Homework 4
COMP 379
Brian Nguyen
11/14/2021

This program is meant to implement k-fold and grid search cross-validation (CV) techniques on a linear regression (LR) model to solve a classification problem in the form of a non-linearly separable dataset, comparing their performances (by accuracies and macro F1 scores) in analyzing the dataset with the holdout method.

The chosen dataset, *wine.data*, was preprocessed by randomly splitting it into a pair of subsets, a training (80%) and testing (15%) subset. These subsets were then standardized (after splitting to prevent data leakage) to ensure all features contribute equally to the models' fitting and learning functions. The chosen metrics to evaluate the prediction performances of the models using k-fold and grid-search CV included the individual, maximum, and mean accuracy rates of each technique's iteration observed. For the analysis of the LR model using grid search CV, macro F1 scores were also calculated to accommodate for potential imbalanced class distributions affecting the learning functions of the models.

To implement and test k-fold CV on an LR model, it was fitted using the standardized training subset and evaluated on the same subset. After testing across a range of five different k-values, it was found that the models with 5 and 10 folds performed the best, yielding both maximum and mean accuracy rates of 100% on the standardized testing subset (see *Fig. 1* of the Appendix).

To implement and test gid search CV on an LR model, it was fitted using the standardized training subset and evaluated on the same subset over a search of the LR model's C-value and penalty hyperparameters. The c-values searched over included 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, and 100000. The penalty parameters searched over included the terms L1, L2, Elastic-Net, and no penalty term. The LR model for all searches also utilized the 'saga' solver (as this is the only solver that supports all four penalty parameters) and an L1 ratio of 1. After analyzing the prediction performances of 40 hyperparameter combinations, it was found that all combinations with either no penalty parameter or at least a C-value of 1 yielded accuracy rates of 100%. Of all the combinations analyzed, there was a mean accuracy rate of 83.00%.

To evaluate the prediction performance the best grid search CV hyperparameter combination found on an LR model, it was adjusted to utilize a C-value of 0.00001 and no penalty parameter, the first hyperparameter combination with perfect metrics. The LR model was fitted using the standardized training subset and evaluated on the standardized testing subset. It was found that this model yielded both an accuracy rate and macro F1 score of 100%. When compared to a model fitted and evaluated using the standardized training subset, they had the same perfect metrics. Theoretically, fitting and evaluating on the same training set will always yield 100% because it is essentially memorizing the dataset, but fitting on a training set and evaluating on a testing set (that the model has not seen before) is not guaranteed to perform as well. Utilizing the holdout method is a way to compare and identify which models have the lowest generalization error and performs the best on future/unseen data. This grid search CV hyperparameter combination on the LR model happened to perform perfectly on the classification dataset it was evaluated upon.

Appendix

LR Model w/ k-fold CV Prediction Performances on Standardized Testing Subset		
k-value	**Max Accuracy Rate**	**Mean Accuracy Rate**
2	100	97.89
3	100	98.58
4	100	99.29
5	100	100
10	100	100

Fig. 1: LR Model w/ k-fold CV Prediction Performances on Standardized Testing Subset