KNN + Confusion Matrix + Iris Data set + Colab

CS550 - Machine Learning and Business Intelligence

Submitted by: Divya Pandey(19665)
Instructor: Dr. Henry Chang

Iris Versicolor   Iris Setosa   Iris Virginica

Table of Content

- Introduction
- Design
- Implementation
- Test
- Enhancement Ideas
- Conclusion
- References

Introduction

★ Confusion Matrix is a performance measurement tool used in the evaluation of classification models.

★ It is used to determine the accuracy of the classification model by comparing the actual class labels and the predicted class labels.

★ The purpose of the Confusion Matrix is to provide a clear picture of the model's performance in a tabular format.

Design

Understanding Confusion matrix and applying it on KNN-Classifier on Iris Data set

		Predicted values		
		Positive	**Negative**	**Totals**
Actual Values	**Positive**	TP	FN	P = (TP + FN) = Actual Total Positives
	Negative	FP	TN	N = (FP + TN) = Actual Total Negatives
	Totals	Predicted Total Positives	Predicted Total Negatives	

Cat

TP Cat ==> Cat 5	FP ! Cat ==> Cat 2
FN Cat ==> ! Cat 3	TN ! Cat ==> ! Cat 17

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN+FP}{2}}$$

Design

Higher F1 Score ==> Better Classifier	Lower F1 Score ==> Worse Classifier
TP FN FP TN	TP FN FP TN

Design

| Low Accuracy | High Accuracy | High Accuracy |
| **High Precision** | Low Precision | **High Precision** |

Implementation

★ Go to **Colab**

```
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

from google.colab import files
uploaded = files.upload()
```

Choose Files Iris_data.csv
- **Iris_data.csv**(text/csv) - 4815 bytes, last modified: 2/10/2023 - 100% done
Saving Iris_data.csv to Iris_data.csv

Implementation

★   Download Iris [Dataset](#)

★   Execute [iris_knn.ipynb](#)

Test

```
5
[0.9400000000000001, 0.9400000000000001, 0.9400000000000001, 0.9400000000000001, 0.9400000000000001]
(1, 0.9400000000000001)
(3, 0.9400000000000001)
(5, 0.9400000000000001)
(7, 0.9400000000000001)
(9, 0.9400000000000001)
```
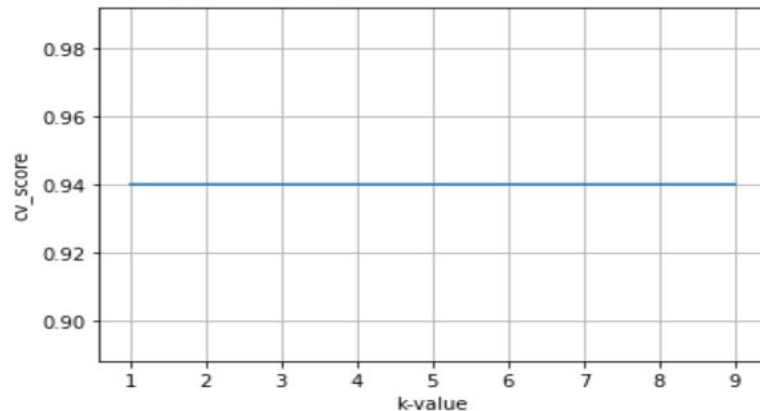


```
0.98
0.96
```

Test -Confusion Matrix for Iris Data set

```python
#confusion matrix and classification_report
#precision = TP/TP+FP
#Recall = TP/TP+FN

print(metrics.confusion_matrix(y_test, cross_predict))
print(metrics.classification_report(y_test, cross_predict))
```

```
[[19  0  0]
 [ 0 15  0]
 [ 0  2 14]]
               precision    recall  f1-score   support

    I. setosa       1.00      1.00      1.00        19
I. versicolor       0.88      1.00      0.94        15
 I. virginica       1.00      0.88      0.93        16

     accuracy                           0.96        50
    macro avg       0.96      0.96      0.96        50
 weighted avg       0.96      0.96      0.96        50
```

Enhancement Ideas

- We can use various data visualization techniques, such as scatter plots, to understand the data and gain insights into the relationships between the features. This can help to decide on the best features to use in our model and also identify any outliers or anomalies in the data.

Conclusion

★ Accuracy is one of the most commonly used performance metrics in machine learning. However, it is not always the best metric to use, especially when the data set is imbalanced. In such cases, other metrics such as True Positive Rate (TPR) or Recall, True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), and Precision are more appropriate.

★ All these metrics can be derived from the confusion matrix and they all play a critical role in evaluating the performance of a machine learning model. The choice of metric depends on the specific business problem and the desired outcome of the model.

GitHubLink

★ https://github.com/divyapandey03/Machine-Learning/tree/main/Supervised%20Learning/KNN%20%2B%20Confusion%20Matrix%20%2B%20Iris%20Data%20set%20%2B%20Colab

References

★ *Beena, V. (2020, May 15). Understanding confusion matrix and applying it on KNN-classifier on Iris dataset. Retrieved February 10, 2023, from https://ai.plainenglish.io/understanding-confusion-matrix-and-applying-it-on-knn-classifier-on-iris-dataset-b57f85d05cd8*

★ *Classification: Precision and recall | machine learning | google developers. (n.d.). Retrieved February 10, 2023, from https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall*