SSIE 637 Advanced Topics in Healthcare
Assignment 03
Fall 2016

Instructor: *Daehan Won*

Due 11/15/2016

Part 1 Short answers
P1-Question 1 (15 %)
In a hospital, some patient features are requiring higher cost to collect (e.g., body or brain scans) whereas others are not (e.g., temperature, blood pressure). Therefore, we have decided to first ask our classification algorithm to predict whether a patient has a disease, and if the classifier is 75% confident that the patient has a disease, then we will do additional examinations to collect additional patient features. In this case, which classification methods do you recommend: logistics regression, decision tree, or naive Bayes? Justify your answer in one or two sentences.

P1-Question 2 (5 %)
Consider a logistic regression model with weights $\beta = (-\ln(4), -\ln(2), \ln(3))$ and label $y \in \{0, 1\}$. A given data has feature vector $x = (1, 1, 1)$. What is the probability that the given data is $y = 1$ i.e., $(P(Y = 1|x))$? (Hint. See the page 19 in our lecture note 7th)

Part 2 Implementation algorithms
P2-Question 1 (30%)
Given data set "random_classification.txt", apply following algorithms.
Note that you should use first 400 samples as training and remaining 100 samples as testing. Note that the last column is class variable
(1) Gaussian Naive Bayes
(2) Decision Tree
(3) Logistic Regression
Report prediction accuracy, sensitivity, and specificity of your prediction results. Which one is the best?

P2-Question 2 (30%)
Apply following algorithms in Q1 to "breast_cancer" data.
(1) Gaussian Naive Bayes
(2) Logistic Regression
(3) SVM
Use first 500 samples as training and remaining 199 samples as testing.
Report prediction accuracy, sensitivity, and specificity of your prediction results.

P2-Question 3 (20%)
Given data set "lasso.txt", only a few features are actually corresponding to the output value (the last column). Applying LASSO and determine; the number of relevant features and their indice. (Hint, train the lasso model and obtain coefficients for each variable)

Part 3 Optional problems
P3-Question 1 (10%)
Consider 5-fold cross validation for P2-Q1 and P2-Q2 and report accuracy, sensitivity, and specificity.