CS 256 Assignment 1

Implementation of k-Nearest Neighbors (kNN) algorithm in Python

by Abhishek Manoj Sharma

Contents

Setting up Python environment:

The kNN algorithm is entirely built using in-built Python functions, however, the program uses 3rd-party libraries for displaying the graph and following are the steps to set that up.

Open Terminal and run the following 3 commands:

Number	Command	Description
1	sudo apt install python-pip	This command installs 'pip' which is a package management system for Python
2	pip install matplotlib	This command installs matplotlib
3	sudo apt-get install python-tk	This command installs the python-tk package

Setting up program components:

The assignment contains the following 4 Python files:

1. abhishek_sharma_knn.py
2. abhishek_sharma_knn_environment.py
3. abhishek_sharma_knn_agent.py
4. abhishek_sharma_knn_graph.py

Please ensure that all the above mentioned 4 files are placed together in the same directory.

Python File Name	Description
abhishek_sharma_knn.py	This is the file which must be executed to run the program. It accepts two command line parameters as inputs: 1. Directory name containing the training and testing files 2. Value of k
abhishek_sharma_knn_environment.py	This file contains the Environment class
abhishek_sharma_knn_agent.py	This file contains the Agent class
abhishek_sharma_knn_graph.py	This file contains the Graph class

How to run:

Kindly execute the program using the following steps:

1. Open terminal and navigate to the path containing the 4 Python (.py) files mentioned in the previous section.

2. Execute the following command:
python abhishek_sharma_knn.py <directory_name> <value_of_k>

<directory_name>: Specifies the name of directory containing the training and testing datasets.
<value_of_k>: Specifies the value of k

Note: For any value of k, the program measures the accuracy for all odd values of k from 1 to k.
For instance, if k=9, the program will display the accuracy for k=1,3,5,7,9.

Example command: **python abhishek_sharma_knn.py hepatitis 11**

Program in execution:

Following are the screenshots of the program running on few datasets with k=11.

```
C:\Users\abhis\PycharmProjects\CS256_HW1_KNN>python abhishek_sharma_knn.py banana 11
| File Name                | Accuracy (k=1)     | Accuracy (k=3)     | Accuracy (k=5)     | Accuracy (k=7)     | Accuracy (k=9)     | Accuracy (k=11)    |
| banana-10-10tst.dat      | 86.60% (459/530)   | 88.30% (468/530)   | 89.25% (473/530)   | 89.62% (475/530)   | 89.43% (474/530)   | 89.81% (476/530)   |
| banana-10-1tst.dat       | 86.04% (456/530)   | 86.60% (459/530)   | 87.55% (464/530)   | 87.55% (464/530)   | 87.92% (466/530)   | 87.92% (466/530)   |
| banana-10-2tst.dat       | 87.17% (462/530)   | 88.68% (470/530)   | 88.30% (468/530)   | 89.25% (473/530)   | 89.81% (476/530)   | 90.19% (478/530)   |
| banana-10-3tst.dat       | 86.60% (459/530)   | 87.36% (463/530)   | 87.55% (464/530)   | 88.11% (467/530)   | 89.06% (472/530)   | 88.68% (470/530)   |
| banana-10-4tst.dat       | 87.55% (464/530)   | 88.49% (469/530)   | 88.87% (471/530)   | 88.11% (467/530)   | 88.30% (468/530)   | 89.06% (472/530)   |
| banana-10-5tst.dat       | 88.68% (470/530)   | 89.06% (472/530)   | 90.00% (477/530)   | 89.62% (475/530)   | 89.81% (476/530)   | 89.81% (476/530)   |
| banana-10-6tst.dat       | 87.55% (464/530)   | 88.68% (470/530)   | 89.81% (476/530)   | 89.62% (475/530)   | 90.19% (478/530)   | 90.75% (481/530)   |
| banana-10-7tst.dat       | 87.36% (463/530)   | 88.49% (469/530)   | 88.87% (471/530)   | 89.81% (476/530)   | 90.38% (479/530)   | 90.57% (480/530)   |
| banana-10-8tst.dat       | 89.43% (474/530)   | 89.43% (474/530)   | 90.00% (477/530)   | 90.57% (480/530)   | 90.38% (479/530)   | 90.38% (479/530)   |
| banana-10-9tst.dat       | 88.87% (471/530)   | 89.81% (476/530)   | 90.19% (478/530)   | 90.75% (481/530)   | 91.13% (483/530)   | 91.51% (485/530)   |
| banana-5-1tst.dat        | 87.17% (924/1060)  | 88.21% (935/1060)  | 88.58% (939/1060)  | 89.25% (946/1060)  | 88.96% (943/1060)  | 89.43% (948/1060)  |
| banana-5-2tst.dat        | 85.28% (904/1060)  | 86.98% (922/1060)  | 87.74% (930/1060)  | 87.55% (928/1060)  | 88.49% (938/1060)  | 88.68% (940/1060)  |
| banana-5-3tst.dat        | 86.79% (920/1060)  | 88.49% (938/1060)  | 89.06% (944/1060)  | 89.81% (952/1060)  | 90.57% (960/1060)  | 90.57% (960/1060)  |
| banana-5-4tst.dat        | 86.98% (922/1060)  | 87.55% (928/1060)  | 87.55% (928/1060)  | 88.11% (934/1060)  | 88.49% (938/1060)  | 88.87% (942/1060)  |
| banana-5-5tst.dat        | 88.30% (936/1060)  | 90.00% (954/1060)  | 89.62% (950/1060)  | 90.38% (958/1060)  | 90.19% (956/1060)  | 90.57% (960/1060)  |
----------------------------------------
| Average Accuracies     |
----------------------------------------
K=1: 87.25%
K=3: 88.37%
K=5: 88.77%
K=7: 89.16%
K=9: 89.49%
K=11: 89.75%
```

