_k_ nearest neighbors

1) Given a dataset as follows:

X1	X2	Class
0.376000	0.488000	0
0.312000	0.544000	0
0.298000	0.624000	0
0.394000	0.600000	0
0.506000	0.512000	0
0.488000	0.334000	1
0.478000	0.398000	1
0.606000	0.366000	1
0.428000	0.294000	1
0.542000	0.252000	1

- Classifying the testset with *1*NN, *3*NN:

X1	X2	Class
0.550000	0.364000	?
0.558000	0.470000	?
0.456000	0.450000	?

0.450000	0.570000	?

2) Implement *k*NN from scratch in Python. The program requires 3 parameters:

- file name of trainset

- file name of testset

- number of nearest neighbors (k)

Dataset with *m* examples, *n* dimensions (attribute), *c* classes (*0, 1, …, c-1*), is in the format:

$val\_i_1\_a_1$ $val\_i_1\_a_2$ … $val\_i_1\_a_n$ $class\_i_1$

$val\_i_2\_a_1$ $val\_i_2\_a_2$ … $val\_i_2\_a_n$ $class\_i_2$

…

$val\_i_m\_a_1$ $val\_i_m\_a_2$ … $val\_i_m\_a_n$ $class\_i_m$

The program reports the classification results (accuracy, confusion matrix) with different trials k=1, 3, etc for 5 datasets:

- Iris (**.trn**: trainset, **.tst**: testset)

- Optics (**.trn**: trainset, **.tst**: testset)

- Letter (**.trn**: trainset, **.tst**: testset)

- Leukemia (**.trn**: trainset, **.tst**: testset)

- Fp (**.trn**: trainset, **.tst**: testset)

http://www.cit.ctu.edu.vn/~dtnghi/ml/data.tar.gz

3) Proof of Cover-Hart's theorem:

For sufficiently large training set size *m*, the error rate of the *1*NN classifier is less than twice the Bayes error rate.