Penguin Species Classification Using Support Vector Machines

Nithin Sai Jalukuru
School of Engineering and Applied Sciences
University at Buffalo, Buffalo, NY, USA
njalukur@buffalo.edu

Abstract : In this present work, Classification of penguin species is done using SVM (Multiclass classification) based on the island, culmen measurements, flipper length, body mass attributes.

Keywords : SVM, Kernels, Multiclass classification, Penguin species, one vs one, one vs rest classifier.

1.INTRODUCTION

Perhaps one of the most well-known and discussed machine learning methods is support vector machines. Support Vector Machines are systems that use the hypothesis space of linear functions in a high-dimensional feature space and are trained using an optimization theory-based learning method that incorporates a learning bias. The data points that are hardest to classify are support vectors, or those that are closest to the hyperplane (decision surface). The margin is maximized by SVMs.

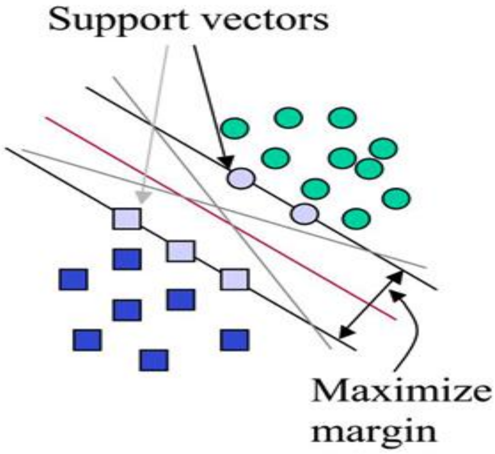


Fig1: Representation of SVM (Source : CSE 474/574 lecture slides ,Prof Chen)

1.1 *SVM Kernels*

The SVM algorithm is implemented in practice using a *kernel.* There are 3 kernels which can be implemented into this

problem, they are Linear Kernel SVM, Polynomial Kernel SVM , Radial Kernel SVM

Linear Kernel SVM : The dot-product is called the kernel and can be re-written as

$$\mathbf{K}(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{X}}_i) = \sum(\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}_i) \qquad \text{--(1)}$$

The kernel defines the similarity or a distance measure between new data and the support vectors.

Because the distance is a linear combination of the inputs, the linear SVM or linear kernel uses the dot product as the similarity metric.

Polynomial Kernel SVM : This kernel can be used to transform the input space into higher dimensions. This is called the *Kernel Trick*

$$\mathbf{K}(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{X}}_i) = \sum(\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}_i)^d \qquad \text{-- (2)}$$

where the learning algorithm needs the polynomial's degree to be manually supplied. Curved lines can exist in the input space thanks to the polynomial kernel.

Radial Kernel SVM: This kernel is more complex than moth linear and polynomial kernels.

$$\mathbf{K}(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{X}}_i) = \boldsymbol{e}^{-\boldsymbol{gamma} \times \sum(\mathcal{X}-\mathcal{X}_i^2)} \qquad \text{--(3)}$$

Gamma is a parameter that the learning algorithm needs to be given. Gamma should be set to a good default of 0.1. Due to its extreme locality, the radial kernel can produce intricate regions in the feature space, much like closed polygons in a two-dimensional environment.

1.2 *multi-class classification SVM*

Although most real-world scenarios entail multiclass classification, SVMs are typically designed to do binary classification. Researchers have suggested a number of ways to create multiclass SVMs from binary SVMs, and this area of study is still active. One versus. One and One vs. Rest multi-class techniques are used in this problem to solve the classification problem using penguin data. Let's see how these functions work.

I. *One vs. one approach* :

By setting the breakdown to a binary classifier for each class, this approach creates SVM classifiers for every pair of classes that could exist. Therefore , for $\mathcal{M}$ classes , there will be $\frac{\mathcal{M} \times (\mathcal{M}-1)}{2}$ binary classifiers/ SVM's. Each classifier produces an output in the form of a class label. However, the primary drawback of this approach is the rise in classifiers as the number of classes rises.

This method ignores the points of the third class and requires a hyperplane to divide every two classes. This signifies that the present split solely takes into account the points of the two classes.

For instance, Chinstrap, Adelle, and Gentoo are three classes that are represented in our scenario. In order to maximize separation only between certain data points, the Chinstrap-Adelle hyperplane is used. The gentoo data points are unrelated to this. The same is true of the Adelle-Gentoo and Chinstrap-Gentoo hyperplanes.

