A comparison of SVM Kernal functions in the use of trinitarian-class classification problems

Sean Daly - sd3191a@gre.ac.uk - ID: 001013392-5

April 9, 2021

Abstract

In this project, the performance of Support Vector Machines, a popular classifier, is analysed. The project will focus on the choice of kernel function to determine the best choice for optimal performance in classifying species of penguin (Horst et al. 2020) using two measurements of the beak. Numerous performance metrics will be examined, both quantitative and qualitative. The project will conclude with a recommendation of the kernel to use in classification problems involving three classes.

1 Introduction

In a majority of data problems and work-flows, organisation of data entries is essential to maintain large databases and data collections. A solution that is commonly used is to organised entries into groups or classes allowing for easier organisation, searching e.t.c. The benefits differ depending on the implementation and context, but they can range from as small as better organisation to something as major as diagnosing cancer. The aim of classification in Machine Learning is to group data items with numerous character features into a specific class that have similar characteristics. The resulting classifier is then used to assign class labels to the testing instances where the values of the predictor features are known, but the value of the class label is unknown (Kotsiantis et al. 2007). Originally, the classification process would of been done by hand or another manual method, however since the rise in scale of data sets and data collection caused by the popularity and demand for Big Data, manually classifying entries has become inefficient and obsolete. Machine Learning has created numerous solutions to various classification problems, providing multiple methods with their own strengths and weaknesses.

2 Literature Review

In this section, previous work is analysed in subject areas that have a relation to our project. The findings from these works will shape the methodology and structure of our experiments to find the optimal hyper-parameters for three class classification using Support Vector Machines.

A study into the recent advances in designing multi-class SVMs was carried out by Angulo et al. (2006), into which the study looks into methods used in classifying multiple classes with an SVM. Support Vector Machines are one of many techniques used in Machine Learning and exploratory data analysis. It is used in classification as well as outlier detection. The Support Vector Machine (SVM) was first proposed in 1995 by Cortes & Vapnik (1995). The idea of an SVM is to map data points of n-dimensions into a higher dimensional domain such that a n-dimensional plane can be used to separate data points from different classes. Like with most Machine Learning models, SVM comes with a host of hyper-parameters that can be changed to tune the model for better results. One of these parameters is the kernel function used by the model. Described by Patle & Chouhan (2013) as a "mathematical trick that allows the SVM to perform a 'two dimensional classification of a set of originally one-dimensional data". They also later on in their paper conclude "Through kernel selection we find the comparison of accuracy between Dataset. Empirical analysis of kernel selection shows that kernel selection reflects the accuracy of classification." In short, the choice of the kernel is the biggest factor in terms of performance of the model. In his paper, Kotsiantis et al. (2007) says "The selection of an appropriate kernel function is important, since the kernel function defines the transformed feature space in which the training set instances will be classified". This further proves how important the choice of kernel is. This is further backed by the findings from Trivedi & Dey (2013). As well as the choice of kernel, the optimisation of hyperparameters can boost performance of the SVM in a particular problem domain (Grama et al. 2017).

Experiments similar to the one carried out in this report have already been conducted. Hussain et al. (2011) carried out a study to compare SVM kernels in breast cancer detection, in which four different kernels were compared on a number of different subsets, each with a different number of features. Overall, they concluded that "sigmoid results in the best specificity whereas polynomial function gives the best sensitivity". In another paper, Yekkehkhany et al. (2014) use SVM on multi-temporal full-polarimetric L-band SAR data to classify crops in agricultural regions. In their conclusion they find that the RBF kernel outperforms linear and 3rd degree polynomial kernels. They also state "RBF is the most frequently used kernel in optical remote sensing data". Their findings back up this statement, providing a valid reason why the RBF kernel is commonly used in this domain. Madheswaran & Dhas (2015) also found RBF the most effective kernel when using SVMs to classify MRI images of the brain. In another study, Ren et al. (2016) conducted a study on how different kernels effected the classification performance on well-logging data. Their results showed that the polynomial kernel performed the best in their domain.

