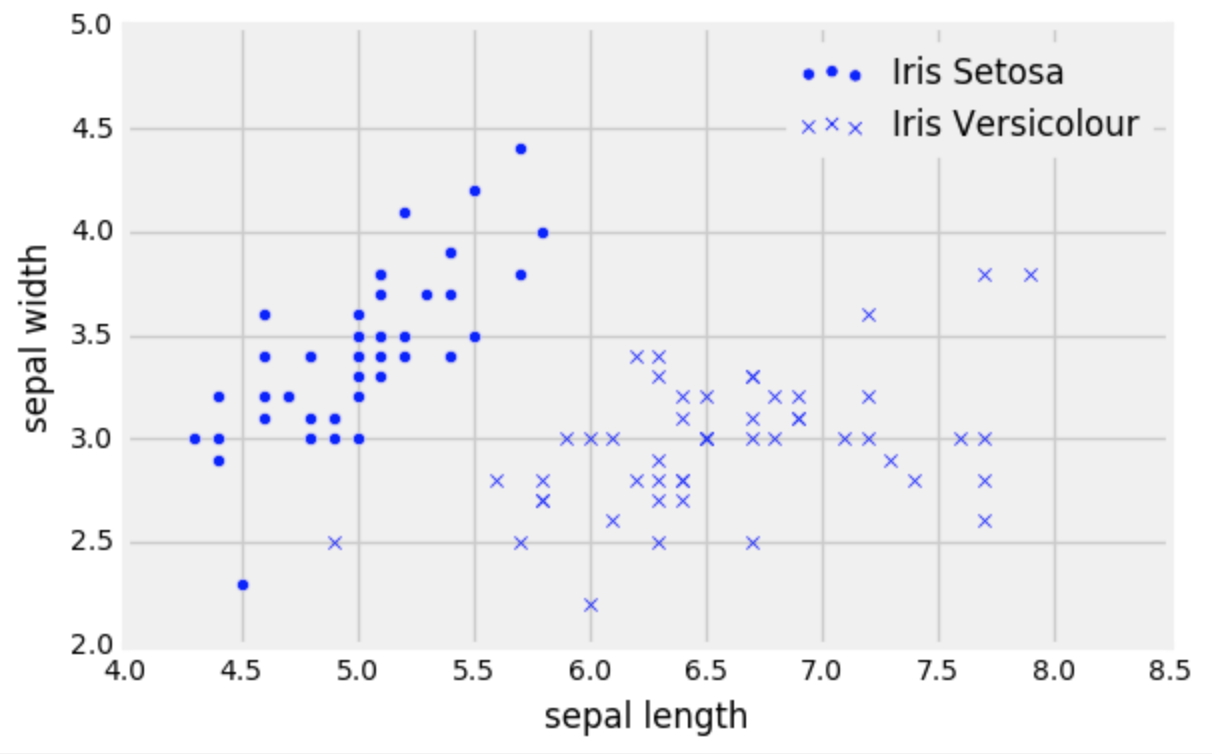
**Data 8 Spring 2020**

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

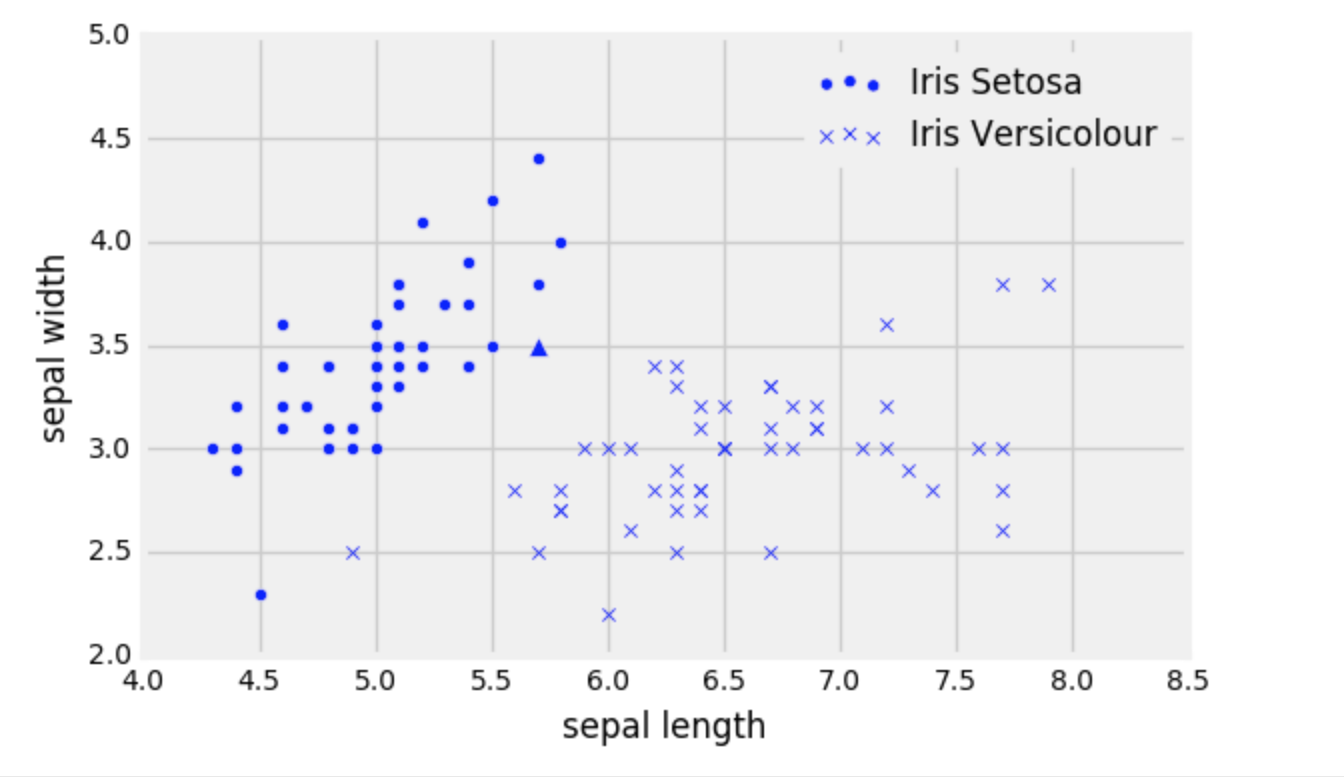
Given the text of an email, how would you determine whether the email is malicious or safe? Perhaps the kinds of words that are used, or the time the email is sent? In this worksheet, we’ll discuss *classification*, a term that describes a set of methods and techniques to answer questions like the one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



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.



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?

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?

