33회 기출문제

데이터 준전문가

ADSP, Advanced Data Analytics semi-Professional

류영표 강사
ryp1662@gmail.com

연습문제

1과목. 데이터 이해

연습문제

1. 데이터 사이언티스트가 갖추어야 하는 역량 중 소프트 스킬에 해당하지 않는 것은?

① 통찰력 있는 분석

② 설득력 있는 전달

③ 다분야 간 협력

④ 빅데이터 이론적 지식

연습문제

2. 데이터 크기를 작은 것부터 큰 것 순서로 올바르게 나열한 것은?

① PB < EB < ZB < YB

② PB < YB < EB < ZB

③ YB < ZB < EB < PB

④ PB < ZB < EB < YB

연습문제

3. 데이터베이스의 특징이 아닌 것은?

① USB 등 컴퓨터가 접근할 수 있는 저장 매체에 저장할 수 있다.

② 데이터는 공동으로 이용된다.

③ 데이터가 중복되어 있지 않다.

④ 정형 데이터만 저장할 수 있다.

연습문제

4. 미래의 빅데이터 관점에서 볼 때 사물인터넷(IoT)과 가장 관련이 큰 것은?

① 모든 사물의 데이터화

② 모든 사물의 독립화

③ 모든 사물의 그래픽화

④ 모든 사물의 정형화

연습문제

5. 빅데이터가 가져온 변화로 맞지 않은 것은?

① 서비스 산업이 확대되고 제조업의 생산성이 감소되었다.

② 빅데이터 시대에는 데이터 획득비용이 기하급수적으로 감소하고 모든 곳에서 데이터가 넘쳐나 사용자 전수조사가 가능해졌다.

③ 가능한 한 많은 데이터를 모으고 그 데이터를 다양한 방식으로 조합해 숨은 정보를 찾아낸다.

④ 데이터의 질보다 양을 강조하게 되었다.

연습문제

6. 다음 중 빅데이터 위기 요인과 해결 방안을 잘못 연결된 것을 고르시오.

가. 사생활 침해 → 동의제를 책임제로 전환

나. 책임훼손의 원칙 → 알고리즘 허용

다. 데이터의 오용 → 결과 기반 책임 원칙

① 가, 나 ② 가, 다 ③ 나, 다 ④ 가, 나, 다

연습문제

7. 데이터의 NoSQL 저장방식과 관련이 없는 도구는?

① MongoDB

② Hbase

③ Redis

④ MySQL

연습문제

8. 빅데이터 특징 중 올바르지 않은 것은?

① 비즈니스 상황에서는 인과관계를 모르고 상관관계 분석만으로 충분한 경우가 많다.

② 사전처리에서 사후처리 시대로 변화하였고, 사전처리의 대표적인 예로는 표준화된 문서 포맷을 들 수 있다.

③ 표본조사의 중요성이 높아졌다.

④ 데이터 수가 증가함에 따라 몇 개의 오류데이터가 대세에 영향을 주지 못하는 경향이 증가하고 있다.

연습문제

9. 조직의 의사결정을 위한 데이터 집합체로 데이터 통합, 시계열성, 비소멸성 등의 특징을 가지고 있는 것은?

연습문제

10. 다양한 유형의 데이터를 다루는 통계학과 마이닝을 넘어서는 학문, 데이터 공학, 수학, 통계학, 컴퓨터 공학 등 해당 분야의 전문 지식을 종합한 학문은?

연습문제

2과목. 데이터 분석 기획

연습문제

1. 빅데이터 분석 방법론에서 분석 기획 단계의 task로 적절하지 않은 것은?

① 비즈니스 이해 및 범위 설정

② 프로젝트 정의 및 계획 수립

③ 프로젝트 위험 계획 수립

④ 필요 데이터 정의

연습문제

2. 빅데이터 분석방법론의 계층적 프로세스 모델에 대한 설명으로 적절하지 않은 것은?

