ANIMESH SHARMA: 19901680

THIS IS MY OWN WORK. I HAVEN'T COPIED FROM ANYONE ELSE

A dataset of Boston houses was provided with all the input variables determining the median value of the owner-occupied homes. The dataset had 506 rows and 14 columns which included the median value of owner-occupied homes.

Regression was run for the given dataset with the model having the first 13 columns as input variables and the median value as the response variable. Three different models were used namely:

1. Linear Regression
2. K nearest neighbours regression
3. Support vector machine regression

For all the models, 20% of the data was allocated to test and the remaining was allocated to train.

RMSE was used as the metric for system performance evaluation with RMSE being the most commonly used and most reliable.

PROCESSING DATA

However, before running regression on the three mentioned models, the data was processed to ensure better accuracy and efficiency. The outliers were removed from all the numerical columns if there were any. Inter quartile range was used as the parameter to detect outliers and eventually, remove them. It turned out that after removing the outliers, the size of the data was reduced from 506 observations to 466 observations.

ANY BENEFITS?

Yes, there was a clear and significant improvement in all the models due to the removal of outliers.

When the outliers were retained, the models didn't perform well. Especially, in case of linear regression.

LINEAR REGRESSION

This model as the name suggests shows the linear relationship between the predictor variable and the target variable usually along with an intercept and an error term.

Linear regression was run without normalization and with normalization. It was observed that it made no difference whatsoever whether the regressors (X) were normalized or not. The results were identical. Root Mean Square Error in both cases was 3.37.

Fit_intercept was kept at default value because it made absolutely no sense to not include the intercept in our model. The houses were always going to have a certain value even if all the input variables had zero values, especially when some of the variables are categorical, not numerical.

"CHAS" i.e the Charles river dummy variable is a categorical variable, so even if it is 0, the median value of the house will not be zero.

SUPPORT VECTOR MACHINE

SVM is a machine learning algorithm that supports both linear and non-linear problems, unlike linear regression. Linear SVM has no kernel and finds a minimum margin linear solution to the problem. SVM with kernels are used when the solution is not linearly separable.

Results for SVM regression model were very similar to linear regression. Epsilon value was kept at 0.1.

However, C i.e the regularisation constant did have a significant effect on the results. As C was increased, the results to improved i.e the mean square error kept on decreasing. When C was increased to 10000 RMSE was 3.317. This was clear evidence that as increase in C improves the performance of the model when kernel used was rbf.

Linear kernel gave an extremely high RMSE in case of high C of 10000. However, when C was decreased to 1, with a linear kernel, The RMSE had a value of 4.2. This is in contrast with having rbf kernel. Regularization constant of 0.1 gave a reduced RMSE of 3.8

Similar results were seen for sigmoid where a high C value gave a high MSE and a low C value produced a low MSE. However, the RMSE was a lot higher in this case.

When kernel was set as poly, a high value of C (more than 1000) produced a low RMSE value of 3.5. This is very similar to rbf kernel.

When the gamma parameter was changed from 'scale' to 'auto', the RMSE almost doubled. The model now was performing significantly worse. Clearly, choice of kernels, and other parameters influenced the performance of the model.

K NEIGHBORS REGRESSOR

In KNN regression, mean of k nearest datapoints is calculated as the output. As a rule of thumb, we select odd numbers as k. KNN is a lazy learning model where the computations happen only runtime.

K neighbours regressor model was run for different number of neighbours. Lowest RMSE of 4.1 was obtained with a low number of neighbours. When the number of neighbours were increased RMSE started increasing. This is suggestive of the fact that it's important to find an optimal value of k.

When the weight parameter was changed from the default 'uniform' to 'distance' there was no change in RMSE value.

All the algorithms were used and none of them showed any significant change in results. Ball tree, kd tree, brute and the default algorithm were tried and the results stayed the same.

COMPARISON

Linear regression and support vector machines showed similar results. K neighbours regressor was the worst performing model.

Its' not recommended to use K neighbours regressor because usually, it is very hard to determine the optimal value of number of neighbours.

Both, linear regression and SVM regression perform well with high number of input variables than K neighbour regression.

Linear and non-linear models

However, k neighbours regressor does have its merit i.e it does support non-linear models unlike linear regression. SVM also supports non-linear models. Both, K neighbour regression and SVM have a wider scope for usage because in real world application with more complex scenarios, linear regression may not be applicable.

Linear regression also assumes predictor variables to be mutually independent which again, is a slight disadvantage when it comes to real world scenario.

Linear regression hence might be simpler to use and train but in case of complex scenarios, linear regression might not be the optimal model to use.

Dealing with outliers

As mentioned initially, outliers affected the performance of the models to a certain extent. However, linear regression is the worst model when it comes to dealing with outliers. SVM regression model and K neighbours regressor deal with outliers better than linear regression. SVM deals with outliers better than K-neighbours regression.

CONCLUSIONS

For the given dataset, linear regression and SVM regression ended up performing better than K neighbours. SVM ended up being the best model with the lowest root mean square error.

It is to be noted that the dataset given had a low level of complexity which resulted in linear regression performing reasonably well.

However, linear regression is not recommended for more complex cases. K neighbour regressor and SVM regression are better equipped to deal with more complex cases.

In this scenario, SVM clearly emerged as the clear winner.