

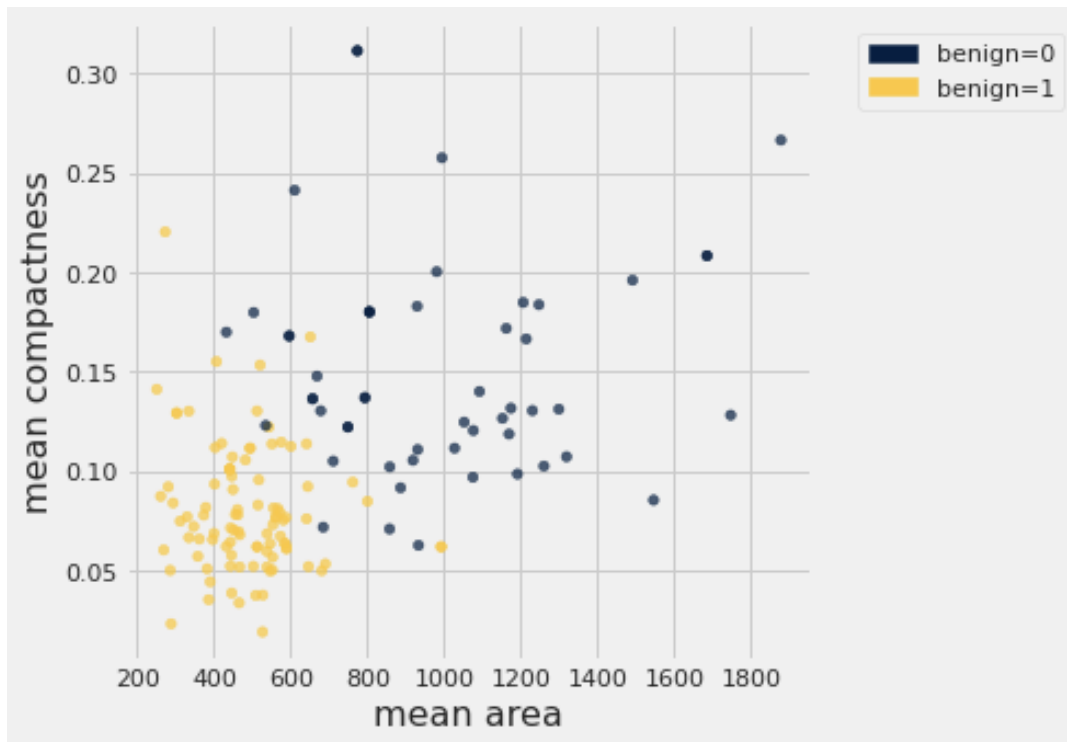
Data 8 Spring 2022

Lab 10: Classification, k-Nearest Neighbors and Conditional Probability

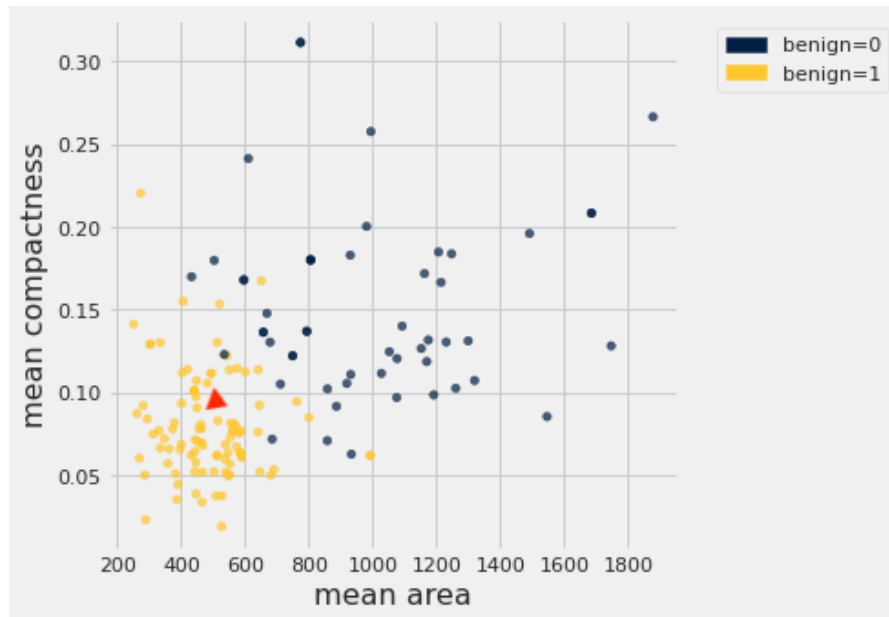
Given the text of an email, how would you determine whether the email is malicious or safe? Perhaps the kinds of words that are used, or the time the email is sent? In this worksheet, we'll discuss *classification*, a term that describes a set of methods and techniques to answer questions like the one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

Question 1. Significant research has been done to understand whether a breast tumor is benign or malignant. Aarushi wants to create a classifier that predicts whether a tumor is benign or not.

- a. Aarushi begins by attempting to classify a new tumor based on the average compactness and average area of the tumor. Draw the decision boundary that the k nearest neighbors algorithm (with $k = 3$) would generate for this problem.



b. Now Aarushi wants to classify a new tumor (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. Prasann suggests that Aarushi should use a different k for his classifier like $k=4$ or $k=8$. Is Prasann's suggestion reasonable?

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 in our breast tumor training dataset we have 30 benign=0 data points and 45 benign=1 data points. What k values are too large?

Question 2. Suppose that all Data 8 students fill out a poll about their podcast listening habits. Interested in starting her own podcast, Sunny is curious about the results of the poll. She finds out the following:

- 80% of students use Spotify and the remaining students use Apple Music
 - 65% of Spotify users listen to podcasts
 - 30% of Apple Music users listen to podcasts
- a. Draw a tree diagram to represent the results of the poll. Assume that students are either Apple Music users *or* Spotify users (i.e. no student can be both).
- b. Assuming that Sunny draws a student uniformly at random from the population, find the following probabilities:
- i. The probability that the student is an Apple Music user
 - ii. Given that the student listens to podcasts, the probability that the student was a Spotify user
 - iii. Given that the student doesn't listen to podcasts, the probability that the student was an Apple Music user
- c. Suppose Rebecca discovers that one of her students interned at Spotify last summer. For this given student, can we still compute probabilities like we did in part b? Why or why not?
- d. (*Just for fun*) What should Sunny's podcast be called?