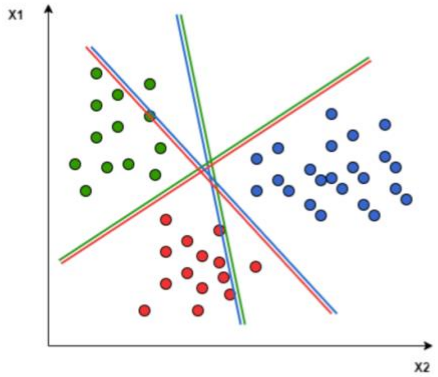


Fig2 : One vs One approach (source : Baeldung CS)

II. *One against Rest approach :* This approach is also called as *winner-take-all classification*. In this method, suppose if the data has $\mathcal{M}$ classes, then $\mathcal{M}$ binary SVM classifiers may be created where each classifier is trained to distinguish one class from the remaining $\mathcal{M} - 1$ classes. The final output is the class that corresponds to the SVM with the largest margin.

A hyperplane is required for the "one to rest" strategy in order to simultaneously divide all classes. As an illustration, think about the classes (Chinstrap, Adelle, Gentoo). The goal of the chinstrap hyperplane is to simultaneously increase the spacing between the chinstrap and all other data points.

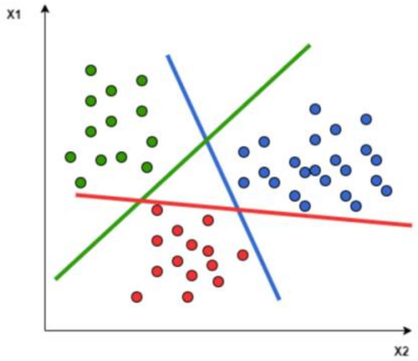


Fig 3 : One vs rest approach (source : Baeldung CS)

There are certain disadvantages, though. First off, a square of the whole number of training samples' worth of memory is needed during the training phase, which is a very high memory demand. Large training data sets may experience issues as a result, and computer memory issues may develop.

1.3 *Learning a SVM Model*:

The SVM model needs to be solved using an optimization procedure. The most popular method for fitting SVM is the Sequential Minimal Optimization (SMO) method that is very efficient. It divides the issue into smaller problems that can be resolved analytically as opposed to by searching or maximizing.

$$\boldsymbol{minimize}_{\mathcal{W},\mathcal{b}} = \frac{1}{2}\|\mathcal{W}\|^2 \textbf{ subjected to}$$
$$\mathcal{Y}_i(\mathcal{W}^T \times \mathcal{X}_i) \geq \mathbf{1} \;\; \boldsymbol{i} = \mathbf{1},\ldots\ldots., \boldsymbol{N}$$

2. EXPERIMENTATION

2.1 Methodology

For the SVM classifier model implementation, a data pipeline structure is constructed as illustrated in the accompanying figure 4.

Data collection and loading into the environment are both parts of data extraction   After loading, the data has been labeled and cleaned so that the ML model can analyze it effectively. . In addition, feature extraction a few features/attributes have been extracted, and the most pertinent ones have been chosen to train the model.

The correct ML model has been selected, trained, and verified during the model selection, training, and validation processes by maximizing the more suitable assessment criteria.

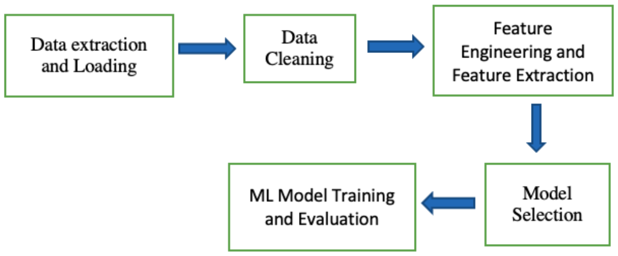

Data extraction and Loading
Data Cleaning
Feature Engineering and Feature Extraction
Model Selection
ML Model Training and Evaluation

Fig 4 : Data-pipeline

2.2 Data Extraction

The **Palmer Archipelago penguin** data , which is perfect data for any classification task also called the new Iris is collected from Kaggle. This data contains 344 rows and 9 attributes. These are

a) *Unnamed :0* which is just representing the ID's from 1 to 4.

b) *Species :* which is the target feature for classification. This data has 3 features namely Adelie, Gentoo, Chinstrap.

