Classification Problem on Iris Dataset

Jasmin Wilson

Abstract

This report gives an account of supervised learning applied to the Iris Dataset. KNN with hyper parameter (k) tuning and a simple neural network was used to identify the species of the Iris flower. With high youden index, neural network with one hidden layer was able to get high accuracy with less false positives. For KNN, lower values of hyper parameter k gives better accuracy. Comparison with SVM and decision tree, KNN and neural network tells us that for categorical data, KNN and SVM is best approach to go with.

1. Introduction

The Iris dataset is a well-known dataset that consists of 150 samples of iris flowers, with measurements of their sepal length, sepal width, petal length, and petal width. The goal of this project was to predict the species of iris flowers based on these measurements using two different machine learning algorithms: K-Nearest Neighbors (KNN) and a simple Neural Network. I implemented two models but for comparisons, used SVN, and Decision Trees as well.

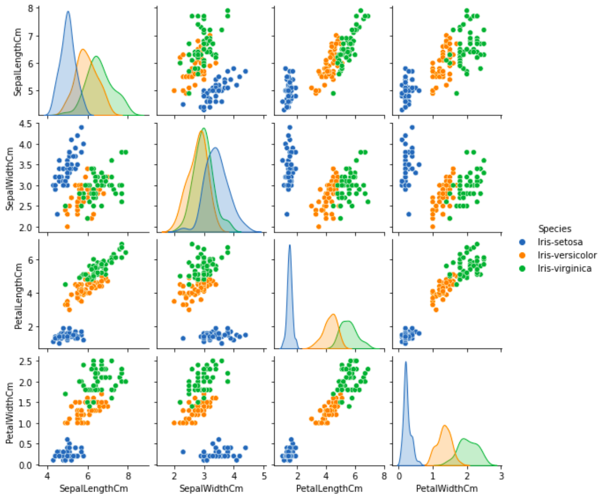

Figure 1. Actual V/s Predict

2. Domain

2.1. History

(1) Supervised learning is a type of machine learning algorithm where the model uses labeled data. Model is given the input as well as output data so it can train itself and predict if new data is thrown at the model.The goal is to predict any unseen data with high accuracy. Supervised models have two kinds of problems, Classification and Regression. Classification deals with categorical data whereas Regression is with continuous target variables. We can see supervised learning in everything from finance to face recognition, robotics, AI. Here we will implement and discuss two important supervised learning models, KNN and Neural Network

2.2. KNN Model

K-Nearest Neighbor model uses 3 step process. First, we find the Euclidean distance between the point to all the points.Then we sort them by distance and return the k neighbors for each point which is then used to predict the class. Because this a classification problem, we need to binarify the class into - Setosa = 0 , Vericolor = 1 and Virginica = 2. Because this is not a probabilistic model but we hard pass calculate the distance,

2.3. Neural Network Model

Neural Netowrks tries to works the same way as of a human brain. A simple neural network might consist of just a few neurons arranged in layers. The first layer takes in the input data, such as an image, and each neuron has a weight associated with it which is randomly generated between -1 and 1. Each layer receives a input and computes output which is the input for the next layer, this is called feed forward. The output is dot product of input and the weights which is then passed to an activation function, here I used Sigmoid function for nonlinearity. Output of the final layer is the predicted values. Now, the model learns from calculating the errors between predicted and true values and pass it down the network and performs corrections. Delta, which is the error of current layer times the Sigmoid derivation of current activation layer, is used to update the weights between previous and current layer by multiplying the delta with activation of previous layer and learning rate. Again, we calculate the error by removing the bias from the weights and multiplying with delta.

3. Approach

I divided the dataset into 85% training and 15% testing. Using the seaborn function, we can get some analysis on the dataset. We can see that Setosa can be easily classified if we draw a line but it will be a challenge and will be bound to false positives on predicting the other two classes.

4. Hypothesis

1. The most important part of KNN model is the k which we call it as hyperparameter. By running the model on different values of k from 0-30, we can see that lower values of k give better accuracy.

2. Neural network with one layer and multiple hidden layer. Both performs wells given the small dataset but higher the hidden layer units, faster is the learning.

3. Comparison on learning KNN v/s neural network.

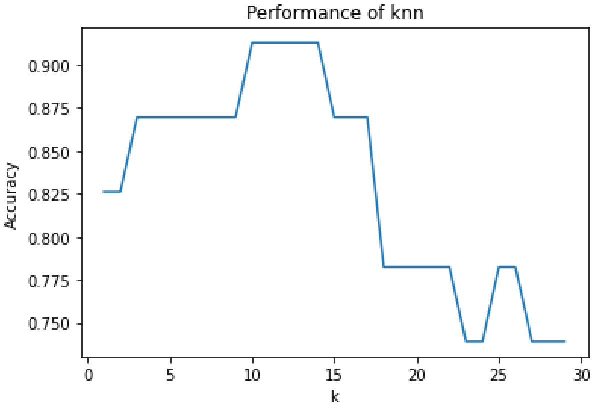

Figure 2. Hyperparameter tuning for K from 1-30

4.1. Hypothesis 1

Determining the optimal k makes the essence of the KNN model. Based on the k, we make predictions and accuracy of that prediction. Lower value of k will always result in low bias but high variance which may be the reason of overfitting and if k is large, it may loosely fit but will have low variance.Since there is no training required, we can find the optimal K and from the fig2. we can see lower K values give us more accuracy. Table 1. also shows a detailed description when K=5. Precision is the ratio of true positives to the sum of true positives and false positives. Recall is the ratio of true positives to the sum of true positives and false negatives. F1-score is the mean of precision and recall value. Because this is a non probabilistic algorithm with multi-class, ROC curves will not work. Either it will be one against all. We can get a better look at the true positives and False positives through confusion Matrix. If we need to calculate the TP, FP, FN, TN using this, we cannot. We need to individually calculate for each class. Like for Setsosa - TP = 4 TN = (9+0+2+8) = 19 FP = (0+0) = 0 FN = (0+0) = 0

	PRICISION	RECALL	F1-SCORE
Setosa	1.00	1.00	1.00
Vericolor	0.88	1.00	0.93
Virginica	1.00	0.80	0.89
Accuracy	0.96		
Macro avg	0.96	0.93	0.94
Weighted avg	0.96	0.96	0.96

Table 1. Detailed classification report based on the test data when K=10

False positives and true negatives is 0 aand 4. It only misclassified 4 of them out of 50, given the data was already much separated for Setsosa from fig 1.

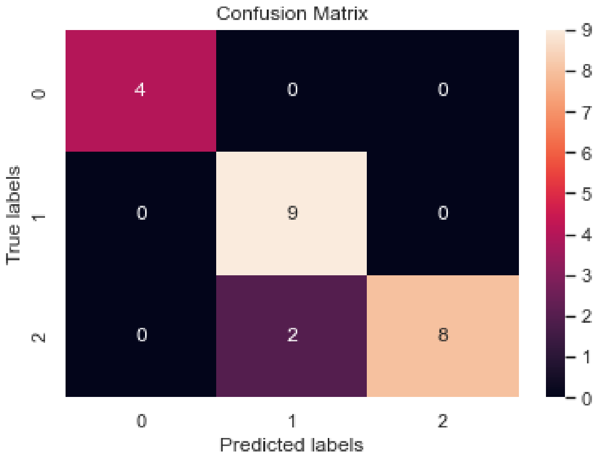

Figure 3. Confusion Matrix for KNN when K=10

4.2. Hypothesis 2

The output of the neural network is probabilistic values between 0-1. We used one-hot encoder so we can get the probablity value for each class. Higher the value, higher the trust it belongs to a certain class like [0.9233,0.04567,0.03456]. Again, since this is a multiclass problem, ROC curve will be one class v/s the rest. Here in fig 4, we can see ROC curve for Vericolor v/s rest. We can see a very high youden index that indicates very less TN's and FP's. Which makes this model reliable. From the ROC curve, we can see one

	One 10 units	Two 10 Units
Testing Accuracy	0.9832	0.956521
0-1loss	0.01695	0.0434782
Youden Index	0.978206	0.9587

Table 2. Detailed classification report based on the Test data with one and two hidden layer for Vericolor class

hidden layer performs better. With 2 layer, something that I noticed is getting stuck at local minimum. If I increase the learning rate, the model make even more mistakes but eventually reaches accuracy of 0.95 on test data.

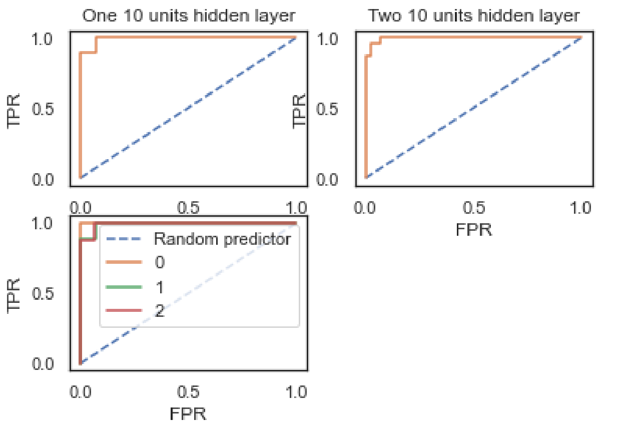

Figure 4. (Top) ROC Curve for Vericolor v/s rest for test data on 1 hidden layer/2 hidden layer, (Bottom) ROC curve for Sesota v/s all,Vericolor v/s rest and Virginica v/s rest

4.3. Hypothesis 3

At last, we can put together all the classifier to this Iris dataset and see the comparison. From fig 5, we see that KNN, SVM and NN which are best for categorical data performed best for Vericolor class but Decision Trees gives us more false positives which is known for overfitting and adding noise to the data.

5. Related Work

(2) One example of related work for KNN is the use of KNN in the field of healthcare. Researchers have explored the use of KNN for predicting patient outcomes based on electronic health record data, as well as for predicting disease progression or treatment response like breast cancer. Other related work for KNN includes the use of KNN in the field of anomaly detection, where it can be used to identify unusual or unexpected events in data. Whereas use of Recurrent Neural Networks (RNNs) for natural language processing tasks, such as machine translation or sentiment analysis, as well as market tracking, forecasting has been a huge deal.

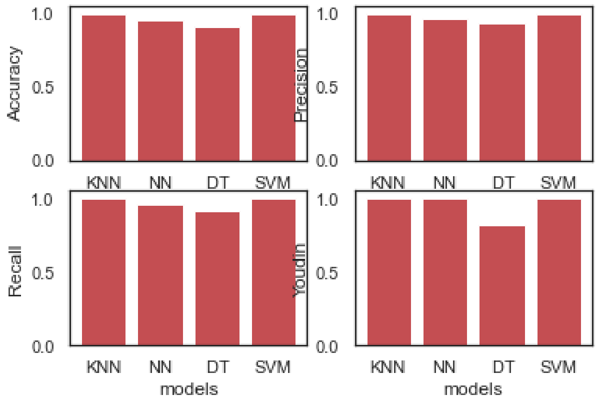

Figure 5. Comparison of performance metrics for all classifiers, (a) accuracy, (b) Precision, (c) Recall, (d) Youden's index

6. Contributions

All the finding in this paper is solely done by myself.

7. Future Work

I would like to implement this neural network with regression data and find trends from dataset specially useful to predict or forecast in the economic side. Also, work with predicting labels from images.

8. Conclusion

I was successfully able to implement KNN and neural network model on Iris Dataset. Both KNN and Neural Networks have their strengths and weaknesses, and the choice between them depends on the specific problem at hand. Using k tuning, we can find the effective accuracy for KNN model and for neural network, single layer gave best result but this can change with the dataset. Ultimately, the best approach is to experiment with different algorithms and techniques to find the one that works best for your specific use case.

References

[1] Qiong Liu and Ying Wu. Supervised learning. 01 2012.

[2] Soumya Prakash Rana, Maitreyee Dey, Gianluigi Tiberi, Lorenzo Sani, Alessandro Vispa, Giovanni Raspa, Michele Duranti, Mohammad Ghavami, and Sandra Dudley. Machine learning approaches for automated lesion detection in microwave breast imaging clinical data. *Scientific reports*, 9(1):1–12, 2019.