SONO YAZILIM TECHNICAL TASK

Name: Duygu Turan
E-mail: dygturan@hotmail.com
Subject: Prediction of a user's rating which is greater than three on a movie
Due Date: 28.08.2017
Language: Python

Contents

1 The Problem

In this task, MovieLens 100K Dataset has given. And when a user and a movie are specified, it is expected to predict the user's rating which is 3 or not. I used first 1000 data instead of all dataset because it takes a long time to get an accuracy. If you want to use all dataset, you should change some of variables.

2 The Method

For the problem, I used K-Nearest Neighbors Algorithm which classifies data according to its kth nearest neighbors' most frequent class.

Firstly, I loaded the dataset and determine k as 3 as a beginning. I concatenated a user, a movie and the user's rating on the movie data with the user's age, sex and occupation and the movie's genres. I repeated this operation for all ratings. Then, I divided data as test and train data.

Afterwards, I found distances between train data and test data by using the Euclidean distance as a distance metric. And I stored those distances and train data's ratings. Then, I sorted the distances by using numpy library's sort[1] function[2] and determined first kth distances' most frequent rating as first test data's rating. I stored first test data's this rating prediction and actual class in predictionsAndReality matrix. Finally, I implemented this operation for all test data and calculated accuracy for the given model by using predictionsAndReality matrix. I showed classification accuracy which is found by using K-Nearest Neighbors Algorithm.

Afterwards, I changed k values and observed results as below table:

k	Accuracy
3	0.47
5	0.50
10	0.525
30	0.485
50	0.49
100	0.445

When k increases, accuracy would be increased. Because if we know about the neighbors of a data more, we can predict its class better. As we can see above table, accuracy increases by k's increment until 10. But sometimes, it doesn't work. Because if we are looking at more neighbor of a data than its class size and the class is too small, we can predict wrong class. To avoid this situation, k value should be well chosen.

3 Discussion

K-Nearest Neighbors Algorithm does not work well enough for this problem. Better results can be obtained with more extensive studies. But factors affecting rating are not enough to predict a rating. It is a hard problem.

4 References

1. https://docs.scipy.org/doc/numpy-1.12.0/reference/generated/numpy.sort.html

2. http://stackoverflow.com/questions/2828059/sorting-arrays-in-numpy-by-column