SVM and Neural Net Evaluation

COSC 423 Project 3 Fall Semester

By Josh Bridges

SVM

Introduction

For this project the mushroom.csv dataset was used to predict whether a particular fungus is poisonous or edible. The task is to test different types of support vector machines (SVMs), to evaluate the best model for classifying this dataset. Three SVM kernels were evaluated: linear, polynomial (poly), and radial basis function (rbf).

Pre-Processing

The dataset required a few steps of pre-processing in order to work with the scikit-learn and Keras modules. Initially, the dataset contained multiple lines with '?' characters where unknown data was. These lines were removed from the set. Additionally, the dataset was presented categorically, with characters representing different values. For the models to function this data had to be converted to ordinal values. The features were encoded using the One-Hot Encoding method which expanded the feature list from 22 to 93 features. Then the targets were encoded using Label Encoding since there were two viable options one set to zero and the other to one. With these changes the data was now usable in the models.

Grid Search

The evaluation process started with a coarse-grid search of the hyperparameter configurations. In this search three different kernels were evaluated: linear, poly, and rbf. Each kernel evaluated a cost hyperparameter from 1-1000. The poly and rbf kernels both had unique hyperparameters. For the poly kernel the degree was varied between three values: two, three, and four, while the rbf kernel's gamma hyperparameter was evaluated at 1e-3 and 1e-4. This broad search of the hyperparameters was evaluated each over a seven-fold cross validation to randomly shuffle the training and test data. The resulting accuracy of each model was averaged over the seven folds and the model with the best accuracy was explored in a fine grid search.

After the coarse grid search identified a model with the polynomial kernel, a cost parameter of one, and a degree of three to be the best performing, the fine-grid search explored the polynomial kernel's hyperparameters further. The degree was evaluated from one to six and the cost hyperparameter was varied from 0.1 to 10000, in a logarithmic scale. The models were again evaluated over a seven-fold cross validation sequence with the resulting accuracies averaged. The fine grid search here again pointed to the polynomial kernel with a degree of three and cost of one to be the best performing model.

Best-Performing Model

The best performing model for the SVM that this study resulted in was the model with a polynomial kernel, a degree of three, and a cost weight of one. This model was then further explored by evaluating the metrics of accuracy, precision, and recall of the model of different K number of folds. K was varied from a range of seven to twelve and generated the precision and

recall plot shown in figure 1. The accuracy was observed to increase past the 93.5% observed at seven folds to an accuracy of 98% at folds nine through twelve. Figure 1 shows the results of a precision and recall curve generated with the best performing model on a single split of the data. The horizontal nature of the figure indicates that the model is performing exceptionally well. However, this could be a fluke and not representative of the general performance. These results require more observation to be further analyzed.

Neural Net

Introduction

The mushroom dataset was used to additionally identify which neural network configuration best works to classify the data. The neural net was set up with one input layer, one hidden layer and one output layer. Four different activation functions were analyzed: linear, sigmoid, rectified linear unit (relu), and hyperbolic tangent (tanh).

Pre-Processing

The same preprocessing procedure that was done on the SVM instance was also performed on the dataset for working with neural nets. The rows with unknown data were removed and the data was converted to ordinal values using one-hot encoding on the feature set and label encoding on the target values.

Grid Search

The coarse grid search for the neural net evaluated each type of activation function for the input and hidden layer. The neurons in those two layers were also varied with neuron counts of 1, 2, 5, 10, 15, 20, 25, 30. For each configuration of activation function and neurons, three-fold cross validation was performed to evaluate the accuracy of the model. More numbers of folds would potentially lead to higher accuracy but were unable to be performed due to hardware limitations. The coarse grid search reported that the best neural net configuration, with an average of 92% accuracy, was a linear activation function for the input layer with a neuron count of 15 and a relu activation function for the hidden layer with a neuron count of 30.

The fine grid search then further explored the neuron counts for the indicated activation functions. The activation functions were set to the linear and relu respectively and the neurons were evaluated from the optimal values. The input layer was tested from a range of 10 to 20 by increments of one and the hidden layer used a range of 25 to 35 by increments of one. Three-fold cross validation was also utilized in this step to find the average of different folds. The fine grid search resulted in the identification of new optimal neuron counts of 17 for the input layer and 34 for the hidden layer that created a model with 95.9% accuracy.

Best-Performing Model

The best performing neural net was run over a single splitting of the dataset to generate Figure 2. This chart showed the same results as Figure 1. These results are not what was expected from models but if correct show extreme success with the mushroom dataset. In observing the epochs of this model, the results were shown to begin with a high accuracy, precision, and recall then quickly converge to 100%.

Conclusion

The model's observed, the best SVM and neural net, showed extreme promise for classifying the mushroom dataset. The best performing SVM reached upwards of 98% accuracy and the neural net had an astonishing 100%. Secondarily, the precision and recall curves show that each model was returning all the results with the correct classification. If possible, it would be useful to further analyze the performance of each model with regards to precision and recall. The analysis of the results from this project indicates that under similar circumstances the best identified neural net outperforms the SVM.

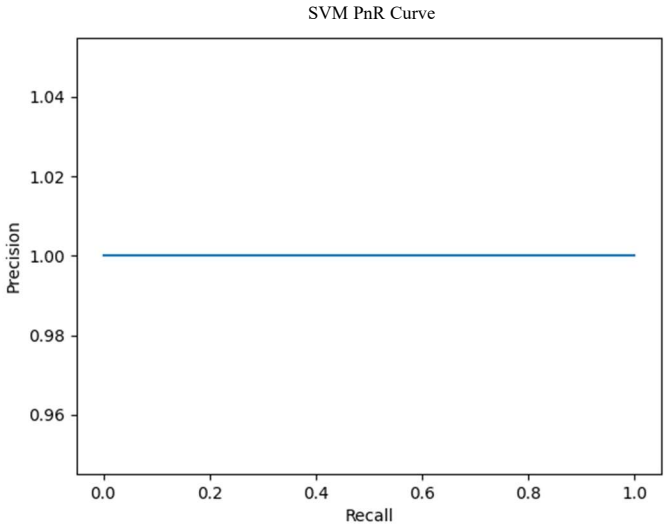

Figure 1. Precision and Recall plot for the best performing SVM

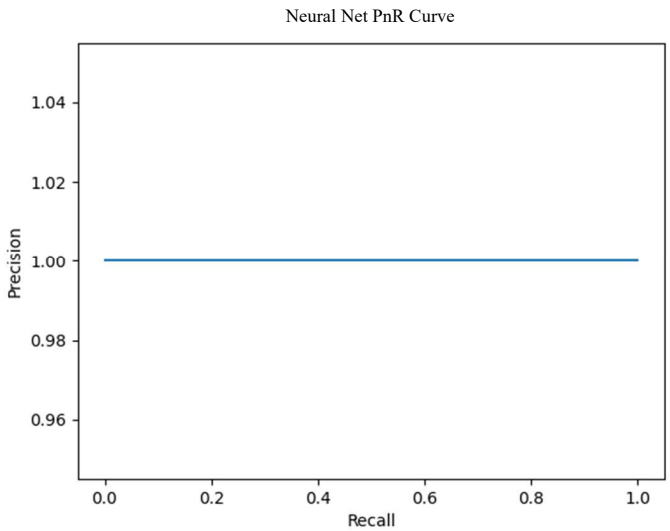

Figure 2. Precision and Recall plot for best performing neural net