c) *Island:* This feature represents to which island foes this penguin species belongs to.

d) *Bill_length_mm and Bill_depth_mm:* These both represent the penguin's bill length and bill depth in mm. These are numerical/continuous features

e) *Flipper_length_mm :* This feature represents the flipper length of the penguin in mm

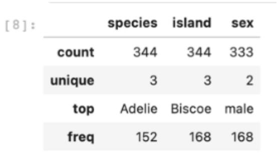
f) *Body_mass_g:* Body mass of penguin species given in grams (continuous feature)
g) *Sex:* A factor denoting the sex of a penguin(male or female) based on the molecular data.
h) *Year :* An integer denoting the year of study.

2.3 Data Cleaning and Feature Engineering

a) Removed unwanted columns , *unnamed:0* and *year* of study. These tow features don't contribute for classifying the penguin species
b) There are some missing values in bill_length and bill_depth , body mass and sex features. So missing data has been handled by *Simple Imputer* with most Frequent strategy.
c) *One hot Encoding* has performed on the categorical attributes for the performance of ML model.
d) Performed *feature standardization* on the numerical data attributes (bill length, bill depth, flipper length, body mass) , where the feature values are re-scaled to have a mean of $\mu = 0\ and\ standard\ deviation\ \sigma\ =\ 1$.

2.4 EDA (Exploratory Data Analysis) and Model selection

a) Description of categorical columns



[8]:

	species	island	sex
count	344	344	333
unique	3	3	2
top	Adelie	Biscoe	male
freq	152	168	168

Fig 5 : Description of categorical attributes

There are 3 unique species with Adelie being the most found.
There are 3 different islands where penguins are found, Biscoe Island has most of the penguins.
The Male penguins are more in number compared to the female penguins

b) Correlation between features

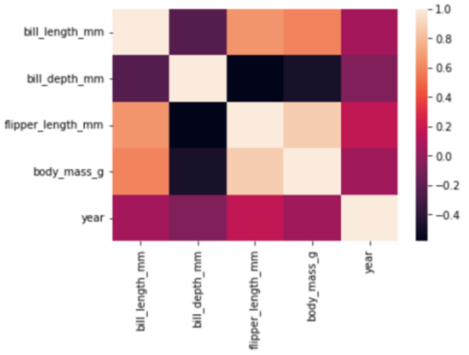


Fig 6 : Correlation Plot

There is absolutely no correlation between the year and other features (so removed as specified earlier). Body mass, flipper length and bill length, bill depth features are highly correlated.

c) Frequency of each species

As shown in the fig , Adelle species penguins are more in number than Gentoo and Chinstrap species.

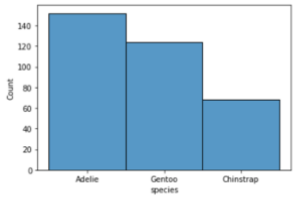


Fig 7 : count of each species

d) Pair plot for the features

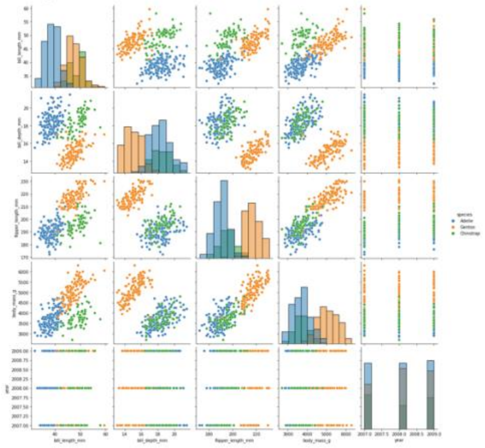


Fig 8 : pair plot of the features representing different species

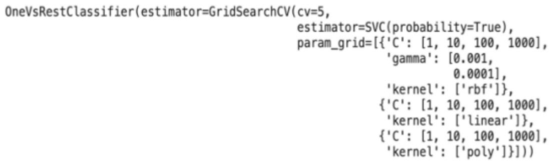
There is a linear trend between the attributes for all the 3 species and distribution of the species across all features also seems good. So, there is no need to go for SMOTE

Finally, the data is split into 2/3rd into training the data and 1/3rd into testing the data. The data is ready SVM machine learning model.

2.5 Support Vector Machine Modelling

In this SVM classification (OnevsRest classifier) model, three kernels (Linear, Radial, Poly) were chosen.

