2021년 패턴인식 설계과제 보고서

32181827 박종기

1. 설계 주제

iris data 의 특성을 바탕으로 Naive Bayes 분류기를 이용한 iris 의 species 예측

2. 설계 소개 및 배경지식

a. 이용하는 데이터 iris.csv 파일에 대한 소개

iris.csv 파일은 R에서 제공하는 데이터셋으로 3가지 종 setosa, versicolor, virginica 에 대하여 150개의 샘플들에 대해 각각의 꽃잎과 꽃받침 에 대한 길이와 너비에 대한 정보가 기술되어있는 파일이다. 길이 및 너비정보에 대한 단위는 cm이다.

b. Naive Bayes 분류기

기계 학습분야에서, 나이브 베이즈 분류(Naïve Bayes Classification)는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종으로 1950 년대 이후 광범위하게 연구되고 있다. 나이브 베이즈 분류는 텍스트 분류에 사용됨으로써 문서를 여러 범주중 하나로 판단하는 문제에 알맞은 방법중 하나이다.

3. 설계 내용

설계는 지도학습 방식을 사용하였다. 설계에 앞서 데이터를 training data 와 test data로 분류해야한다. training data와 test data는 서로 교집합이 없도록 구성한다. training data와 test data는 다음과 같은 방식으로 구성하였다.

```
15  set.seed(1005)
16  train_index <-sample(1:nrow(iris),size=0.8*nrow(iris),replace=FALSE)
17  test_index <-(-train_index)
18
19  train_data <-iris[train_index, ]
20  test_data <-iris[test_index, ]
21
22  View(train_data)
23  View(test_data)
```

iris.csv파일에는 같은 종별로 정렬되어있기 때문에 data들을 고르기 위해서 sampling하는 방식을 택해야 했다. training data는 위의 코드에서 볼 수 있듯이 전체 데이터 중 80%인 120개의 데이터를 무작위로 비복원 추출하였고 test data는 전체 데이터 150개에서 training data 120개를 제외한 나머지 30개로 구성하였다. training data와 test data가 겹치지 않고 무작위로 잘 추출된 것을 확인할 수 있었다.

training data와 test data 를 알맞게 구성한 이후 나이브 베이지안 분류기를 이용해 model1 과 model2를 구성하였다. model1은 꽃받침의 길이와 너비정보들을 바탕으로 iris의 종을 예측하는 모델이고 model2는 꽃잎의 길이와 너비정보들을 바탕으로 irsi의 종을 예측하는 모델이다.

```
27  install.packages("e1071")
28  library(e1071)
29
30  model1<-naiveBayes(train_data[,1:2],train_data$Species,laplace=0)
31
32  model2<-naiveBayes(train_data[,3:4],train_data$Species,laplace=0)
33
```

4. 설계 제한요소

iris의 데이터셋은 크기를 측정한 데이터들에 대한 정보이기 때문에 모든 iris데이터들에 대해 측정도구, 성장시기와 장소, 시간 등에 대한 변인통제가 어렵다. 또한 모집단에 대한 전수조사가 이루어지기는 어렵겠지만 iris데이터들의 전체개수는 150개이고 각각의 종들 에 대한 데이터가 50개씩 주어진 상황이다. 표본에 대한 개수가 부족하였다.

5. 성능평가

성능평가는 model1 에대한 성능평가와 model2에 대한 성능평가를 나누어 진행하였다. model1에 대한 성능평가는 result1에, model2에 대한 성능평가는 result2에서 확인할 수 있다.

```
34  result1<-predict(model1,test_data[,1:2])
35
36  result2<-predict(model2,test_data[,3:4])
37
38  correct_table<-table(iris$Species[test_index], result1)
39  View(correct_table)
40
41  correct_table2<-table(iris$Species[test_index], result2)
42  View(correct_table2)
43
```

분류현황을 명확하게 확인하기 위해 표 correct_table과 correct_table2 를 작성하였다. result1에

대한 표는 correct_table, result2에 대한 표는 correct_table2에 나타내었다. Var1은 정답에 대한 정보이고 result1 과 result2는 분류이후 결과를 보여주는 항목이다. Freq는 해당 항목에 대한 빈도 수이다. Var1과 result1또는 result2의 종이 일치하면 분류가 잘 이루어진 사례라고 볼 수 있다.

	Var1	result1	Freq
1	setosa	setosa	10
4	setosa	versicolor	1
7	setosa	virginica	0
2	versicolor	setosa	0
5	versicolor	versicolor	7
8	versicolor	virginica	6
3	virginica	setosa	0
6	virginica	versicolor	2
9	virginica	virginica	4

	Var1	result2	Freq
1	setosa	setosa	11
4	setosa	versicolor	0
7	setosa	virginica	0
2	versicolor	setosa	0
5	versicolor	versicolor	12
8	versicolor	virginica	1
3	virginica	setosa	0
6	virginica	versicolor	0
9	virginica	virginica	6

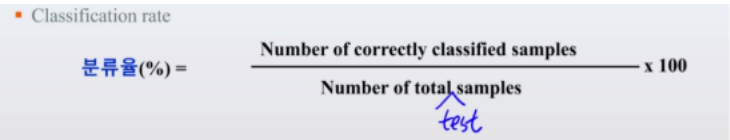
예를들어 result1의 두번째 행을 보면

4	setosa	versicolor	1

원래 정답은 setosa종이지만 분류기를 사용한 결과 versicolor로 오분류된 샘플 1개가 존재한다는 의미로 해석할 수 있다.

