CS244 Exercise 1

Task 1: Good practices with training and test data

Suppose you have a file of data where the examples are classified into two possible classes, 0 and 1. In the file, the first half of examples belong to class 0 and the last half of examples belong to class 1. Before applying your learning algorithm, you split the data so that the first 70% of examples from the file correspond to your training data and the last 30% of examples from the file correspond to your test data. Why might this be problematic?

Suppose we repeatedly build a decision tree based on training data and assess its performance using test data, and each time we build the tree, we try a different approach for pruning the tree. Ultimately, we choose the pruning approach that yielded the highest accuracy on the test data. Why might this be problematic?

Task 2: Aspects of decision trees

True or False. Decision trees cannot be used on features that can take on multiple discrete values.

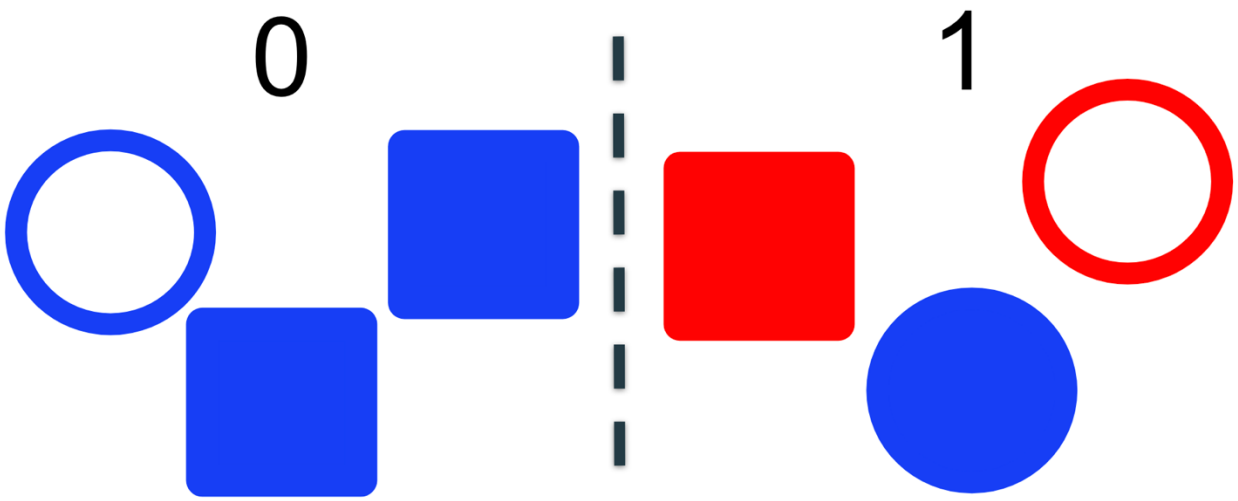
True or False. Decision trees cannot be used on continuous, real-valued features.

True or False. Pruning a large decision tree may help address the problem of overfitting.

True or False. With ensemble methods, multiple classifiers are trained and then used to make predictions.

Task 3: Building decision trees

Consider the following training data, consisting of 6 shapes classified as either 0 or 1 (if you cannot make out the red and blue colors in your hardcopy exercise, check the online version of the exercise available from the course website).
There are three binary features: isCircle?, isFilled?, isRed?

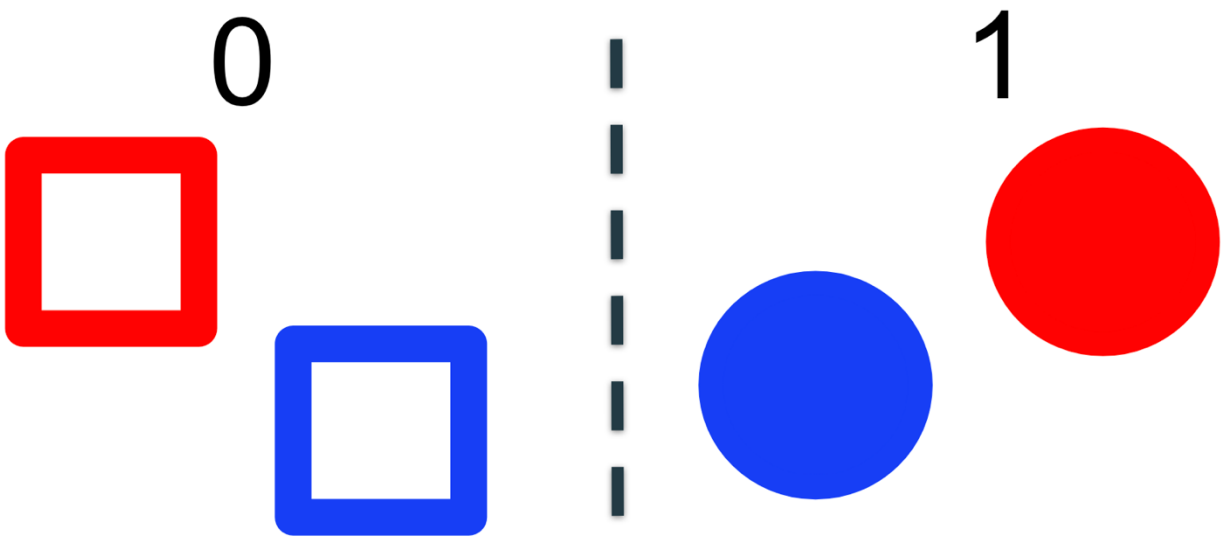


Compute the entropy of this training data. What is the entropy?

For each of the three features, what is the information gain when that feature corresponds to the root node of a decision tree?

Draw the complete decision tree (both internal nodes and leaves) that would result from the above data using a greedy algorithm based on information gain.

Consider the following test data:



What is the accuracy of predictions from the tree for the above test data, i.e., what percent of examples from the test data does the tree correctly classify?

Task 4: Using decision trees for classification

Download the Jupyter Notebook for Exercise 1 from the course website. Open the Notebook in your web browser and work through it. Once you have worked through the Notebook, answer the following questions.

In the complete set of banknote data, how many banknotes are forged (class 0)?

In the complete set of banknote data, how many banknotes are genuine (class 1)?

In the test data, how many banknotes are forged (class 0)?

In the test data, how many banknotes are genuine (class 1)?

What is the accuracy of the decision tree on the test data?

What is the accuracy of the random forest classifier, an ensemble of decision trees, on the test data?

What is the accuracy of the gradient boosting classifier, sequentially learned trees that put more weight on the more challenging points, on the test data?