Cancer Diagnosis

The Data

In this lab, you will train and evaluate a classification algorithm to determine whether or not a fine needle aspiration biopsy is cancerous (malignant) or non-cancerous (benign). The data for this lab are a lightly processed version of the data found at the UC Irvine Machine Learning Repository.

Questions

Question 1

What is the unit of observation in the data frame?

Question 2

We will be fitting models to output a diagnosis ("M" or "B", for "malignant" and "benign", respectively). Will the model predict the probability of being benign or malignant by default? *Hint: remember the concept of the reference level!*

Question 3

part a

Write `ggplot2` code to visualize the distributions of `radius_mean` for malignant vs. benign biopsies.

part b

Describe the relationship between the two distributions in at least one sentence.

Question 4

part a

Write `ggplot2` code to visualize the association between `radius_mean` and `area_mean`.

part b

Modify your code to color the points on your visualization by `diagnosis`.

part c

Describe how the malignant and benign biopsies differ as seen on the visualization.

Question 5

Split the data set into a roughly 80-20 train-test set split (80 percent of the data to the training set, 20 percent of the data to the testing set).

Question 6

Using the training data, fit a simple logistic regression model that predicts the diagnosis using the mean of the texture index.

Question 7

Consider a new biopsy with a mean texture of 15.

part a

What does your model assign as the probability that the cancer is malignant?

part b

Using a probability threshold of 0.5, what *outcome* would your model predict for this biopsy (benign or malignant)?

Question 8

Calculate and report two misclassification rates for your simple model: first on the training data and then on the testing data.

Question 9

part a

What can you change about your classification rule to lower the number of false negatives?

part b

Make the change you identified in the previous question and calculate the new number of false negatives.

part c

Calculate the testing misclassification rate using your new classification rule.

Question 10

In many realms of medicine, classification algorithms can be more accurate than the most well-trained medical doctors. What is gained and what is lost by shifting to algorithmic diagnoses (when whatever diagnosis the model outputs is treated as official)? Answer in at least three sentences.