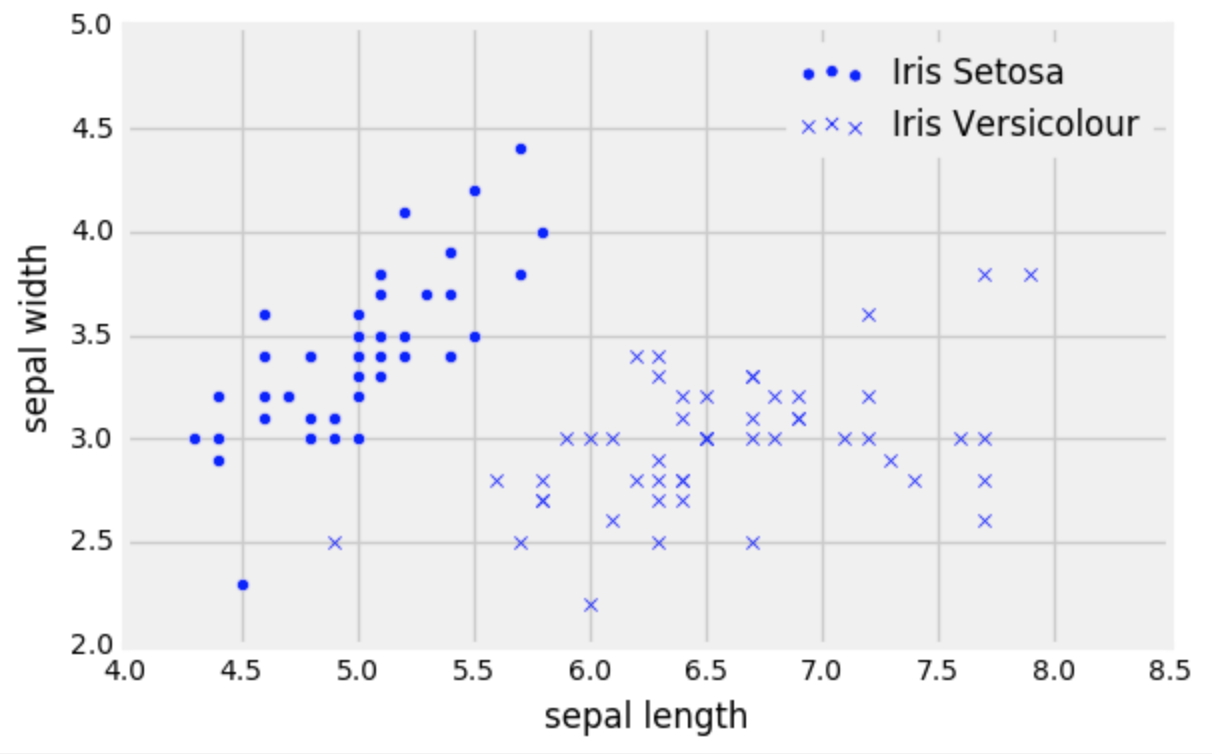
**Data 8 Spring 2020**

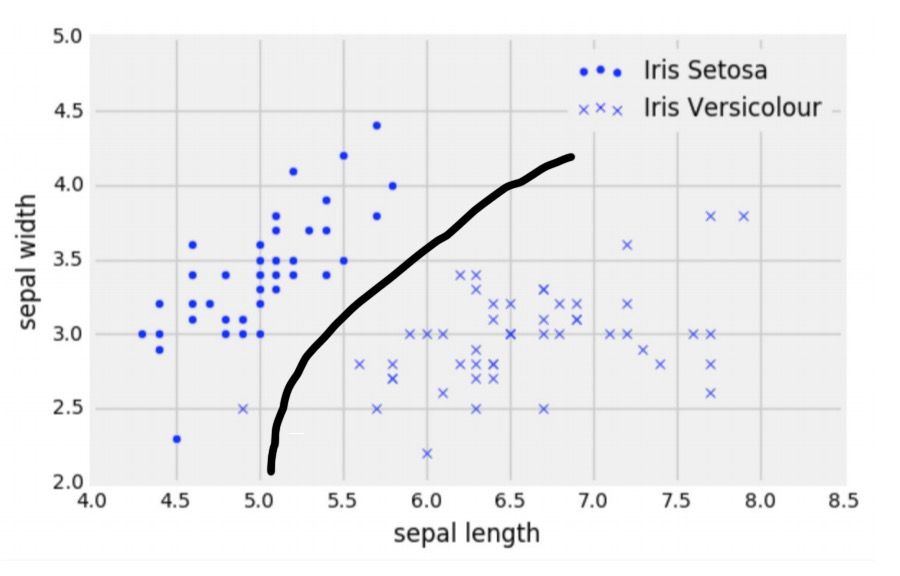
**Discussion: Classification, k-Nearest Neighbors and Conditional Probability (Lab10)**