① Task는 단계를 구성하는 단위 활동으로 input, output로 구성된 단위 프로세스이다.

② Phase(단계)는 최상위 단계로 프로세스 그룹을 통하여 완성된 단계별 산출물을 생성한다.

③ Step(스텝)은 마지막 계층으로 WBS(Work Breakdown Structure)의 워크패키지에 해당한다.

④ Phase, Task, Step 계층이 있다.

연습문제

3. 분석 과제 우선순위 선정 매트릭스에 관한 설명 중 가장 적절하지 않은 것은?

① 시급성의 판단기준은 전략도 중요도와 비용범위에 따라 난이도는 분석수준과 복잡도 평가로 구분한다.

② 데이터 분석 과제를 추진할 때, 우선 고려해야 하는 요소는 전략도 중요도에 따른 시급성이 가장 중요한 요소이다.

③ 난이도는 해당 기업의 현 상황에 따라 조율할 수 있다.

④ 사분면 영역에서 가장 우선적인 분석 과제 적용이 필요한 영역은 3사분면 영역이다.

연습문제

4. 분석 마스터플랜 수립 시 우선 순위 고려사항에 해당하지 않는 것은?

① 전략적 중요도

② 비즈니스 성과 및 ROI

③ 실행 용이성

④ 데이터 필요 우선 순위

연습문제

5. 하향식 접근 방식의 순서로 올바른 것은?

(가) 문제탐색 단계 (나) 문제 정의 (다) 해결방안탐색 (라) 타당성 검토

① 나 -가- 다- 라

② 가 -나- 다- 라

③ 가 -나- 라- 다

④ 나 -가- 라- 다

연습문제

6. 다음 중 상향식 접근방식의 특징으로 적절하지 않은 것은?

① 상향식 접근 방식의 데이터 분석은 지도학습 방법에 의해 수행된다.

② 문제의 정의 자체가 어려운 경우 사용하는 방식이다.

③ 디자인 싱킹(Design Thinking)의 발산 단계에 해당한다.

④ 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식이다.

연습문제

7. 다음 중 ROI 관점에서의 분석 과제에 대한 우선순위 평가 기준 관련 설명 중 적절하지 않은 것은?

① 분석 난이도는 분석 준비도와 성숙도에 따라 해당 기업의 분석 수준을 파악하고 이를 바탕으로 결정된다.

② 시급성이 높고 난이도가 어려운 분석 과제는 우선 순위 기준이 높다.

③ 난이도에 우선 순위 기준을 놓으면 시급성 높고 난이도 쉬운 과제가 가장 먼저 수행되어야 한다.

④ 시급성이 높고 난이도가 어려운 분석과제는 경영진에 의해 조정 가능하다.

