STAT3612 Homework 3: Binary Classification

Date: November 24, 2020

Sumbit (in the ipynb format) via Moodle before 11:59pm December 6, 2020.

Use the `load_breast_cancer()` from `sklearn.datasets` to get a copy of the breast cancer (diagnostic) data with 569 instances and 30 numeric predictive attributes. The binary responses include 212 Malignant and 357 Benign cases. Use `train_test_split` to divide the data into 80% training data and 20% testing data, then perform the following machine learning tasks.

Step 1. (20%) Fit a decision tree classifier with `max_depth =3`. Visualize the fitted tree by `export_graphviz`. Report the training and testing accuracy.

Step 2. (20%) Fit the random forests and gradient boosting machines. Report the training and testing accuracy for both models.

Step 3. (20%) Fit support vector classifiers with linear and RBF kernels. Report the training and testing accuracy for both models.

Step 4. (20%) Fit a neural network with two hidden layers each having 40 nodes. Report the training and testing accuracy.

Step 5. (20%) From the above model fits, pick the one with the best testing accuracy. Run post-hoc analysis for model interpretation in terms of a) feature importance and b) partial dependence plots of the 5 leading important features.