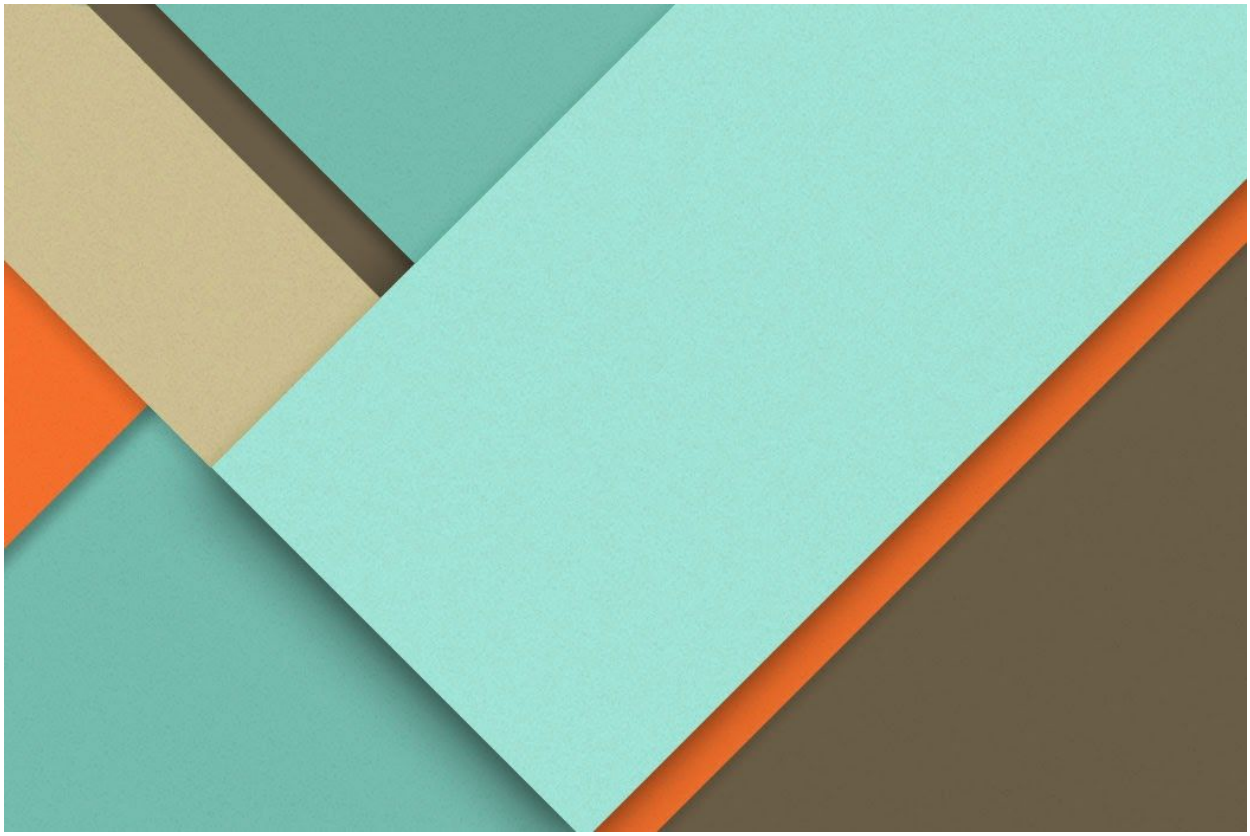


Assignment 1

Take 4 labeled datasets (multiclass) from UCI repository and find 10-fold cross-validation accuracy using NB Classifier.

Nitesh Yadav(35), Rohit Rana(40), Rohit Shakya(41)

M.Sc. Computer Science
Department of Computer Science
University of Delhi

Title of Database

Blocks Classification

Source

Donato Malerba
Department of Information
University of Bari

Overview

The problem consists in classifying all the blocks of the page layout of a document that has been detected by a segmentation process. This is an essential step in document analysis in order to separate text from graphic areas. Indeed, the five classes are:

1. **text**
2. **horizontal line**
3. **picture**
4. **vertical line**
5. **graphic**

Specifications

- The 5473 examples come from 54 distinct documents.
- All attributes are numeric.
- **Number of Instances:** 5473
- **Number of Attributes :**

  height: integer. | Height of the block.

  length: integer. | Length of the block.

  area: integer. | Area of the block (height * length);

  eccen: continuous. | Eccentricity of the block (length / height);

  p_black: continuous. | Percentage of black pixels within the block (blackpix / area);

p_and: continuous. | Percentage of black pixels after the application of the Run Length Smoothing Algorithm (RLSA)* (blackand / area);

mean_tr: continuous. | Mean number of white-black transitions (blackpix / wb_trans);

blackpix: integer. | Total number of black pixels in the original bitmap of the block.

blackand: integer. | Total number of black pixels in the bitmap of the block after the RLSA.

wb_trans: integer. | Number of white-black transitions in the original bitmap of the block.

***RLSA -** RUN LENGTH SMOOTHING ALGORITHM(RLSA) is a method mainly used to extract out the ROI(region of interest) with applied heuristics.

- **Missing Attribute Values**: No missing value

- **Class Distribution:**

Class	Frequency
text	4913
horiz. line	329
graphic	28
vert. line	88
picture	115

* Data is highly imbalanced.

