Anthony Schroeder

Chadd Frasier

KNN Analysis

Data set 1: spam

Here is the matrix of loss values, based on the error between testing and training data, as we can see the error on training data is slightly less than the error for test. For this example, the there is no difference in the validation data used as they both suggest a NN of 1.

L1, Test Data

```
[1] 0.1651503 0.1913427 0.2082529 0.2198098 0.2262476 0.2321478 0.2390011 0.2
467604 0.2516313 0.2561621 0.2607464 0.2658993 0.2693050 0.2722636
[15] 0.2750482 0.2780707 0.2814674 0.2838434 0.2855775 0.2880066 0.2904801 0.
2921022 0.2934273 0.2950901 0.2964544 0.2980637 0.2994210 0.3009177
[29] 0.3018644 0.3033545
```

L1,Training Data

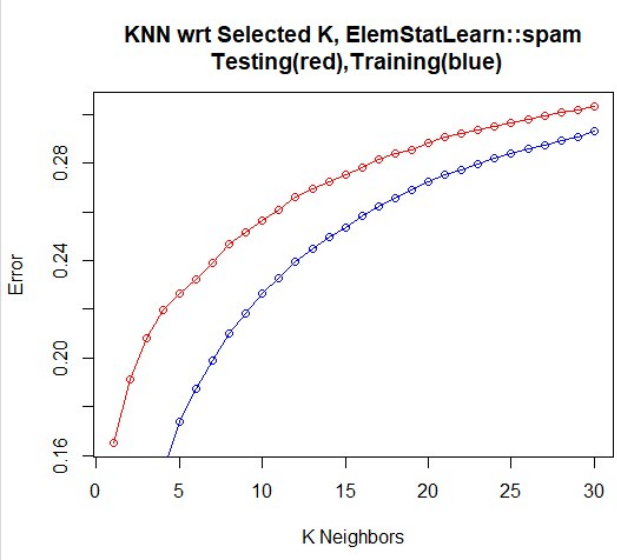
```
 [1] 0.0005514959 0.0836895078 0.1260168206 0.1537984282 0.1738039432 0.18732
47851 0.1987748912 0.2100337791 0.2181472801 0.2263890804 0.2325558076
[12] 0.2394066823 0.2448854055 0.2494435800 0.2537157038 0.2583069075 0.26196
05680 0.2654074176 0.2690429369 0.2721908176 0.2751045544 0.2773460512
[23] 0.2796923612 0.2818431454 0.2838604715 0.2856908017 0.2874366163 0.28928
42371 0.2908713160 0.2928489361
```

print out and/or plot the matrix.

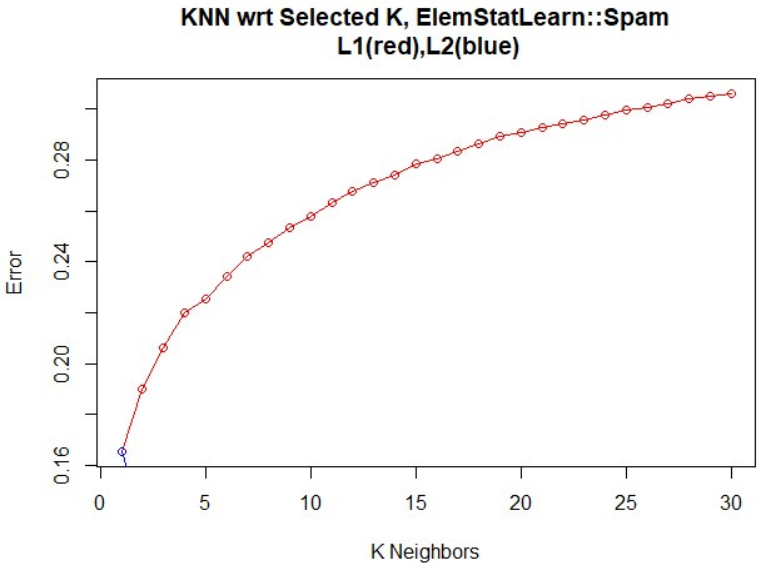
Comment on difference between NN and baseline.

