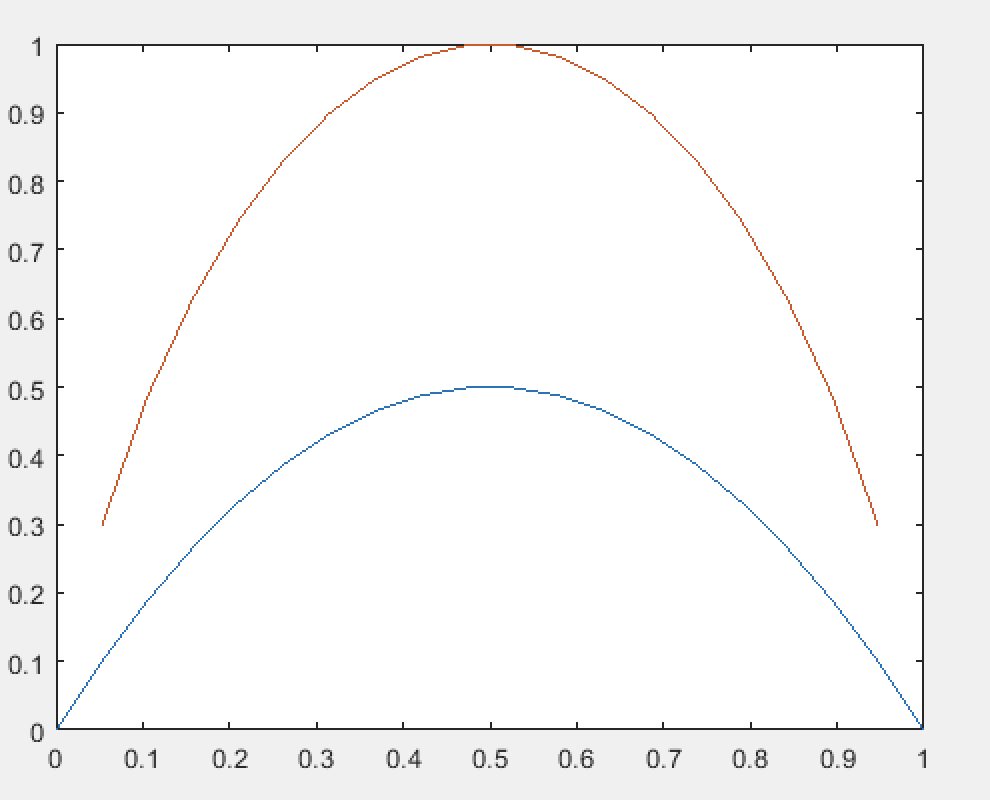
**3(a)**



Comment on similarities:

The blue curve represents the gini index and the red curve represents the entropy. We can see that they both have a bell shape with the maximum value at p = 0.5 and the minimum values at p = 0 and p = 1.

**6(a)**

Training data:

0 – 437

1 – 273

The majority in the training data is 0 with a fraction of 0.6155.

Testing data:

0 – 108

1 – 69

Class 0 in the testing data has a fraction of 0.6102. The simple baseline is fairly accurate with a percent difference of 0.8%.

**6(b)**

Training and testing accuracy for decision tree classification:

Training: 0.0775

Testing: 0.2765

**6(c)**

k = 1

Training: 0.0155

Testing: 0.2698

k = 3

Training: 0.1042

Testing: 0.2309

k = 5

Training: 0.1296

Testing: 0.2199

**6(d)**

knn1 = fitcknn(x\_train, y\_train , 'NumNeighbors', 1)

cv\_k1 = kfoldLoss(crossval(knn1))

|  |  |
| --- | --- |
|  | Classification loss |
| Decision tree | 0.1789 |
| KNN, k =1 | 0.2239 |
| KNN, k = 3 | 0.1915 |
| KNN, k = 5 | 0.1718 |

**6(e)**

for k = 1:15

knn = fitcknn(x\_train, y\_train , 'NumNeighbors', k)

t\_err(k) = loss(knn, x\_test, y\_test)

v\_err(k) = kfoldLoss(crossval(knn))

end

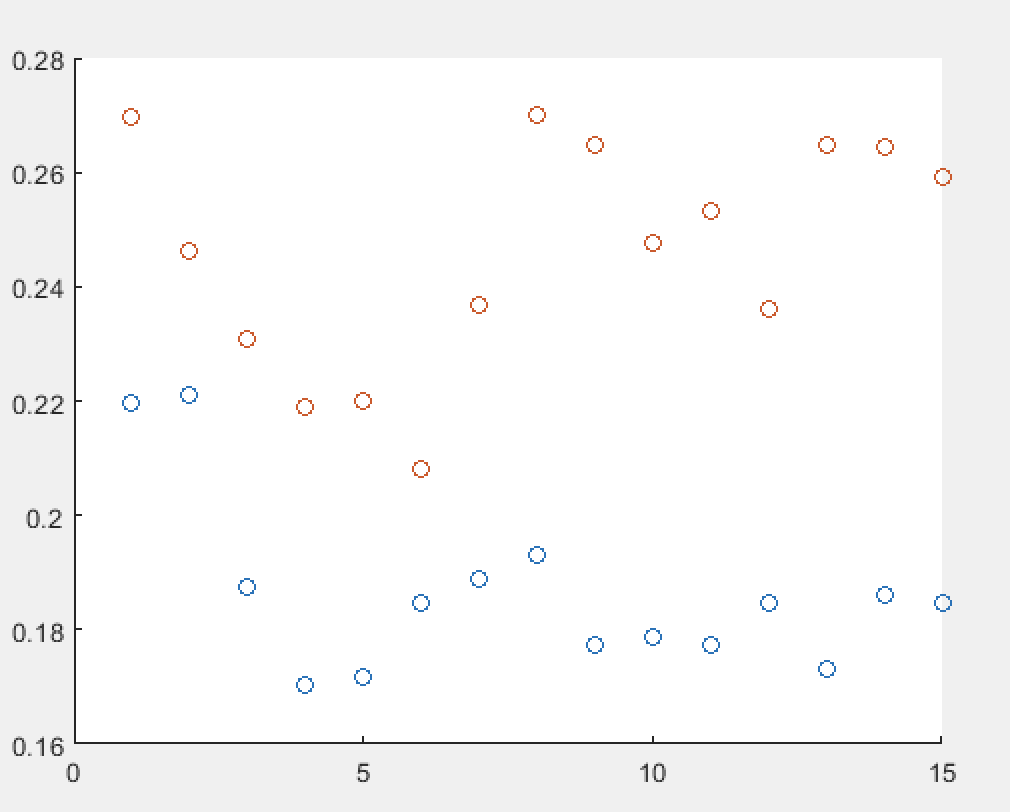
x = transpose(linspace(1,15,15))

scatter(x, v\_err)

hold on

scatter(x, t\_err)

hold off



**6(f)**

We should use cross validation because the cross-validation error is generally lower than the classification tree.