Business Intelligence
Lab Report
On
Classification and Clustering

HÖGSKOLAN
DALARNA

By
Manthri Bala Kiran
H19baman@du.se

Task1: Descriptive Analysis, Unsupervised Learning – IKEA

Unsupervised Learning:

Unsupervised Machine Learning algorithms infer patterns from a dataset without reference to known or labeled outcomes. Unliked Supervised Machine Learning, unsupervised machine learning methods cannot be directly applied to a regression or a classification problem. The best time to use unsupervised machine learning is when you do not have data on desired outcomes [1]. Some applications of unsupervised machine learning techniques include:

Clustering:

Clusters are sets of data points that share similar attributes and clustering algorithms are the methods that group these data points into different cluster based on their similarities. One of example, Clustering algorithms used for disease classification in medical science [2].

Principal Component Analysis:

It is a technique for feature extraction- so it combines our important variables in a specific way, we can drop the "Least Important" variables while still retaining the most valuables parts of all the variables.

IKEA- Problem Dataset:

Using the clustering we would be finding a new location for IKEA to open a new store. We have been given txt file with the important features of the Municipalities in Sweden. We would be clustering our data and finding the best new location for IKEA store.

Dataset:

In the dataset we have 207 observation with 12 variables, and we would be considering important variables from them. There are 8 cities which have Border and 10 cities with Infrastructure.

Implications:

- We use PCA for feature extraction which would be selecting the important variables from the dataset.
- Now we are going to calculate a matrix that summarizes how our variables would be related to each other. And then we would be breaking the matrix into separate components.
- We would be using 95% of the features to fit into the PCA doing that we would be having 5 PCA components.
- Now, we would be using the PCA components and trying to find number of clusters by elbow method.
- From the elbow method we can conclude to use 5 clusters for the KMeans method.
- We can plot the points in each cluster and see the grouping of the clusters.
- We can also check with the count of municipalities which have been divided into each cluster.

Results:

Now we would be checking existing IKEA stores with the dataset with the new IKEA locations and we can see calculate the mean for all the columns in the dataset with respective to the clusters. Now, to make conclusions to choose the best municipality for which is good to have a new IKEA store. We can conclude the following:

- Cluster 1 has 32 observations. And we have 7 municipalities already have IKEA stores, while rest of them do not have IKEA. We can see that Huddinge Municipality have a highest sales index but do not have an IKEA store. We can choose Huddinge would be best location to open a new IKEA store.
- Cluster 2 has 157 observations. And we can see that it has only one IKEA store where the IKEA store has started which is Älmhult and it is the center for many locations. And it would be not a best option to open a new IKEA store here as all the municipalities have low sales index.
- Cluster 3 has only Stockholm municipality which have an IKEA store already.
- Cluster 4 has 9 observations. And we have 2 municipalities already have IKEA store. And we can see that there are no next best municipalities for having a new IKEA store because of the low sales index.
- Cluster 5 has 8 observations. And Haparanda already have an IKEA store and we see next best city to have an IKEA store would be Strömstad because it has a good relative sales index and it is sharing the border with Norway.

Task2: Predictive Analysis, Supervised Learning – Titanic

Supervised Learning:

Supervised Learning is where you have input variables(x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(x)$$

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data [3].

Supervised learning problems can be further grouped into regression and classification problems.

- Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “No disease”.
- Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Titanic – Problem Dataset:

The dataset has two different parts train and test. Where we would be training the machine learning algorithm with the train data set and making the algorithm learn to predict the passengers who survived or did not survive the Titanic disaster for the test dataset.

Dataset:

The train dataset has 891 observations and 12 variables, the important variables in the train data which would be necessary to train the Algorithm is Sex, Age, Survival, Cabin, Fare. The test dataset has 418 observations and 11 variables, the important variables in the test data which would be necessary to predict the survival of the passengers would be Sex, Age, Cabin, Fare.

Using the Pandas, we have read the datasets as train and test for analysis and predicting the survival.

Data Analyze:

Now, will try to understand and analyze the dataset to see how the variables would be affecting the survival of passenger.

- When we compare Pclass column with Survived, the “Pclass 1” which is upper class has the highest percentage of Survival with 62%.
- When we compare the Survival with the Sex and we see that female has the highest survival ratio of 74%.
- When we compare with relation of the passenger who having at least one Spouse or Sibling have survived with 53%.
- When we investigate the passengers, who have parent or child with a count of three have survived with the percentage of 60%.

- When we investigate Embarked with survival, people who have boarded at port of Cherbourg has the highest percentage of survival with 55%. And we can also say that the port of Cherbourg has the highest Male survival percentage compare to the rest ports.
- When we check with the age, we see that passengers who have the age less than 16 have the highest percentage of survival with 49%.
- The fare of the passengers would be affecting the survival. If the Fare is more than 31 currency people has the survival percentage of 58%.
- When we check with the titles impacting the survival of the passengers. The "Mrs" title has the highest survival percentage of 79%.

Data Pre-Processing:

Now let us try to clean our data and which would be helping us to find the survival passengers accurately.

- Firstly, I tried to extract the title from the names column and replace the similar title with appropriate titles and no frequent titles as the rare combinations of the titles. Next, we would be categorizing them with respective to the titles of the passengers and fill the null values with 0.
- Next let us try to change the Sex observations to numeric category(int) saying Female as "1" and Male as "0".
- Age of the passengers would be affecting with the survival of the passengers. So, let us try to fill the null values of the age column with the mean of the ages. We have calculated the mean of the ages with the title and let us try to fill the null values of the age column with the mean ages. And we can change the datatype from float to int.
- Now let us try to categorize the age column with replacing the values of ages accordingly with the values from 0 to 4.
- Let us replace the null values in the Embarked column with the frequent Embarked port. And let us categorize the Embarked column with integers ranging from 0 – 2.
- The null values in the Fare column can be replaced with the median values. And let's divided the fares in the fare column into four categories and replace them with the categorical values from 0-3 and convert the data types of the column into int.
- The cabin has many null values in the dataset but the cabin position of the passenger would be affecting much with the survival of the passenger. So, the null values in cabin can be replaced. All the variables in the column can be categorized from 1-8 and then converted to the datatype int.
- We can see the correlation of the data and we can see that the survival has the highest correlation with Sex, Fare and Title and negative correlation with Pclass and Deck.

Decision Tree:

Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including change event outcomes, resource costs, and utility. Tree bases learning algorithms are considered to be one of the best and mostly used supervised learning methods [4].

Using the above columns, we have the prediction accuracy of 91.58 accuracy for the decision tree.

References

[1] Datarobot, "Datarobot," Datarobot, [Online]. Available: https://www.datarobot.com/wiki/unsupervised-machine-learning/ [Accessed 17 05 2020].

[2] L. Pierson, "Dummies," Dummies, [Online]. Available: https://www.dummies.com/programming/big-data/data-science/the-importance-of-clustering-and-classification-in-data-science/ [Accessed 17 05 2020].

[3] J. Brownlee, "machinelearningmastery," machinelearningmastery, [Online]. Available: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ [Accessed 17 05 2020].

[4] R. S. Brid, "Medium," Medium, [Online]. Available: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb [Accessed 17 05 2020].