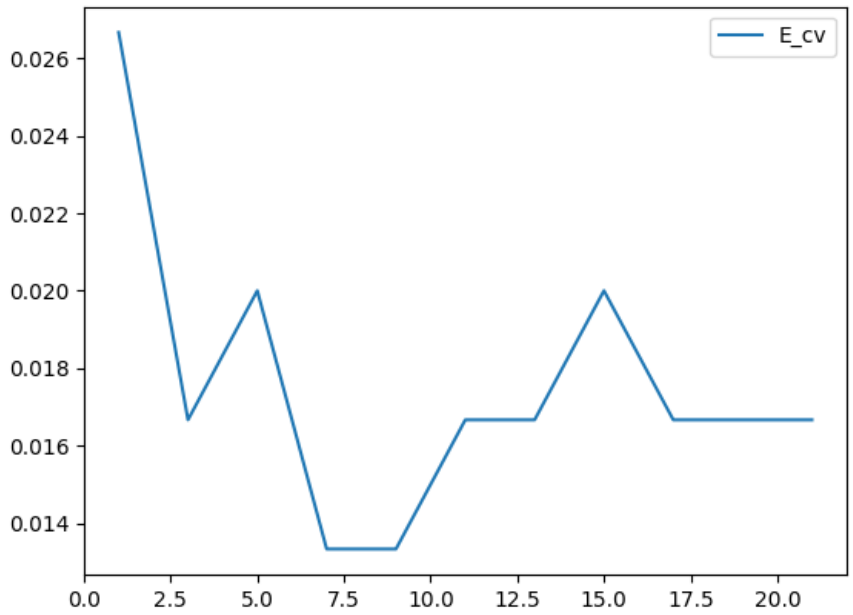
Homework 11

Runmin Lu

November 20, 2021

1

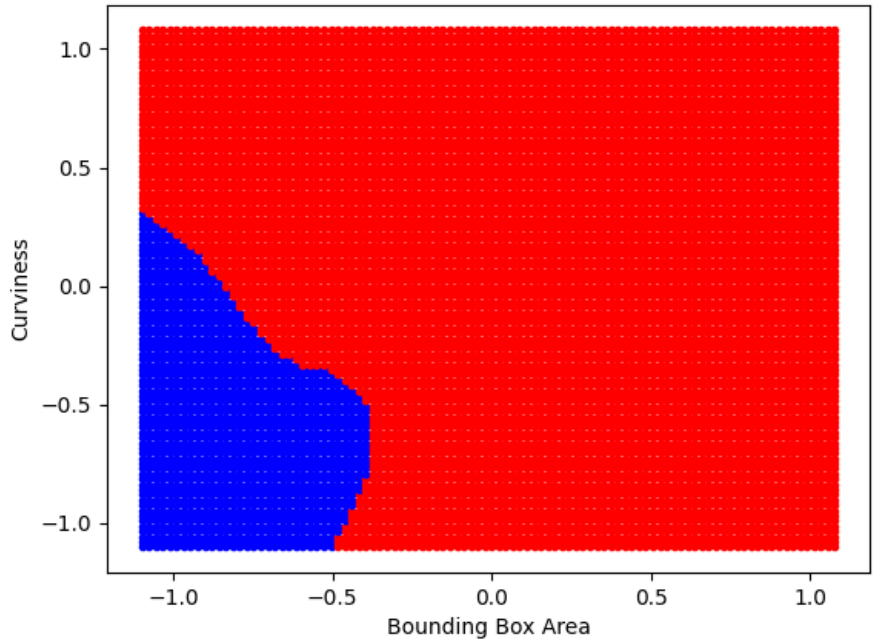
(a)



The optimal k is 9.

(b)

Note: blue is 1, red is not 1.



$E_{\text{in}} \approx 0.0167$

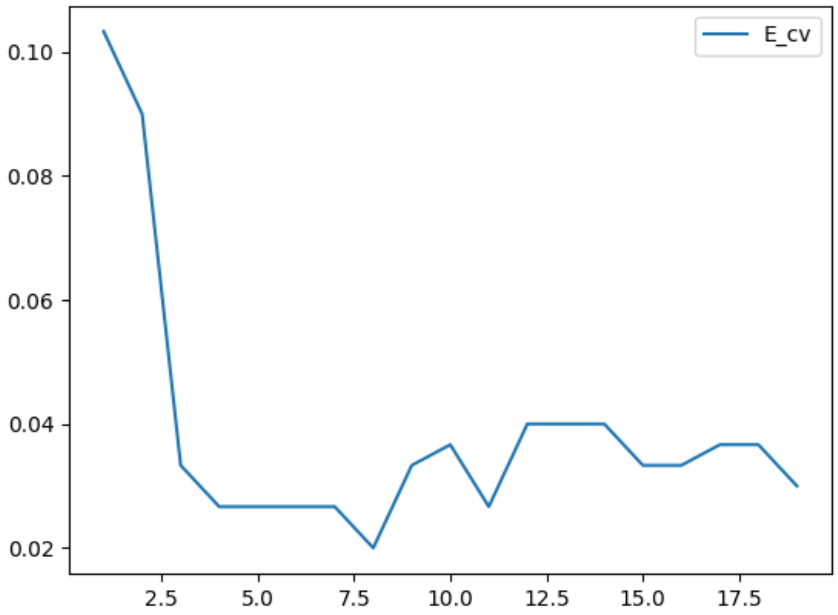
$E_{\text{cv}} \approx 0.0133$

(c)