Given the text of an email, how would you determine whether the email is malicious or safe? Perhaps the kinds of words that are used, or the time the email is sent? In this worksheet, we’ll discuss *classification*, a term that describes a set of methods and techniques to answer questions like the one above.

**Question 1.** R. A. Fisher collected a dataset of Iris flowers, which contains two types of iris flowers (Setosa or Versicolor) and the measurements for the sepal width and sepal length. Krista wants to create a classifier that predicts the correct flower type given a new flower.

a. Krista begins by attempting to classify a new flower as an Iris Setosa or an Iris Versicolor based on the sepal length and sepal width of the flower. Draw the decision boundary that the k nearest neighbors algorithm (with k = 3) would generate for this problem.

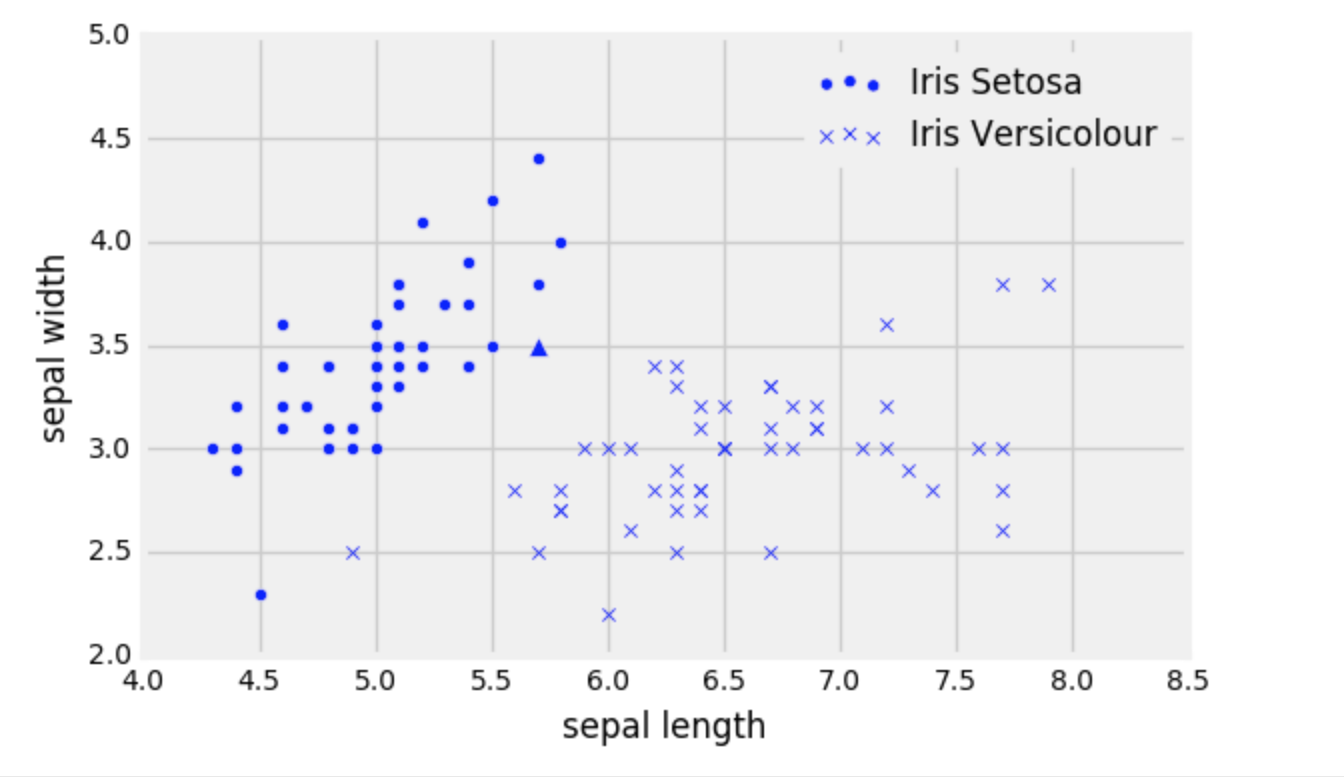


****

**Answer:**

The important thing here is to convey the main idea of what a decision boundary is, aka the boundary such that if a new point falls on the right side of the boundary, it will be classified as x and o if it falls on the left side of the boundary. The boundary in this case would be almost linearly separating the two classes. For areas where the split is not so well defined, urge students to move an imaginary point across the plot and try to notice when its classification would change.

b. Now Krista wants to classify a new flower (represented as a triangle in the scatter plot on the next page). Describe the steps she would take to classify this new point based on a k nearest neighbors classifier with k=3.



**Answer:**

1. Compute the euclidean distance between the new point and all the points in our dataset.
2. Sort all the data based on the calculated distance.
3. Take the top 3 neighbors and take a majority vote.

In this particular case we can eyeball that new point should be classified as o.

c. Deven suggests that Krista should use a different k for her classifier because he says 3 is too small. What values of k should she avoid?

**Answer:**

If we choose k to be even, we run the danger that both classes will get the same number of votes. In this case it is unclear how we should decide how to classify the new point.

d. When trying to develop a classifier, we split our original dataset into a training and a test set. We don’t look at or use the test set until we have finished training. Why is that a good idea in general? What might happen if we didn’t?

**Answer:**

The role of the test set is to have a way of understanding how well our classifier would perform in a real world scenario with data it has not seen before. Emphasize the fact that we should only run our algorithm on the test data once, after we are done selecting the number of neighbors. The main idea behind the train-test split is that of generalization. We want our algorithm to be able to generalize so having a test set is exactly testing whether our algorithm can do that.