연습문제

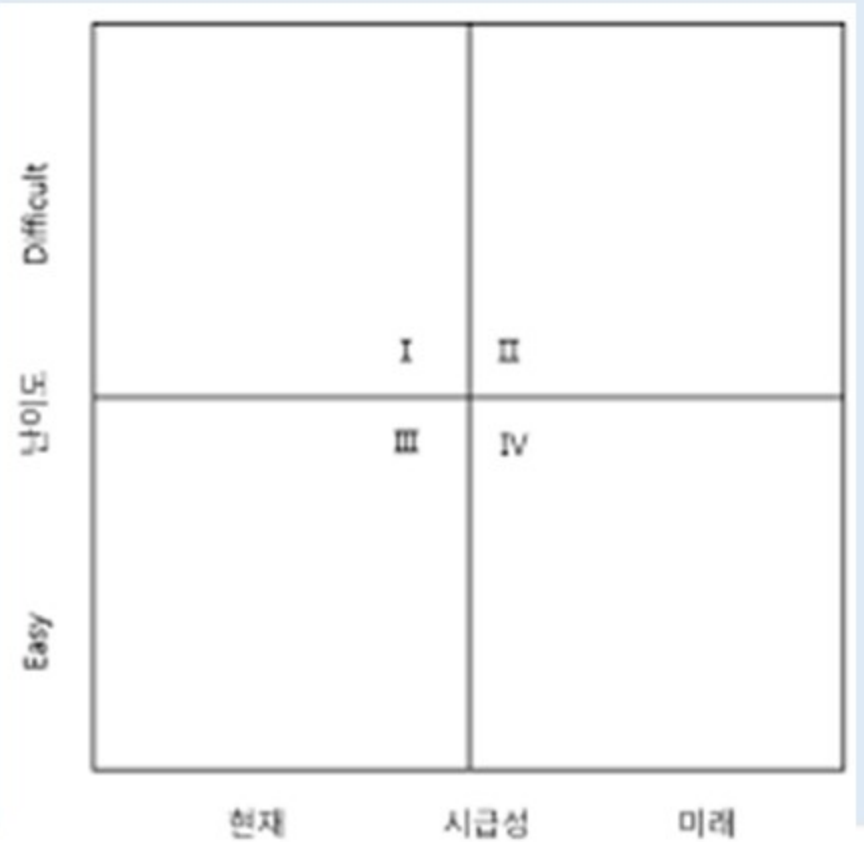
8. 포트폴리오 사분면 분석을 통한 과제 우선순위를 선정하는 기법 중 분석 과제의 적용 우선순위를 '시급성'에 둔다면 결정해야 할 우선순위는?

① III - IV - II

② I - II - III

③ II - IV - I

④ III - I -II



연습문제

9. 반복을 통하여 점증적으로 개발하는 방법으로써 처음 시도하는 프로젝트에 적용이 용이하지만, 반복에 대한 관리 체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있다.

연습문제

10. 다음 중 빈칸에 들어갈 올바른 용어는 무엇인가?

분석방법 (how) \ 분석대상 (what)	Known	Un-Known
Known	최적화(Optimization)	
Un-Known	솔루션(Solution)	발견(Discovery)

연습문제

3과목. 데이터 분석

연습문제

1. Accuracy와 Precision에 대한 설명으로 틀린 것은?

① Accuracy는 실제가 True인 것 중 예측도 True인 것을 의미한다.

② Precision은 예측이 True인 것 중 실제도 True인 것을 의미한다.

③ Precision은 TP / (TP+FP)으로 구할 수 있다.

④ Accuracy는 (TP+TN) / (TP+FP+FN+TN)으로 구할 수 있다.

연습문제

2. 확률변수 X가 확률질량함수를 갖는 이산형 확률변수인 경우 그 기댓값으로 옳은 것은?

① $E(X) = \sum xf(x)$

② $E(X) = \int xf(x)$

③ $E(X) = E([x-\mu]^2)$

④ $E(X) = x^3 - \mathrm{x}^2$

연습문제

3. 아래 오분류표를 이용하여 구한 F1값은 얼마인가?