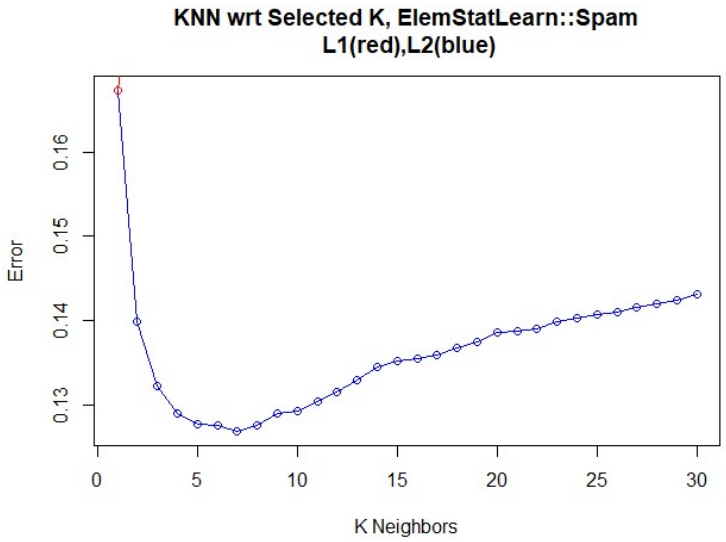
As we can see, the way to

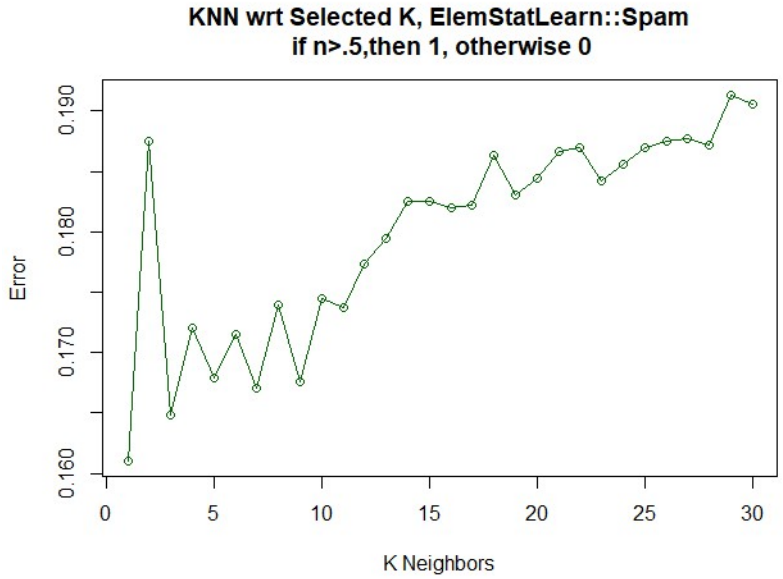
Train/validation loss plot



Plot the two loss functions. (For Testing Data)







What is the optimal number of neighbors?

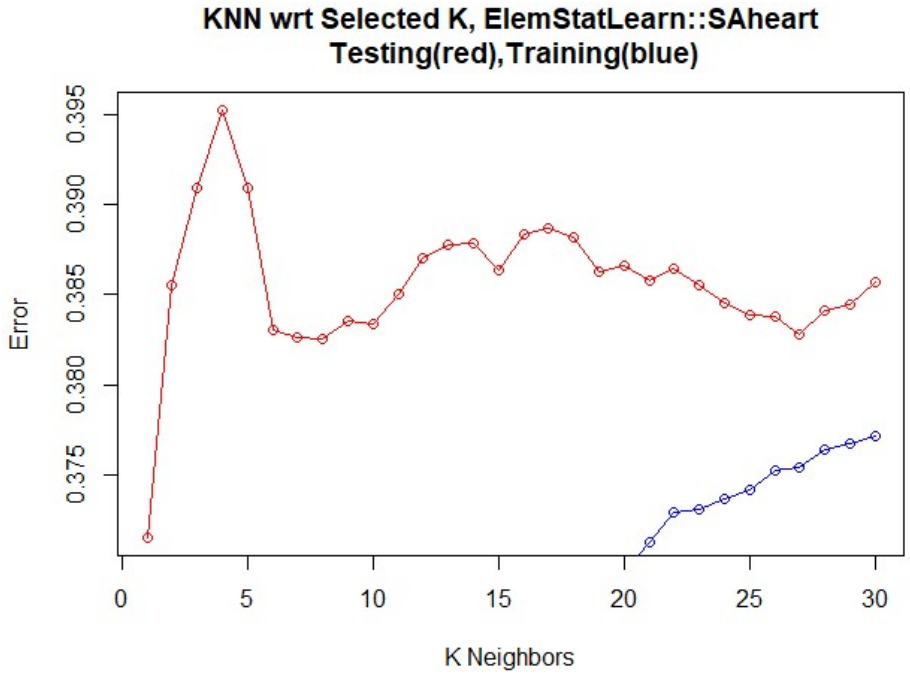
The optimal Number of neighbors Depending on your loss function, as L1's best KNN = 1, and L2's best KNN = 8.

Data set 2: SAheart

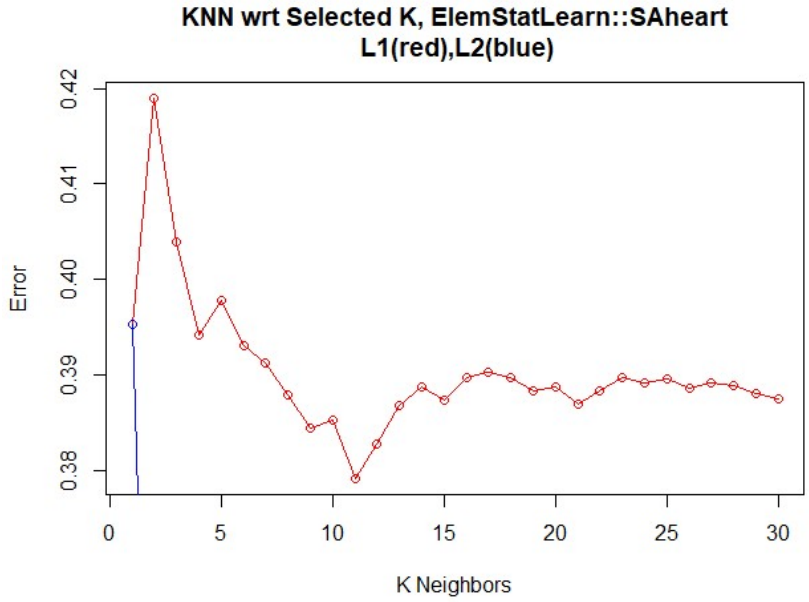
Here is the matrix of loss values, based on the error between testing and tra ining data, as we can see the error on training data is slightly less than th e error for test.

