Summary And Knowledge Derived From Support Vector Machines, Practical Session

Abdulrasaq, Surajudeen.* and Ezukwoke, Ifeanyi Kenneth.**

*MLDM M1

Introduction

Support vector Machine (SVM) are founded on the idea of a plane that separate and define a decision boundaries, this plane or line need to be the optimal so that objects can be mapped efficiently, now SVM try to gives us this optimal plane or line using a set of Mathematical functions called kernels, this kernels are rbf, polynomial , linear and sigmoid, and they are dot product of input data points mapped into the higher dimensional feature space by transformation. In addition to this we have whats called hyper-parameter C which help in controlling the effect of the soft-margin, in a nutshell, the sets of exercises has come with different set of knowledge and the overall summary of what was learned in each task is depicted below.

Exercise 2

Knowledge gained

C parameter regulate the margin, and also the quantity that measure the error, the higher the C the tighter the model fit during training and the better our prediction, lower C results in flat outputs, while gamma parameter has no relationship with C from our observation.

Dataset Generation(Exercise 3)

Knowledge gained

Trying different SVC kernels including linear, rbf, sigmoid and poly, we observed the kernel that performed the best on the random dataset at train time is the sigmoid kernel. With a best score of 70 and a C value of 2.

Multiple calls of this method on the same dataset with Random fixed at False will still produce similar result except streetwise, RBF, SIGMOID AND LINEAR all gave the same score while polynomial seems to give the lowest results, but the model choose Best Kernel: sigmoid, Best Gamma: auto

Moon dataset: The moon random dataset however produced a different result after iterating over the different kernel. The best kernel chosen after training was the linear kernel with a best score of 86 and a C value of 100. The decision boundary plotted indicates this result.

Iris datasets: Analyzing the iris dataset using SVC of Svm on sklearn yielded following result

The best score : 96 Best C: 100 Best kernel: Linear Best Gamma: auto

On plotting the decision boundaries for each kernel, we observed the linear kernel indeed performed the best amongst other kernel. Followed closely by the rbf kernel and lastly the poly kernel. Sigmoid performed the worst on the iris dataset.

Existing Dataset (Exercise 4)

Knowledge gained

Ozone dataset(ozone.dat)

Observation: We iterated over the different kernel which took a bit of time. In the end got the following result and the conclusions The best score : 89 Best C: 5 Best kernel: Linear Best Gamma: auto The best score : 81 Best C: 1 Best kernel: sigmoid Best Gamma: auto The best score : 84 Best C: 1 Best kernel: rbf Best Gamma: auto The best score : 82 Best C: 1 Best kernel: poly Best Gamma: auto. From the above result, Linear kernel gave the best result as seen above followed next by the rbf kernel.

Using Lasso Regression for Prediction

BONUS FOR CASTING INTO LASSO REGRESSION ===BEST RESULT FOR THE TARGET LABEL===
Best alpha: 0.05 Best score: 0.24176487171826797

PEAK PREDICTION OF OZONE: *194.12238851360664*

CONCLUSION

Using Lasso regression we predicted that our ozone level decreased by more than 35 from almost 350 to around 194. 122.

The meaning of this is that the activities of industries have increased environmental pollution emission of greenhouse gases that have plagued the ozone. Consequently leading to a significant reduction in the ozone layer.