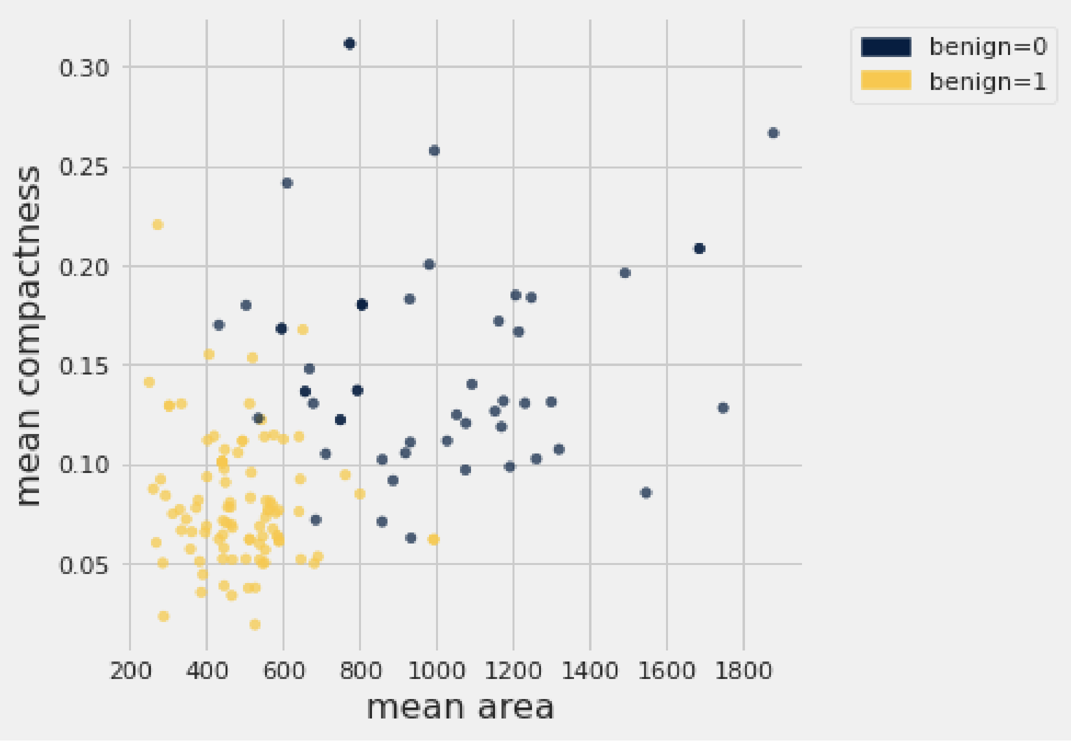
Data 8 Spring 2022

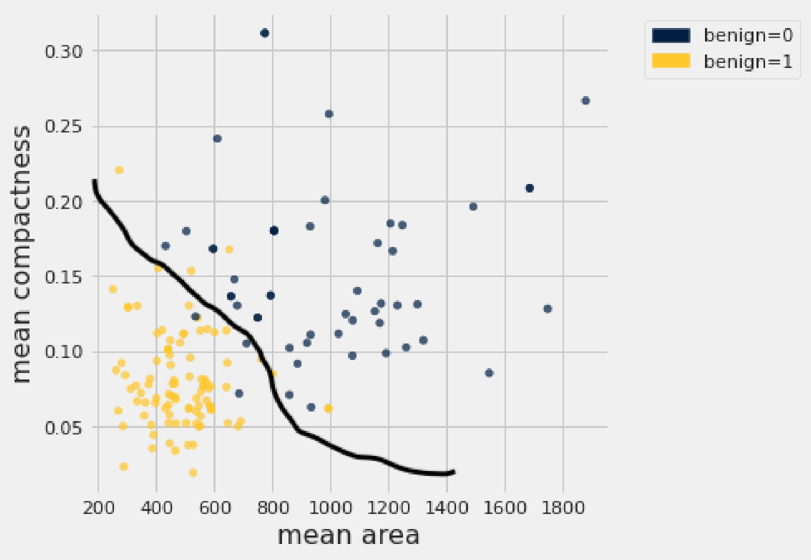
Lab 10: Classification, k-Nearest Neighbors and Conditional Probability

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.

