Michael Bass

Challenge 5 Report

Introduction

In this challenge the goal was to create an SVM classifier using a generated data set. It was our task to evaluate different support vector machine (SVM) kernels and select the most appropriate kernel for our dataset. I applied kernels contained in the Scikit Learn API, as well as experimenting with my own kernel. Ultimately Scikit Learn's radial basis function (RBF) kernel performed best, and I used it to evaluate the test dataset.

Generated Data

In order to get a better understanding of the generated data, I first graphed the data, shown in Fig. 1. It is easily seen that the data is not linearly separable, and that it will require SVM to extract higher dimensional features to provide more accurate classification.

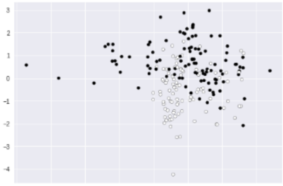

Fig. 1 - Visualizing the Generated Data

Applying SVM

As with many other machine learning models, applying an SVM classifier using Scikit-Learn (SL) is very strait forward. I looked through the SL documentation on the support vector classifier (SVC), their SVM implementation, but it did not seem that altering many of the API parameters would be effective. Therefore, I experimented with different kernels using their default settings. The kernels I used were: RBF, linear, polynomial, and sigmoid. Out of these, RBF provided the best accuracy with 40 mispredictions. The visualization of each kernel's classification boundary is shown in the following images.

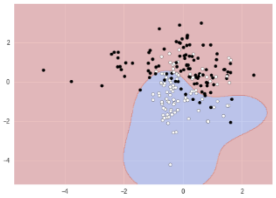

Fig. 2 – RBF Kernel

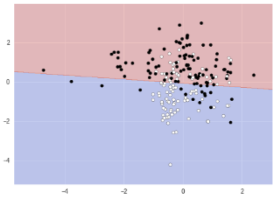

Fig. 3 – Linear Kernel

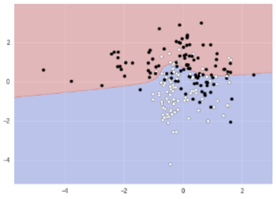

Fig. 4 – Polynomial Kernel

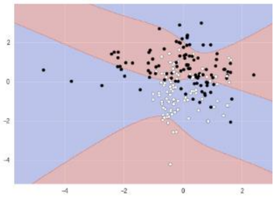

Fig. 5 – Sigmoid Kernel

Experimental Kernel

To learn more about SVM kernels, I watched a YouTube video from Andrew Ng [1]. In this video he discussed the role of the kernel and discussed a Gaussian based landmark kernel. In this kernel, the data scientist creates a set of landmarks and creates a higher dimension feature set as the gaussian distance from each of the landmarks. The new feature set has one feature for each landmark, and is made given the following formula:

$$e^{-\frac{\|x_i - l_j\|^2}{2}}$$

Using this approach, I created a new feature set, and set the SVC to a linear kernel. Therefore, this allowed me to hard code the kernel. The resulting decision boundaries are shown in Fig. 6. I used four landmarks which are shown as green X's in the figure. This kernel had 47 mispredictions, only a little worse than RBF.

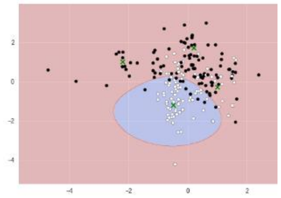

Fig. 6 – My own kernel

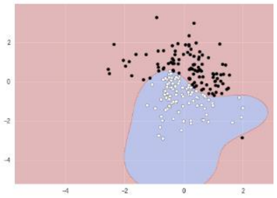

Fig. 7 – Applying the SVC to the Test Data

Applying Test Data

The final task was to apply the trained kernel to the test data. Since the RBF kernel had the best accuracy, this is the kernel I selected. The result of the prediction is shown in Fig. 7.

Conclusion

In this challenge I used SVM to classify a generated dataset. I initially used the SL SVC API, and then created my own kernel using landmarks and a Gaussian curve. The kernel that performed best was SL's RBF kernel. I applied this kernel to the test data and visualized the results.

This has been my first experience with SVM. I think that creating my own kernel helped me gain greater understanding about the translation to a higher dimension feature set. For my learning experience, this was the most valuable part of the assignment.

References

[1] - https://www.youtube.com/watch?v=mTyT-oHoivA&t=301s