Classifiers and Regressors

Diego A. Batista Cruz

CAP4613 Deep Learning for Computer Graphics

Video Link: https://youtu.be/kpnr5wM2rBQ

General Implementation

Classifiers

This repository contains K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and Linear Support Vector Machine (SVM) classifiers.

KNN.py

Functions and outputs necessary for KNN classification.

MLP.py

Functions and outputs necessary for MLP classification.

SVM.py

Functions and outputs necessary for linear SVM classification.

Regressors

This repository contains K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and Linear Support Vector Machine (SVM) regressors.

KNNRegressor.py

Functions and outputs necessary for KNN regression.

MLPRegressor.py

Functions and outputs necessary for MLP regression.

LinearRegression.py

Functions and outputs necessary for linear regression using normal equations.

Datasets

Repository containing the datasets provided for the project. These are:

- tictac_final.txt
- tictac_multi.txt

- tictac_single.txt

Classifiers_Regressonrs.py

Script serving as the Main Program that runs the analysis of all the different classifiers and regressors on the datasets needed to be tested.

Game.py

Script serving as the Main Program that runs the Human vs. Computer component.

TrainandPickle.py

Script serving as the main program that trains and pickled the model into the "trained_model.mdl" file. "trained_model.mdl" is used as the foundation of the tic-tac-toe game computer moves.

trained_model.mdl

Binary file storing the trained model.

Models

Linear SVM

An SVM is a supervised machine learning model used for classification tasks. This model constructs a hyperplane that separates the data into classes and classifies depending on proximity to this dataset. The test data set has been set to 20% of original dataset. A k-fold split was generated after to prevent the model from overfitting the data. From here the model was defined, fit, and evaluated

KNN

An KNN is a supervised learning model used for both classification and regression tasks by labeling the samples according to the nearest neighbors. To find the perfect number of neighbors different models were validated to determine several neighbors. A k-fold split was generated after to prevent the model from overfitting the data. From here the model was defined, fit, and evaluated.

Classifier: 10-fold split used to prevent overfitting. 1 neighbor used.

Regressor: 10-fold split used to prevent overfitting. 1 neighbor used.

MLP

MLP is a type of artificial neural network. For both the regressor and the classifier the data was split using a 10-fold split to prevent the data from being overfitted and increasing prediction accuracy. The following parameters were used for each of the models. From here the model was defined, fit, and evaluated.

Classifier: Two hidden layers were used with 256 nodes and one with 128 nodes.

Regressor: Two hidden layers were used with 256 nodes and one with 128 nodes.

Linear Regression

Linear Regression aims to fit a set of data points by drawing a plane through the data that generalized the values to an accurate regressor that can predict output variables. The train split was set to a 80/20 split and a 10-fold split to prevent overfitting.

$$Y_i = x_i\,\theta + e_i$$
$$\theta = (X^T X)^{-1} X^T y$$

Theta was calculated using the training dataset (X = x_train, Y = y_train). This parameter fits the slope of the model to the data. MSE is then calculated to evaluate the model.

Evaluation

Accuracy

	Accuracy		
Model	Tictactoe_final.txt	Tictactoe_single.txt	Tictactoe_multi.txt
SVM Classifier	0.9792	0.8162	N/A
KNN Classifier	1.0	0.8520	N/A
MLP Classifier	0.8802	0.9786	N/A
Linear Regressor	N/A	N/A	MSE = 0.1858
KNN Regressor	N/A	N/A	0.6224
MLP Regressor	N/A	N/A	0.7917

Best Model

The best model appears to be the KNN Classifier which successfully classified all the data for the final dataset. The single dataset was best tackled by the MLP Classifier. This will be

the recommended model for the tic-tac-toe computer. The MLP regressor appears to be the best model for the multi dataset.

Evaluation on $1/10^{th}$ of Data

	Accuracy		
Model	Tictactoe_final.txt	Tictactoe_single.txt	Tictactoe_multi.txt
SVM Classifier	0.9867	0.6651	N/A
KNN Classifier	0.9312	0.6875	N/A
MLP Classifier	0.5497	0.7894	N/A
Linear Regressor	N/A	N/A	MSE = 0.1756
KNN Regressor	N/A	N/A	0.3569
MLP Regressor	N/A	N/A	0.3476

Discussing Data Volume

The largest increase in accuracy when increasing the amount of data is most noticeable on the MLP. The other models do not change as much as the data decreases, especially relating classification. The regression models were more impacted by lower data volume than the classification models.

Certain models scale better to larger datasets than others. Small amounts of data trained model that lacks the ability to accurately predict values. Statistically, larger datasets will generate smaller variance and better predictive performance. The architecture of KNN and SVM/Linear Regression noise is very disruptive, however, as long as the data follows a pattern that is not disruptive to the overall shape of the data will still generate predictions that are accurate. The structure of an MLP model depends largely on the number of connections made between layers which increase the accuracy of the predictions generated by the model.

Diego A. Batista Cruz (UFID: 9487-1119)
dbatistacruz@ufl.edu