Studies of SVM performance variations dependant on kernel and problem domain allow researchers to save time on testing different kernels and select specific kernels based on results. Abakar & Yu (2014) shows that the PUK kernel has similar performance to that of the RBF kernel in terms of predicting yarn tenacity. Although these experiments seem similar to each other, they all produce useful results for future researchers to use, hence the basis for experiments and studies like this one.

In terms of evaluating a models performance, Sokolova et al. (2006) and Tao (2021) provides us with the formulae for precision, recall and F-score:

$$Precision(p) = \frac{TP}{TP + FP}$$

$$Recall(r) = \frac{TP}{TP + FN}$$

$$F - score(F) = \frac{(B^2 + 1)pr}{B^2p + r}$$

Precision and recall can be used to validate the accuracy of a classifier. The F-score acts as an mean, created from the precision and recall values. Another heuristic to use for comparison could be to use the area under the respective ROC curve. This area represents the probability that a randomly chosen positive example is correctly rated (ranked) with greater suspicion than a randomly chosen negative sample (Bradley 1997).

3 Methodology

In this project, we want to compare a number of different kernels against the same dataset to analyse the best kernel for three-way classification. Multiple SVMs will be trained on the same training data, each with a different kernel, and then tested on the same test dataset. The predicted labels from each machine will then be analysed through the calculation of a confusion matrix ans well as precision, recall and F-score results. Graphical displays of how the models perform will also be produced, allowing for a more visual comparison of kernel performance. This section explains the methods used in the experiments carried out in this paper as well as the state of variables used.

3.1 Models

There will be four separate instances of SVMs: one with the RBF kernel, one with the sigmoid kernel and one with a 3 degree polynomial kernel. There will also be a Linear SVC instance as well, which can be considered a more efficient version of the "linear" kernel. These four classifiers are all independent from one another and don't effect each others performance.

3.2 Experiments

There will be only one experiment taking place: a single train/test cycle of the model to return some performance metrics that can be analysed. In the case where the metrics are too close or similar (caused by over-fitting) the test/train ratio will be changed to allow for more variation in results. Once the performance metrics are available, a comparison of all the models will be conducted. The metrics that will be used will be the accuracy, precision, recall and f-score for each subclass. On top of this, confusion matrices will be generated for each model as well as a visual plot of all the data points and their representative sub species.

3.3 Dataset

As mentioned in the introduction, the dataset being used is of a collection of 344 observations of different penguins (Horst et al. 2020). Each observation has 8 variables of information collected about it ranging from sex, subspecies (Gentoo, Adeline and Chinstrap) weight and measurements of its beak. Prior to this experiment, the dataset was imported into R and the dataset analysed. Upon inspection, it was observed that the length and width of the beak had a strong correlation when compared to subspecies. A scatter plot of the dimensions of the beak produce clusters of each particular subspecies. It is evidential that this can be used on a machine learning model to predict the subspecies of a particular penguin.

The dataset is reduced down from 8 variables to 3 which included subspecies, beak length and beak width. A new variable was added to each observation, a numerical representation of the subspecies, acting as the label (-1 for Adeline, 0 for Chinstrap and 1 for Gentoo). Then the data was separated into two subsets: training and testing data. Originally the share ratio was 80:20 train/test, however it was observed that a lot of the models produced very similar results, evidence of over-fitting. To variate the results, the train/test ratio was changed to 30:70. Although in Machine Learning this is considered bad practice to have more testing points than training points, this will test the learning efficiency of the models with different kernels. This should not matter too much due to the strong correlation we saw in the analysis of the data, therefore the models should easily fit.

