

Math 5364 Homework 4

1. Produce a data set called `exdata` similar to the one on p. 42 of the Chapter 4 slides.
 - There are 1800 black points, whose x and y coordinates were generated from a uniform distribution on $[0, 20]$. (Hint: `runif` command.)
 - There are three sets with 400 red points each that were obtained from normal distributions, with these parameters: $(\mu_X, \mu_Y) = (10, 5)$, $(\mu_X, \mu_Y) = (5, 15)$, and $(\mu_X, \mu_Y) = (15, 15)$. All of these normal distributions have $\sigma_X = \sigma_Y = 2$. (Hint: `rnorm` command.)
 - The dataframe `exdata` should have three columns, `x`, `y`, and `class`.
 - Class labels for the black points are 0, and the labels for the red points are 1.
2. This problem reconstructs some of the results from the slides for `exdata`.
 - (a) Create a scatterplot for `exdata` like the one on p. 42 of the slides.
 - (b) Randomly split `exdata` into a training set containing 30% of the records, and a test set containing 70% of the records.
 - (c) Use `rpart` to fit a decision tree called `extree` to the training data, and find the training error and testing error for this tree. Also, plot `extree` with the `plot` command.
 - (d) Construct trees with `maxdepth = 1, 2, \dots, 6`. For each tree, store its training error, test error, and number of nodes in a matrix (note that you can expedite this task with a `for-loop`).
 - (e) Construct trees with `minsplit = 1` and `cp = 10^{-2.0}, 10^{-2.1}, 10^{-2.2}, \dots, 10^{-2.9}, 10^{-3.0}`, and store their training error, test error, and number of nodes in a matrix.
 - (f) Use the information in this matrix to reproduce the plot of training/test error vs. number of nodes, as given on p. 54 of the slides.
 - (g) Let `extree2` be the tree with `minsplit = 1` and `cp = 10^{-3.0}`, and plot this tree.
 - (h) Use McNemar's test to determine if there is a statistically significant difference between the accuracies of `extree` and `extree2`.
3. Write a function called `zcritical` that accepts inputs `alpha` and `numtails` and returns z_α when `numtails=1` and $z_{\alpha/2}$ when `numtails=2`. Verify that `zcritical(0.05, 2)=1.96` and `zcritical(0.05, 1)=1.645`. The `qnorm` function will be helpful on this problem.
4. Write a function called `accuracyconfint` that accepts inputs `accuracy`, `n`, and `alpha` and returns the confidence interval for accuracy given on p. 62 of the slides. Use this function to find a confidence interval for the accuracy of `extree`.
5. Randomly split `kyphosis` into about 70% training data and about 30% test data.
 - (a) Use `rpart` to fit a tree called `ktree` to the training data, and find the training and test error rates.
 - (b) Find an exact binomial confidence interval for the accuracy of `ktree`.