2021년도 2학기 바이오빅데이터와데이터마이닝
과제2 보고서

이화여자대학교 컴퓨터공학과
1871056 한지수

1번 문제 분석

새로운 자료를 X라 가정하면, 새로운 자료는 다음과 같이 정의할 수 있다.

X = (키=6, 몸무게=130, 발 크기=8)

세 가지 속성 모두 독립이며 연속형 속성이고, 다음과 같이 정규분포를 따른다.

#Attribute	Domain
1. 키(feet)	연속형 속성
2. 몸무게(lbs)	연속형 속성
3. 발 크기(inches)	연속형 속성
4. 성별	분류형 속성 (분류 목표)

위의 속성을 바탕으로 새로운 자료 X의 성별을 구분해야 한다.

이를 위해 X의 사후 확률을 구분해야 하며, 사후 확률이 큰 집단으로 X를 분류하게 된다.

순수 베이즈 분류기 공식에 따라 P(X|성별 = 여자)P(성별 = 여자), P(X|성별=남자)P(성별=남자) 크기를 비교해야 한다.

훈련 자료로 보아 P(성별=여자) = P(성별=남자) = 0.5이므로, 이를 제외한 P(X|성별=여자), P(X|성별=남자)를 비교하면 된다.

각 확률을 구하기 위한 공식은 다음과 같이 구할 수 있다.

P(X | 성별 = 여자) = P(키=6 | 성별 = 여자) × P(몸무게=130 | 성별 = 여자) × P(발 크기=8 | 성별 = 여자)

P(X | 성별 = 남자) = P(키=6 | 성별 = 남자) × P(몸무게=130 | 성별 = 남자) × P(발 크기=8 | 성별 = 남자)

이를 위해 앞선 속성의 평균과 분산 데이터로부터 확률을 구해야 하며, 본 문제 풀이를 위해 소수점 다섯째자리에서 반올림하여 나타냈다.

1871056 한지수

1번 문제 풀이

For 키

P(키=6| 성별 = 남자) = $\frac{1}{\sqrt{2\pi}(3.5033e-02)} e^{-\frac{(6-5.855)^2}{2(3.5033e-02)^2}} = 1.5789$

P(키=6| 성별 = 여자) = $\frac{1}{\sqrt{2\pi}(9.7225e-02)} e^{-\frac{(6-5.4175)^2}{2(9.7225e-02)^2}} = 0.2235$

For 몸무게

P(몸무게=130 | 성별 = 남자) = $\frac{1}{\sqrt{2\pi}(1.2292e+02)} e^{-\frac{(130-176.26)^2}{2(1.2292e+02)^2}} = 0.0000$

P(몸무게=130 | 성별 = 여자) = $\frac{1}{\sqrt{2\pi}(5.5833e+02)} e^{-\frac{(130-132.5)^2}{2(5.5833e+02)^2}} = 0.0168$

For 발 크기

P(발 크기=8| 성별 = 남자) = $\frac{1}{\sqrt{2\pi}(9.1667e-01)} e^{-\frac{(8-11.25)^2}{2(9.1667e-01)^2}} = 0.0013$

P(발 크기=8| 성별 = 여자) = $\frac{1}{\sqrt{2\pi}(1.6667e+00)} e^{-\frac{(8-7.5)^2}{2(1.6667e+00)^2}} = 0.2867$

앞선 사후확률 정의에 따라 구한 확률들을 모두 종합해보면 다음과 같이 각 확률이 나타난다.

P(X | 성별 = 남자) = P(키 = 6 | 성별 = 남자) × P(몸무게=130 | 성별 = 남자) × P(발 크기=8 | 성별 = 남자) = 0

P(X | 성별 = 여자) = P(키 = 6 | 성별 = 여자) × P(몸무게=130 | 성별 = 여자) × P(발 크기=8 | 성별 = 여자) = 0.0011

Since P(X | 성별 = 남자)P(남자) < P(X | 성별 = 여자)P(여자), Therefore P(남자|X) < P(여자|X) ➔ 즉, 새로운 자료는 여자로 분류된다.

1871056 한지수

1번 문제 분석

RandomForest.R 코드 파이프라인 및 결정트리 결과

1. Dataset Preperation
    a. 데이터 셋의 column name 수정 (2-11행: a-i, class)
    b. 결측치 처리: “?”, NA 가진 instance
2. Split training set and test set (train: test = 0.7:0.3)
3. Training set의 Random ForestTree: ctree() "conditional inference tree", Random Forest Tree인 bio_ctree 생성
4. Classification with test set and check accuracy