Created the parameter grid based on the results of random search and Performed Grid search CV to fine tune parameters for best SVM fit. GridSearch CV helps to identify the parameters that will improve the performance for this particular model.



```
OneVsRestClassifier(estimator=GridSearchCV(cv=5,
                                           estimator=SVC(probability=True),
                                           param_grid=[{'C': [1, 10, 100, 1000],
                                                        'gamma': [0.001,
                                                                  0.0001],
                                                        'kernel': ['rbf']},
                                                       {'C': [1, 10, 100, 1000],
                                                        'kernel': ['linear']},
                                                       {'C': [1, 10, 100, 1000],
                                                        'kernel': ['poly']}]))
```

Fig 9 : One vs Rest classifier svm fit

After preforming the operations as shown above, the values for best parameters are, Best C is 100, Best Kernel is rbf (radial basis kernel) and best gamma is 0.001 with an Best score of 0.9869 in training data.

2.6 Model Evaluation Metrics

Overall accuracy, recall, precision, f1 score, and AUC of the ROC curve are the most common evaluation measures for classification models.

False positives (FP) and false negatives (FN) are outcomes that were mistakenly classified by the model, while true positives (TP) and true negatives (TN) are outcomes of the positive class and negative class, respectively.

a) Overall Accuracy (OA): This is defined by the following the equation

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

The model could attain almost perfect overall accuracy if it consistently predicts the majority of classes.

b) The issue is more severe the more unbalanced the data. So, we require additional measurements. include Recall. It measures the proportion of accurately predicted positive classes to all positively categorized items.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Recall is important when we believe False Negatives are more important than False Positives

c) Precision: It is the ratio of correctly predicted positive classes to all items predicted to be positive

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

It tells us how correct or precise that our model's positive predictions are. When we think False Positives are more significant than False Negatives, precision is crucial.

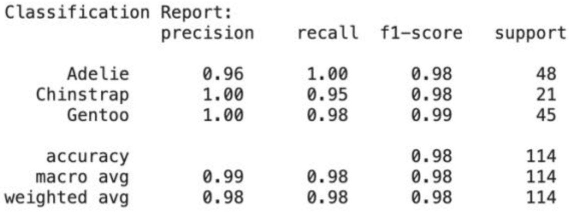
d) F1- Score : The F1-score is a single performance statistic that considers both recall and precision. It is also frequently referred to as the F-Measure. It is calculated by averaging the two metrics harmonically.

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

with values closer to one indicating better performance, and values closer to zero indicating poorer performance

2.7 Test Results

Let's look at the *classification report*



```
Classification Report:
              precision    recall  f1-score   support

      Adelie       0.96      1.00      0.98        48
   Chinstrap       1.00      0.95      0.98        21
      Gentoo       1.00      0.98      0.99        45

    accuracy                           0.98       114
   macro avg       0.99      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114
```

Fig10 : Classification report

SVM model with rbf kernel (best fine-tuned parameters) has classified each species almost perfectly for the test data. *With an recall, precision and F1 score of 0.98*

It is always possible to achieve a perfect score. This based on kernel chosen and the data modelled.

Also look at *Confusion Matrix*

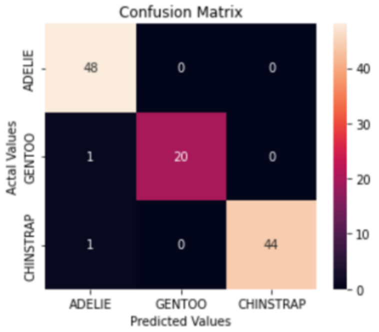


Fig 11 : Confusion Matrix

SVM model has wrongly predicted Gentoo and Chinstrap species into Adelie species. Other than that one vs rest svm classifier did an excellent job in classification.

3. CONCLUSION

SVM classification (One vs Rest classifier with rbf kernel and fine-tuned parameters) precision, recall and F1 score to 0.98. ROC AUC of our model also near to perfect one. So, we can conclude that our classifier did a Excellent job in classifying the penguin species.

4. REFERENCES

1.CSE 474/574 lecture slides (Prof. Chen)
2. Jason Brownlee, Master Machine Learning Algorithms

3. “Multiclass Approaches for Support Vector Machine Based Land Cover Classification”
4. https://www.baeldung.com/cs/svm-multiclass-classification
5. Pier Francesco at.el, “Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal Pump in the Oil and Gas Industry”.