Data Mining
팀 프로젝트 보고서

노트북 제품 추천

과목 교수님	노상욱 교수님
전공	컴퓨터정보공학부
학년	4학년
팀원	201620727 송민아 201621132 이정은 201621504 최윤지
제출 일자	2020년 12월 03일 (목)

목 차

▣ 주제

대부분의 사람들은 노트북을 하나 이상씩은 가지고 있다. 노트북은 다양한 용도와 이유로 앞으로도 계속 사용될 것이다. 노트북을 바꾸거나 처음 사는 사람들은 자신에게 어떤 노트북이 맞는 것인지, 유용하게 사용할 수 있을지 가늠하기 어렵다. 따라서 이 주제를 통해 그런 어려움을 겪고 있는 사람들에게 도움이 될 수 있을 것 같아 선정하게 되었다.

▣ 속성

→ 성별 : 남, 여
→ 연령대 : 10대, 20대, 30대, 40대 이상
→ 문서 작업 : YES, NO
→ 디자인 & 영상 작업 : YES, NO
→ 고사양 게임 : YES, NO
→ 터치스크린, 펜의 사용여부 : YES, NO
→ 무게 : 1kg 미만, 1kg~1.3kg, 1.3kg이상
→ 예산 : 100만원 미만, 100~160만원, 160만원 이상
→ **class** : **Samsung Galaxy Book Flex, Samsung Always 9, Apple MacBook Pro, Apple MacBook Air, LG gram 14, LG 울트라 PC**

▣ 데이터의 수집 방법

→ 데이터는 설문조사를 통해 수집하였습니다.

	gender	age	Document	Design &	High-end	Writing	weight	budget	product
2	Male	20~29	YES	YES	NO	NO	1.3kg ~	160~	Apple MacBook
3	Male	20~29	YES	NO	NO	YES	1.1kg ~ 1.	160~	Samsung Galaxy
4	Female	20~29	YES	NO	NO	YES	1.1kg ~ 1.	160~	Samsung Galaxy
5	Female	30~39	NO	NO	YES	NO	1.1kg ~ 1.	~100	LG UltraPC
6	Male	20~29	YES	YES	YES	NO	1.1kg ~ 1.	100~160	Apple MacBook
7	Male	20~29	YES	NO	YES	NO	1.1kg ~ 1.	~100	Samsung Always
8	Female	20~29	YES	YES	NO	NO	1.3kg ~	160~	Apple MacBook
9	Male	20~29	YES	NO	NO	NO	~ 1kg	100~160	LG gram14
10	Female	20~29	YES	NO	NO	YES	1.1kg ~ 1.	160~	Samsung Galaxy
11	Female	40~	YES	NO	NO	YES	1.1kg ~ 1.	160~	Samsung Galaxy
12	Female	20~29	YES	NO	YES	YES	1.1kg ~ 1.	160~	Samsung Galaxy
13	Female	20~29	YES	NO	NO	YES	~ 1kg	100~160	Samsung Always

→ 설문조사를 통해 얻은 데이터들을 .csv파일로 변환한 후 Weka 실습을 실행하였습니다.

▣ 알고리즘

1) 1R

```
=== Classifier model (full training set) ===

weight:
        1.3kg ~ -> Apple MacBook pro
        1.1kg ~ 1.3kg   -> Samsung Galaxy Book Flex
        ~ 1kg   -> Samsung Always 9
(134/242 instances correct)


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         134               55.3719 %
Incorrectly Classified Instances       108               44.6281 %
```

→ instances의 개수 : 242개

→ 기준이 되는 속성 : 무게 **~1kg : Samsung Always 9**
1.1kg ~ 1.3kg : Samsung Galaxy Book Pro
1.3kg~ : Apple MacBook Pro

→ 정확도 : 55.3719%

→ 정밀도 : 31.083%

→ 재현율 : 55.4%

▶ 가장 높은 정확도 = 171개일 때, 55.6%이다.

▶ 가장 낮은 정확도 = 24개일 때, 29.2%이다.