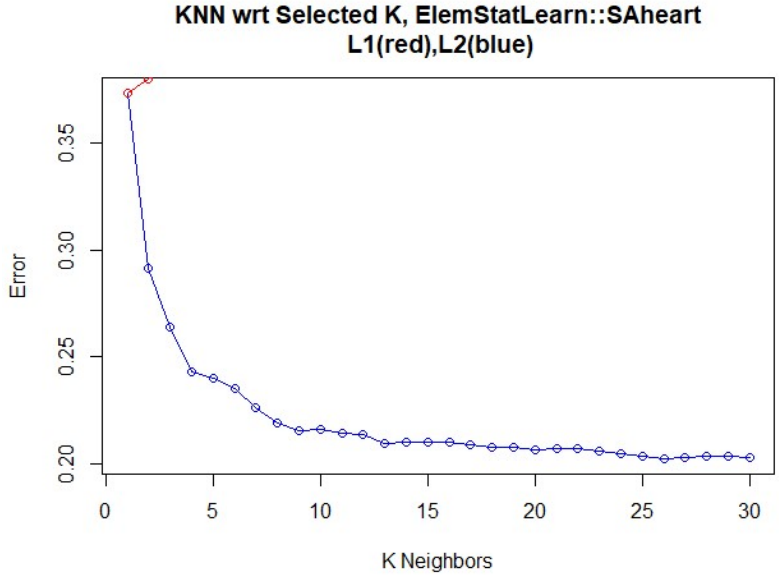
```
L1, Testing Data Error

[1] 0.3714903 0.3855292 0.3909287 0.3952484 0.3909287 0.3830094 0.3825980 0.3
825594 0.3834893 0.3833693 0.3850383 0.3869690 0.3877721 0.3878433 0.3863211
0.3883639
[17] 0.3886418 0.3881689 0.3862680 0.3866091 0.3857863 0.3864127 0.3854822 0.
3845392 0.3838445 0.3837847 0.3827694 0.3841407 0.3844492 0.3856731

L1, Training Data Error

[1] 0.0000000 0.1978378 0.2659459 0.2940541 0.3141622 0.3254054 0.3334363 0.3
398649 0.3466667 0.3496216 0.3526290 0.3560360 0.3591684 0.3606178 0.3631712
0.3656757
[17] 0.3664865 0.3675676 0.3681935 0.3696757 0.3712741 0.3729238 0.3731140 0.
3736486 0.3741405 0.3752183 0.3754154 0.3763707 0.3767381 0.3771171
```

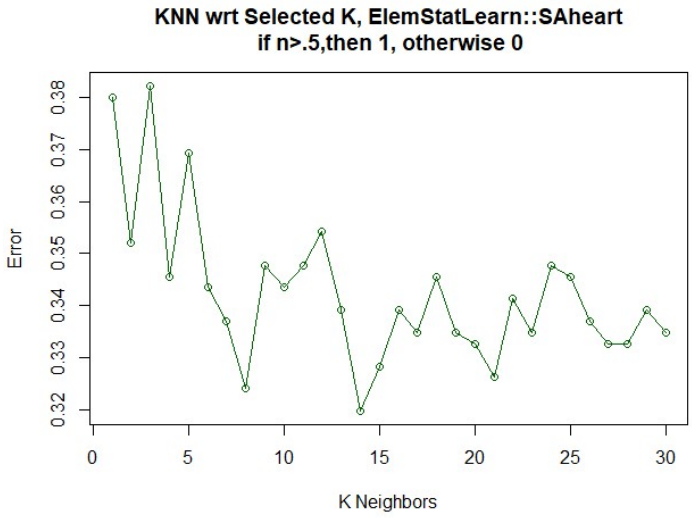
Here is the Training Data and Testing Data loss plot. This is the error with respect to the different NN's used. And shows the different



plot the two loss functions.


KNN wrt Selected K, ElemStatLearn::SAheart
L1(red),L2(blue)
Error
0.38
0.39
0.40
0.41
0.42
0
5
10
15
20
25
30
K Neighbors


KNN wrt Selected K, ElemStatLearn::SAheart
L1(red),L2(blue)
Error
0.20
0.25
0.30
0.35
0
5
10
15
20
25
30
K Neighbors



What is the optimal number of neighbors?

The optimal Number of neighbors Depending on your loss function, as L1's best KNN = 1, and L2's best KNN = 30.