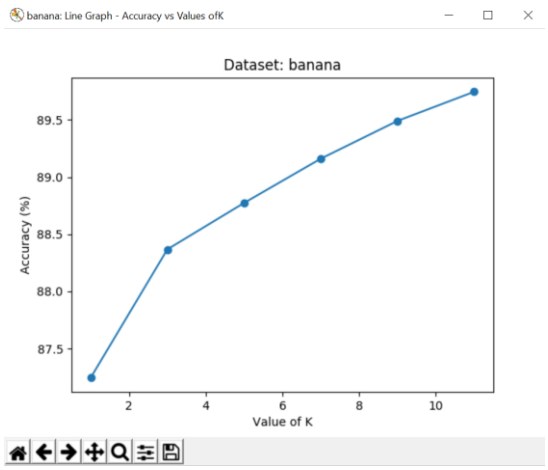
```
C:\Users\abhis\PycharmProjects\CS256_HW1_KNN>python abhishek_sharma_knn.py dermatology 11
| File Name                | Accuracy (k=1)     | Accuracy (k=3)     | Accuracy (k=5)     | Accuracy (k=7)     | Accuracy (k=9)     | Accuracy (k=11)    |
| dermatology-10-10tst.dat | 82.86% (029/035)   | 85.71% (030/035)   | 80.00% (028/035)   | 82.86% (029/035)   | 82.86% (029/035)   | 82.86% (029/035)   |
| dermatology-10-1tst.dat  | 88.57% (031/035)   | 88.57% (031/035)   | 88.57% (031/035)   | 88.57% (031/035)   | 88.57% (031/035)   | 88.57% (031/035)   |
| dermatology-10-2tst.dat  | 91.67% (033/036)   | 94.44% (034/036)   | 97.22% (035/036)   | 97.22% (035/036)   | 97.22% (035/036)   | 97.22% (035/036)   |
| dermatology-10-3tst.dat  | 81.08% (030/037)   | 83.78% (031/037)   | 78.38% (029/037)   | 78.38% (029/037)   | 72.97% (027/037)   | 70.27% (026/037)   |
| dermatology-10-4tst.dat  | 94.44% (034/036)   | 94.44% (034/036)   | 91.67% (033/036)   | 91.67% (033/036)   | 91.67% (033/036)   | 91.67% (033/036)   |
| dermatology-10-5tst.dat  | 94.59% (035/037)   | 97.30% (036/037)   | 97.30% (036/037)   | 97.30% (036/037)   | 97.30% (036/037)   | 97.30% (036/037)   |
| dermatology-10-6tst.dat  | 94.44% (034/036)   | 94.44% (034/036)   | 94.44% (034/036)   | 91.67% (033/036)   | 91.67% (033/036)   | 83.33% (030/036)   |
| dermatology-10-7tst.dat  | 88.57% (031/035)   | 88.57% (031/035)   | 91.43% (032/035)   | 91.43% (032/035)   | 91.43% (032/035)   | 94.29% (033/035)   |
| dermatology-10-8tst.dat  | 94.29% (033/035)   | 94.29% (033/035)   | 91.43% (032/035)   | 85.71% (030/035)   | 82.86% (029/035)   | 82.86% (029/035)   |
| dermatology-10-9tst.dat  | 88.89% (032/036)   | 91.67% (033/036)   | 88.89% (032/036)   | 86.11% (031/036)   | 86.11% (031/036)   | 86.11% (031/036)   |
| dermatology-5-1tst.dat   | 88.89% (064/072)   | 91.67% (066/072)   | 90.28% (065/072)   | 90.28% (065/072)   | 90.28% (065/072)   | 87.50% (063/072)   |
| dermatology-5-2tst.dat   | 92.96% (066/071)   | 92.96% (066/071)   | 91.55% (065/071)   | 91.55% (065/071)   | 88.73% (063/071)   | 87.32% (062/071)   |
| dermatology-5-3tst.dat   | 88.57% (062/070)   | 88.57% (062/070)   | 88.57% (062/070)   | 88.57% (062/070)   | 87.14% (061/070)   | 85.71% (060/070)   |
| dermatology-5-4tst.dat   | 91.55% (065/071)   | 94.37% (067/071)   | 92.96% (066/071)   | 94.37% (067/071)   | 88.73% (063/071)   | 87.32% (062/071)   |
| dermatology-5-5tst.dat   | 90.54% (067/074)   | 90.54% (067/074)   | 87.84% (065/074)   | 85.14% (063/074)   | 83.78% (062/074)   | 79.73% (059/074)   |
----------------------------------------
| Average Accuracies     |
----------------------------------------
K=1: 90.22%
K=3: 91.48%
K=5: 90.08%
K=7: 89.53%
K=9: 87.99%
