Name: Byung Hun lee

1.Use \*\*logistic + lasso regression\*\* with \*\*10-fold cross-validation\*\* to

identify useful predictors. Plot a lasso plot with readable tick labels on the X and

Y coordinates in your plot for easy visualization and verification. Missing clear

and readable tick labels in your plot will cost you significant points for this

assignment.

Instead of fitting line to the data, logistic regression fits an “S” shaped “Logistic Function”. If each x (predictors) is large enough number after standardization and converting binary value of target variable, there is high probability that the target indicate the bacterium (rows) that are interesting enough for further study. Likewise, there are only a small probability that a small number of predictors response to the target indicating the bacterium that are interesting enough for further study. Although logistic regression tells that probability that a mouse is obese or not, it is usually used for classification. For example, if the probability whether the bacterium interesting enough for further study or not is > 50%, then we will classify it as interesting enough for further study, otherwise we will classify it as not interesting enough for further study.

For this assignment 5, target variable coverted to binary variable was predicted by x0~x12 predictors.

With lasso regression, we also tried to find best lasso regression model using different alpha and these below.

for ind, i in enumerate(lambdas):

reg = Lasso(alpha = i)

reg.fit(x\_train, y\_train)

coeff\_a[ind,:] = reg.coef\_

train\_r\_squared[ind] = reg.score(x\_train, y\_train)

test\_r\_squared[ind] = reg.score(x\_test, y\_test)

2. Which top \*\*THREE (3)\*\* predictors are you going to select to explain why a

bacteria is an “interesting” candidates for further study?

X1, x10, x11 are top three predictors because i tested to see if a variable’s effect on the prediction is significantly different from 0. If not, it means the variable is not helping the prediction. In this case, these top 3 predictors are useful for classifying samples. The z-value is the number of standard deviations the estimate is from 0 on standard normal curve. The estimate is less than 1 standard deviation from 0 so the other predictor is not statistically as significant as top 3 predictors.

array([[-0.1119313 , 0.00544534, -0.11908434, 1.95024794, -3.09050253,

1.64511823, -0.15444238, -0.14349965, 0.44362345, 0.7913768 ,

0.09720706, -0.11849826, 0.43719503]])

print(log\_reg.coef\_)

print(log\_reg.intercept\_)

[[-0.28788025 0.06454795 -0.10178408 2.24973594 -3.07328242 1.61082795

-0.22587435 -0.31696034 0.57513313 1.05870398 0.10624406 0.02116713

0.49229669]]

[-4.56225241]

3. What is the lambda (l) value you choose in order to select the top 3 predictors

you identified in the last question?

Lambda l = 0.05

4. What are the q values for the 3 selected predictors at the lambda (l) value you

identified in the last question?

0.31, 0.036 according to my confusion matrix

# True positive: 245 (upper-left) – Number of positives we predicted correctly

# True negative: 36(lower-right) – Number of negatives we predicted correctly

# False positive: 15 (top-right) – Number of positives we predicted wrongly

# False negative: 9(lower-left) – Number of negatives we predicted wrongly