Jingning Zheng

920440765

SFSU-MATH449-01

Professor:Dr. Alexandra Piryatinska

Math449 project

Identification of Cancer Cells

The data comes from Kaggle, which is 570 cancer cells and 30 features to determine whether the cancer cells in the data are benign or malignant. Contains a response variable, a cell id, and 30 independent variables.

First, organize the data, remove the id and a meaningless variable, and then add a variable y that is the same as the response variable, replacing it with 1 and 0, 1 represents evil, and 0 represents benign. Check whether the data contains NA or null values.

Fit three models, namely the null model, the full model, and the model to be used. From the results of the full model, it can be known that the independent variable concave.points_mean has no meaning, so it is preferentially excluded. The new model contains all variables except concave.points_mean, the result is that all variables are meaningful, and this model is used as the final model.

Leave one out cross-validation to test the accuracy of the model. From the results, the accuracy rate is 0.94, which is about 94%, which shows that the model's accuracy for the data used is relatively high. Use K-fold cross-validation to verify the model, set k to 10, and the results are very close after five tests.

From the ROC graph, the applicability of the model is extremely high, which also shows that the model is correct. The Correlation value is also relatively close to 1.

From the results of the model, whether cancer cells are malignant or benign is related to almost all variables; only concave.points_mean has a poor correlation with cancer cells. Therefore, the error rate of this data is low.