$E_{\text{test}} \approx 0.0127$

2

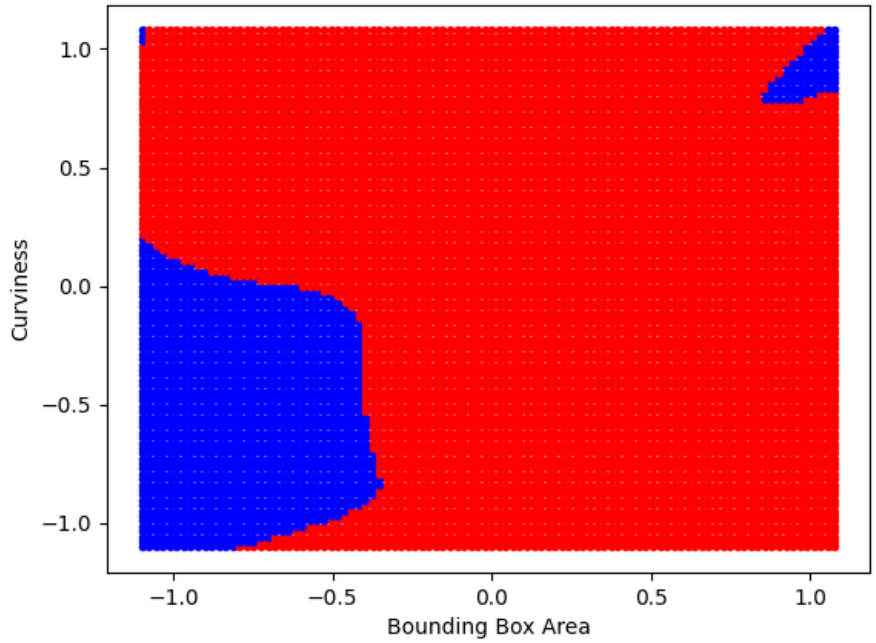
(a)



The optimal k is 8.

(b)

Note: blue is 1, red is not 1.



$E_{\text{in}} \approx 0.0267$

$E_{\text{cv}} \approx 0.0200$

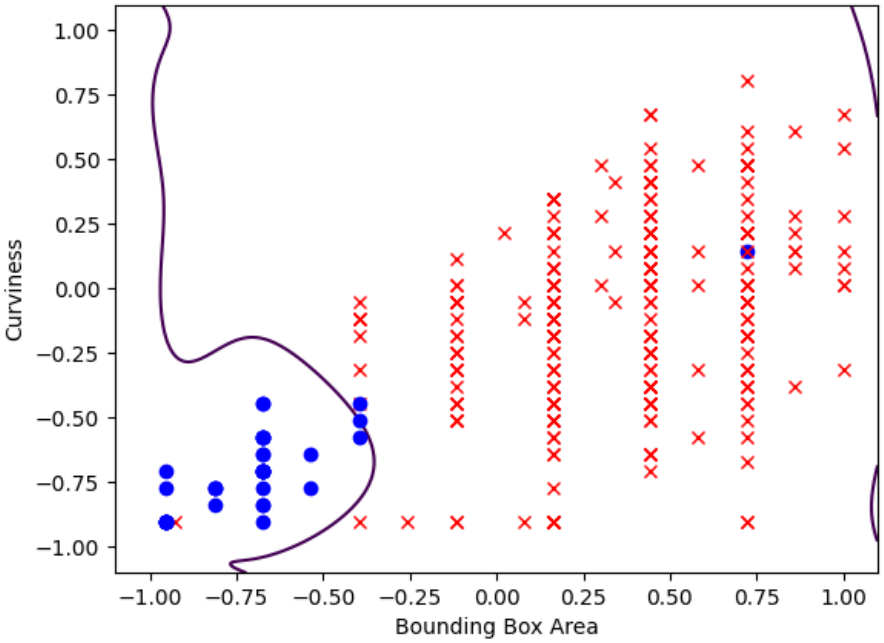
(c)

$E_{\text{test}} \approx 0.0156$

3

Recall from the linear model, $E_{\text{test}} \approx 0.0111$, which is less than k-NN (0.0127), which is less than RBF (0.0156).

However, I believe that k-NN is still the best model for this problem becaues the shape of the linear model's boundary (shown below) looks like there's still some overfitting going on, while k-NN's boundary looks very smooth and matches my expectation.



This is the result of the 2 features I picked, which are not particularly good for this problem, but k-NN resists this flaw by taking the majority of the votes so those outliers don't affect the classification of the test points when k is not too small.

For the same reason, RBF is also not as good as k-NN because outliers also more or less affect the classifications. As you can see in the plot in 2(b), the blue region on the upper right corner is probably the result of an outlier 1 being one of the centers.