Group

AAssignmentNo:

6

TitleoftheAssignment:

1. Implement Simple Naïve Bayes classification algorithm using Python/R on iris.csvdataset.

2. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision,Recall on the given dataset.

ObjectiveoftheAssignment: StudentsshouldbeabletodataanalysisusingNaïveBayesclassification algorithm usingPython forany open source dataset

---

Prerequisite:

1. BasicofPythonProgramming

2.ConceptofJoinandMarginalProbability.

Contents forTheory:

1. **Concepts used inNaïve Bayes classifier**
2. **Naive Bayes Example**
3. **Confusion Matrix Evaluation Metrics**

---

1. Concepts used inNaïve Bayes classifier

- Naïve Bayes Classifier can be used for Classification of categorical data.
  - Let there be a 'j'number of classes. C={1,2,….j}
  - Let,inputobservationisspecifiedby'P'features.Thereforeinputobservationxis given ,x = {F1,F2,…..Fp}
  - TheNaïveBayesclassifierdependsonBayes'rulefromprobabilitytheory.
-

Priorprobabilities:Probabilitieswhicharecalculatedforsomeeventbasedonnootherinformation are called Prior probabilities.

For example, P(A), P(B), P(C) are prior probabilities because while calculating P(A),occurrencesofeventBorCarenotconcernedi.e.noinformationaboutoccurrenceofany other event is used.

ConditionalProbabilities:

$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)} \quad if\ P(B) \neq 0 \qquad \ldots\ldots(1)$$

$$P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)} \qquad \ldots\ldots\ldots(2)$$

From equation (1) and (2) ,

$$P(A \cap B) = P\left(\frac{A}{B}\right).P(B) = P\left(\frac{B}{A}\right).P(A)$$

$$\therefore \qquad P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right).P(A)}{P(B)}$$

Is called the Bayes Rule.

2. Example of Naive Bayes

WehaveadatasetwithsomefeaturesOutlook,Temp,Humidity,andWindy,andthetargethere is to predict whether a person orteam will play tennis or not.

Outlook	Temp	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

$$X = [\underbrace{Outlook}_{x_1}, \underbrace{Temp}_{x_2}, \underbrace{Humidity}_{x_3}, \underbrace{Windy}_{x_4}]$$

$$C_k = [\underbrace{Yes}_{C_1}, \underbrace{No}_{C_2}]$$

ConditionalProbability

Here, we are predicting the probability of class1 and class2 based on the given condition. If I tryto write the same formula in terms of classes and features, we will get the following equation

$$P(C_k \mid X) = \frac{P(X \mid C_k) * P(C_k)}{P(X)}$$

Now we have two classes and four features, so if we write this formula for class C1, it will besomething like this.

$$P(C_1 \mid x_1 \cap x_2 \cap x_3 \cap x_4) = \frac{P(x_1 \cap x_2 \cap x_3 \cap x_4 \mid C_1) * P(C_1)}{P(x_1 \cap x_2 \cap x_3 \cap x_4)}$$

Here, we replaced Ck with C1 and X with the intersection of X1, X2, X3, X4. You might have aquestion,It's because we are taking the situation when all these features are present at the sametime.

why only C1

The Naive Bayes algorithm assumes that all the features are independent of each other or in otherwords all the features are unrelated. With that assumption, we can further simplify the aboveformula and write it in this form

$$P(C_1 \mid x_1 \cap x_2 \cap x_3 \cap x_4) = \frac{P(x_1 \mid C_1) * P(x_2 \mid C_1) * P(x_3 \mid C_1) * P(x_4 \mid C_1) * P(C_1)}{P(x_1) * P(x_2) * P(x_3) * P(x_4)}$$

This is the final equation of the Naive Bayes and we have to calculate the probability of both C1and C2.For this particular example.

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes \mid X) = P(Rainy \mid Yes) \times P(Cool \mid Yes) \times P(High \mid Yes) \times P(True \mid Yes) \times P(Yes)$$

$$P(Yes \mid X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No \mid X) = P(Rainy \mid No) \times P(Cool \mid No) \times P(High \mid No) \times P(True \mid No) \times P(No)$$

$$P(No \mid X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

P(N0|Today)>P(Yes|Today)So,thepredictionthatgolfwouldbeplayedis‘No’.

Algorithm (IrisDataset):

Step1:ImportlibrariesandcreatealiasforPandas,NumpyandMatplotlibStep 2:

Import the Iris dataset by calling URL.

Step 3: Initializethe data

frameStep4:PerformDataPreproces

sing

- ConvertCategoricaltoNumericalValuesifapplicable

- `CheckforNullValue`

- Divide the dataset into Independent(X) andDependent(Y)variables.
- Splitthedatasetintotrainingandtestingdatasets
- ScaletheFeaturesifnecessary.

Step5:UseNaiveBayesalgorithm(TraintheMachine)toCreateModel

```
#importtheclass
fromsklearn.naive_bayesimportGaussianNBgaussian
=
GaussianNB()gaussian.fit(X_train,y_train)
```

Step6:Predictthe y_predforall valuesoftrain_xand test_x

```
Y_pred=gaussian.predict(X_test)
```

Step 7:Evaluate the performance of Model fortrain_y and test_y

```
accuracy=accuracy_score(y_test,Y_pred)
precision=precision_score(y_test,Y_pred,average='micro')recall=recall_sco
re(y_test,Y_pred,average='micro')
```

Step8:Calculatetherequiredevaluationparameters

```
fromsklearn.metricsimportprecision_score,confusion_matrix,accuracy_score,
recall_scorecm=confusion_matrix(y_test,Y_pred)
```

Conclusion:

InthiswaywehavedonedataanalysisusingNaiveBayesAlgorithmforIrisdatasetandevaluated the performance of the model.