CS760_Fall_2016_hw2_part2&3

author: Ying Li
li528@wisc.edu

Part2:

For the yeast data set, draw a plot showing how test-set accuracy varies as a function of k. Your plot should show accuracy for k = 1, 5, 10, 20, 30.

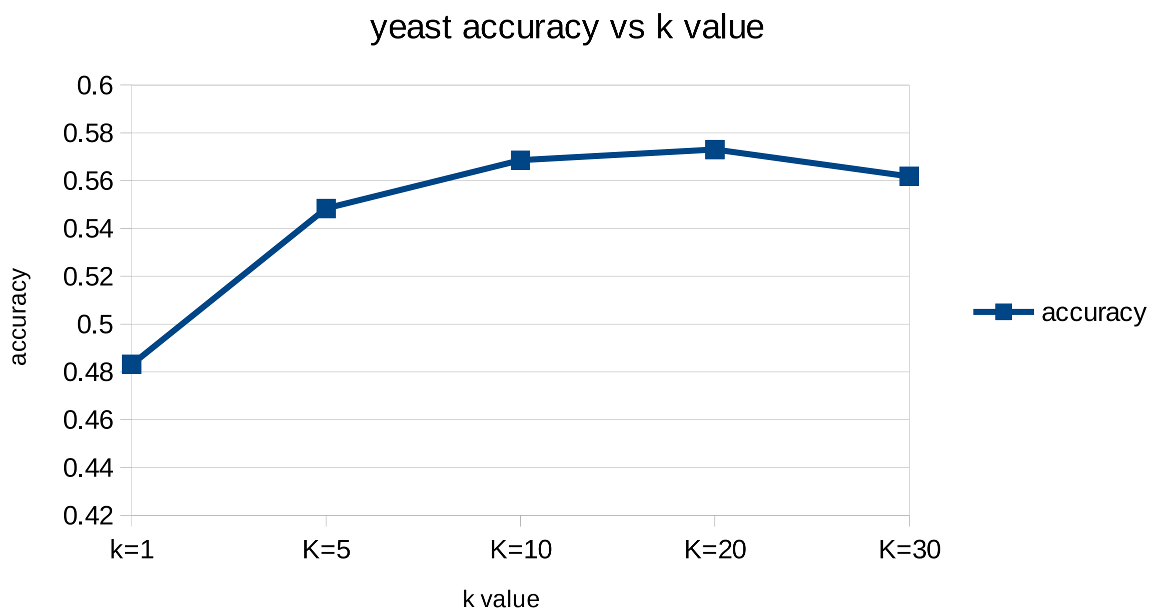


Figure 1: Yeast data-set accuracy vs. k value

For the wine data, draw a similar plot showing test-set mean absolute error as a function of k, for k = 1, 2, 3, 5, 10.

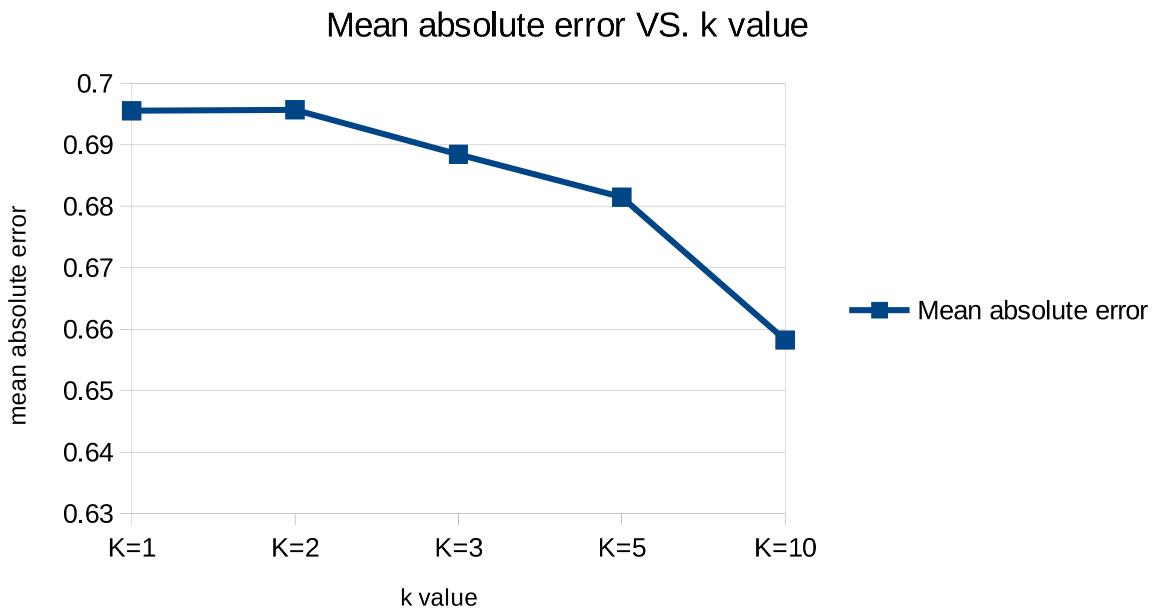


Figure 2: Mean absolute error VS. k value

CS760_Fall_2016_hw2_part2&3

author: Ying Li
li528@wisc.edu

For the yeast data set, construct confusion matrices for the k = 1 and k = 30 test-set results. Show these confusion matrices and briefly discuss what the matrices tell you about the effect of k on the misclassifications.

K=1											
actual	CYT	69	39	21	4	1	0	1	2	1	0
	NUC	46	57	13	6	0	0	0	1	0	0
	MIT	26	12	34	2	3	0	2	0	2	0
	ME3	3	9	2	32	1	0	0	0	0	0
	ME2	3	0	2	3	6	1	2	0	1	0
	ME1	0	0	1	0	3	6	3	0	1	0
	EXC	0	0	1	0	1	3	5	0	0	0
	VAC	3	0	1	0	0	1	0	1	0	0
	POX	1	1	1	0	0	0	0	0	5	0
	ERL	0	0	0	0	0	0	0	0	0	0
		CYT	NUC	MIT	ME3	ME2	ME1	EXC	VAC	POX	ERL
		pridicted									
Number of correctly classified instances : 215											
Total number of instances : 445											

Table 3: confusion matrix for K=1 for yeast data-set

K=30											
actual	CYT	89	36	11	1	0	1	0	0	0	0
	NUC	48	59	13	3	0	0	0	0	0	0
	MIT	25	5	45	1	1	3	1	0	0	0
	ME3	3	2	2	40	0	0	0	0	0	0
	ME2	4	2	3	3	3	3	0	0	0	0
	ME1	0	0	1	0	1	9	3	0	0	0
	EXC	0	0	1	0	0	4	5	0	0	0
	VAC	2	3	1	0	0	0	0	0	0	0
	POX	6	0	2	0	0	0	0	0	0	0
	ERL	0	0	0	0	0	0	0	0	0	0
		CYT	NUC	MIT	ME3	ME2	ME1	EXC	VAC	POX	ERL
		pridicted									
Number of correctly classified instances : 250											
Total number of instances : 445											

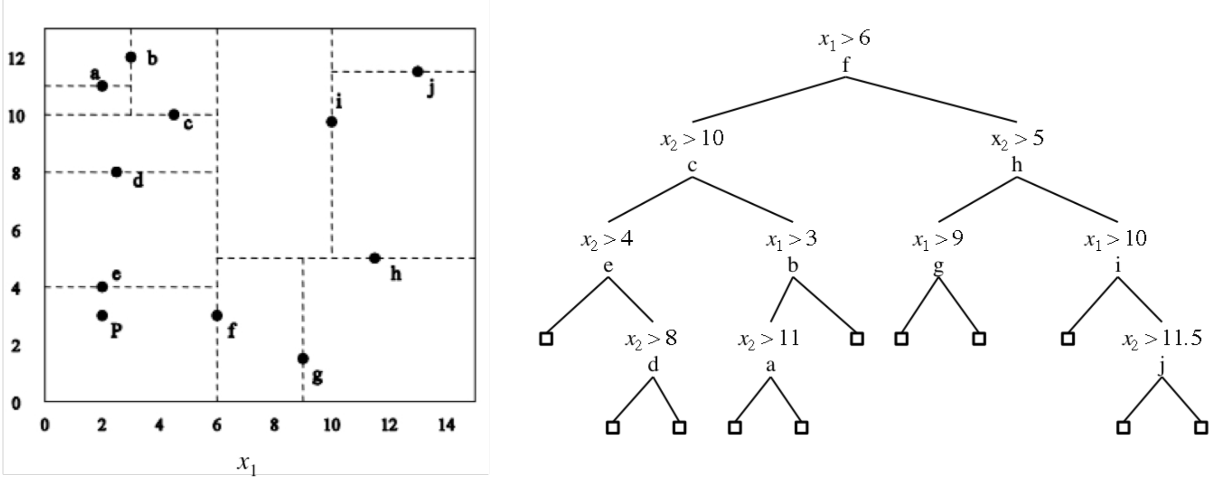
Table 4: confusion matrix for k=30 for yeast data-set

Based on the observation of comparing table 3 and table 4, our confusion matrix for k=1 and k=30 showed that with the increase of k, the more samples moved to the left and diagnal for this specific case. This is very obviously to think about, because with the increase of k, the majority vote will give us the result which could be favored to classes with more samples in the training set. With the k increase, the majority vote will play a more significant role during classification which could be good or not good for classification based on the specific situation.

CS760_Fall_2016_hw2_part2&3

author: Ying Li
li528@wisc.edu

Part3:



Using the k-d tree and the training set displayed in the figure below, show how the nearest neighbor for $\mathbf{x}^{(q)} = (7, 10)$ would be found. Assume that, for all instances in the figure, the features (i.e. coordinates) have integer values. For each step in the search, show the distance to the current node, the best distance found so far, the best node found so far, and the contents of the priority queue. You should use Euclidean distance and assume that the coordinates of the instances in the figure below are those shown in the table below.

Instance	***x1***	***x2***
a	2	11
b	3	12
c	5	10
d	2	8
e	2	4
f	6	3
g	9	2
h	12	5
i	10	10
j	13	11.5

Distance	Best Distance	Best Node	Priority queue
	∞		(f,0)
7.0710678119	7.0710678119	f	(h,0), (c,1)
7.0710678119	7.0710678119	f	(i,0), (c,1), (g,5)
3	3	I	(c,1), (j,3), (g,5)
2	2	c	(e,0), (b,0), (j,3), (g,5)
7.8102496759	2	c	(d,0), (b,0), (j,3), (g,5)
5.3851648071	2	c	(b,0), (j,3), (g,5)
4.472135955	2	c	(j,3), (a,4), (g,5)
return C			