Data set 3: Train Zip

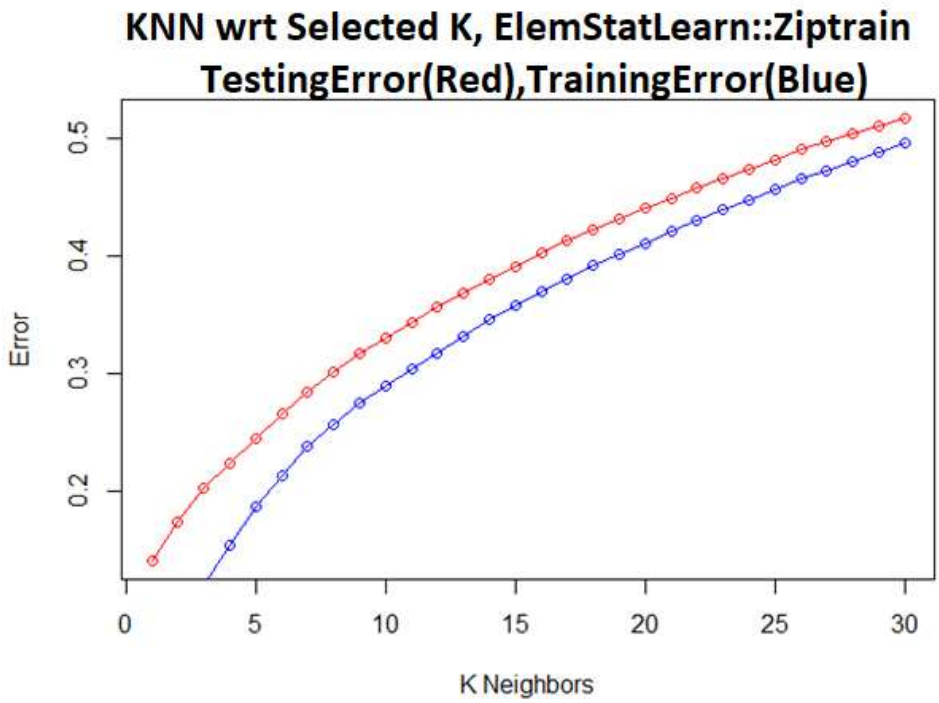
Here is the matrix of loss values, based on the error between testing and training data, as we can see the error on training data is slightly less than the error for test. For this example, the there is no difference in the validation data used as they both suggest a NN of 1.

```
 [1] 0.1560966 0.1846652 0.2086524 0.2304634 0.2527783 0.2728909 0.2891644 0.
3021058 0.3185854 0.3342627 0.3459115 0.3583186 0.3699195 0.3822053 0.3928006
0.4030900
[17] 0.4142595 0.4235007 0.4316348 0.4408306 0.4504128 0.4586331 0.4665824 0.
4751374 0.4820734 0.4899787 0.4984401 0.5043337 0.5107280 0.5177499
 [1] 0.0000000 0.0724595 0.1173949 0.1555965 0.1870987 0.2137457 0.2378007 0.
2559278 0.2719140 0.2881296 0.3028875 0.3173049 0.3305917 0.3438530 0.3558697
0.3672496
[17] 0.3786075 0.3902362 0.4006563 0.4103780 0.4194824 0.4292275 0.4378519 0.
4466331 0.4558154 0.4644349 0.4722704 0.4800056 0.4872818 0.4941941
```

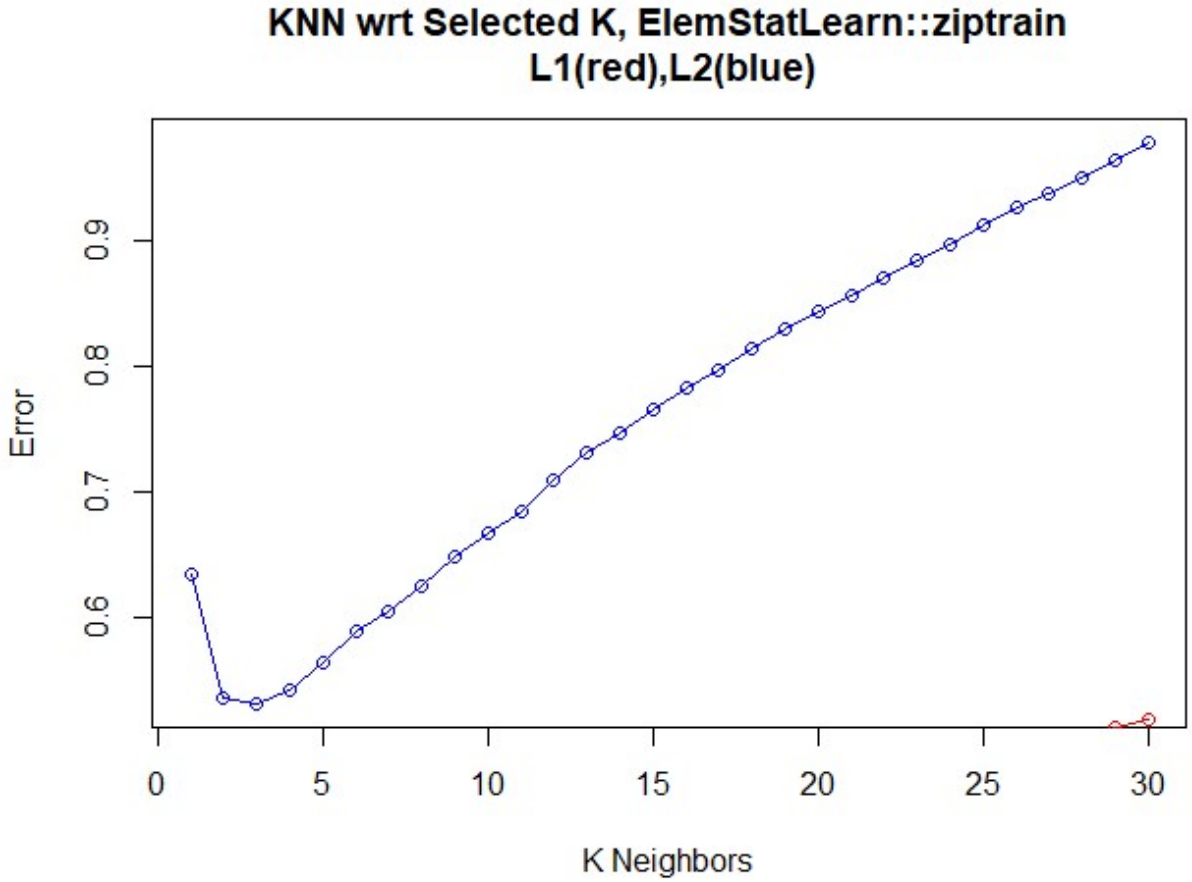
print out and/or plot the matrix.

