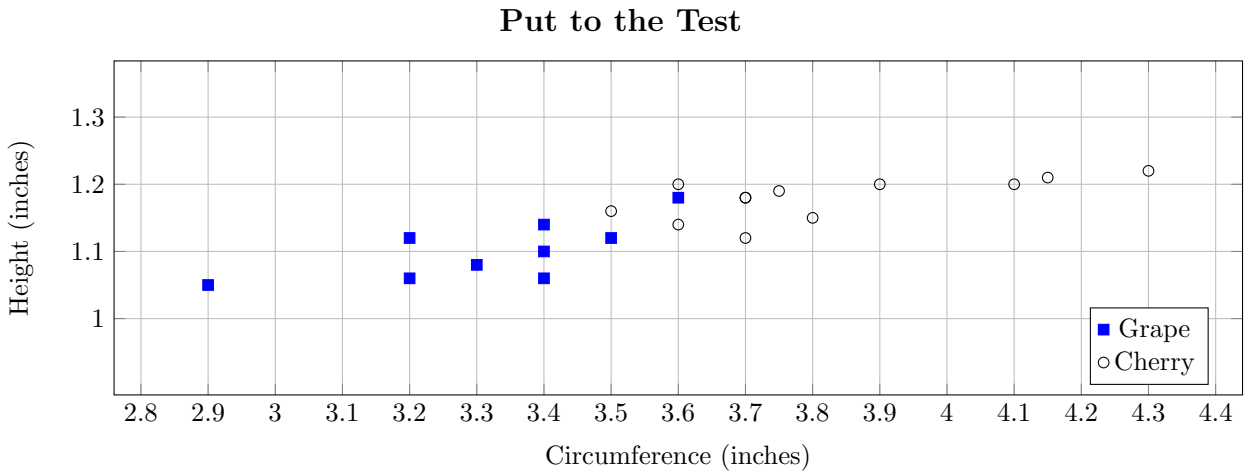
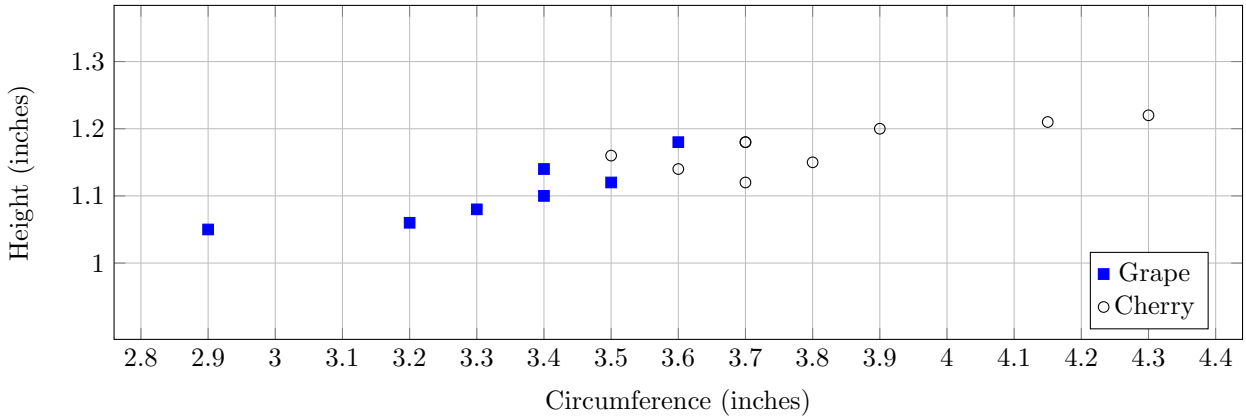


Intuitively, we saw that the 3-nearest neighbor algorithm seemed to be better at classifying grapes and cherries than the 1-nearest neighbor algorithm. In order to verify this, we'll divide our given data into a *training set* and a *testing set*.

The *training set* will be used for classification. So, we'll figure out the cherry and grape regions based only on the data from the training set. It's called the "training set" because it's used to *train* the algorithm - it's how the algorithm learns what to do!

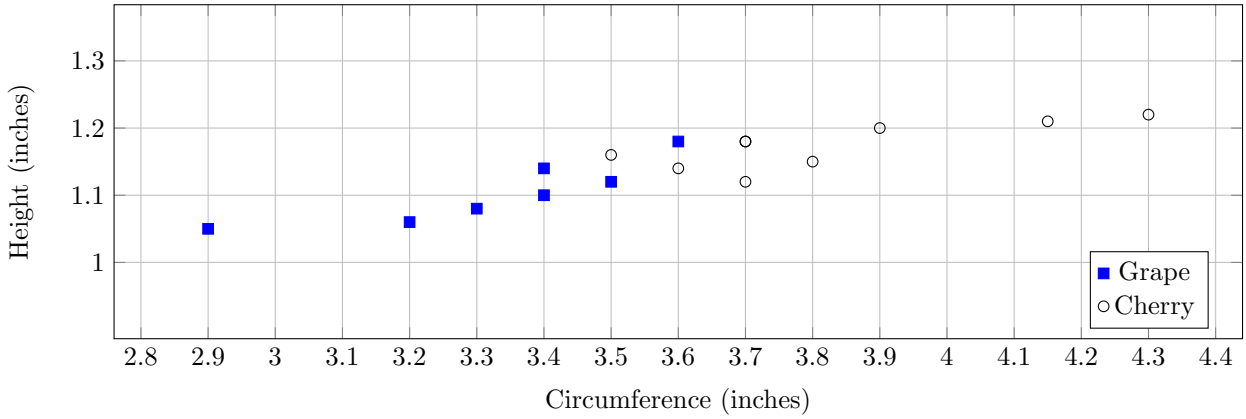
The *testing set* will be used to evaluate how well our classification algorithm is performing. For each data point P from the testing set, we'll use k-nearest neighbor algorithm to predict whether P is a grape or a cherry. Then, we'll check if the algorithm was correct.

Below, we've separated our data into a training set and a testing set. The points from the training set are plotted above, and the points from the testing set are in the table below.



Circumference (inches)	Height (inches)	Fruit
3.2	1.12	Grape
4.1	1.2	Cherry
3.4	1.06	Grape
3.75	1.19	Cherry
3.6	1.20	Cherry

In order to compare 1-nearest neighbor classification and 3-nearest neighbor classification, we'll measure how accurately each algorithm is able to classify the testing set.
Let's start with 1-nearest neighbor classification. Using the training set below, shade in the Cherry region for the 1-nearest neighbor algorithm.



For each point in the testing set, plot it on the graph above, and use the 1-nearest neighbor algorithm to predict if it's a grape or a cherry. Then check if your prediction is correct, filling in "Yes" or "No".

Circumference (inches)	Height (inches)	Predicted Fruit	Actual Fruit	Correct?
3.2	1.12		Grape	
4.1	1.2		Cherry	
3.4	1.06		Grape	
3.75	1.19		Cherry	
3.6	1.20		Cherry	

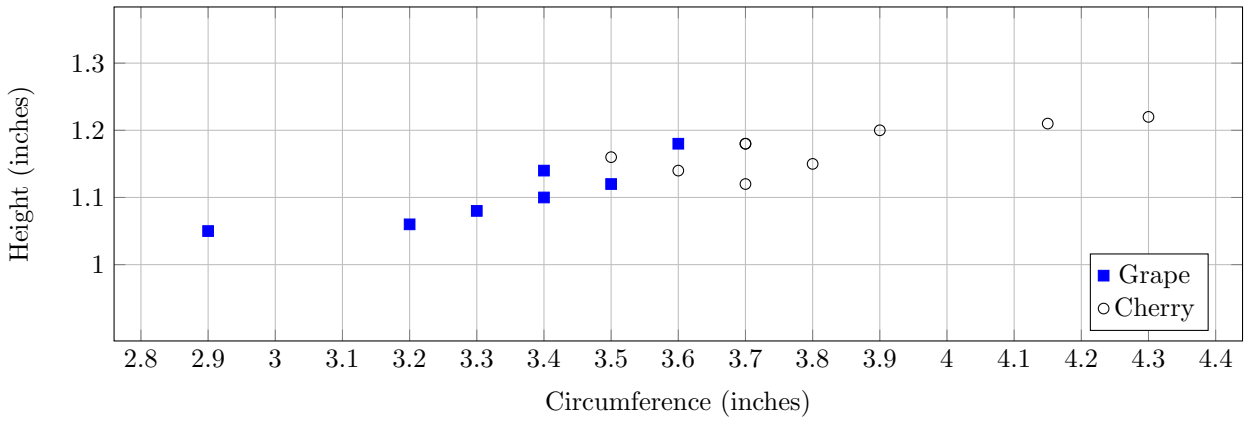
What percent of your predictions were correct?

We can also organize our results into a *confusion matrix*, by counting how many times each combination of predicted fruit and actual fruit occurs.

$$\left(\begin{array}{l|l} \text{number of cherries that were predicted to be cherries} & \text{number of cherries that were predicted to be grapes} \\ \hline \text{number of grapes that were predicted to be cherries} & \text{number of grapes that were predicted to be grapes} \end{array}\right)$$

What is the confusion matrix for your results?

Next, let's evaluate the precision of 3-nearest neighbor classification. Using the training set below, shade in the Cherry region for the 3-nearest neighbor algorithm.



For each point in the testing set, plot it on the graph above, and use the 3-nearest neighbor algorithm to predict if it's a grape or a cherry. Then check if your prediction is correct, filling in "Yes" or "No".

Circumference (inches)	Height (inches)	Predicted Fruit	Actual Fruit	Correct?
3.2	1.12		Grape	
4.1	1.2		Cherry	
3.4	1.06		Grape	
3.75	1.19		Cherry	
3.6	1.20		Cherry	

What percent of your predictions were correct?

What is the confusion matrix for your results?

Which was more accurate, classification using the 1-nearest neighbor algorithm or classification using the 3-nearest neighbor algorithm?

How does the choice of the split between training and testing sets affect the algorithm? What are some "bad" ways to split the data?