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?

**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.

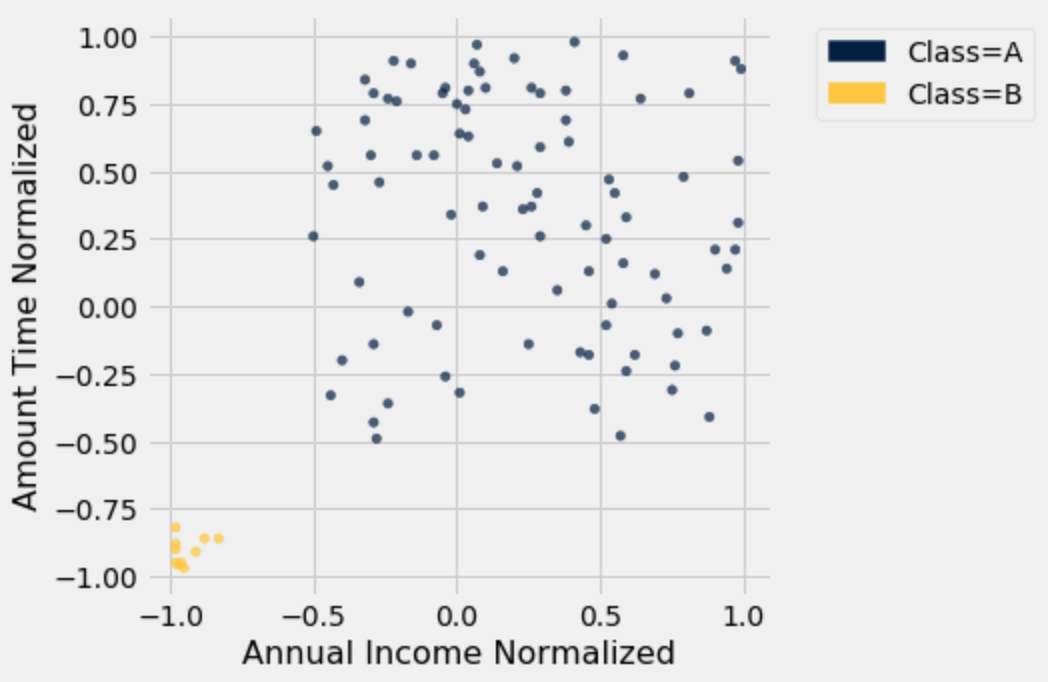
* Annual income of each customer (in dollars)
* The average amount they spend every time they visit his website.
* Their age

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?

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?

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
2. The probability that a customer is an A type customer
3. The probability that a customer is classified correctly
4. The probability that a customer is classified correctly given that they are an A type customer.