2) Decision Tree

```
budget = 160~
|   Writing = NO: Apple MacBook pro (38.0)
|   Writing = YES
|   |   weight = 1.3kg ~: Samsung Galaxy Book Flex (2.0)
|   |   weight = 1.1kg ~ 1.3kg: Samsung Galaxy Book Flex (51.0/2.0)
|   |   weight = ~ 1kg: LG gram14 (2.0)
budget = ~100
|   Design & media = YES: Apple MacBook Air (2.0)
|   Design & media = NO: LG UltraPC (26.0/2.0)
budget = 100~160
|   weight = 1.3kg ~: Samsung Always 9 (0.0)
|   weight = 1.1kg ~ 1.3kg
|   |   Design & media = YES: Apple MacBook Air (33.0/3.0)
|   |   Design & media = NO
|   |   |   age = 20~29: LG gram14 (3.0/1.0)
|   |   |   age = 30~39
|   |   |   |   Writing = NO: Apple MacBook Air (2.0)
|   |   |   |   Writing = YES: Samsung Galaxy Book Flex (2.0)
|   |   |   age = 40~: Samsung Galaxy Book Flex (0.0)
|   |   |   age = 10~19: Samsung Always 9 (1.0)
|   weight = ~ 1kg
|   |   Writing = NO
|   |   |   age = 20~29
|   |   |   |   High-end games = NO
|   |   |   |   |   gender = Male: Samsung Always 9 (7.0/2.0)
|   |   |   |   |   gender = Female: LG gram14 (29.0/11.0)
|   |   |   |   High-end games = YES: LG gram14 (8.0/2.0)
|   |   |   age = 30~39: Samsung Always 9 (13.0/5.0)
|   |   |   age = 40~: Samsung Always 9 (10.0/5.0)
|   |   |   age = 10~19: Samsung Always 9 (9.0)
|   |   Writing = YES: Samsung Always 9 (4.0)

Number of Leaves  :     20

Size of the tree :      32


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         186               76.8595 %
Incorrectly Classified Instances        56               23.1405 %
```

→ instances의 개수 : 242개

→ 정확도 : 76.8595%

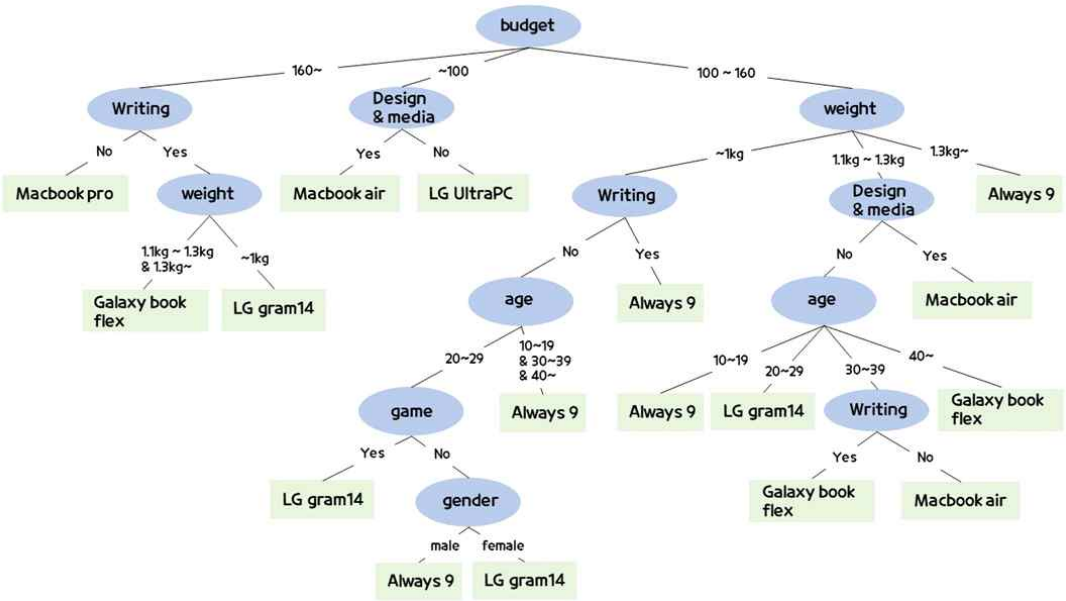
→ 정밀도 : 76.8%

→ 재현율 : 76.9%

▶ 가장 높은 정확도 = 122개일 때, 77.9%이다.

