43.

a). Bayes optimal classifier.

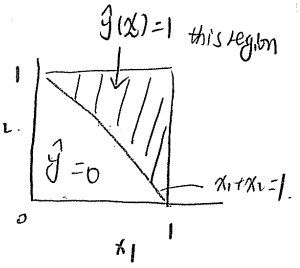
$$\hat{g}(x) = \mathbb{I}(g(1)f(x|1) > g(0)f(x|0))$$

where $f(x|1) = x_1 + x_2$ $\quad f(x|0) = 1$

$g(1)f(x|1) > g(0)f(x|0)$

$\Leftrightarrow 0.5 \cdot (x_1 + x_2) > 0.5 \cdot 1$

$\Leftrightarrow x_1 + x_2 > 1$. i.e. $\hat{g}(x) = \begin{cases} 1 & x_1 + x_2 > 1 \\ 0 & \text{o.w.} \end{cases}$



Error rate would be the expectation of loss function.

theoretical error rate $= P(y=1)P(\hat{g}=0|y=1) + P(y=0)P(\hat{g}=1|y=0)$

$$= 0.5 \cdot \iint_{\{x:\hat{g}(x)=0\}} f(x|1)\,dx_1 dx_2 + 0.5 \iint_{\{x:\hat{g}(x)=1\}} f(x|0)\,dx_1 dx_2$$

$$= 0.5 \cdot \int_0^1 \int_0^{1-x_1} (x_1 + x_2)\,dx_2\,dx_1 + 0.5 \int_0^1 \int_{1-x_2}^1 1 \cdot dx_1\,dx_2$$

$$= 0.5 \cdot \int_0^1 \left(x_1 x_2 + \frac{x_2^2}{2}\right)\Big|_{x_2=0}^{1-x_1} dx_1 + 0.5 \cdot 0.5$$

$$= 0.5 \cdot \int_0^1 \left(x_1(1-x_1) + \frac{(1-x_1)^2}{2}\right) dx_1 + \frac{1}{4}$$

$$= \frac{1}{2} \cdot \int_0^1 \left(\frac{1}{2} - \frac{x_1^2}{2}\right) dx_1 + \frac{1}{4}$$

$$= \frac{1}{4} \cdot \frac{2}{3} + \frac{1}{4}$$

$$\approx 0.41667$$

b) In the R code, changed *N<- 400* into *N<-100*. Fig1 is the 5 nearest neighbor classifier classifies plot.

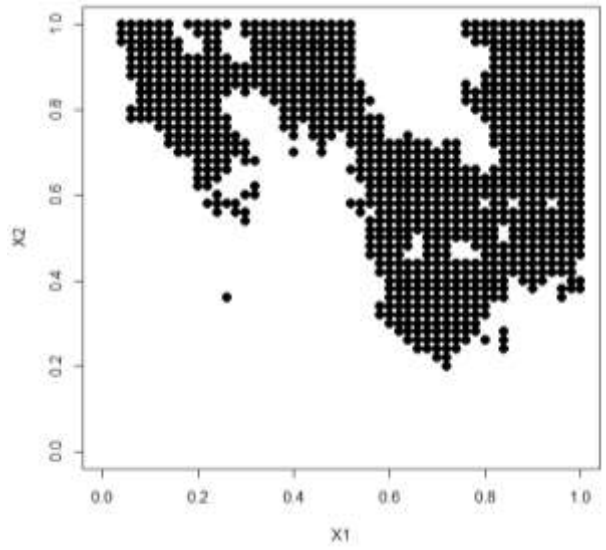


Fig1: Plot of 5 nearest neighbor classifier

The error rate of 5 nearest neighbor classifier classifies is

$\frac{1}{N}\{\#\ of\ points\ with\ classified\ value = 1\ using\ 5\ nearest\ neighbor\ classifier\ method\ \} = \frac{960}{N} = 0.3691$

In the R code, change *K<-10* into *K<-5.* The K = 5 fold cross validation error rate for 5nn classifier is found to be 0.4.

Fig 2 shows the plot of 9 nearest neighbor classifier by changing *k=5* into *k<-9* in the R code.

