Chapter 8
Airbnb (Air Bed and Breakfast) Listing Analysis Through Machine Learning Techniques

Xiang Li
Cornell University, USA

Jingxi Liao
University of Central Florida, USA

Tianchuan Gao
Columbia University, USA

ABSTRACT

Machine learning is a broad field that contains multiple fields of discipline including mathematics, computer science, and data science. Some of the concepts, like deep neural networks, can be complicated and difficult to explain in several words. This chapter focuses on essential methods like classification from supervised learning, clustering, and dimensionality reduction that can be easily interpreted and explained in an acceptable way for beginners. In this chapter, data for Airbnb (Air Bed and Breakfast) listings in London are used as the source data to study the effect of each machine learning technique. By using the K-means clustering, principal component analysis (PCA), random forest, and other methods to help build classification models from the features, it is able to predict the classification results and provide some performance measurements to test the model.

INTRODUCTION

Nowadays, machine learning (ML) is well-known and can be used in solving different types of problems such as probability, convex analysis and approximation theory. It is a type of artificial intelligence (AI) and it mainly focuses on letting the computer learn by itself without the control from humans (Expert.ai Team, 2020). It may look difficult to some beginners, but the method we mentioned here is about classification from supervised learning, clustering, and dimensionality reduction which is easy to explain and

DOI: 10.4018/978-1-7998-8455-2.ch008

understand. Moreover, we want to show not only the effect of machine learning but also how close this technique can be applied to our daily life, so we use the dataset from Airbnb listings to do the analysis.

Airbnb which stands for Air Bed and Breakfast, a famous online marketplace for lodging, is often used by a large number of travelers and landlords. It provides a platform between tenants and renters and helps them match each other easily and conveniently. It was built in 2008 and started in San Francisco, California USA before spreading to all over the world (Bivens, 2019). Based on some statistics, the Airbnb covers 220 countries and regions with active listings, has nearly 500 million guests since its creation and was joined by 14,000 new hosts in each month of 2021 (Deane, 2021). In order to keep our data source comprehensive and multifarious, we select the Airbnb listing from London as a dataset which contains 76,619 numbers of listings information and over 8 features. Then, we use K-means clustering, hierarchical clustering, Principal Component Analysis, random forest to analyze the date we choose and we will use the decision tree to predict the data after the analysis process.

Firstly, We are going to introduce K-means clustering. K-means clustering is one of the unsupervised learning which is easy to explain. Cluster is a common type of data analysis and it is used to separate the original data to different subgroups or clusters, so the data with the same group will be very similar (Dabbura, 2018). Furthermore, K-mean, which is a kind of algorithm of the centroid-based and distance-based, is used to assign different data points to different clusters through the calculation of the distance from the point to the cluster centroid which is randomly selected in the beginning (Sharma, 2019). After that, we will reselect new cluster centroids and redo the assigned process again and again, so our goal in the K-mean cluster is to repeat the select and assign process and find suitable clusters with minimal distance from each data point to the cluster centroid (Sharma, 2019).

The second method we use is hierarchical clustering which is similar to the first method that is a type of unsupervised learning and is to cluster data points but with different standards. In hierarchical clustering, we initially consider each of the data points as different clusters and then find the closest two clusters and merge them together (Patlolla, 2018). Hierarchical clustering is similar to the K-mean cluster in that those processes will run cyclically but it is different that all the data points will be in a single cluster. Compare K-mean clustering with hierarchical clustering, we have the assumption that if the dataset has a large number of variables, it is better to use K-mean clustering and if we want the result explicable and structured, hierarchical clustering is more suitable (Das, 2020).

Moreover, we also mention Principal Component Analysis (PCA) during analyzing the dataset. Principal Component Analysis which is also called PCA is a method to reduce or refine the dimension of a dataset and the smaller dataset which we transferred from the original dataset still contains the important information (Jaadi, 2021). Therefore, our goal in the PCA is to make the dataset concise and effective.

The above method we have introduced is all used in the analysis process. Like we previously said, we will do the data prediction after the analysis process and the method we used is the decision tree. Nowadays, the decision tree usually appears in machine learning and it is a type of supervised machine learning. As its name decision tree, it is used to build a model like a tree and use the tree truck to present all the possible consequences in different kinds of data. Specifically, the tree is made of nodes, leaf and branches with respect to test, class label that is decision and conjunctions that are connected to class labels (Yadav, 2018).

Furthermore, the second method we used in the process of data prediction is random forest which is a type of supervised learning. Just like its named "forest", it is made of many decision trees and these trees will be merged to get an accurate predicted result (Donges, 2019). Compare the decision tree and the random forest, we can get that the decision tree can be explained easily but it is hard for us to pick