▶ 가장 낮은 정확도 = 24개일 때, 58.3%이다.

→ Decision Tree는 다음과 같이 나타낼 수 있다.



3) Naive Bayes

```
=== Classifier model (full training set) ===

Naive Bayes Classifier
```

Attribute	Class Apple MacBook pro (0.17)	Samsung Galaxy Book Flex (0.22)	LG UltraPC (0.1)	Samsung Always 9 (0.2)	LG gram14 (0.17)	Apple MacBook Air (0.14)
gender						
Male	25.0	31.0	14.0	26.0	10.0	15.0
Female	19.0	25.0	10.0	24.0	32.0	21.0
[total]	44.0	56.0	26.0	50.0	42.0	36.0
age						
20~29	27.0	17.0	2.0	25.0	29.0	13.0
30~39	9.0	15.0	8.0	10.0	6.0	8.0
40~	5.0	13.0	2.0	6.0	8.0	6.0
10~19	5.0	13.0	14.0	11.0	1.0	11.0
[total]	46.0	58.0	28.0	52.0	44.0	38.0
Document						
YES	34.0	46.0	19.0	47.0	40.0	32.0
NO	10.0	10.0	7.0	3.0	2.0	4.0
[total]	44.0	56.0	26.0	50.0	42.0	36.0
Design & media						
YES	41.0	10.0	1.0	5.0	6.0	33.0
NO	3.0	46.0	25.0	45.0	36.0	3.0
[total]	44.0	56.0	26.0	50.0	42.0	36.0
High-end games						
NO	25.0	28.0	13.0	40.0	35.0	24.0
YES	19.0	28.0	13.0	10.0	7.0	12.0
[total]	44.0	56.0	26.0	50.0	42.0	36.0
Writing						
NO	41.0	1.0	25.0	45.0	39.0	29.0
YES	3.0	55.0	1.0	5.0	3.0	7.0
[total]	44.0	56.0	26.0	50.0	42.0	36.0
weight						
1.3kg ~	39.0	3.0	3.0	1.0	1.0	2.0
1.1kg ~ 1.3kg	5.0	53.0	23.0	5.0	3.0	34.0
~ 1kg	1.0	1.0	1.0	45.0	39.0	1.0
[total]	45.0	57.0	27.0	51.0	43.0	37.0
budget						
160~	41.0	52.0	1.0	1.0	3.0	1.0
~100	1.0	1.0	25.0	3.0	1.0	3.0
100~160	3.0	4.0	1.0	47.0	39.0	33.0
[total]	45.0	57.0	27.0	51.0	43.0	37.0

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         197               81.405 %
Incorrectly Classified Instances        45               18.595 %
```

→ instances의 개수 : 242개

→ 정확도 : 81.405%

→ 정밀도 : 81.3%

→ 재현율 : 81.4%

▶ 가장 높은 정확도 = 146개일 때, 82.2%이다.

▶ 가장 낮은 정확도 = 24개일 때, 54.2%이다.

4) Association Rule

```
Best rules found:

