CS031 Fall 2023

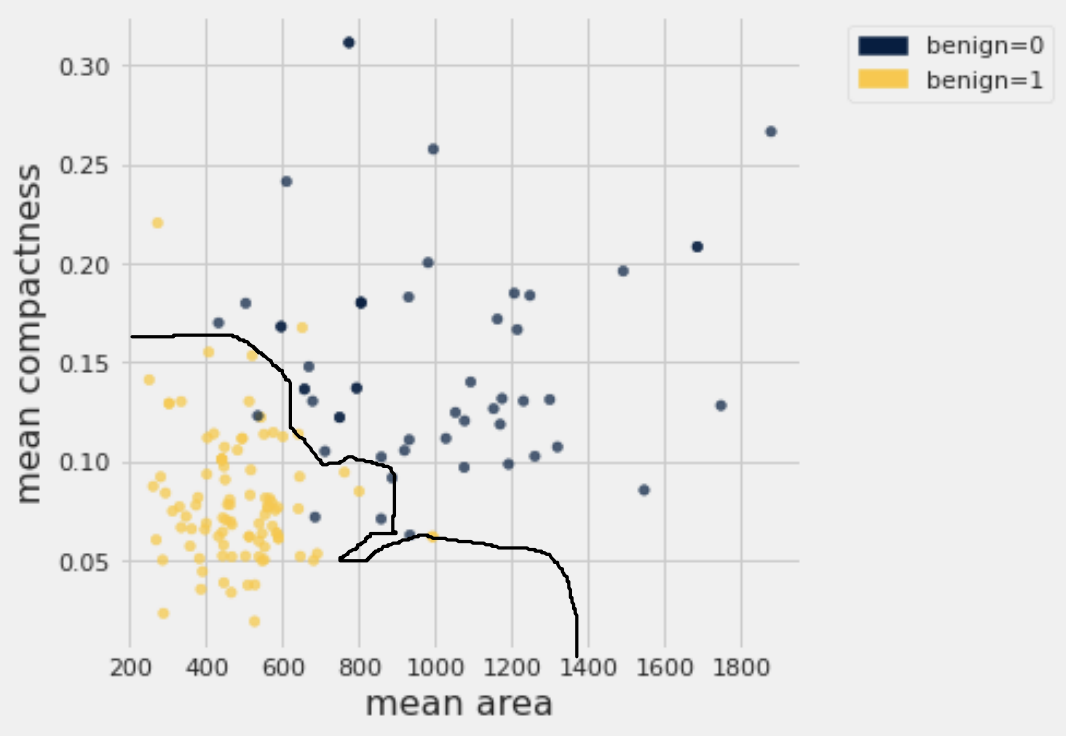
Lab 10: Classification, k-Nearest Neighbors and Conditional Probability

---

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.

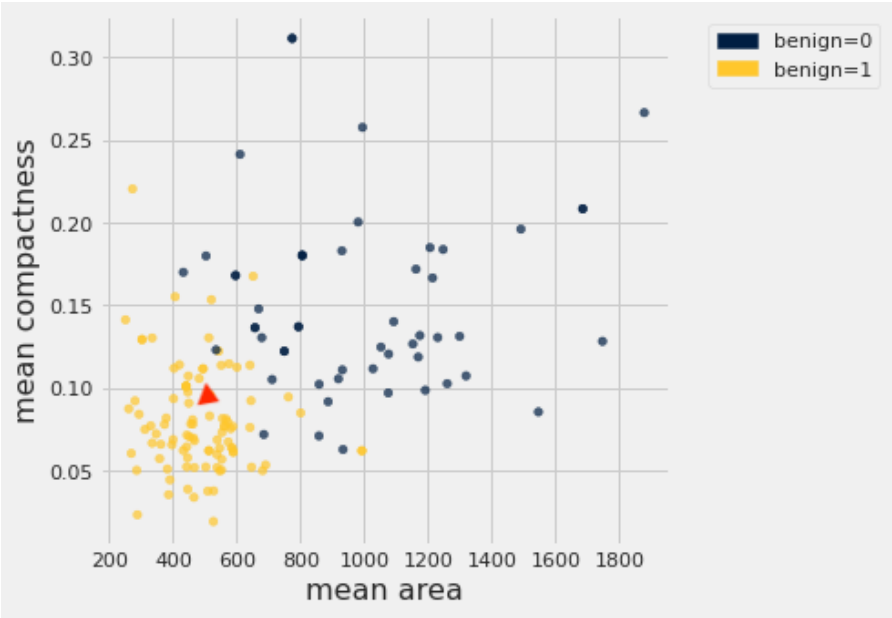
Question 1. Significant research has been done to understand whether a breast tumor is benign or malignant. Aarushi wants to create a classifier that predicts whether a tumor is benign or not.

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.



something like this

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.



map out the x and y coordinates of each point, use pythagorean's theorem and apply the distance formula to the entire table, sort by distance ascending, pick top 3 examples & determine whether a majority of those selected are benign or not; side with the majority in classifying the current point. This one is would be benign, as its 3 closest neighbors are.

c. Prasann suggests that Aarushi should use a different k for his classifier like k=4 or k=8. Is Prasann's suggestion reasonable?

no because then one would have to account for ties

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?

that is a good idea because it gives us "free" training data, it doesn't guarantee that our classifier will always be right (after all, if we train classifiers on training data, they will be right 100% of the time, which is not a good test of the classifier), and it prevents overfitting & other anomalies in some cases that could jeopardize the accuracy & consistency of the classifier.

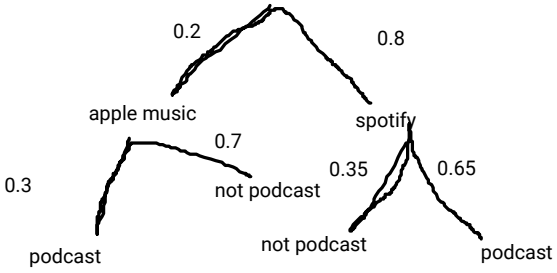
e. Suppose in our breast tumor training dataset we have 30 benign=0 data points and 45 benign=1 data points. What k values are too large?

A k value of 61 will ensure that tumors are always classified as benign. A k value greater than 30 would be too large because it would lose the ability to represent the minority class. In general, a k value of between 9 or less is a good idea because that is about the square root of 30 + 45.

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).



same thing as a normal tree but vertical

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

   0.2

   ii. Given that the student listens to podcasts, the probability that the student was a Spotify user

   (0.65*0.8)/(0.65*0.8 + 0.3*0.2) = 0.89655172413

   iii. Given that the student doesn't listen to podcasts, the probability that the student was an Apple Music user

   (0.7*0.2)/(0.7*0.2 + 0.35*0.8) = 0.333333333333

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?

No, because their internship at Spotify is a subjective prior that might increase the probability of them using Spotify, whereas Bayes's rule is based on random selection devoid of subjective priors.

d. (*Just for fun*) What should Sunny's podcast be called?

The Data.