Caroline Rinks
COSC 423

Support Vector Machine:

I. Introduction

This project involves a binary classification problem with the mushrooms dataset, found in the file 'mushrooms.csv', which uses a Support Vector Machine to classify mushrooms as edible (e) or poisonous (p) given 22 features: class, cap shape, cap surface, cap color, bruises, odor, gill attachment, gill spacing, gill size, gill color, stalk shape, stalk root, stalk-surface above ring, stalk surface below ring, stalk color above ring, stalk color below ring, veil type, veil color, ring number, ring type, spore print color, population, and habitat.

II. Pre-Processing

The dataset is read into a pandas dataframe and then scanned for any invalid instances that need to be removed from the dataframe (e.g. question-marks). Then the textual labels for each of the features are encoded into a numerical format by creating a dictionary that maps numeric labels to the textual labels. Finally, the data is shuffled and split into 80% training data and 20% testing data.

III. Grid Search

The Coarse Grid Search explores the following hyperparameters: the Kernel hyperparameter with Linear, Sigmoid, Poly, and RBF configurations; the Soft-Margin Constant C with values 10 and 1000; the degree hyperparameter with values 2, 3, and 4; and the gamma hyperparameter with values 0.001 and 0.0001. The table below represents the different hyperparameter configurations and their observed accuracies (highlighted in yellow). Each model was evaluated during the grid search by K-Fold Cross Validation, where k=5.

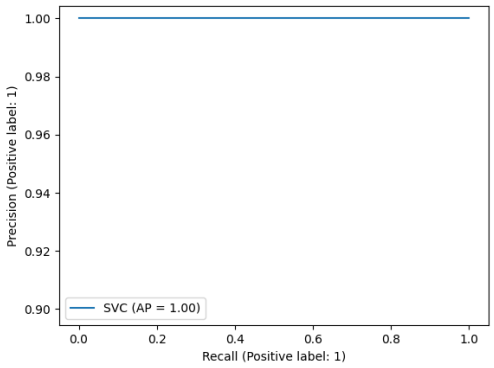
			Degree			**Gamma**	
			2	3	4	0.001	0.0001
Linear	C=10	0.994906					
	C=1000	1.000000					
Sigmoid	C=10	0.567220					
	C=1000	0.566777					
Poly	C=10		1.000000	1.000000	1.000000		
	C=1000		1.000000	1.000000	1.000000		
RBF	C=10					0.957032	0.924031
	C=1000					1.000000	0.980509

The Fine Grid Search explores Linear and Poly Kernel configurations; 1, 100, 250, 500, 750, and 1000 for the C constant; and 3 and 4 for the degree. The resulting accuracies for all possible combinations of these configurations were equal to 1.000000, except for the configurations {'C': 1, 'kernel': 'linear'} and {'C': 1, 'degree': 3, 'kernel': 'poly'} which resulted in an accuracy of 0.987597 and 0.999557, respectfully.

IV. Best-Performing Models

Several models tied with an accuracy of 1.000000. Only one will be examined now, so see section III for other models that performed with the highest accuracy. The Linear kernel model with a C of 100 performed with an accuracy, precision, and recall of 1.000000. A Precision-Recall Plot for the model is shown below.

Caroline Rinks
COSC 423



Neural Networks

I. Introduction

The same binary classification problem and dataset described above in the Support Vector Machine Introduction applies but with a different model: a Neural Network.

II. Pre-Processing

The dataset is read into a pandas dataframe and then scanned for any invalid instances that need to be removed from the dataframe (e.g. question-marks). The textual labels for each feature are encoded into a numerical format by creating a dictionary that maps numeric labels to the textual labels. Finally, the data is shuffled and split into 80% training and 20% testing.

III. Grid Search

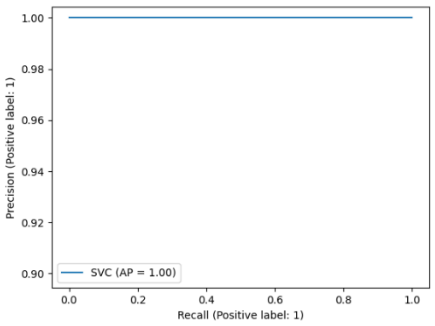
The Coarse Grid Search explores a Sequential Keras model with an input layer, one hidden layer, and an output layer. The other hyperparameters explored are the number of neurons (1, 10, 20, or 30) and the activation function (Linear, Sigmoid, Relu, and Tanh). The table below represents the different hyperparameter configurations and their observed accuracies. K-Fold Cross Validation is implemented with k=3 to reduce computation time.

	Neurons=1	**Neurons=10**	**Neurons=20**	**Neurons=30**
Linear	0.950609	0.968771	0.973643	0.978516
Sigmoid	0.952159	0.968992	0.972757	0.979402
Relu	0.835659	0.999779	1.000000	1.000000
Tanh	0.969435	0.999779	1.000000	1.000000

The Fine Grid Search explores a Sequential Keras model with an input layer, two hidden layers, and an output layer. The other hyperparameters explored are the number of neurons (20, 25, or 30) and the activation function (relu or tanh). All configurations explored in the Fine Grid Search resulted in an accuracy of 1.00000.

IV. Best-Performing Models

One Best Performing Model will be examined. A model with two hidden layers, each with 20 neurons and a relu activation function performed with an accuracy, precision, and recall of 1.00000. A Precision-Recall plot for the model is shown to the right.



V. Conclusion

NN's appear to be more accurate than SVM's for this binary classification problem. However, accuracy can be improved by exploring various configurations for each model's hyperparameters.