3.4 Experimental settings

In terms of Hyper-parameters, none were changed during or after the experiment, these were all set before the experiment. Their values are based on previous studies seen in the Literature Review or what their default values are as set by the scikit-learn library. Different kernels have different parameters, however all have one common parameter: C. C is the regularisation parameter for the SVM. For all models, this is set to 1. The sigmoid model does not have any other hyper parameters set. The Linear SVC has a mximum iteration of 10000, RBF has gamma of 0.7 and the polynomial kernel has 3 degrees and gamma set to "auto".

4 Results

There are three groups of results to show in this section: the graphical representations, the statistics and the confusion matrices. We will discuss these in this order with each kernel.

Firstly the graphs. Each graph is a scatter plot of each data point with a color designated to each subspecies of penguin. Red for Adeline, green for Chinstrap and blue for Gentoo. The dots represent the training points of the data and the triangles are the testing data, whose color is being decided by the SVM. What we are observing here is if the colors of the points are all clustered or mixed together, which shows how the SVM decision boundaries differentiate with each kernel.

Firstly, the sigmoid kernel clearly struggles to learn all but one species. All testing points have been labelled as Adeline, clearly problematic. The Linear SVC seems to perform fine, with testing points matching the cluster shapes and three clear groups appearing. RBF performs similar, however labels some Chinstrap and Gentoo outliers as Adeline. Lastly the polynomial kernel looks the same as Linear SVC with some points being a diferent color. Its clear to see that sigmoid has struggled to fit to the training data. This could be caused by a bias in the training set or just poor performance on the kernels side. Since sigmoid performed so poorly, it feels reasonable enough to rule it out.

Next we look at the performance metrics. Precision, recall and F1-score is calculated for each class of penguin and then a macro-average is calculated, whilst the accuracy is calculated overall for each kernel (see the table below). The confusion matrices reflect a similar outcome, they confirm the sigmoid kernels inability to fit to three classes, only fitting to one.

Kernel	Accuracy	Precision	Recall	F1-score
Sigmoid	0.4538	0.15	0.33	0.21
Linear SVC	0.9328	0.94	0.89	0.91
RBF	0.9412	0.95	0.92	0.93
Polynomial	0.9370	0.94	0.90	0.92

4.1 Discussion

It is clear to see that the poor sigmoid performance is backed up with poor metrics. RBF looks to be the best overall, scoring high in all areas with the highest scores across the board. In the graphs, RBF, Linear SVC and polynomial kernels seem to perform similar, with some points having different colors depending on the kernel. Overall, it seems RBF performs the best. However more work is needed to clarify this as with hyperparameter tuning, other kernels could perform just or even better. Also the dataset in question is rather small meaning that time complexity of the kernels is not much of an issue here, however as the data increases, the speed of the kernels will begin to become a deciding factor.

5 Conclusion

To conclude, in this experiment we found that the RBF kernel is the best performing kernel in tri-class classification. It had the best accuracy, F1-score, recall and precision when compared to sigmoid, polynomial and Linear SVC. We have seen graphically that polynomial, Linear SVC and RBF perform similar to each other. Further work is needed in hyperparameter tuning and also data scaling to see how the kernels compare with larger amounts of data.

References

Abakar, K. A. & Yu, C. (2014), 'Performance of svm based on puk kernel in comparison to svm based on rbf kernel in prediction of yarn tenacity'.

Angulo, C., Ruiz, F. J., González, L. & Ortega, J. A. (2006), 'Multi-classification by using tri-class svm', *Neural Processing Letters* **23**(1), 89–101.

Bradley, A. P. (1997), 'The use of the area under the roc curve in the evaluation of machine learning algorithms', *Pattern recognition* **30**(7), 1145–1159.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Grama, L., Tuns, L. & Rusu, C. (2017), On the optimization of svm kernel parameters for improving audio classification accuracy, *in* '2017 14th International Conference on Engineering of Modern Electric Systems (EMES)', IEEE, pp. 224–227.