		예측값		합계
		True	False	합계
실제값	True	200	300	500
	False	300	200	500
합계		500	500	1000

① 0.5

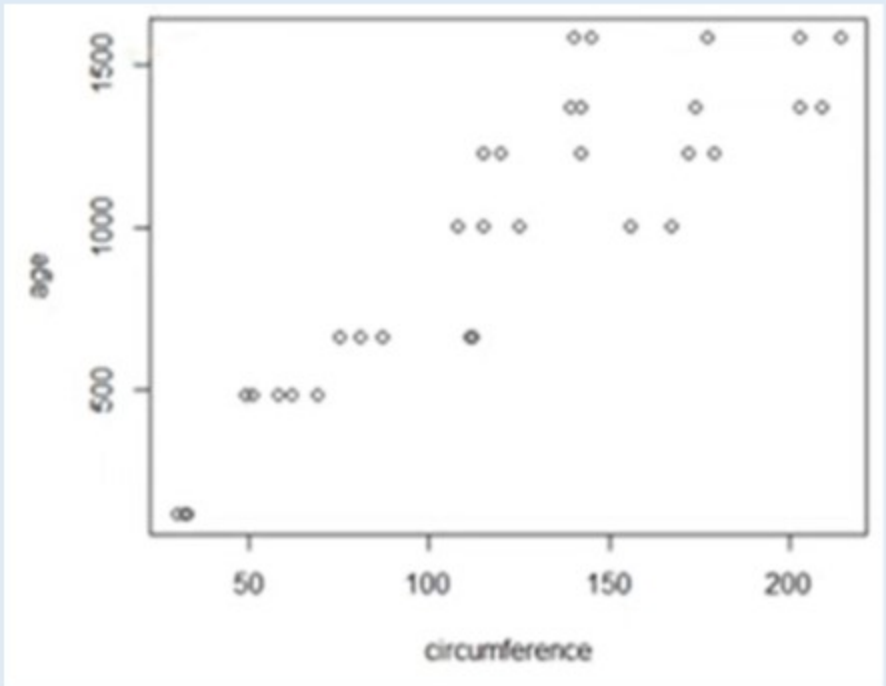
② 0.4

③ 0.2

④ 0.3

연습문제

4. 오렌지 데이터에 대한 산점도 관련 설명으로 틀린 것은?



① 양의 상관 관계임을 알 수 있다.

② Circumference 데이터는 10~230 정도의 범위 인 것을 알 수 있다.

③ 종별로 관계를 나타낼 수 있다.

④ Age 데이터는 10 ~ 1600 정도의 범위인 것을 알 수 있다.

연습문제

5. 다음 중 K-means 군집의 단점으로 가장 부적절한 것은?

① 볼록한 형태가 아닌 군집이 존재하면 성능이 떨어진다.

② 사전에 주어진 목적이 없으므로 결과 해석이 어렵다

③ 잡음이나 이상값에 영향을 많이 받는다.

④ 한 번 군집이 형성되면 군집내 객체들은 다른 군집으로 이동할 수 없다.

연습문제

6. 다음 설명 중 연관 규칙의 단점이 아닌 것은?

① 분석 품목 수가 증가하면 분석 계산이 기하급수적으로 증가한다.

② 너무 세분화된 품목을 가지고 연관규칙을 찾으려면 의미 없는 분석 결과가 도출된다.

③ 품목 간에 구체적으로 어떠한 영향을 주는지 해석하기 어렵다

④ 상대적 거래량이 적으면 규칙 발견 시 제외되기 쉽다.

연습문제

7. 아래 데이터셋 A,B 간의 유사성을 유클리드 거리로 계산하면 얼마인가?

	키	몸무게
A	165	65
B	170	70

① 50

② 25

③ $\sqrt{50}$

④ 10

연습문제

8. 스피어만 상관계수에서 사용하는 척도는?

① 명목척도

② 서열척도

③ 등간척도

④ 비율척도

연습문제

9. SOM은 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도 형태로 형상화하는 방법이다. 다음 중 SOM 방법에 대한 설명으로 부적절한 것은?

① SOM은 입력변수의 위치 관계를 그대로 보존한다는 특징이 있다. 이러한 SOM의 특징으로 인해 입력변수의 정보와 그들의 관계가 지도상에 그대로 나타난다.

② SOM을 이용한 군집분석은 인공신경망의 역전파 알고리즘을 사용함으로써 수행속도가 빠르고 군집의 성능이 매우 우수하다.

③ SOM 알고리즘은 고차원의 데이터를 저차원의 지도 형태로 형성화하기 때문에 시각적으로 이해하기 쉬울뿐 아니라 변수의 위치 관계를 그대로 보존하기 때문에 실제 데이터와 유사하면 지도상 가깝게 표현된다.