 1. Design & media=NO High-end games=NO 106 ==> Document=YES 106    <conf:(1)> lift:(1.14) lev:(0.05) [13] conv:(13.14)
 2. High-end games=NO budget=100~160 96 ==> Document=YES 92    <conf:(0.96)> lift:(1.09) lev:(0.03) [7] conv:(2.38)
 3. High-end games=NO Writing=NO 122 ==> Document=YES 116    <conf:(0.95)> lift:(1.09) lev:(0.04) [9] conv:(2.16)
 4. gender=Female High-end games=NO 98 ==> Document=YES 93    <conf:(0.95)> lift:(1.08) lev:(0.03) [7] conv:(2.02)
 5. High-end games=NO 159 ==> Document=YES 150    <conf:(0.94)> lift:(1.08) lev:(0.04) [10] conv:(1.97)
 6. budget=100~160 121 ==> Document=YES 114    <conf:(0.94)> lift:(1.08) lev:(0.03) [8] conv:(1.88)
 7. Writing=NO budget=100~160 110 ==> Document=YES 103    <conf:(0.94)> lift:(1.07) lev:(0.03) [6] conv:(1.7)
 8. gender=Female Writing=NO 92 ==> Document=YES 86    <conf:(0.93)> lift:(1.07) lev:(0.02) [5] conv:(1.63)
 9. Design & media=NO Writing=NO 101 ==> Document=YES 93    <conf:(0.92)> lift:(1.05) lev:(0.02) [4] conv:(1.39)
10. gender=Female 125 ==> Document=YES 115    <conf:(0.92)> lift:(1.05) lev:(0.02) [5] conv:(1.41)
```

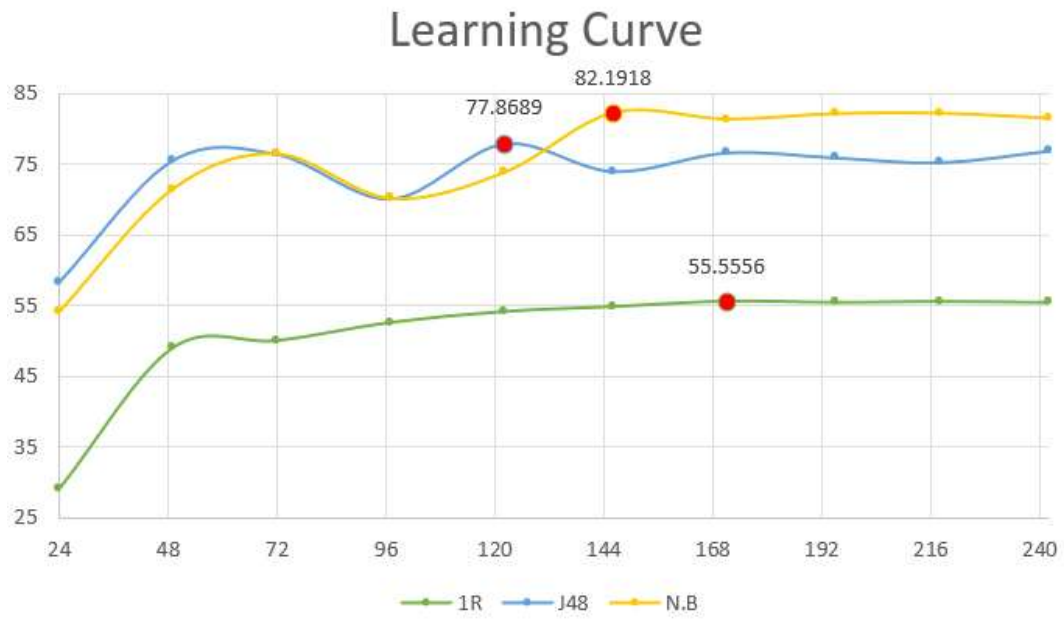
→ 242개의 데이터를 적용하여 Association Rule을 구현하였다.

→ 정확도 100%의 연관 규칙은 1가지

디자인 & 영상 작업을 하지 않고 고 사양 게임을 하지 않으면 문서 작업을 한다.

→ 더 높은 정확도를 내는 기계 학습 알고리즘을 사용하는 것이 유리하다.

▣ Learning Curve



→ 1 Rule, DECISION TREE (J48), NAÏVE BAYES (N.B)의 학습 곡선

→ 1 Rule : 171개의 인스턴스에서 가장 높은 정확도 55.5556%가 나온 것을 확인할 수 있다.

→ DECISION TREE (J48) : 122개의 인스턴스에서 가장 높은 정확도 77.8689%가 나온 것을 확인할 수 있다.