Horst, A. M., Hill, A. P. & Gorman, K. B. (2020), *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0.
URL: *https://allisonhorst.github.io/palmerpenguins/*

Hussain, M., Wajid, S. K., Elzaart, A. & Berbar, M. (2011), A comparison of svm kernel functions for breast cancer detection, *in* '2011 Eighth International Conference Computer Graphics, Imaging and Visualization', IEEE, pp. 145–150.

Kotsiantis, S. B., Zaharakis, I. & Pintelas, P. (2007), 'Supervised machine learning: A review of classification techniques', *Emerging artificial intelligence applications in computer engineering* **160**(1), 3–24.

Madheswaran, M. & Dhas, D. A. S. (2015), 'Classification of brain mri images using support vector machine with various kernels.'.

Patle, A. & Chouhan, D. S. (2013), Svm kernel functions for classification, *in* '2013 International Conference on Advances in Technology and Engineering (ICATE)', IEEE, pp. 1–9.

Ren, Y., Hu, F. & Miao, H. (2016), The optimization of kernel function and its parameters for svm in well-logging, *in* '2016 13th International Conference on Service Systems and Service Management (ICSSSM)', IEEE, pp. 1–5.

Sokolova, M., Japkowicz, N. & Szpakowicz, S. (2006), Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation, *in* 'Australasian joint conference on artificial intelligence', Springer, pp. 1015–1021.

Tao, C. (2021), 'How to evaluate a classification machine learning model'.
URL: *https://towardsdatascience.com/how-to-evaluate-a-classification-machine-learning-model-d81901d491b1*

Trivedi, S. K. & Dey, S. (2013), 'Effect of various kernels and feature selection methods on svm performance for detecting email spams', *International Journal of Computer Applications* **66**(21).

Yekkehkhany, B., Safari, A., Homayouni, S. & Hasanlou, M. (2014), 'A comparison study of different kernel functions for svm-based classification of multi-temporal polarimetry sar data', *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **40**(2), 281.

