Introduction to Machine Learning Program assignment #2

Abstract

Implement Kd-Tree and use K-NN classifier to analyze a data set.

Problem(80%)

- Construct K-NN classifier with Kd-Tree and use Kd-Tree to find the designated nearest instance.
- Distance metric: Euclidean distance

Data set

'train.csv' is an ecoli data with 300 instances and 9 attributes without column 0 (name of ecoli). Column 10 is the class of ecoli. There are 8 classes: cp, im pp, imU, om, omL, inL, imS.

Attibute information:

1. mcg: McGeoch's method for signal sequence recognition.
2. gvh: von Heijne's method for signal sequence recognition.
3. lip: von Heijne's Signal Peptidase II consensus sequence score. Binary attribute.
5. chg: Presence of charge on N-terminus of predicted lipoproteins. Binary attribute.
6. aac: score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.
7. alm1: score of the ALOM membrane spanning region prediction program.
8. alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence.

'test.csv' is the testing data that TAs will test on your model. The format is same with 'train.csv' and it has 36 instances. We won't provide this file. You can generate your own test.csv from train.csv.

index	0	1	2	3	4	5	6	7	8	9	10
0	AMY2_EC	0.21	0.34	0.48	0.5	0.51	0.28	0.39	0.97	0.2	cp
1	PHOH_EC	0.29	0.47	0.48	0.5	0.41	0.23	0.34	0.63	0.12	cp
2	PTHP_EC(	0.3	0.45	0.48	0.5	0.36	0.21	0.32	0.61	0.26	cp
3	DEOC_EC	0.34	0.33	0.48	0.5	0.38	0.35	0.44	0.81	0.42	cp
4	PTCA_EC(	0.5	0.51	0.48	0.5	0.27	0.23	0.34	0.75	0.27	cp
5	CSPA_EC(	0.31	0.23	0.48	0.5	0.73	0.05	0.14	0.77	0.21	cp
6	NLPB_EC(	0.66	0.49	1	0.5	0.54	0.56	0.36	0.5	0.39	omL
7	PTHA_EC(	0.41	0.43	0.48	0.5	0.5	0.24	0.25	0.46	0.05	cp
8	PSTC_EC(	0.58	0.34	0.48	0.5	0.56	0.87	0.81	0.05	0.45	im
9	ECOT_EC(	0.67	0.81	0.48	0.5	0.25	0.42	0.25	0.49	0.52	pp
10	DSBA_EC(	0.71	0.71	0.48	0.5	0.4	0.54	0.39	0.47	0.61	pp
11	EMRB_EC	0.71	0.52	0.48	0.5	0.64	1	0.99	0.55	0.65	im
12	CAFA_EC(	0.39	0.32	0.48	0.5	0.46	0.24	0.35	0.08	0.52	cp
13	GALP_EC(	0.63	0.5	0.48	0.5	0.59	0.85	0.86	0.16	0.46	im
14	DGAL_EC	0.63	1	0.48	0.5	0.35	0.51	0.49	0.68	0.37	pp
15	THGA_EC	0.27	0.3	0.48	0.5	0.71	0.28	0.39	0.7	0.46	cp
16	TORA_EC	0.43	0.59	0.48	0.5	0.52	0.49	0.56	0.88	0.14	pp
17	SYM_ECO	0.61	0.36	0.48	0.5	0.49	0.35	0.44	0.74	0.07	cp
18	SYGA_EC	0.51	0.49	0.48	0.5	0.53	0.14	0.26	0.29	0.15	cp

Notice: You need to read the training and testing data as a csv file with command line.

Execution

```
$ unzip 0416001.zip && chmod +x 0416001/run.sh && cd 0416001/ && ./run.sh ../train.csv ../test.csv > output.txt
Archive:  0416001.zip
  inflating: 0416001/0416001.py
  inflating: 0416001/knn.py
 extracting: 0416001/run.sh
```

Output format as above. Please output the k nearest neighbors' index (sort from nearest to farthest) of first three testing data and accuracy of the testing data as a .txt file. **(k = 1, 5, 10, 100)**, see the output.txt example on e3.

Submit your code and report **(20%)**.

The report should include the K you use and explained why, language and explanation of your code.

Bonus (20%)

Implement PCA in this data set (Can use library to calculate eigenvalue and eigenvector) and also explain your idea and the explain the result(whether better or worse) in report.

Environment

Your program will be executed on the following environment:

- Ubuntu 16.04.3 LTS
- gcc 5.4.0

- openjdk 1.8.0_131
- python 2.7.12
- python 3.5.2