* Traverse through the nodes until you get to a leaf corresponding to a prediction
* Root node at the top and then each internal node has a binary query function
* Recursively call the query function
* State variable that changes during recursion is which node you’re at
* Implemented like any tree data structure
* Basic form of the query function is if a certain dimension is greater than 1 – indicator function
* Decision trees may not always be balanced
* Within a partition, we make a static prediction that isn’t changing
* There are arrangements like the triangle with opposite in middle that linear models don’t cover but linear models can
* Decision trees generally axis aligned, but may not be more convenient than linear model
* End up with wasted boundaries
* Decision trees are fairly interpretable too
* All training data has to end up in one of the leaf nodes
* The number of possible queries is the number of possible splits, which is the number of features times the number of training points
* S prime is a subset of the training set
* Choose the split that has the largest impurity reduction – considers all possible splits
* The problem with 0/1 Loss is that it’s discontinuous
* We want a more continuous measure – Bernoulli and entropy measures are two such measures
* They all basically have the same curve; they have a maximal value at 50-50 and then fall off
* Loop through ALL possible partitions across all dimensions
* Sometimes you save the computation you did
* In top down training, every intermediate step is a decision tree, you don’t backtrack, and you don’t reconsider earlier splits
* Stopping condition – one is to not split if resulting children are smaller than a minimum size
* Another is setting a maximize set
* Bagging: we’re trying to minimize the test error of the true distribution
* The training set is drawn from the true distribution
* Decision trees are low bias but high variance models
* In practice we’ll resample S’ with replacement and then we average the decisions
* Bootstrap aggregation
* This causes the variance to decrease
* Every example is sampled with replacement and S’ is the same size as S
* Averaging for classification is a majority vote
* Random forests restricts which dimensions (features) you look at