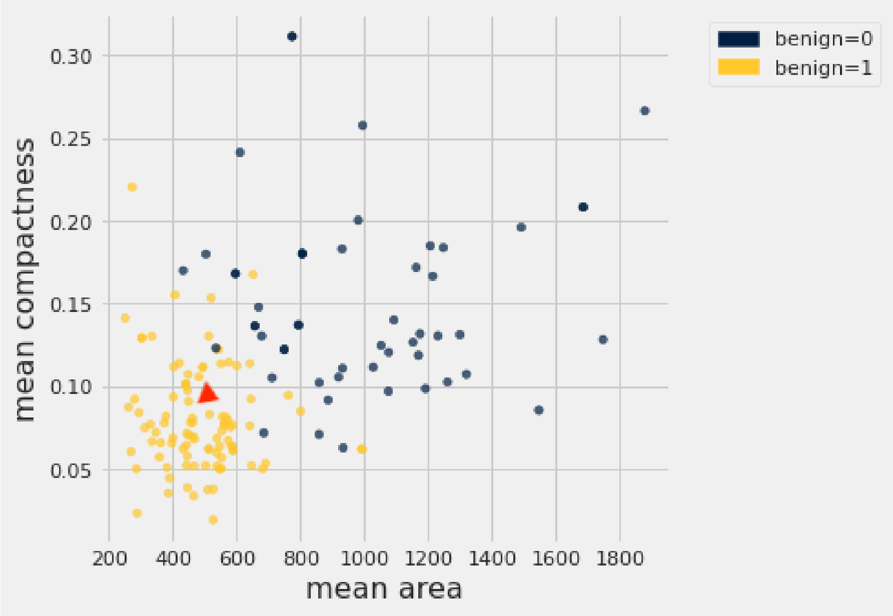




Answer:

A decision boundary is the plane or line that separates the classification of one class from another. It is a boundary such that if a new point falls on the right side of the boundary, it will be classified as 0 and 1 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 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.



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 1.

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

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.

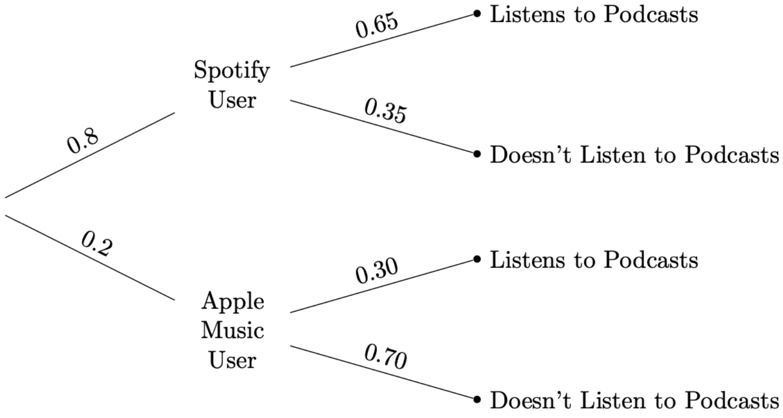
e. Suppose in our breast tumor training dataset we have 30 benign=0 data points and 45 benign=1 data points. What k values are too large?

Answer:
Using overly large values of k will result in issues such as always predicting the same value. In this example, any k greater than or equal to 61 will *always predict* benign=1 no matter what. In this course though, we will not go into detail as to how to choose an optimal k value.

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

P(Apple User) = 0.2; we can read this directly off of the tree diagram. Since the student is drawn at random from the population, there is a 20% chance we select an Apple user.

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

P(Spotify User | Listens to Podcasts) = (0.8 * 0.65) / [(0.8 * 0.65) + (0.2 * 0.3)] = 0.896551724; We use Bayes' rule to calculate the probability that the student was a Spotify user given that they listened to podcasts. Our prior probability is **0.8**, likelihood is **0.65**, and total probability of listening to podcasts is **(0.8)*(0.65) + (0.2)*(0.3).**

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

P(Apple User | Doesn't Listen to Podcasts) = (0.2 * 0.7) / [(0.2 * 0.7) + (0.8 * 0.35)] = ⅓; The calculation for part (iii) is identical to that of part (ii), only switching the values of our prior, likelihood, and total probabilities.

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, these calculations assume that we select a member of the population **uniformly at random**, which is not the case here. We would expect the probability that she uses Spotify to be higher than the rest of the population.

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

It's Always Sunny Everywhere I Go