K=11: 86.45%
```

```
C:\Users\abhis\PycharmProjects\CS256_HW1_KNN>python abhishek_sharma_knn.py sonar 11
| File Name                | Accuracy (k=1)     | Accuracy (k=3)     | Accuracy (k=5)     | Accuracy (k=7)     | Accuracy (k=9)     | Accuracy (k=11)    |
| sonar-10-10tst.dat       | 75.00% (015/020)   | 75.00% (015/020)   | 60.00% (012/020)   | 55.00% (011/020)   | 60.00% (012/020)   | 60.00% (012/020)   |
| sonar-10-1tst.dat        | 90.48% (019/021)   | 90.48% (019/021)   | 90.48% (019/021)   | 85.71% (018/021)   | 85.71% (018/021)   | 80.95% (017/021)   |
| sonar-10-2tst.dat        | 85.71% (018/021)   | 95.24% (020/021)   | 95.24% (020/021)   | 90.48% (019/021)   | 85.71% (018/021)   | 85.71% (018/021)   |
| sonar-10-3tst.dat        | 80.95% (017/021)   | 95.24% (020/021)   | 90.48% (019/021)   | 90.48% (019/021)   | 90.48% (019/021)   | 85.71% (018/021)   |
| sonar-10-4tst.dat        | 85.71% (018/021)   | 85.71% (018/021)   | 80.95% (017/021)   | 80.95% (017/021)   | 80.95% (017/021)   | 80.95% (017/021)   |
| sonar-10-5tst.dat        | 61.90% (013/021)   | 61.90% (013/021)   | 66.67% (014/021)   | 71.43% (015/021)   | 76.19% (016/021)   | 71.43% (015/021)   |
| sonar-10-6tst.dat        | 80.95% (017/021)   | 80.95% (017/021)   | 80.95% (017/021)   | 80.95% (017/021)   | 80.95% (017/021)   | 80.95% (017/021)   |
| sonar-10-7tst.dat        | 95.24% (020/021)   | 95.24% (020/021)   | 95.24% (020/021)   | 95.24% (020/021)   | 95.24% (020/021)   | 90.48% (019/021)   |
| sonar-10-8tst.dat        | 85.71% (018/021)   | 85.71% (018/021)   | 90.48% (019/021)   | 90.48% (019/021)   | 85.71% (018/021)   | 76.19% (016/021)   |
| sonar-10-9tst.dat        | 85.00% (017/020)   | 85.00% (017/020)   | 85.00% (017/020)   | 85.00% (017/020)   | 85.00% (017/020)   | 80.00% (016/020)   |
| sonar-5-1tst.dat         | 80.95% (034/042)   | 80.95% (034/042)   | 78.57% (033/042)   | 76.19% (032/042)   | 76.19% (032/042)   | 73.81% (031/042)   |
| sonar-5-2tst.dat         | 85.71% (036/042)   | 85.71% (036/042)   | 85.71% (036/042)   | 85.71% (036/042)   | 88.10% (037/042)   | 88.10% (037/042)   |
| sonar-5-3tst.dat         | 78.57% (033/042)   | 76.19% (032/042)   | 76.19% (032/042)   | 71.43% (030/042)   | 73.81% (031/042)   | 73.81% (031/042)   |
| sonar-5-4tst.dat         | 80.49% (033/041)   | 82.93% (034/041)   | 82.93% (034/041)   | 82.93% (034/041)   | 80.49% (033/041)   | 78.05% (032/041)   |
| sonar-5-5tst.dat         | 73.17% (030/041)   | 75.61% (031/041)   | 75.61% (031/041)   | 78.05% (032/041)   | 78.05% (032/041)   | 70.73% (029/041)   |
----------------------------------------
| Average Accuracies     |
----------------------------------------
K=1: 81.25%
K=3: 82.69%
K=5: 81.73%
K=7: 80.77%
K=9: 81.01%
K=11: 78.12%
```