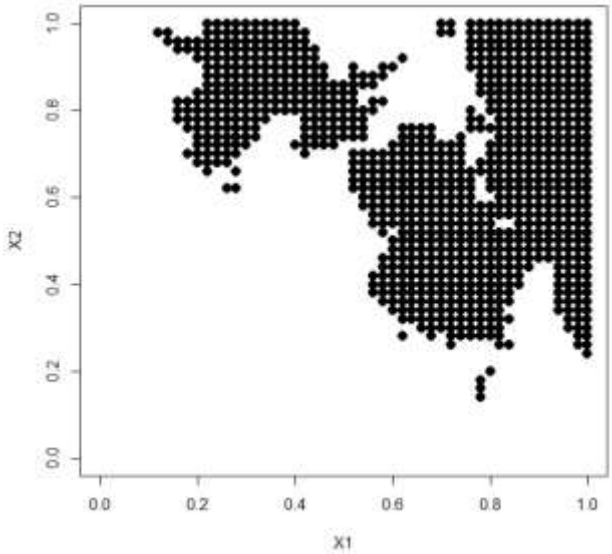


Fig2. 9 nearest neighbor classifier

c) training set of size N = 400. Fig3 is the 5 nearest neighbor classifier classifies plot.

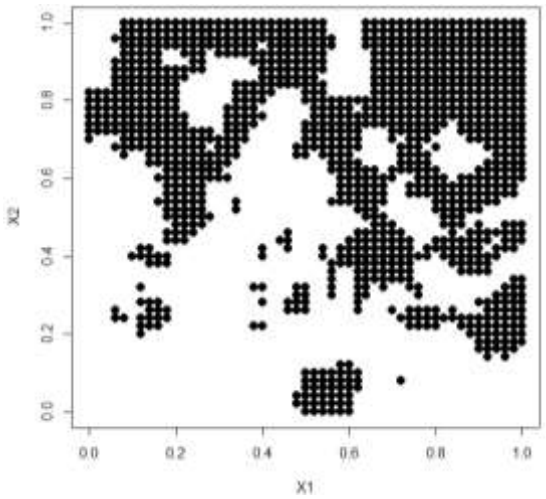


Fig3: 5 nearest neighbor classifier with training set of size N=400

The K = 5 fold cross validation error rate for this classifier is found to be 0.44

Fig4 is the 9 nearest neighbor classifier classifies plot

The error rate of 9 nearest neighbor classifier classifies is 0.4175

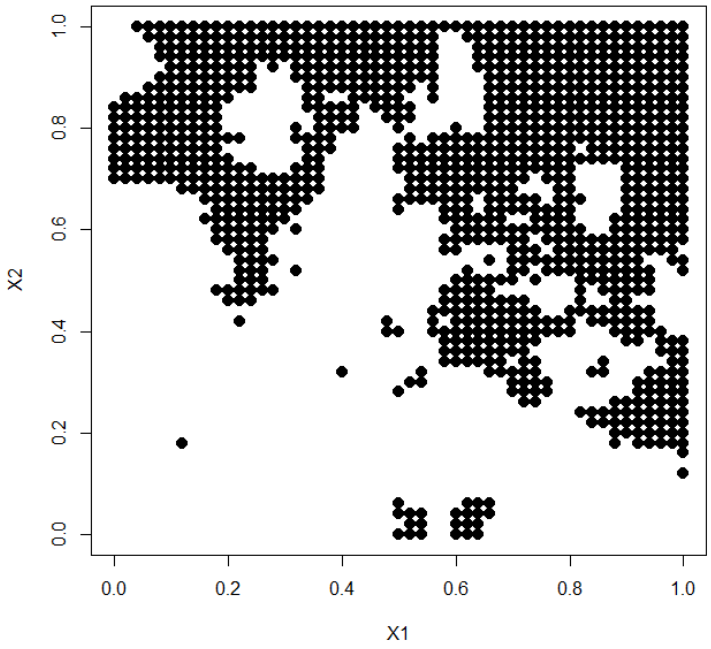


Fig4: 9 nearest neighbor classifier with training set of size N=400

d) Fig5 plot the change of error rate for the K = 10 cross-validation error rate, when k changes from 1 to 21.