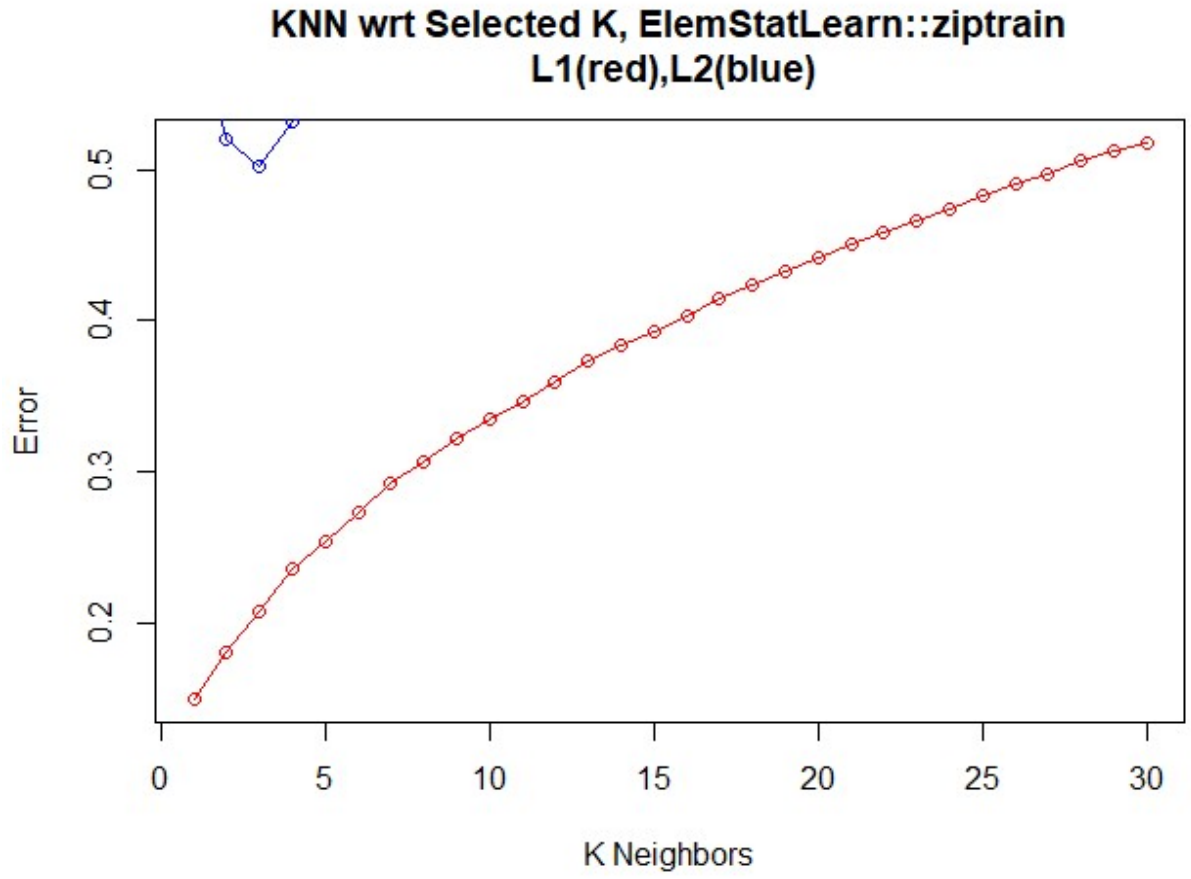
comment on difference between NN and baseline.

Train/validation loss plot



plot the two loss functions.


KNN wrt Selected K, ElemStatLearn::ziptrain
L1(red),L2(blue)
Error
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30
K Neighbors



What is the optimal number of neighbors?

The optimal Number of neighbors Depends on the loss function used, as L1's best KNN = 1, and L2's best KNN = 3.