6 Appendix

Figure 1: Scatterplot of the Sigmoid kernel

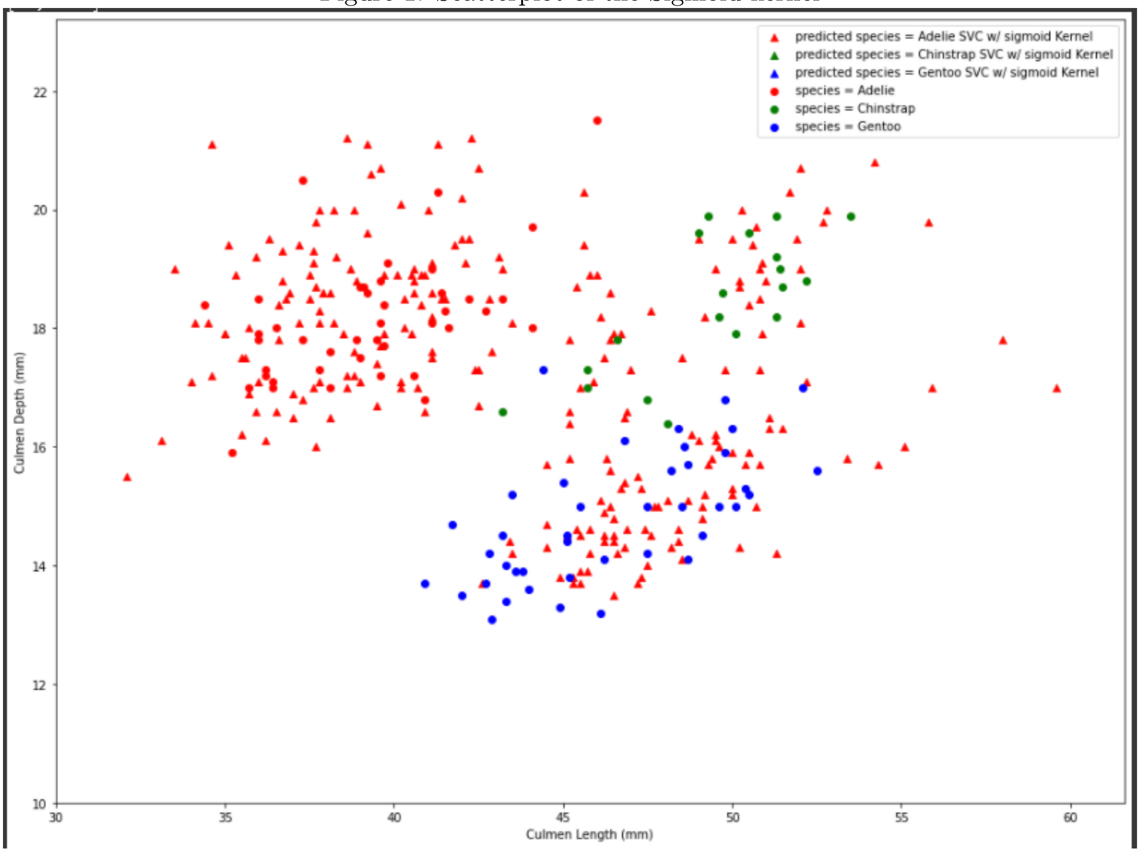

Figure 2: Scatterplot of the Linear SVC

