Stone Barrett

CSE 590: Python Data Analytics

Homework 4

1)

For this part, I chose to use the breast cancer dataset from sklearn. The three classifiers I chose to use are Linear Support Vector Machine (LSVM), K-Nearest Neighbors (KNN), and Random Forest (RF). The average of cross-validation scores for each classifier + k-fold value combination are printed in this order:

LSVM, cv = 2

KNN, cv = 2

RF, cv = 2

LSVM, cv = 20

KNN, cv = 20

RF, cv = 20

Both LSVM and KNN saw improvements in their average score when increasing the k-fold splits from 2 to 20. RF, however, stayed the same at 96%.

The most accurate of the models in this case was RF, so that is what I used for the confusion matrix. The matrix showed that there were 50 true positives, 5 false negatives, 3 false positives, and 85 true negatives. This means 90.9% of positives were detected and 96.6% of negatives were detected.

NOTE: the script suppresses warnings because LSVM was throwing warnings for failed convergence. Normally, I would just increase max\_iter by a few orders of magnitude but for this dataset that seemingly was not working. The scores should still be accurate.

2 – option 2)

For this part, I used mglearn’s extended Boston housing dataset. The three regression methods I chose are Linear Regression (LR), Ridge, and Lasso. The average of cross-validation scores for each regression method + scoring metric are printed in this order:

LR, r2

Ridge, r2

Lasso, r2

LR, -MSE

Ridge, -MSE

Lasso, -MSE

All three regressors performed poorly on this dataset, but the highest r2 score was from Ridge. That also happens to be the highest -MSE value, meaning Ridge performed the best by both metrics.

R2 acts as a “goodness of fit” score to show how accurately the predicted scores line up with the true values. MSE acts as a score tallying the penalties, or in this case the differences between predicted and true values. MSE exaggerates the faults of the model such that scores less than ideal are heavily scrutinized.