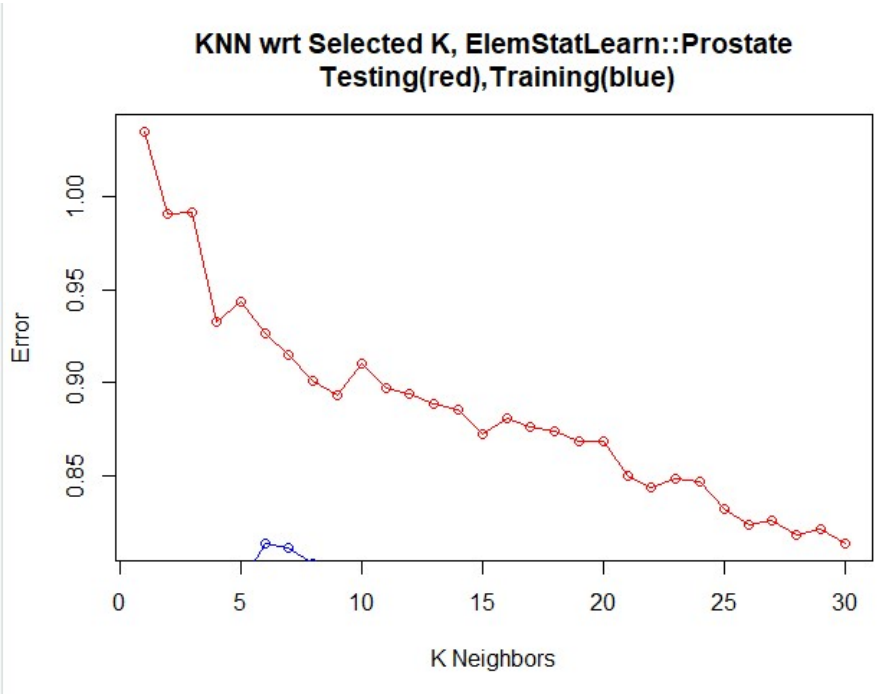
** Data set 4: Prostate

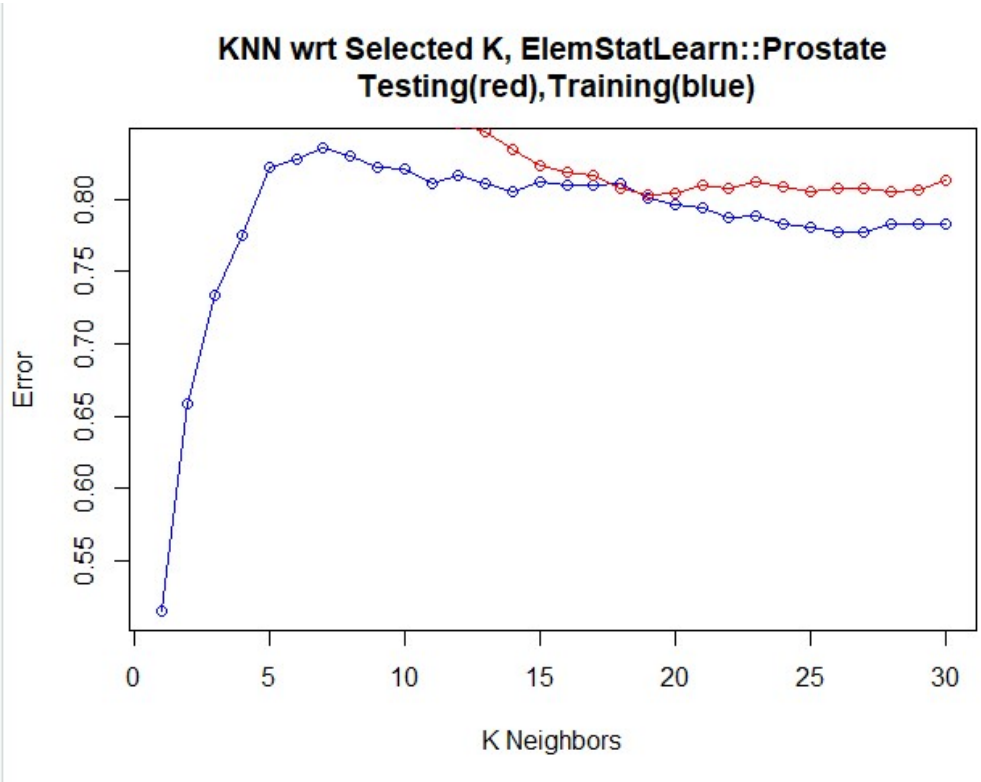
Matrix of loss values

print out and/or plot the matrix.

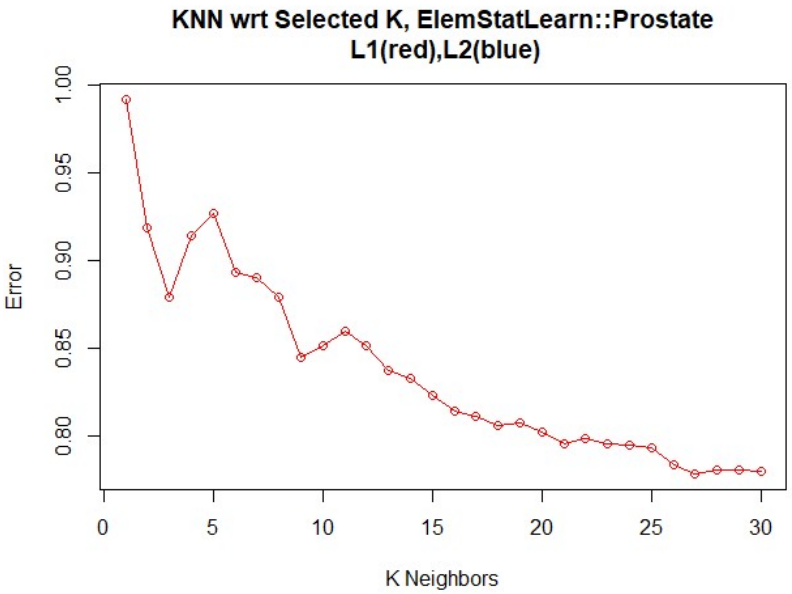
comment on difference between NN and baseline.