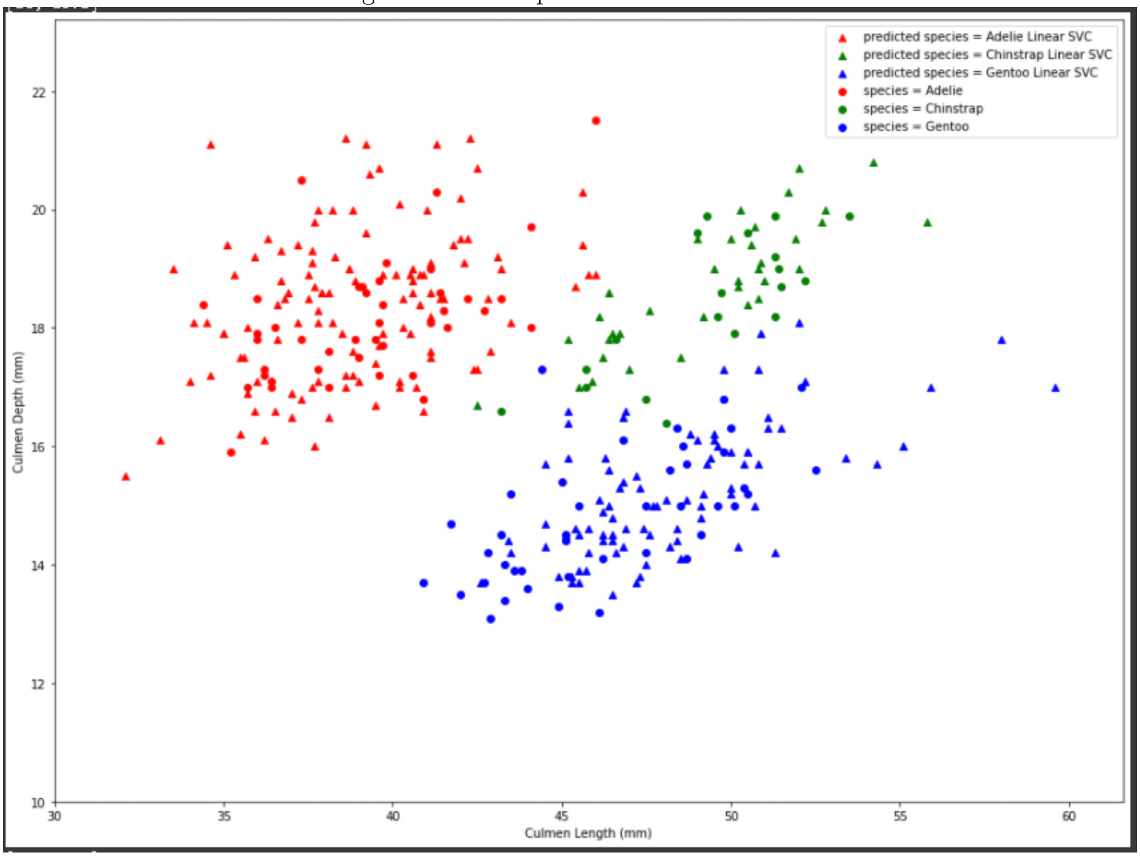

Figure 3: Scatterplot of the RBF kernel

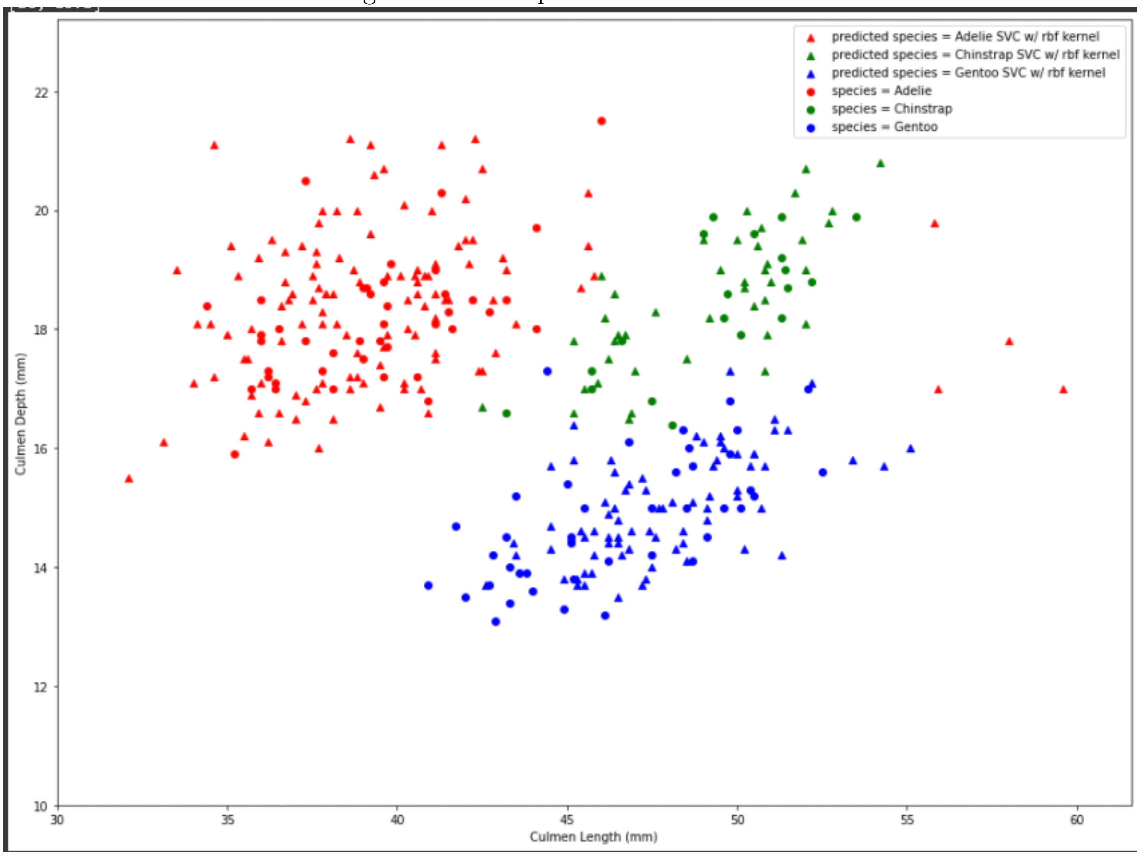

Figure 4: Scatterplot of the polynomial kernel