6. 결론

표의 내용을 바탕으로 Classification rate 분류율 및 분류오차를 계산해볼 수 있다.



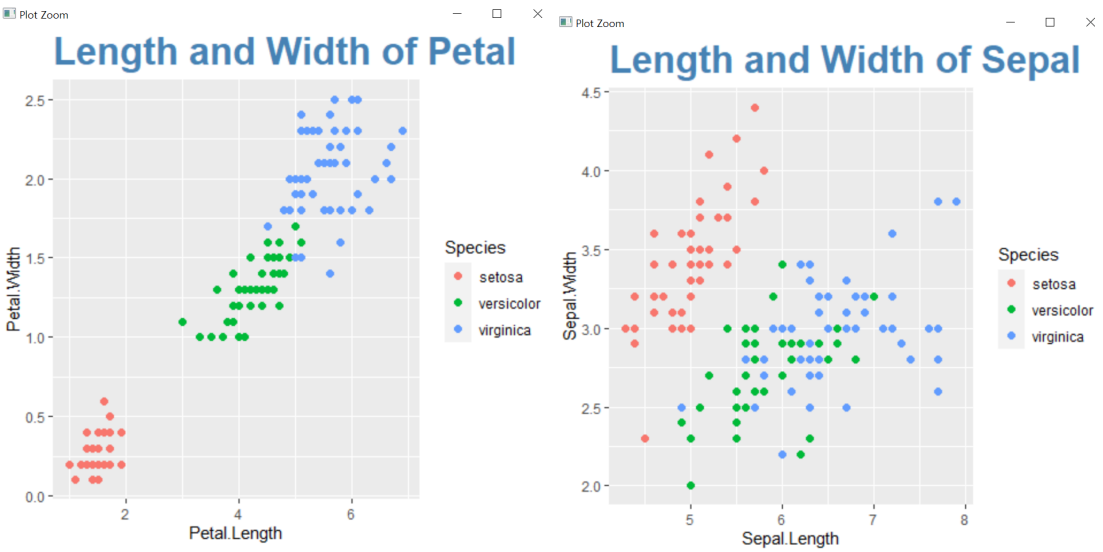
분류율은 전체 test_data 의 개수중 올바르게 분류된 데이터의 개수의 비율을 구하는 과정이고 분류오차는 전체 중 분류율을 뺀 값으로 다음과 같이 구할 수 있다.

꽃받침의 길이와 너비정보를 바탕으로 종을 예측한 model1의 경우 분류율은 70%, 분류오차는 30%이고 꽃잎의 길이와 너비정보를 바탕으로 종을 예측한 model2의 경우 분류율은 약 96.7%, 분류오차는 3.3%라는 수치를 얻을 수 있다. 꽃잎의 길이와 너비를 바탕으로 종을 예측하는 방법이 꽃받침의 길이와 너비정보를 바탕으로 종을 예측하는 방법보다 분류가 잘 이루어진다고 해석해볼 수 있다.

분포도를 이용하여 확인해볼수도 있었다. 꽃받침의 길이와 너비에 따른 종의 분포도와 꽃잎의 길이와 너비에 따른 종의 분포도를 나타내보았다. 분포도는 다음과 같은 코드를 이용하여 나타내었다.

```
.libPaths("c:/Rpackages")
install.packages('ggplot2')
library(ggplot2)

View(iris)

ggplot(data=iris, aes(x=Petal.Length,y=Petal.Width,color=Species))+
  geom_point(size=2)+ggtitle('Length and Width of Petal')+
  theme(plot.title = element_text(size=25, face='bold',
                                  colour = 'steelblue'))

ggplot(data=iris, aes(x=Sepal.Length,y=Sepal.Width,color=Species))+
  geom_point(size=2)+ggtitle('Length and Width of Sepal')+
  theme(plot.title = element_text(size=25, face='bold',
                                  colour = 'steelblue'))

```

R studio의 ggplot 을 사용하여 시각화하는 방식을 사용하였다. 두개의 창을 만들었는데 첫번째는 꽃잎의 길이에 따른 너비가 어떻게 되는지를, 두번째는 꽃받침의 길이에 따른 너비가 어떻게 되는지를 종별로 색상을 구분하여 분포도를 구성하였다.



빨간색 점은 setosa, 초록색 점은 versicolor, 파란색 점은 virginica 를 나타낸다. 꽃잎 또는 꽃받침의 길이 및 너비정보를 가지고 decision boundary 를 설정할 때에도 꽃잎의 정보를 사용하는 것이 유리해보인다.

설계 제한요소에 기술하였던 test data의 개수가 적다는 단점을 보완하기 위해 결과로 도출한 분류율 및 분류오차를 신뢰하기 힘들다. 따라서 교차검증방법 class validation 방식을 사용하여 분류율 및 분류오차를 구해볼수도 있다.