csv) dataset we have to find out for the information of a connection whether the connection was normal or anomalous.