④ 입력층과 2차원의 격자 형태의 경쟁층으로 이루어져 있다.

연습문제

10. 과대적합에 대한 설명으로 가장 부적절한 것은?

① 생성된 모델이 훈련 데이터에 너무 최적화되어 학습하여 테스트데이터의 작은 변화에 민감하게 반응하는 경우는 발생하지 않는다.

② 학습데이터가 모집단의 특성을 충분히 설명하지 못할 때 자주 발생한다.

③ 변수가 너무 많아 모형이 복잡할 때 생긴다.

④ 과대적합이 발생할 것으로 예상되면 학습을 빠르게 종료하는 방법으로 과대적합을 방지할 수 있다.

연습문제

11. 아래는 피자와 햄버거의 거래 관계를 나타낸 표로, Pizza/Hamburgers는 피자/햄버거를 포함하는 거래수를 의미하고, (Pizza) / (Hamburgers)는 피자/햄버거를 포함하지 않은 거래 수를 의미한다. 아래 표에서 피자 구매와 햄버거 구매에 대해 설명한 것으로 가장 적절한 것은?

	Pizza	(Pizza)	합계
Hamburgers	2,000	500	2,500
(Hamburgers)	1,000	1,500	2,500
합계	3,000	2,000	5,000

① 지지도가 0.6로 전체 구매 중 햄버거와 피자가 같이 구매되는 경향이 높다.

② 정확도가 0.7로 햄버거와 피자의 구매 관련성이 높다.

③ 향상도가 1보다 크므로 햄버거와 피자 사이에 연관성이 높다고 할 수 있다.

④ 연관규칙 중 “햄버거->피자” 보다 “피자->햄버거“의 신뢰도가 더 높다.

연습문제

12. 앙상블 모형(Ensemble)이란 주어진 자료로부터 여러 개의 예측 모형을 만든 후 이러한 예측 모형들을 결합하여 하나의 최종 예측 모형을 만드는 방법을 말한다. 다음 중 앙상블 모형에 대한 설명으로 적절하지 않은 것은?

① 배깅은 주어진 자료에서 여러 개의 붓스트랩(Boostrap) 자료를 생성하고 각 붓스트랩 자료에 예측모형을 만든 후 결합하여 최종 모형을 만드는 방법이다.

② 부스팅은 배깅의 과정과 유사하여 재표본 과정에서 각 자료에 동일한 확률을 부여하여 여러 모형을 만들어 결합하는 방식이다.

③ 랜덤 포레스트(Random Forest)는 의사결정나무모형의 특징인 분산이 크다는 점을 고려하여 배깅보다 더 많은 무작위성을 추가한 방법으로 약한 학습기들을 생성하고 이를 선형 결합해 최종 학습기를 만드는 방법이다.

④ 앙상블 모형은 훈련을 한 뒤 예측을 하는데 사용하므로 교사학습법(Supervised Learning)이다.

연습문제

13. 정규분포 신뢰수준 95%일 때에 대한 설명으로 가장 적절하지 않은 것은?

① 표본크기가 커질수록 신뢰구간이 좁아진다. 이는 정보가 많을수록 추정량이 더 정밀하다는 것을 의미한다.

② 99% 신뢰수준에 대한 신뢰구간이 95% 신뢰수준에 대한 신뢰구간보다 길다.

③ 신뢰수준은 모수값이 정해져 있을 때, 다수 신뢰구간 중 모수값을 포함하는 신뢰구간이 존재할 확률을 말한다.

④ 신뢰수준 95% 의미는 추정값이 신뢰구간에 존재할 확률이 95%라 할 수 있다.

연습문제

14. 회귀방정식의 변수선택 방법에 대한 설명으로 적절하지 않는 것은?

① 전진선택법은 변수를 추가해도 영향을 받지 않는다.

② 후진제거법은 독립변수 후보 모두를 포함한 모형에서 시작한다.