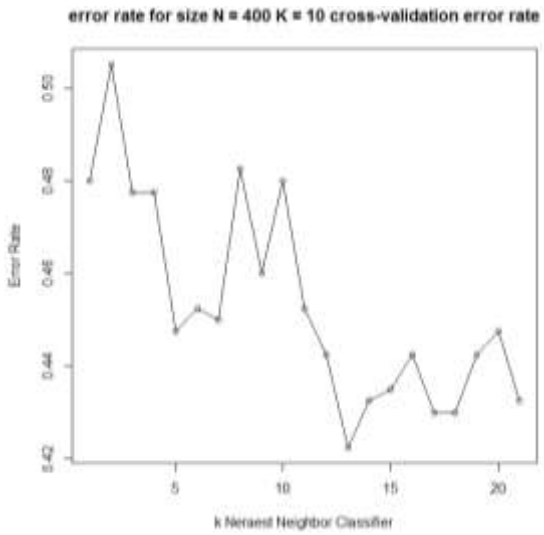


Fig 5: Error rate plot for cross-validation

```
> summary(tune.outCV)

Parameter tuning of 'knn.wrapper':

- sampling method: 10-fold cross validation

- best parameters:
  k
 14

- best performance: 0.4075 this is corresponding error rate

- Detailed performance results:
    k  error dispersion
1   1 0.4625 0.08270597
2   2 0.4525 0.05329426
3   3 0.4900 0.05797509
4   4 0.4475 0.08032054
5   5 0.4400 0.07655789
6   6 0.4525 0.07115125
7   7 0.4500 0.06009252
8   8 0.4425 0.05143766
9   9 0.4300 0.06749486
10 10 0.4400 0.06992059
11 11 0.4300 0.06213784
12 12 0.4100 0.04440971
13 13 0.4200 0.07527727
14 14 0.4075 0.07550754
15 15 0.4150 0.07564537
16 16 0.4325 0.06129392
17 17 0.4100 0.07472171
18 18 0.4225 0.08032054
19 19 0.4200 0.06213784
20 20 0.4300 0.05374838
21 21 0.4175 0.06566963
```

e) bootstrap estimate of error rate method to find the k

```
> summary(tune.outB)

Parameter tuning of ‘knn.wrapper’:

- sampling method: bootstrapping

- best parameters:
  k
 18

- best performance: 0.4325373

- Detailed performance results:
    k     error dispersion
1   1 0.4833547 0.03324709
2   2 0.4873835 0.03586695
3   3 0.4794762 0.03563394
4   4 0.4709533 0.03823128
5   5 0.4630987 0.03806964
6   6 0.4566878 0.03648453
7   7 0.4513516 0.03670744
8   8 0.4475896 0.03732373
9   9 0.4442186 0.03529113
10 10 0.4417733 0.03618796
11 11 0.4379539 0.03485343
12 12 0.4367804 0.03576552
13 13 0.4350803 0.03638944
14 14 0.4338508 0.03565449
15 15 0.4331470 0.03538290
16 16 0.4331827 0.03551297
17 17 0.4326344 0.03506605
18 18 0.4325373 0.03563902
19 19 0.4326702 0.03499044
20 20 0.4328132 0.03440708
21 21 0.4326057 0.03476990
```

44. a) Bayes optimal classifier error rate is

44 Bayes classifies.

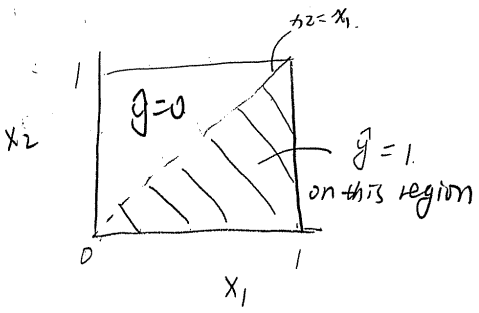
$g(1) f(\mathbf{x}|1) > g(0) f(\mathbf{x}|0)$

$\Leftrightarrow 0.5 \cdot (x_1 - x_2 + 1) > 0.5 \cdot 1$

$\Leftrightarrow x_1 - x_2 + 1 > 1.$

$\Leftrightarrow x_1 - x_2 > 0$

i.e. $\hat{g}(\mathbf{x}) = \begin{cases} 1 & x_1 - x_2 > 0 \\ 0 & \text{o.w.} \end{cases}$



theoretical error rate.

$P(y=1) P(\hat{g}=0 \mid y=1) + P(y=0) P(\hat{g}=1 \mid y=0)$

$= 0.5 \cdot \iint_{\{\mathbf{x}:\, \hat{g}(\mathbf{x})=0\}} f(\mathbf{x}|1)\, dx_1 dx_2 + 0.5 \iint_{\{\mathbf{x}:\, \hat{g}(\mathbf{x})=1\}} f(\mathbf{x}|0)\, dx_1 dx_2$

$= 0.5 \cdot \int_0^1 \int_0^{x_2} (x_1 - x_2 + 1)\, dx_1\, dx_2 + 0.5 \cdot 0.5$

$= \frac{1}{2} \cdot \int_0^1 \left(x_2 - \frac{x_2^2}{2}\right) dx_2 + \frac{1}{4}$

