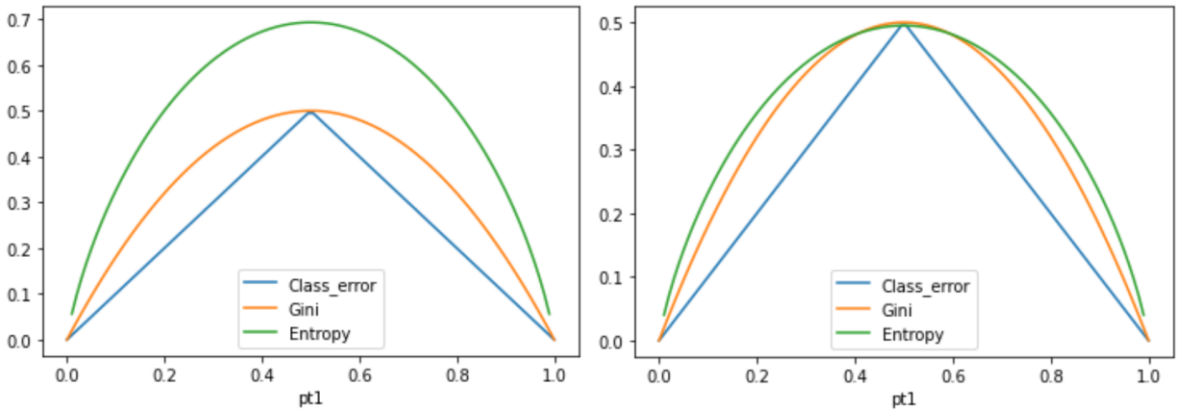
SDSC5001 Statistical Machine Learning I
Assignment #3

1. Consider the Gini index, classification error, and entropy in a simple classification setting with two classes (0 and 1). Create a single plot that displays each of these quantities as a function of $\hat{p}_{t1}$, the proportion of training observations in node *t* that are from class 1. The *x* axis should display $\hat{p}_{t1}$, ranging from 0 to 1, and the *y* axis should display the value of the Gini index, classification error, and entropy. You can make the plot by hand or software.

The plot is shown below. The right graph is when the maximal of entropy scaled to 0.5.
gini is shown in orange, entropy is shown in green, and classification error is shown in blue.



2. Suppose we produce 10 bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of *X*, produce 10 estimates of *P*(Class is Red | *X*):

0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach, and the other one is to classify based on the average probability.

(a) What is the final classification under the majority vote approach?

(b) What is the final classification under the average probability approach?

(a) The number of red predictions is greater than the number of green predictions based on a 50% threshold, thus RED.
(b) The average of the probabilities is less than the 50% threshold, thus GREEN.