Homework Assignment 2

Due: 10 26, 2020

1. (3 pt. each) True/false questions.  
(a) The meaning of *k* in *k*-means and *k*-nearest neighbors are identical. Circle one: True / False

False

Reason: k in k-means means number of cluster, k in knn means number of nearest neighbor.

(b) Distance metric functions can be shared between the *k*-means and *k*-nearest neighbors algorithms. Circle one: True / False

True

Reason: both k-means and knn have to get distance. The method of getting distance can be shared.

(c) *k*-nearest neighbors can only model linear decision boundaries. Circle one: True / False

False

Reason: It can model non linear decision boundaries. (circle shape)

(d) The *k*-nearest neighbors algorithm produces a decision model as a result of training. Circle one: True / False

False

Reason: it produce data to determine the neighbor of the test data

(e) For the *k*-nearest neighbors algorithm, the training data is required only when the model is being trained. After training, the algorithm does not need to keep the training data.  
Circle one: True / False

False

Reason: training data is used to find which data is neighbor of test data.

2.

(a) Submit your complete code in both your report and *Jupyter* notebook.

텍스트이(가) 표시된 사진

자동 생성된 설명

I used for loop to calculate all distance between vector a and matrix B

(b) In your own words, explain what code blocks (1), (2), and (3) do, respectively in your report.

Code block(1) : compute the Euclidean distance between 1 row of test data and every row of training data. Do this for all test data.

Code block(2) : for each test data, get the index of k neighbors. Select k neighbors in ascending order of distance.

Code block(3) : check neighbor’s label of test data, and set label with the most frequency as the attribute of the test data

(c) Write your script in both your report and *Jupyter* notebook. 텍스트이(가) 표시된 사진

자동 생성된 설명

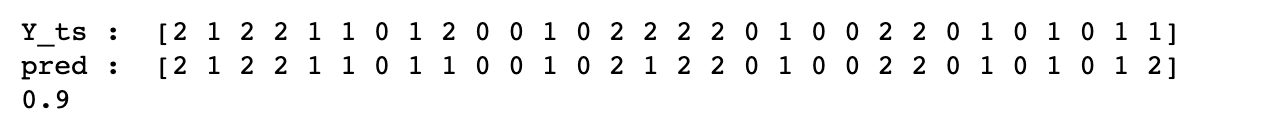
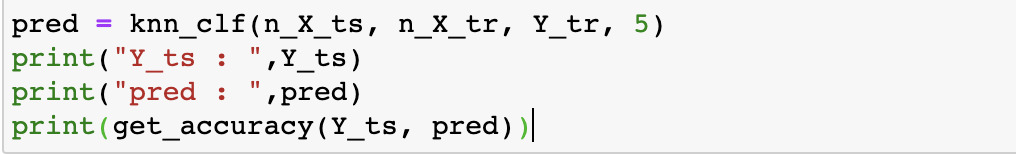
(d) getAccuracy : Submit your implementation in both your report and *Jupyter* notebook. 텍스트이(가) 표시된 사진

자동 생성된 설명

If prediction result is same with real label, add 1 to count variable. After iteration, divide with vector size

