AI 프로그래밍 059분반

HW12 보고서

2020년 12월 06일

정보컴퓨터공학과

201824481 박지우

목차

1. Main.py – kNN

```
def kNN(self, query):
    distance = []
    for i in range(len(self._trainDy)):
        calc = 0
        for j in range(len(query)):
            calc += (self._trainDX[i][j] - query[j]) ** 2
        calc = calc ** 0.5
        distance.append([i, calc])
    distance.sort(key=lambda x: x[1])

    avg = 0
    for i in range(self._k):
        avg += self._trainDy[distance[i][0]]
    avg /= self._k
    return avg
```

kNN 함수는 query로 들어온 test 값 1개에 대하여 저장된 trainDX와의 Euclidean distance를 구한 후, query와 거리가 제일 짧은 trainDX의 trainDy k개의 평균을 구하여 반환해주는 함수이다. Distance는 각 trainDX에 따라 [index, 거리 값]을 저장하는 2차원 리스트이다. 거리 값은 1개의 trainDX의 i번째 값과 query의 i번째 값을 빼고 제곱한 것을 모두 더한 후 루트를 씌우는 방식으로 계산하였다. Distance에 정보를 입력한 후, 거리 값이 작은 것부터 순서대로 정렬한다. 처음부터 k개까지가 query와 제일 가까운 trainDx이므로, 그 값들의 평균을 구해 반환한다.

2. 출력 결과

2-1. Linear Data

2-1-1. Linear Regression

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\ve
Enter the file name of training data: lin_train.txt
Enter the file name of test data: lin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 1

RMSE:  0.3
```

RMSE는 0.3으로 매우 낮았다. 즉, 실제 값과 굉장히 유사했다. Linear Regression이 데이터 값들을 토대로 Linear한 함수 식을 만드는 구조이기 때문에 Linear Data를 학습하고 테스트했을 때 오차가 매우 낮음을 알 수 있었다.

2-1-2. k-NN

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\ven
Enter the file name of training data: lin_train.txt
Enter the file name of test data: lin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 3

RMSE:  0.98
```

<k = 3, RMSE = 0.98>

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\
Enter the file name of training data: lin_train.txt
Enter the file name of test data: lin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 5

RMSE:  0.91
```

<k = 5, RMSE = 0.91>

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\v
Enter the file name of training data: lin_train.txt
Enter the file name of test data: lin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 7

RMSE:  0.91
```

<k = 7, RMSE = 0.91>

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\
Enter the file name of training data: lin_train.txt
Enter the file name of test data: lin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 8

RMSE:  0.95
```

<k = 8, RMSE = 0.95>

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\
Enter the file name of training data: lin_train.txt
Enter the file name of test data: lin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 10

RMSE:  1.03
```

<k = 10, RMSE = 1.03>

k의 값이 5나 7일 때 RMSE가 0.91로 제일 낮았다. Linear Data이기 때문에 Data가 들쑥날쑥하지 않아 k값이 매우 커지지 않는 이상 RMSE는 큰 차이를 보이지 않았다. Linear Regression에 비해 Linear Data에 특화되어 있지 않아 Linear Regression보다 RMSE가 높게 나왔지만 그럼에도 비교적 낮게 나와 실제 값과 유사함을 알 수 있었다.

2-2. Nonlinear Data

2-2-1. Linear Regression

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\venv
Enter the file name of training data: nonlin_train.txt
Enter the file name of test data: nonlin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 1

RMSE:  82.33
```

Linear Data와 비교하였을 때 RMSE가 굉장히 크게 나왔다. Nonlinear Data를 Linear한 함수 식으로 계산하다 보니 당연히 추정 값이 실제 값과 큰 차이를 보일 수밖에 없었다.

2-2-2. k-NN

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\ve
Enter the file name of training data: nonlin_train.txt
Enter the file name of test data: nonlin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 2

RMSE:  24.42
```

<k = 2, RMSE = 24.42>

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\ver
Enter the file name of training data: nonlin_train.txt
Enter the file name of test data: nonlin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 3

RMSE:  21.61
```

<k = 3, RMSE = 21.61>

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\ven
Enter the file name of training data: nonlin_train.txt
Enter the file name of test data: nonlin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 4

RMSE:  22.72
```

<k = 4, RMSE = 22.72>

```
"C:\Users\Park Jiwoo\PycharmProjects\HW12_201824481\ven
Enter the file name of training data: nonlin_train.txt
Enter the file name of test data: nonlin_test.txt

Which learning algorithm do you want to use?
 1. Linear Regression
 2. k-NN
Enter the number: 2
Enter the value for k: 5

RMSE:  23.3
```

<k = 5, RMSE = 23.3>

k의 값이 3일 때 RMSE가 21.61로 나왔다. 정돈되어 있는 Linear Data와 달리 Nonlinear Data이기 때문에 앞보다 RMSE가 비교적 높게 나왔지만, Linear Regression에 비해 훨씬 낮은 RMSE가 나왔다. 현재 훈련 데이터의 개수는 500개로 훨씬 더 많은 훈련 데이터가

있다면 RMSE가 낮아질 것으로 보인다.

결론적으로, Linear Regression은 Linear Data에 한정적으로 적용할 수 있는 알고리즘이지만 k-NN은 데이터의 종류에 상관없이 매우 유용하게 적용할 수 있는 알고리즘이며 훈련 데이터의 개수가 많을수록 RMSE를 훨씬 더 낮출 수 있을 것이다.