⇒ bio_ctree: 노드 1,2,4,7에서 4개의 분할 생성

Train dataset = 480 instances, Test dataset = 202 instances

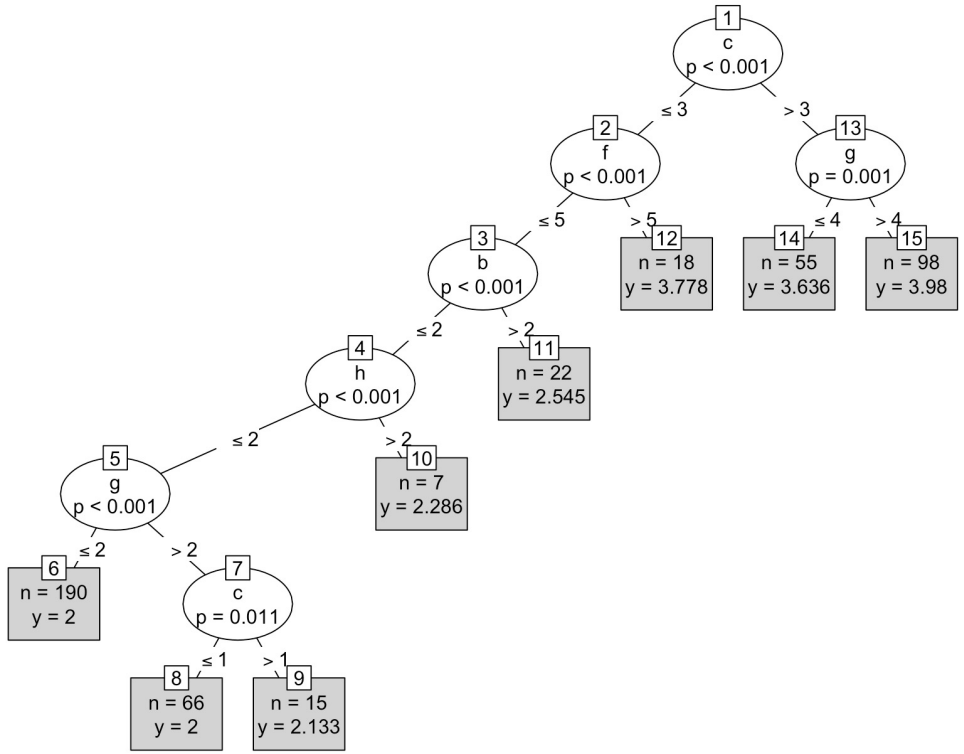
Test set accuracy: 72.4444%

➢ 본 문제는 중간고사 대체 과제와 동일하게 2-10행 중 3,4,5,7 행(Column b,c,d,f) 속성이 결정트리 분류 속성에 들어가 문제를 해결하려 하였으나 다음과 같은 Warning 메시지를 받았습니다. 이 메시지가 전체 Random Forest 생성에 있어 영향을 미칠 수 있다고 판단하여 중간고사 대체 과제와 과제 2 모두 2-10행 중 3,4,5,7,8,9,10 행(Column b,c,d,f,g,h,i)으로 입력데이터를 수정하여 모델을 만들었습니다.(Decision Tree: bio_ctree, Random Forest 모델: bio_rf)

2번 문제 풀이 및 결과 분석

Decision Tree 생성 결과 ⇒ bio_ctree

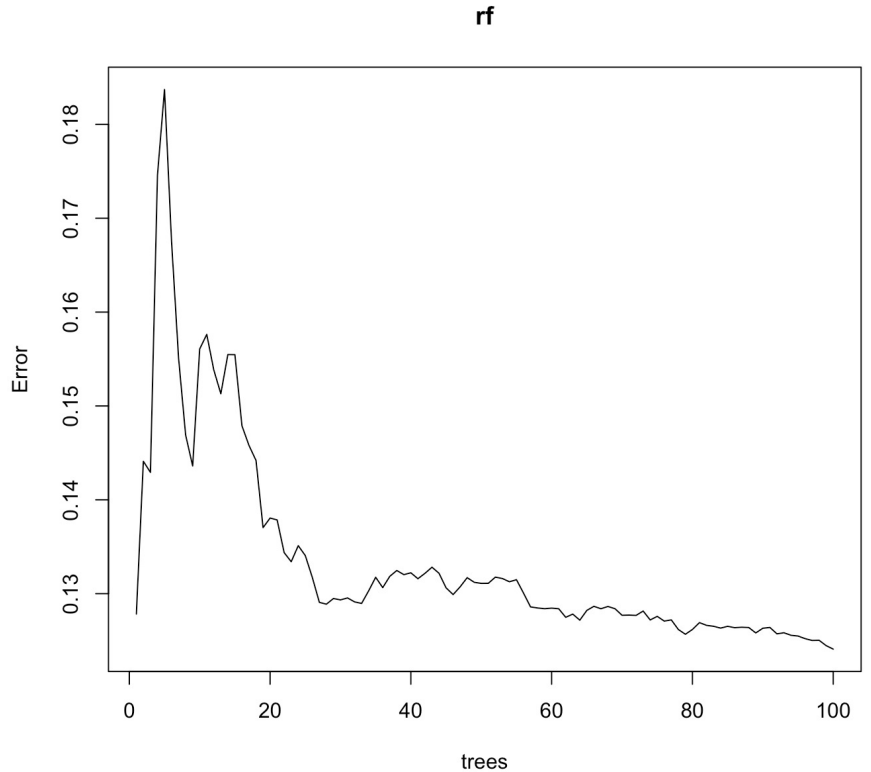
error.rpart(오분류율): 0.9763033, Test set accuracy: 54.9763%



〔최종적으로 생성된 bio_ctree 결정트리〕

Random Forest 생성 결과 ⇒ bio_rf

error.rpart(오분류율): 0.5948718, Test set accuracy: 45.12821%



〔최종적으로 생성된 bio_rf 모델〕

➔ Decision Tree가 오분류율, 예측정확도 모두 더 높은 경향을 볼 수 있다.