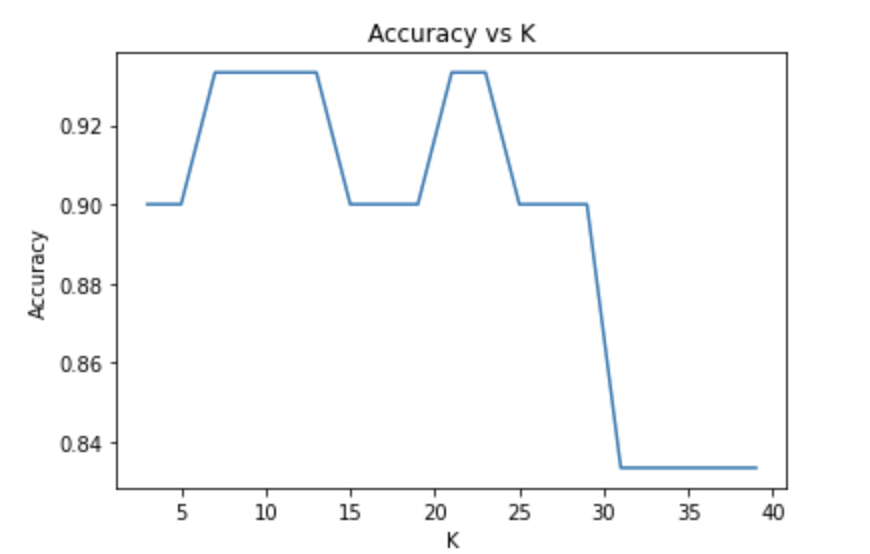
(e) Write a code snippet that performs classification on the datasets loaded in Problem 2 (c), using *k*NN with *k* = 5;

Submit your code and results in both your report and *Jupyter* notebook.



I used knn\_clf function, and get\_accuracy function.

(f) How does the accuracy change with respect to *k*? Submit your code and results in both your report and *Jupyter* notebook 텍스트이(가) 표시된 사진

자동 생성된 설명

I draw graph to show change. When K= 7, 9, 11, 13, 21, 23, the accuracy is high, but when K is bigger than 31, the accuracy is very low.

(g) now implement knn\_reg() that performs regression using the *k*nn approach. Submit your implementation in both your report and *Jupyter* notebook.텍스트이(가) 표시된 사진

자동 생성된 설명

I used mean information to perform regression. I calculate the average of k neighbor’s charge, and use to regression.

(h) Write your script in both your report and *Jupyter* notebook. 텍스트이(가) 표시된 사진

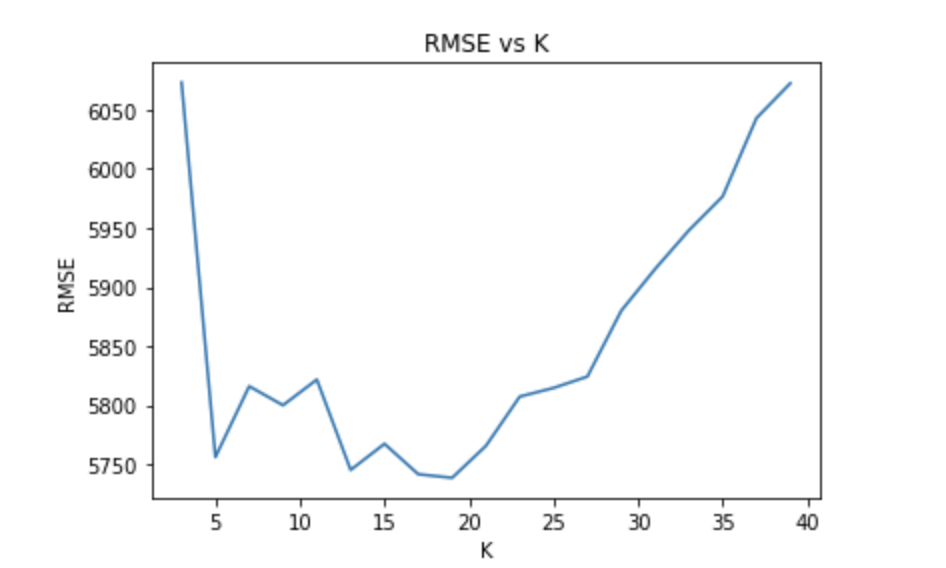
자동 생성된 설명

First, load the data.

Second, I calculate mu and sigma, and put this information to normalize function as parameter.

Third, excute knn\_reg and check error by using\_rmse function.

(i) How does the RMSE change with respect to *k*? Submit your code and results in both your report and *Jupyter* notebook. Include your short analysis in the report.



When k= 19, the RMSE is the lowest. The RMSE decreases until k is 19 and then the RMSE increases.