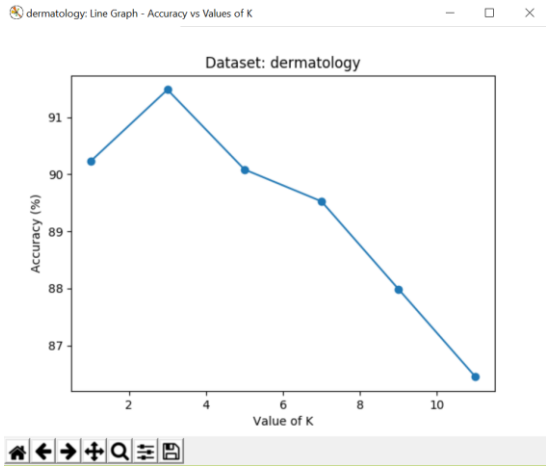
Analysis of datasets:

As visible from the two graphs below (Graph 1 and Graph 2), there is no certainty that increasing or decreasing the value of k will always increase the accuracy. Once the optimal value of k is reached, increasing or decreasing it further can lead to a reduction in accuracy.

Graph 1 is of the dataset *banana*. In this case, as k increases from 1 to 11, the accuracy also increases. However, in *dermatology* dataset (Graph 2), the accuracy increases from k=1 to k=3, but, starts decreasing when the value of k is increased further.

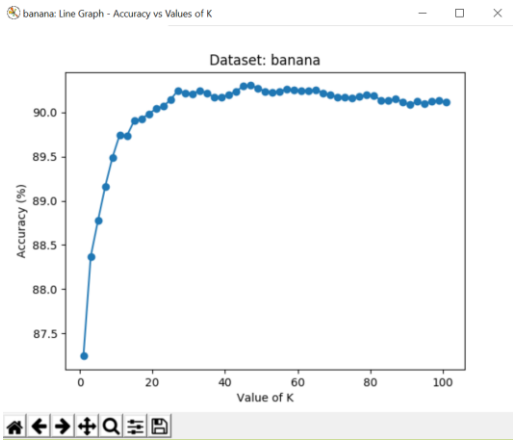


Graph 1



Graph 2

Graph 3 is for the same dataset *banana* shown in Graph 1, but this has k's value equivalent to all odd values from 1 to 101. Unlike the previous graph where k ranged from 1 to 11, this time the accuracy does not keep increasing but falls for the first time when k goes from 11 to 13. Accuracy peaks at k=47, and does not increase again for all odd values of k till 101. This shows that just because the accuracy of a dataset increased with the value of k (like in Graph 1), it may not increase further once the optimal k is reached (Graph 3).



Graph 3

Following are the graphical displays for some datasets generated by the program, showcasing how the accuracy depends on the value of k: