K-NN+CONFUSION MATRIX

DONE BY:SAHITI EMANI
STUDENT ID:19556
GUIDED BY : PROF.HENRY CHANG

TABLE OF CONTENTS:

INTRODUCTION

K-NN:

- The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.
- The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful.
- KNN captures the idea of similarity sometimes called distance, proximity, or closeness.

CONFUSION MATRIX:

- Confusion matrix is one such important tool which helps us evaluate model's performance.
- It is a matrix of size n x n .where 'n' is the number of class.

STEPS TO CALCULATE K-NN:

- The KNN Algorithm: STEPS
- Load the data.
- Initialize K to your chosen number of neighbors.
- For each example in the data: Calculate the distance between the query example and the current example from the data. Now, Add the distance and the index of the example to an ordered collection.
- Sort it in ordered collection of distances and indices from smallest to largest in ascending order by the distances.
- Pick the first K entries from the sorted collection.
- Get the labels of the selected K entries.
- If regression, return the mean of the K labels.
- If classification, return the mode of the K labels.

PHASES OF K-NN:

- **Training Phase:**
  - **KNN** : the data sort of indexing process in order to find the closest neighbors efficiently during the inference phase.
  - Else, it would have to compare each new case during inference with the whole dataset making it quite inefficient.
- **Validation Phase:**

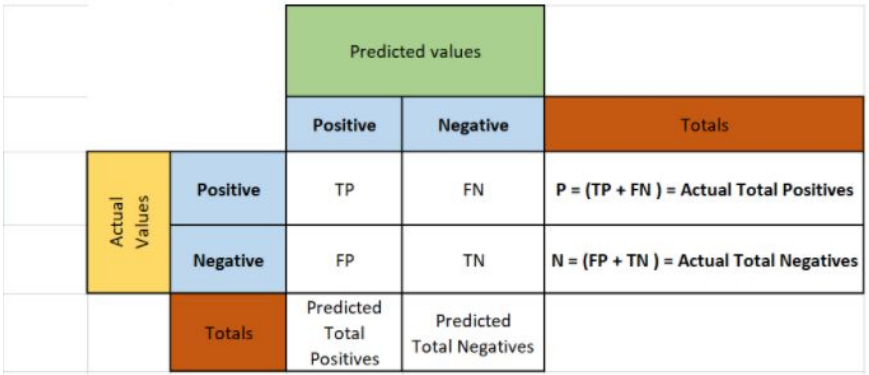
  Measures model accuracy against the training data as a function of iteration count training progress. Overfitting is by the upward movement of this empirical curve and indicates the point at which training should cease.

- **Testing phase** :
  - This phase finds its optimal solution of parameters **K** value , Distance calculating technique etc.

CONFUSION MATRIX:

- Confusion matrix, rows show actual values and columns indicate predicted values:

1. TRUE POSITIVE: Actual positives in the data, which have been correctly predicted as positive by our model. Hence,it's a True Positive(TP).
2. TRUE NEGATIVE: Actual Negatives in the data, which have been correctly predicted as negative by our model. Hence,its a True negative(TN).
3. FALSE POSITIVE: Actual Negatives in data, but our model has predicted them as Positive(FP).
4. FALSE NEGATIVE: Actual Positives in data, but our model has predicted them as Negative(FN).



Python's Representation of Confusion Matrix

WHAT IS ACCURACY AND PRECISION:

- ACCURACY:

  Accuracy refers to the closeness of a measured value to a standard or known value. For example, if in lab you obtain a weight measurement of 3.2 kg for a given substance, but the actual or known weight is 10 kg, then your measurement is not accurate. In this case, your measurement is not close to the known value.

  Is generally not the preferred performance measure for classifiers, especially when you are dealing with skewed datasets i.e., when some classes are much more frequent than others.

Accuracy:

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FN + TN + FP}$$

Accuracy formula

- PRECISION:

  Precision refers to the closeness of two or more measurements to each other. Precision is independent of accuracy.

  Precision is typically used along with another metric named recall, also called sensitivity or the true positive rate (TPR):It is the ratio of positive instances that are correctly detected by the classifier.

- RECALL:
  - The probability that a truly fraudulent transaction is caught by the model.
  - A better way to estimate performance than Accuracy.

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

Precision formula

$$TPR = \frac{TP}{TP + FN}$$

EXAMPLE: K-NN+CONFUSION MATRIX

- Q29)https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/exercise_algorithm.html

K=3		K=5	
Correct Assessment	Predicted Assessment	Correct Assessment	Predicted Assessment
-	-	-	-
-	+	-	-
+	+	-	+
-	-	+	-
-	-	+	-
+	+	-	+
+	+	-	+
-	-	+	-
-	-	-	-
+	+	-	+
-	-	-	-
+	-	+	-
+	+	-	-
+	+	+	+
-	-	-	-
-	-	+	-
-	-	+	-
+	+	-	+
+	+	-	-
+	+	+	+
+	+	+	+
-	-	-	-
+	+	-	+
+	+	+	+
-	-	-	-

- If the objective is to determine the "+" class, please fill this table

K=	TP	FN	FP	TN	Precision	Accuracy	Recall	F1 score
3								
5								

- Which K value represents the better model? Please explain your assessment.

SOLUTION OF GIVEN EXAMPLE:

K=	TP	FN	FP	TN	Precision	Accuracy	Recall	F1 score
3								
5								

CALCULATING FOR K=3;

CS550_W3_HW_Q29_19556_SAHITI_EMANI

Q) Calculation of Accuracy, Precision, Recall, F1-Score.

According to the given data:-

For K=3;

TP = ('++') ⇒ 12
TN = ('--') ⇒ 11
FP = 1
FN = 1

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{12+11}{12+11+1+1} = \frac{23}{25} = 0.92$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{12}{12+1} = \frac{12}{13} = 0.9230 \cong 0.92$$

$$\text{Recall (True PR)} = \frac{TP}{TP+FN} = \frac{12}{12+1} = \frac{12}{13} \cong 0.92.$$

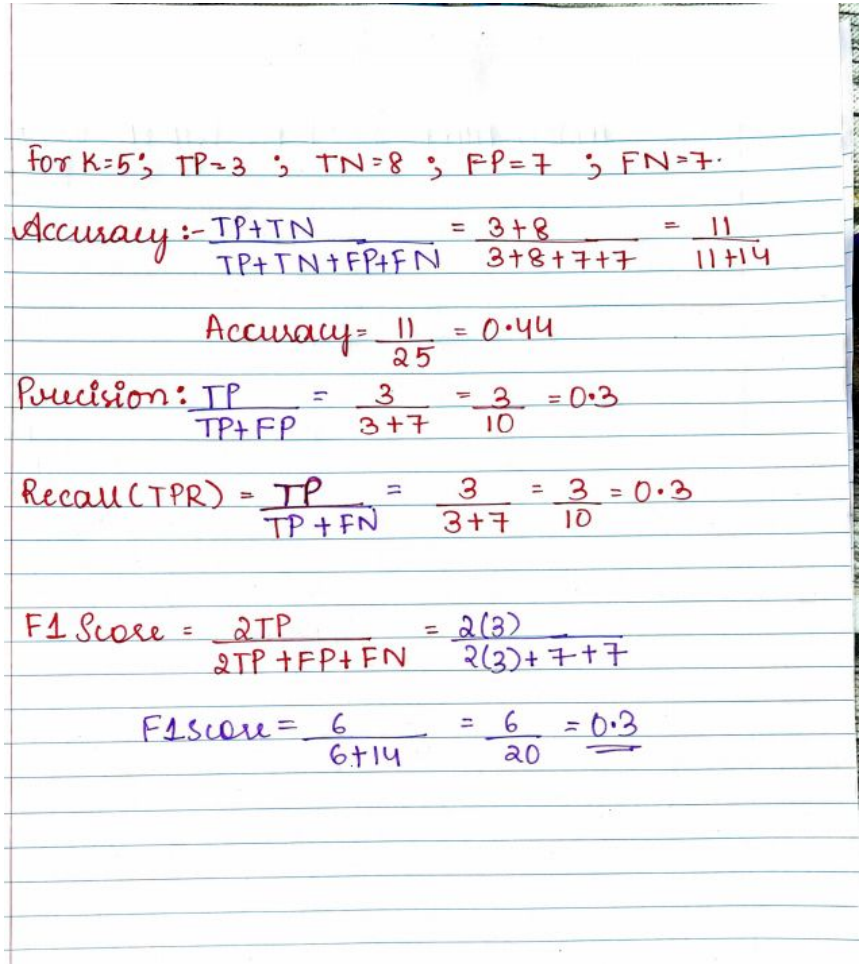
F1 Score = Harmonic Mean of Precision and Sensitivity.

$$= \frac{2TP}{2TP+FP+FN} = \frac{2(12)}{2(12)+1+1} = \frac{24}{24+1+1}$$

$$\text{F1 score} = \frac{24}{26} = 0.9230 \cong 0.$$

SOLUTION:CONT'D

Calculating for k=5;



for K=5; TP=3 ; TN=8 ; FP=7 ; FN=7.

Accuracy :- $\frac{TP+TN}{TP+TN+FP+FN} = \frac{3+8}{3+8+7+7} = \frac{11}{11+14}$

$Accuracy = \frac{11}{25} = 0.44$

Precision: $\frac{TP}{TP+FP} = \frac{3}{3+7} = \frac{3}{10} = 0.3$

Recall(TPR) = $\frac{TP}{TP+FN} = \frac{3}{3+7} = \frac{3}{10} = 0.3$

F1 Score = $\frac{2TP}{2TP+FP+FN} = \frac{2(3)}{2(3)+7+7}$

$F1score = \frac{6}{6+14} = \frac{6}{20} = 0.3$

TAKEAWAYS:

- Learnt about k-nn algorithm and usage.
- Learnt what is confusion matrix and its usage.
- Learnt what is precision and accuracy their role in confusion matrix.
- Learnt how to understand confusion matrix and its key importance associated with TP,FP,TN,FN.
- Calculated the Accuracy,precision,Recall and f1 score for the provided example.

BIBLIOGRAPHY:

- https://ai.plainenglish.io/understanding-confusion-matrix-and-applying-it-on-knn-classifier-on-iris-dataset-b57f85d05cd8
- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/Pick_an_evaluation_metric.html