##############KNN###################################

#k=1

library(class)

knn.pred=knn(traindata, testdata, usedata[(1:(n/2)),m],k = 1)

table(knn.pred ,salary[15082:30162])

table(knn.pred ,salary[15082:30162])

knn.pred <=50K >50K

1 9059 2136

2 2242 1644

knn.misrate.test

[1] 0.1451495

table(knn.pred ,salary[1:15081])

knn.pred <=50K >50K

1 8391 2804

2 2962 924

> knn.misrate.train=(2804+2962)/n

> knn.misrate.train

[1] 0.1911677

So for k=1, 14.5% of the observations in the test set are incorrectly predicted. Of course, it may be that K = 1 results in an overly flexible fit to the data. Below, we repeat the analysis using K = 3.

#k=3

knn.pred=knn(traindata, testdata, usedata[(1:(n/2)),m],k = 3)

table(knn.pred ,salary[15082:30162])

knn.pred <=50K >50K

1 9961 2482

2 1340 1298

> table(knn.pred ,salary[1:15081])

knn.pred <=50K >50K

1 9344 3099

2 2009 629

> knn.misrate.test=(2482+1340)/n

> knn.misrate.test

[1] 0.1267157

> knn.misrate.train=(3099+2009)/n

> knn.misrate.train

[1] 0.1693522

The results have improved and the mis-classsified rate of train set and test set have decreased both. We will try K=5 next step.

#k=5

knn.pred=knn(traindata, testdata, usedata[(1:(n/2)),m],k = 5)

> table(knn.pred ,salary[15082:30162])

knn.pred <=50K >50K

1 10403 2673

2 898 1107

> table(knn.pred ,salary[1:15081])

knn.pred <=50K >50K

1 9844 3232

2 1509 496

> knn.misrate.test=(2673+898)/n

> knn.misrate.test

[1] 0.118394

> knn.misrate.train=(3232+1509)/n

> knn.misrate.train

[1] 0.1571845

The result of k=5 even becomes better. But does it mean the misrate keeps on decreasing with increasing K further? We will try K=10.

#k=10

> knn.pred=knn(traindata, testdata, usedata[(1:(n/2)),m],k = 10)

> table(knn.pred ,salary[15082:30162])

knn.pred <=50K >50K

1 10839 2878

2 462 902

> table(knn.pred ,salary[1:15081])

knn.pred <=50K >50K

1 10339 3378

2 1014 350

> knn.misrate.test=(2878+462)/n

> knn.misrate.test

[1] 0.1107354

> knn.misrate.train=(3378+1014)/n

> knn.misrate.train

[1] 0.1456137

The reason why results get better with the increase of K is that there only two outcomes of K. (whether or not “<=50K)