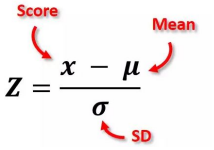
Preprocessing :

Why preprocessing? : It is useful to normalize the values of a normal distribution because they are easy to deal with using **z-score**.

Definition:

A z-score describes the position of a raw score in terms of its distance from the mean when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

Formula



Gaussian NB

In **Gaussian Naive Bayes**, continuous values associated with each feature are assumed to be distributed according to a **Gaussian** distribution. A **Gaussian** distribution is also called Normal distribution.

Gaussian Naive Bayes is an algorithm having a Probabilistic Approach. It involves prior and posterior probability calculation of the classes in the dataset and the test data given a class respectively.

$$Prior\ Probability(c) = \frac{No.of\ instances\ of\ class\ c}{Total\ No.of\ instances\ in\ the\ dataset}$$

Prior probabilities of all the classes are calculated using the same formula.

$$Posterior\ Probability(x \mid c) = P(x_1 \mid c) * P(x_2 \mid c) * P(x_3 \mid c) * \ldots * P(x_n \mid c)$$

K-Fold :

That method is known as "k-fold cross-validation". It's easy to follow and implement. Below are the steps for it:

1. Randomly split your entire dataset into k"folds"
2. For each k-fold in the dataset, build a model on k – 1 folds of the dataset. Then, test the model to check the effectiveness for kth fold
3. Record the error on each of the predictions
4. Repeat this until each of the k-folds has served as the test set
5. The average of k recorded errors is called the cross-validation error and will serve as a performance metric for the model

Stratified K-Fold

Stratification is the process of rearranging the data so as to ensure that each fold is a good representative of the whole.

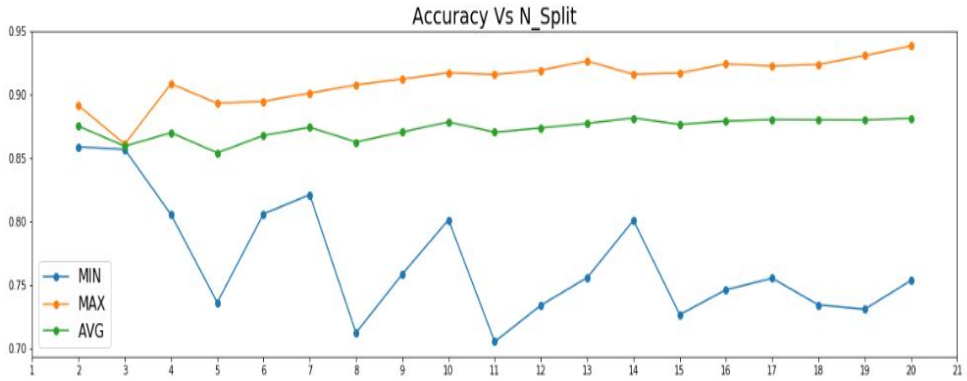
It is generally a better approach when dealing with both bias and variance. A randomly selected fold might not adequately represent the minor class, particularly in cases where there is a huge class imbalance.

Findings:

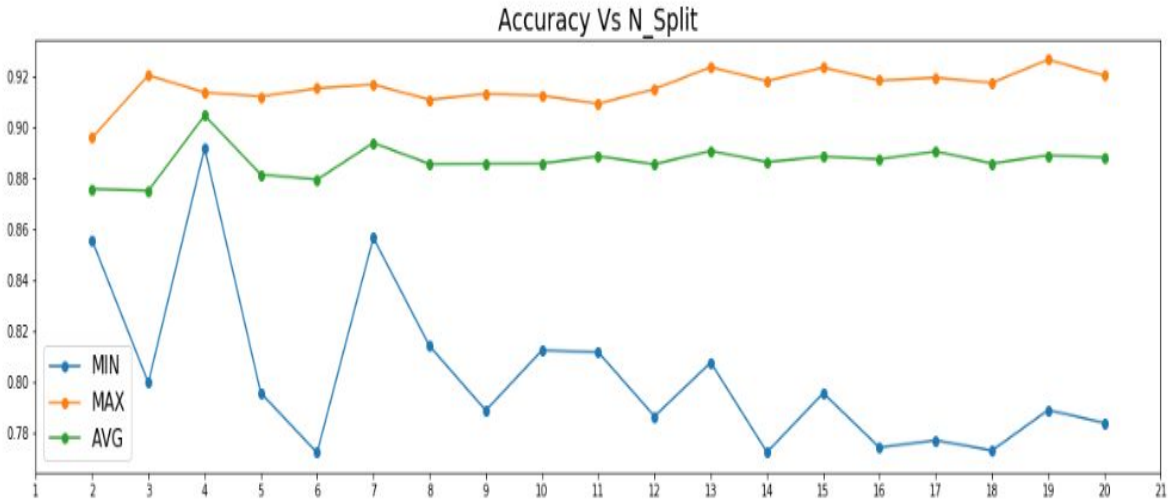
1. **5-Fold Cross-validation vs 10-Fold Cross-validation**

   The average accuracy in 10-fold cross-validation is less as compared to 5-fold cross-validation.

2. **cross-validation score on different split point in k-fold**



3. **cross validation score on different split point in stratified k-fold**



Conclusion:

We can clearly conclude that accuracy improves in stratified k-fold as compared to simple k-fold.