e. Suppose Krista chooses k=1 and calculates the accuracy on the training set. Assume that she does **not** remove the point she’s trying to classify from the training set when calculating the accuracy. What will the accuracy be on the training set? Will it be representative of the accuracy on the test set?

**Answer:**

She will get a training accuracy of 1 (the closest neighbor of each point will be itself). The algorithm will not generalize well to unseen data!

**Question 2.** After seeing how successful Krista’s K-NN classifier is, Gregory, the owner of an e-commerce store, wants to classify all customers in one of two classes A or B. To do that he will use the following features.

* Annual income of each customer (in dollars)
* The average amount they spend every time they visit his website.
* Their age

1. Gregory wants to run a k nearest neighbors classifier but his friend Roshan claims that he may need to preprocess your data somehow before doing that. What could the problem be and how should he resolve it?

**Answer:**

The problem here is that the different features live on very different scales. On the one hand, annual income will be in the 100,000 range while age is just a number between 0-100. This means that the value we calculate for the euclidean distance between two points will primarily reflect the difference in the income of the customers. This means that effectively, our algorithm is not using all the information that we are giving it. To deal with this issue, we could normalize (convert to standard units!!)

b. Suppose the training set has 100 customers and has the following distribution:

* A: 90% of customers
* B: 10% of customers

We produce the following scatterplot of the training set:



Gregory builds a k-NN classifier for this data with k = 21. What would the accuracy of the classifier be in this scenario?

Because k=21, we classify every point as class A. 90% of our data is class A, so the accuracy would be 90%. At face value, this would be a good accuracy but our classifier is not very useful. This is the danger of using a very large k with an imbalanced class. To better measure our accuracy, we should use different metrics such as (confusion matrix, precision, recall, F1 score).

After implementing his classifier with a different k, Gregory runs the classifier on 1000 customers and finds that:

* 501 of the A customers were classified correctly
* 208 of the B customers were classified correctly
* 104 of the A customers were classified incorrectly
* 187 of the B customers were classified incorrectly

c. Find the following probabilities:

1. Given that a customer was classified incorrectly, the likelihood that they are a B type customer  
   187/(187+104) = 187/291
2. The probability that a customer is an A type customer  
   (501+104)/1000 = 605/1000
3. The probability that a customer is classified correctly  
   (501+208)/1000 = 709/1000
4. The probability that a customer is classified correctly given that they are an A type customer.  
   501/605