

Compsci 571 HW2

Yilin Gao (yg95)

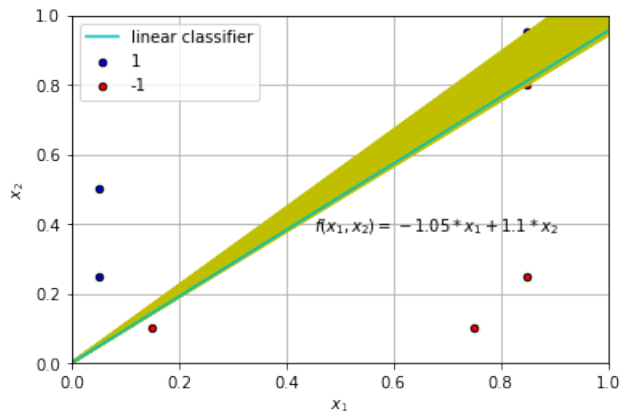
January 29, 2018

1 Classifier for Basketball Courts

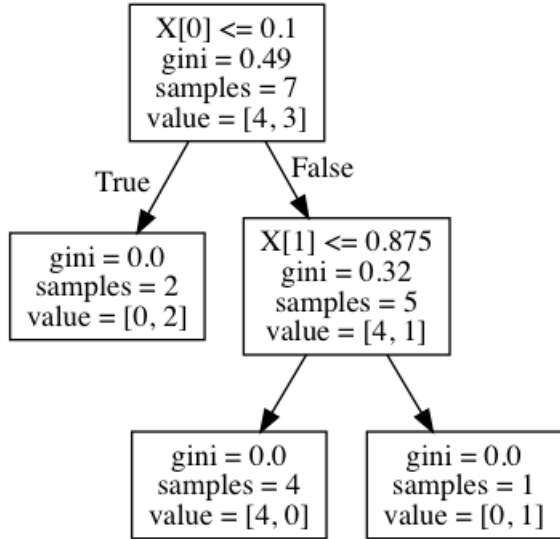
- (a) When running Perceptron algorithm on the dataset, it takes 7 iterations (updates) to converge. The decision boundary is $f(x_1, x_2) = -1.05 * x_1 + 1.1 * x_2$. Because after it converges, all training points are correctly classified, the error rate is 0.

Assume another linear classifier that goes through origin and achieves the same training error rate (0) as the perceptron classifier is $f(x_1, x_2) = w_1 * x_1 + w_2 * x_2$. Set $f(x_1, x_2) = 0$, we get the slope of the boundary is $-\frac{w_1}{w_2}$. From the plot of training data, we know that the boundary should go above point $[0.85, 0.80]$, and go below point $[0.85, 0.95]$. So $\frac{0.80}{0.85} < -\frac{w_1}{w_2} < \frac{0.95}{0.85}$. If we set $w_2 = 1.1$ as the perceptron boundary, we get $-1.229 < w_1 < -1.035$.

The plot of observed data, the perceptron decision boundary (the light blue line), and all other linear boundaries that achieve the same training error (the yellow area) is:



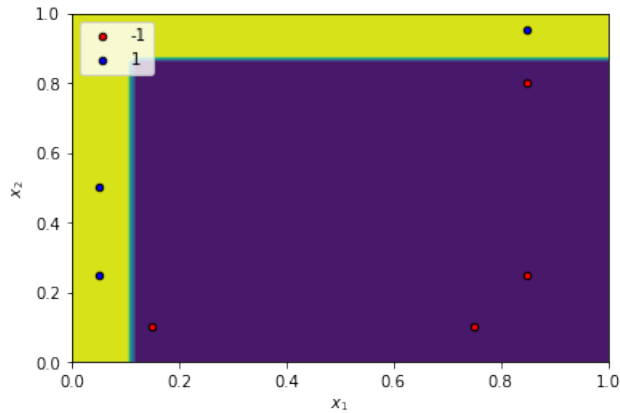
- (b) The fully-grown decision tree using Gini index as splitting criterion on the observed data is:



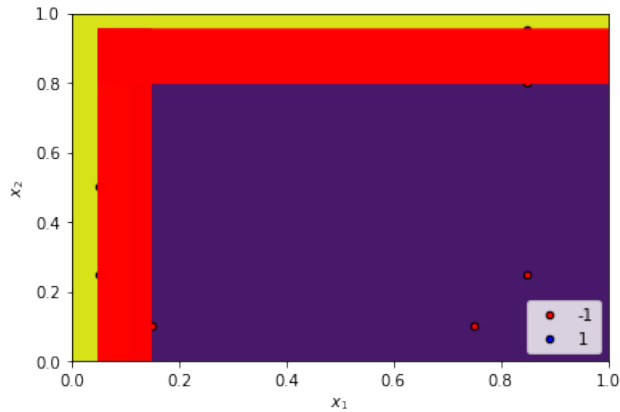
Because all training points are correctly classified by this tree, its training error is 0.

Assume another decision tree with same training error (0) splits on the same feature order but different splitting threshold (v_1 for x_1 and v_2 for x_2). Then the threshold of the first split on x_1 should be able to separate points $[0.05, 0.25], [0.05, 0.5]$ (+1) with $[0.15, 0.1]$ (-1). So v_1 should be $\in (0.05, 0.15)$. The threshold of the second split on x_2 should be able to separate points $[0.85, 0.8]$ (-1) with $[0.85, 0.95]$ (+1). So v_2 should be $\in (0.8, 0.95)$.

The plot of observed data, and the calculated decision boundary is:



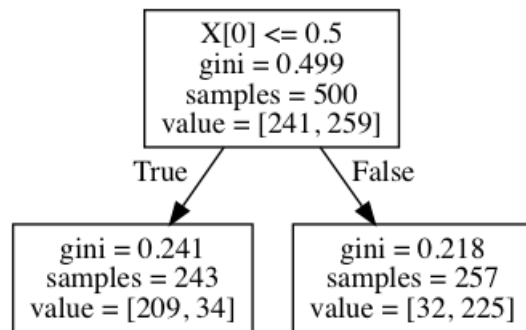
The plot of observed data, the calculated decision boundary, and all other decision boundaries that achieve the same training area (the red area) is:



- (c)
- (d)
- (e)
- (f)
- (g)
- (h)

2 Variable Importance for Trees and Random Forest

- (a) (i) The decision stump based on the best split is:



The decision stump based on the best surrogate split is:

- (ii)
- (iii) The mean least-squares error of predictions on the test data of the decision stump based on the best split is 10.
The mean least-squares error of predictions on the test data of the decision stump based on the best surrogate split is
- (b) (i)
- (ii)
- (iii)
- (c) (i)
- (ii)