→ NAÏVE BAYES (N.B) : 146개의 인스턴스에서 가장 높은 정확도 82.1918%가 나온 것을 확인할 수 있다.

▣ ANOVA

1R	J48	N.B
58.2781	79.4118	80.9524
58.3851	75	82.3529
55.5556	77.8689	82.1918
55.8011	78.0303	80.7692
55.4974	74.6479	81.3253

→ 각각의 Rule에 대해 Learning Curve로 확인한 가장 높은 정확도를 가지는 인스턴스의 주변 값들의 정확도

→ ANOVA TEST를 하기 위해서 Learning Curve로 확인한 가장 높은 정확도를 보인 인스턴스를 중앙값으로 각 인스턴스의 차이를 10으로 두고 N을 5로 설정하였다. 분산 분석: 일원 배치법을 사용하여 평균 및 분산 값을 확인하고 F 비와 90%, 95%, 99%의 F 기각치를 계산하였다.

분산 분석: 일원 배치법

요약표

인자의 수준	관측수	합	평균	분산
1R	5	283.52	56.70	2.22
J48	5	384.96	76.99	4.29
N.B	5	407.59	81.52	0.52

분산 분석

변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치(90)	F 기각치(95)	F 기각치(99)
처리	1746.47	2	873.24	372.54	1.59E-11	2.81	3.89	6.93
잔차	28.13	12	2.34					
계	1774.60	14						

→ $X \sim F(k-1, k(n-1), \alpha)$인 F 분포를 따라 k = 3, n = 5로 자유도 $X \sim F(2,12)$ 와 F 비 값을 구하였다.

$$F_{0.1}(2,12) = 2.81 \qquad F_{0.05}(2,12) = 3.89 \qquad F_{0.01}(2,12) = 6.93$$

→ F 비의 값이 372.54로 F 기각치(90, 95, 99)보다 큰 것으로 보아 유의한 차이 값을 지니고 있다. 따라서 더 높은 정확도를 제공하는 NAÏVE BAYES 알고리즘을 사용하는 것이 더 유리하다.

▣ Bernoulli Distribution

$$P = (f + \frac{z^2}{2N} \pm Z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}})/(1 + \frac{z^2}{N})$$

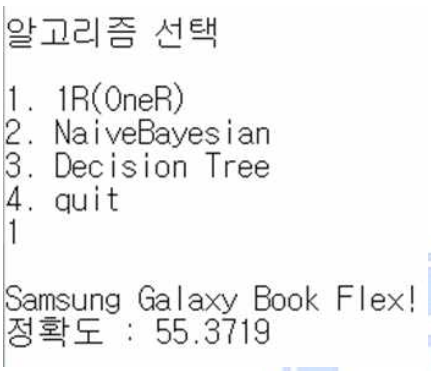
→ 다음과 같은 공식을 이용하여 P의 범위 값을 구하였습니다.
이때의 f는 각각의 Rule들의 정확도, N은 instances의 수이므로 242개, z=1.65(90%), z=1.96(95%), z=2.58(99%)로 계산하였습니다.

	z=1.65(90%)	z=1.96(95%)	z=2.58(99%)
1R(f=0.554)	[0.5071<P<0.6120]	[0.4997<P<0.6240]	[0.4864<P<0.6490]
NB(f=0.814)	[0.7784<P<0.8608]	[0.7730<P<0.8708]	[0.7635<P<0.8919]
DT(f=0.769)	[0.7300<P<0.8191]	[0.7240<P<0.8297]	[0.7134<P<0.8520]