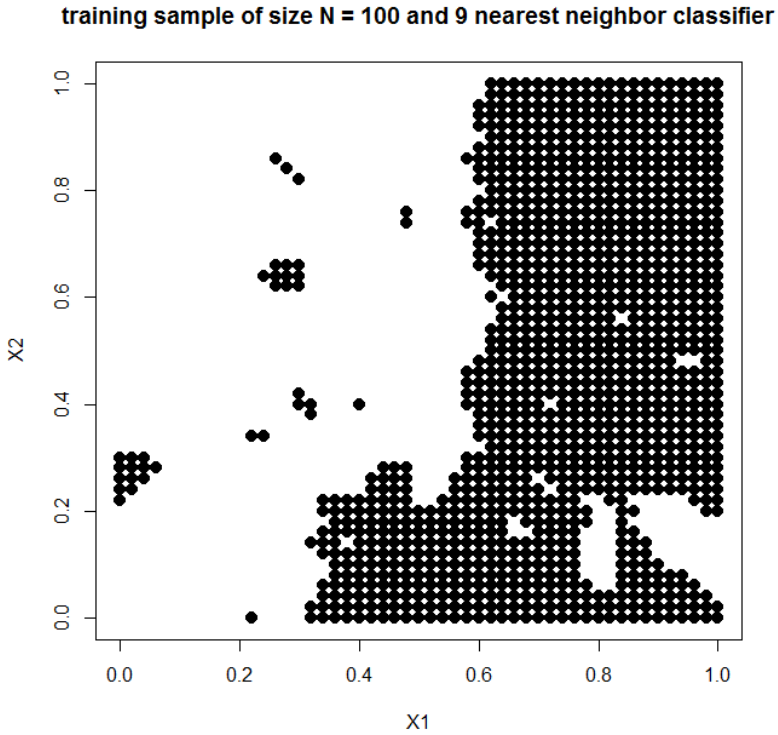
$= \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{4}$

$\approx 0.4167.$

b)

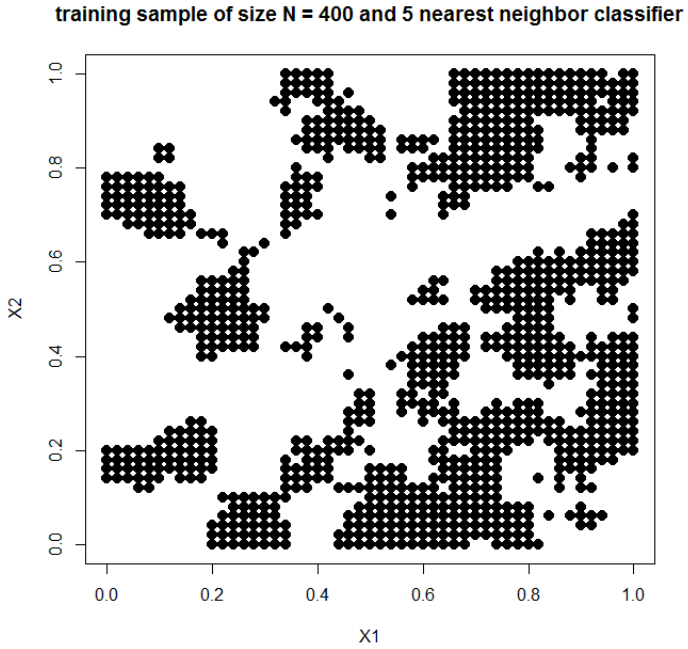
Fig 6. Training sample of N = 100, k =5, Error rate = 0.4602076

