26 BMIN 7054 + 20 CS 7053 + 26 BE 9063 001

Data Science for Biomedical Informatics

Spring 2018

MB Rao

Homework Sheet No. 3 Due Date: February 22, 2018 Maximum points: 30

Theme: Cross-validation in Logistic Regression.

Download the data “BreastCancer” from the “mlbench” package.

There is one response variable and ten predictors. The response variable is binary (benign, malignant). The data came from University of Wisconsin Medical Hospital. The goal is to develop a simple diagnostic tool for breast cancer. There are some missing observations.

Questions

1. Write down the dimensions of the data.
2. Show the top six rows of the data.
3. Describe the data.
4. Get the summary statistics of the data.
5. Postulate a logistic regression model.
6. Fit the model.
7. Comment on the coefficients.
8. Check goodness-of-fit.
9. Obtain the confusion matrix.
10. Calculate the misclassification rate.
11. Check model accuracy using 10-fold cross validation as well as LOOCV.
12. Comment on what you did for Question 11.