Practice problems for logistic regression with partial answers

1. Read Swets et al. (*Scientific American*, 2000)
2. This problem builds on the data simulation problem from the discriminant analysis homework.
	1. Simulate 100 observations each from two populations with the following distributions:

 and 

Set a seed number of 1181 before using the first distribution and a seed number of 1218 before using the second distribution. Plot the simulated data in a scatter plot where the plotting symbols correspond to the population. Based on this plot, how well do you think logistic regression will separate out the observations into the correct populations?

> library(mvtnorm)

> set.seed(1181)

> pop1 <- rmvnorm(n = 100, mean = c(0,0), sigma = matrix(data = c(1, 0, 0, 1), nrow = 2, ncol = 2))

> head(pop1)

 [,1] [,2]

[1,] -0.2962019 -1.02229182

[2,] -0.2277220 2.09802300

[3,] 1.8393558 0.29361539

[4,] -0.7641729 -1.16671752

[5,] -0.2279988 0.05907115

[6,] -0.5433259 0.61151723

> set.seed(1218)

> pop2 <- rmvnorm(n = 100, mean = c(6,0), sigma = matrix(data = c(1, 0, 0, 1), nrow = 2, ncol = 2))

> head(pop2)

 [,1] [,2]

[1,] 6.005221 -0.264053889

[2,] 6.013727 0.594869118

[3,] 7.351746 0.003988692

[4,] 8.637141 0.778251779

[5,] 6.447020 0.211541760

[6,] 6.298675 -0.635608118

> dev.new()

> par(pty = "s")

> plot(x = pop1[,1], y = pop1[,2], xlab = "Variable #1", ylab = "Variable #2", pch

 = 1, col = "blue", xlim = c(-5, 10), ylim = c(-5, 10), panel.first = grid())

> points(x = pop2[,1], y = pop2[,2], xlab = "Variable #1", ylab = "Variable #2",

 pch = 2, col = "red")

> legend(x = 0, y = 8, legend = c("Pop. #1", "Pop. #2"), col = c("blue", "red"),

 pch = c(1, 2), bty = "n")



The two groups of points are fairly well separated so I would expect logistic regression to have difficulty estimating the model due to complete separation!

* 1. Combine the two sets of data into one data frame and estimate a logistic regression model with both variables in it. Use cross-validation with a cutoff probability of 0.5 to assess the accuracy of the classification method.

In order for glm() to estimate the model, I changed the population numbers to 0 and 1 from 1 and 2, respectively.

> set1 <- rbind(data.frame(pop1, pop = 0), data.frame(pop2, pop = 1))

> head(set1)

 X1 X2 pop

1 -0.2962019 -1.02229182 0

2 -0.2277220 2.09802300 0

3 1.8393558 0.29361539 0

4 -0.7641729 -1.16671752 0

5 -0.2279988 0.05907115 0

6 -0.5433259 0.61151723 0

Note that the model can not be estimated due to compete separation! Therefore, one cannot assess the accuracy of the logistic regression classification method here.

* 1. Repeat a) and b) with the variable #1 mean for population #2 being decreased from 6 to 5, 4, 3, and 2. Describe trends you see as the mean decreases. Also, examine the ROC curve for these cases too.

I received warnings messages for means of 4 and 5. For means of 2 and 3, I did not receive any warning messages. Not surprisingly, the classification accuracy decreases as the mean decreases from 3 and 2 (the two samples overlap more as the mean decreases).

Note that I always used a seed number of 1218 before simulating my data for the second population. The same seed number was used here only to keep this simulations simple. In practice, one should always use a different seed number before simulating different data sets.

ROC curves for mean = 3 (left) and mean = 2 (right)

 

As the samples overlap more (means get smaller), the ROC curves become closer to the line of chance accuracy. This is to be expected because it becomes more difficult to differentiate between the populations.

* 1. For the mean of 2 case given in c), construct a scatter plot showing the observations with their original and classified populations. Discuss which observations are misclassified.

Here’s my plot for a mean of 2 (smaller symbols denote the original populations and larger symbols denote the classifications):



* 1. Examine what happens if the cutoff probability is changed to 0.8 for the mean = 2 case.

The specificity greatly increases at a cost of a lower sensitivity.



Notice that more of the larger blue circles are on the right side than were before.

* 1. Compare your results to those obtained from the DA and NNC homework.

For mean = 2: The overall accuracy for the logistic regression models (cutoff = 0.5) is 0.845 (using my seed numbers and cross-validation). This is similar to the other methods (for NNC - depends on K chosen).