Error rate for N = 100, k =5, K =5 is 0.51

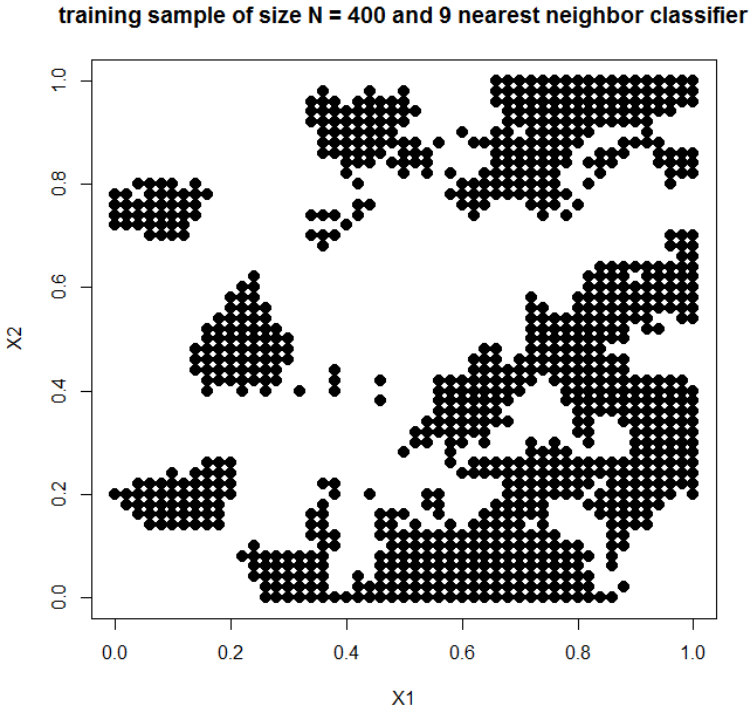


Error rate for K = 5 is 0.45

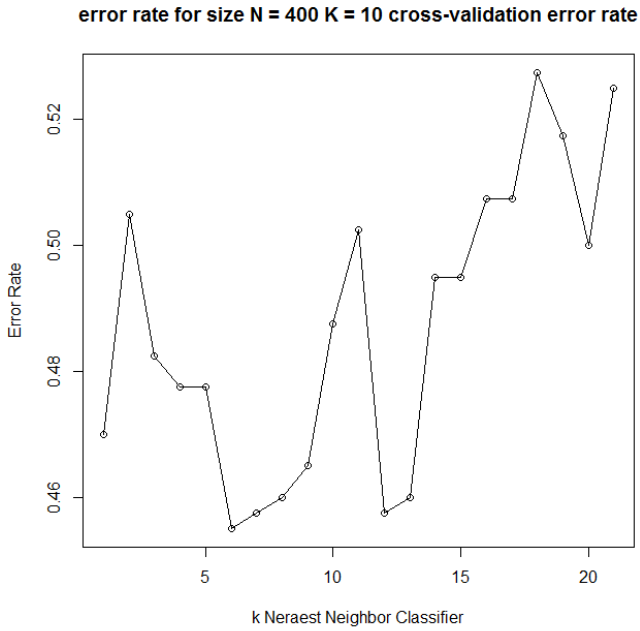
c) error rate is 0.4018



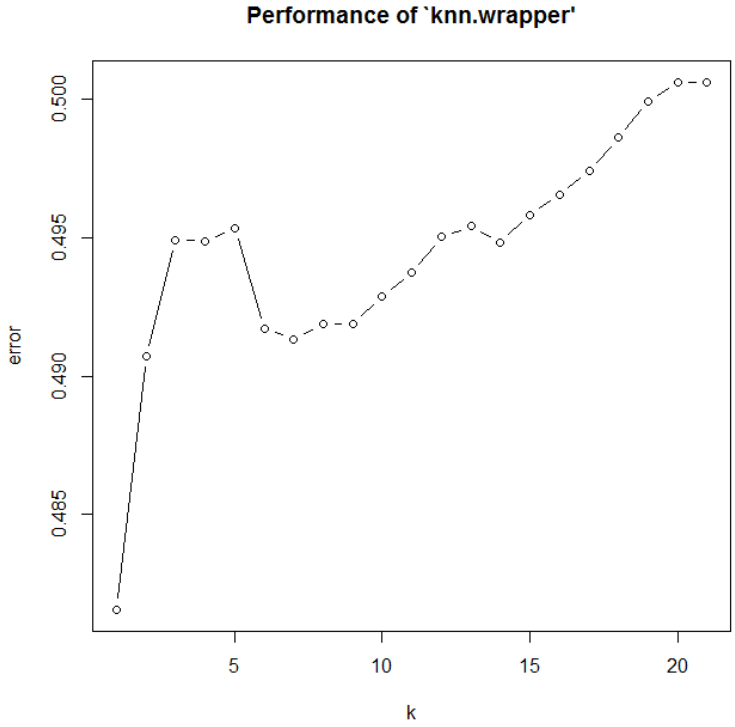
k = 9 K = 5 error rate is 0.47



d)


error rate for size N = 400 K = 10 cross-validation error rate
Error Rate
0.46
0.48
0.50
0.52
5
10
15
20
k Neraest Neighbor Classifier

e)


Performance of `knn.wrapper'
error
0.485
0.490
0.495
0.500
5
10
15
20
k