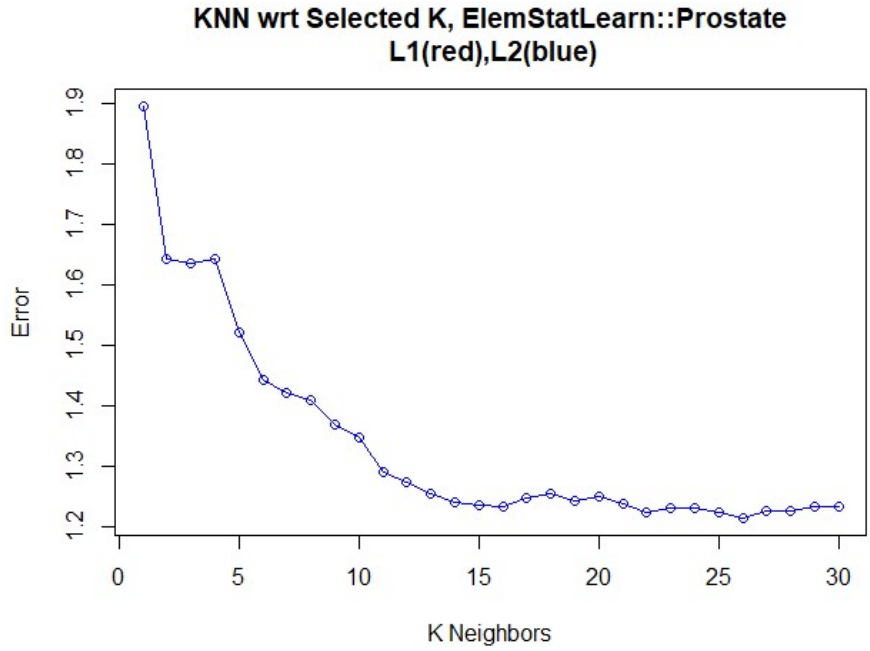
Train/validation loss plot





plot the two loss functions.





What is the optimal number of neighbors?

For this problem the best solution is to use the L1 Loss function, with 30 or more K samples.

Data set 5: Ozone

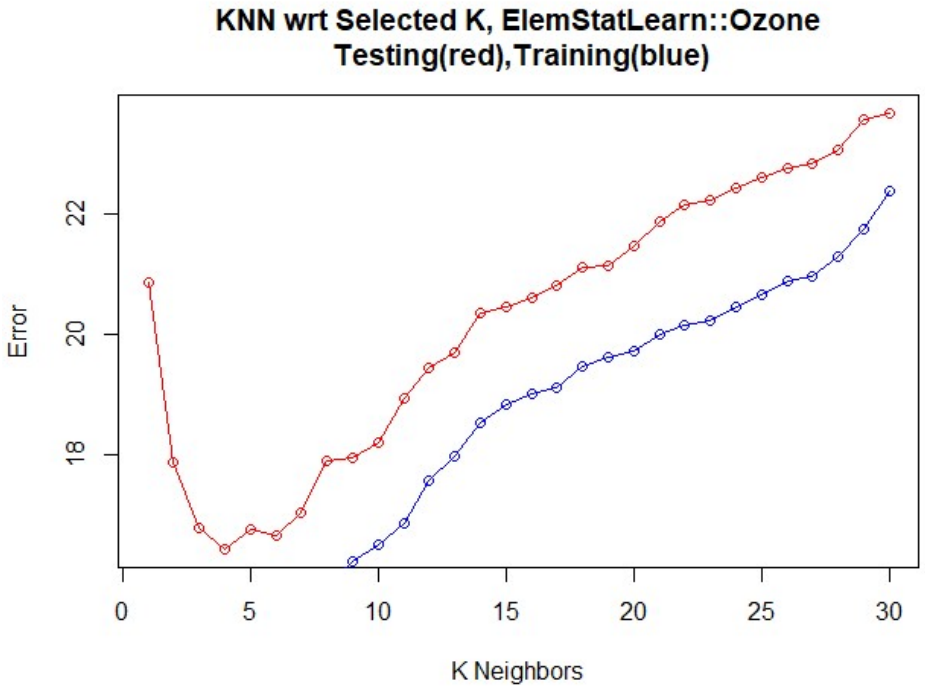
Matrix of loss values

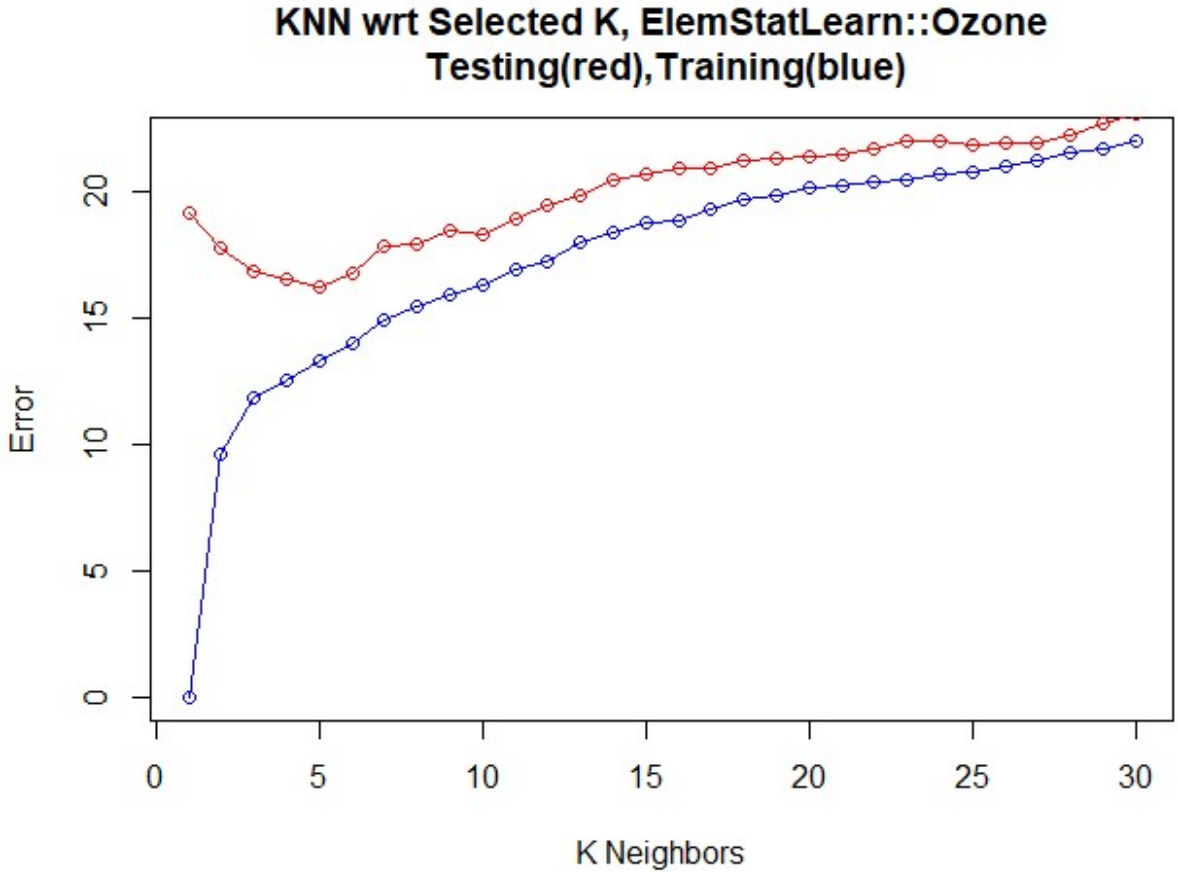
```
[1] 20.84821 17.87946 16.78571 16.42634 16.75893 16.66667 17.03444 17.89844 1
7.95238 18.20536 18.93101 19.42485 19.70124 20.35459 20.44821 20.59877 20.790
44 21.11161
[19] 21.12218 21.47143 21.87457 22.13312 22.21002 22.41183 22.59036 22.74897
22.82937 23.04369 23.55203 23.66637
 [1]  0.00896861  9.46860987 11.44693572 12.31502242 13.55156951 14.25560538
15.01985906 15.85145740 16.23716991 16.51614350 16.86302487 17.56726457 17.96
895481
[14] 18.53619475 18.81853513 18.99887892 19.09601688 19.45440957 19.60797734
19.72331839 19.99423447 20.13473298 20.22304543 20.45702541 20.65668161 20.86
598827
[27] 20.95914300 21.26985906 21.73140560 22.38071749
```