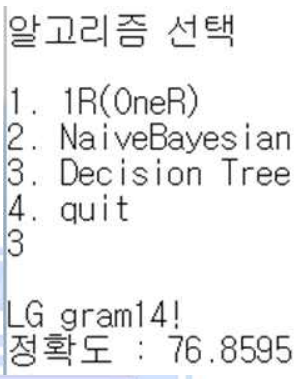
▣ 새로운 시험 데이터

▶ 첫 번째 시험 데이터

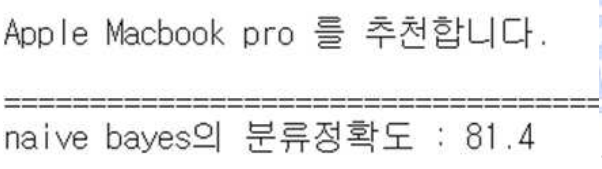
Gender	Age	Document	Design & Media	High-end Game	writing	weight	Budget
Female	20 ~ 29	YES	NO	NO	NO	1.1kg ~ 1.3kg	100 ~ 160



<1 Rule>



<Decision Tree>

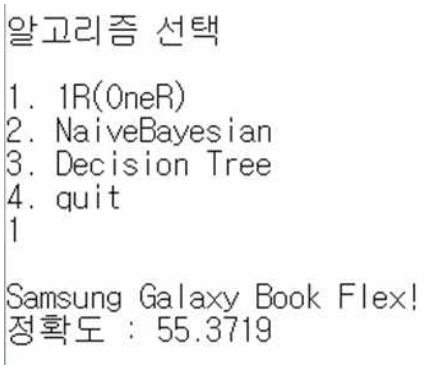


<Naive Bayes>

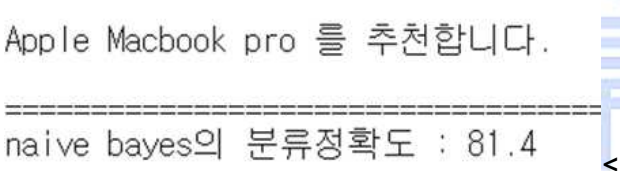
→ 1 Rule의 경우 무게로 class를 추천하기 때문에 Samsung Galaxy Book Flex로 추천하고, Decision Tree는 앞서 보았던 트리 그래프를 참고하면 LG gram14를 추천한다. Naive Bayes는 Apple Macbook pro를 추천하는 것을 볼 수 있다.

▶ 두 번째 시험 데이터

Gender	Age	Document	Design & Media	High-end Game	writing	weight	Budget
Male	40~	NO	YES	YES	YES	1.1kg ~ 1.3kg	100 ~ 160



<1 Rule> **<Decision Tree>**



<Naive Bayes>

→ 1 Rule의 경우 무게로 class를 추천하기 때문에 Samsung Galaxy Book Flex로 추천하고, Decision Tree는 앞서 보았던 트리 그래프를 참고하면 Apple Macbook air를 추천한다. Naive Bayes는 Apple Macbook pro를 추천하는 것을 볼 수 있다.

▣ 결론 & 고찰

→ Learing Curve를 통해 1 Rule은 171개의 데이터에서 55.6%, Decision Tree는 122개의 데이터에서 77.9% Naive Bayes는 146개의 데이터에서 82.2%인 가장 높은 정확도를 확인할 수 있었다.

→ 설정한 도메인의 class는 6개인데 속성은 4개가 최대라서 1 Rule에서는 6개의 클래스를 모두 구분할 수 없어서 정확도가 현저히 떨어졌다. 이러한 결과로 인해 1 Rule을 사용하여 class를 구분할 때는 class 개수 이상의 속성값이 필요하다는 것을 깨달았다.

→ ANOVA Test를 통해 1 Rule과 Naive Bayes의 차이가 커서 우연하다는 가설을 기각하고 통계적으로 유의하여 가장 높은 정확도를 보이는 기계학습 알고리즘 사용이 유리하다는 결과를 얻었다.

▣ 역할 분담

▶ 최윤지 : 설문조사, 엑셀 작업, Learning Curve, ANOVA Test, 보고서 작성

▶ 송민아 : 설문조사, 엑셀 작업, Naive Bayes 알고리즘 작성, PPT 작성

▶ 이정은 : 설문조사, Bernoulli Distribution, 1 Rule, Decision Tree 알고리즘 작성, 보고서 작성

▣ 참고문헌/자료

▶ 데이터 마이닝 강의자료