print out and/or plot the matrix.

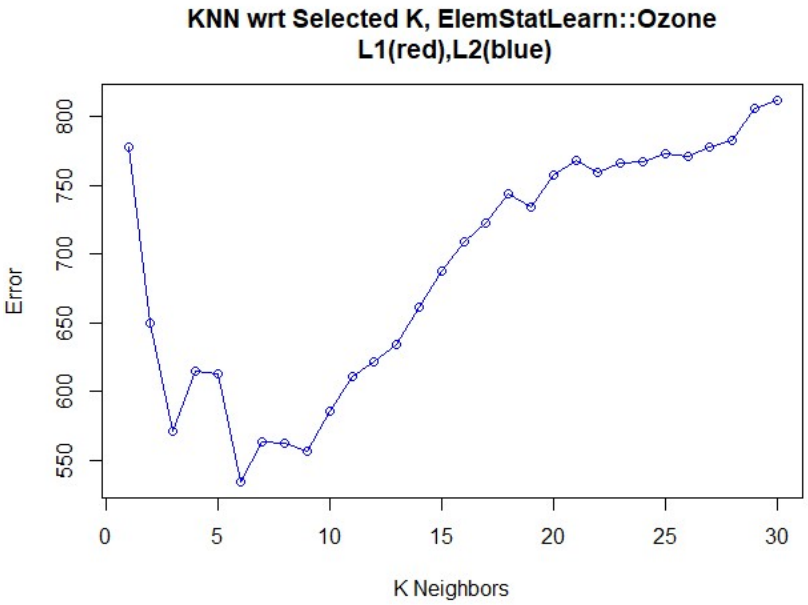
comment on difference between NN and baseline.

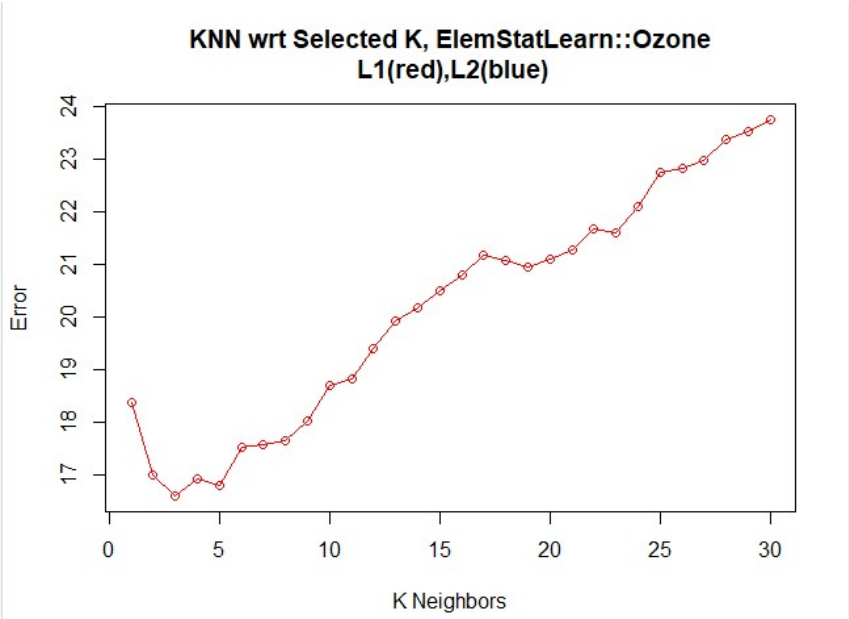
Train/validation loss plot





Plot the two loss functions. (Based on the Testing Data)





What is the optimal number of neighbors?

For this problem the best solution is to use the L1 Loss function, with 3 k samples.