③ 단계별 선택법은 기준 통계치에 가장 도움이 되지 않는 변수를 삭제하거나, 모델에서 빠져 있는 변수중에서 기준 통계치를 가장 개선시키는 변수를 추가한다.

④ 회귀모델에서 변수 선택을 위한 판단 기준에는 Cp, AIC, BIC 등이 있으며 값이 작을수록 좋다.

연습문제

15. 주성분분석에 대한 설명 중 적절하지 않은 것은?

```
> data_1 <- prcomp(data, scale=TRUE)
> data_1
Standard deviations (1, .., p=4):
[1] 1.4154072  1.3086525  0.4377899  0.3039594

Rotation (n • k) = (4 • 4)
                         PC1        PC2        PC3        PC4
                x1  0.2388128 -0.6895993  0.5325178  0.4287728
                x2  0.4604720 -0.5393126 -0.5603653 -0.4278997
                x3  0.6038420  0.3514805 -0.3277028  0.6359616
                x4  0.6052345  0.3317472  0.5431634 -0.4781303

> summary(data_1)
Importance of components:
                                 PC1      PC2      PC3      PC4
Standard deviation            1.4154   1.3087  0.43779  0.3040
Proportion of Variance        0.5008   0.4281  0.04791  0.0231
Cumulative Proportion         0.5008   0.9290  0.97690  1.0000
```

① 제 3변수까지 사용하면 97.69%의 누적 비율을 갖게 된다.

② 제 2변수는 42.81%의 분산 비율을 갖는다.

③ 변수들의 scale이 많이 다른 경우 특정 변수가 전체적인 경향을 좌우하기 때문에 상관계수 행렬을 사용하여 분석하는 것이 좋다.

④ PC2 로딩벡터는 모두 양의 방향을 가지고 있다.

연습문제

16. 표본추출시 발생하는 오차에 관한 설명 중 잘못된 설명은?

① 표본 오차(Sampling Error)는 모집단의 일부인 표본에서 얻은 자료를 통해 모집단 전체의 특성을 추론함으로써 생기는 오차를 의미한다.

② 비표본 오차(non-sampling error)는 표본크기가 증가함에 따라 증가한다.

③ 표본 편의(Sampling Bias)는 표본추출방법에서 기인하는 오차를 의미하고, 표본 추출 방법에 의해 최소화하거나 없앨 수 있다.

④ 표본 오차는 표본의 크기는 증가시키고, 표본 선택 방법을 엄격히 하여 줄일 수 있다.

연습문제

17. 양성 나온 사람 중에 실제 질병이 있는 사람의 확률은 무엇인가?

이때, 양성인 사람은 0.2, 실제 질병이 있는 사람은 0.1, 검사 결과 양성인 사람은 0.9이다.

① 0.09

② 0.45

③ 0.18

④ 0.5

연습문제

18. 다음 표를 사용한 재현율에 관련된 설명으로 틀린 것은?

Confusion Matrix		Predict	
		FALSE	TRUE
Actual	FALSE	30	70
	TRUE	10	40

① 재현율(Recall)을 민감도(Sensitivity)라고도 한다.

② 재현율 3/100이다.

③ 재현율 TP / (TP+FN)이다.

④ 재현율과 정밀도(Precision)을 사용해 F1-score를 구한다.

연습문제

19. 주성분 분석에 대한 설명 중 올바른 것은?

① 독립변수들과 주성분과의 거리인 '정보손실량'을 최대화하거나 분산을 최소화 한다.

② 상관관계가 있는 변수들을 선형 결합에 의해 상관관계가 있는 새로운 변수(주성분)를 만들고 분산을 최소화하는 변수로 축약한다.

③ 여러 개의 양적 변수(Quantitative variable)들 사이의 분산-공분산 관계를 이용하여 변수들의 선형 결합(linear combination)으로 표현하는 기법이다.

④ 정규화 전후의 주성분 결과는 동일하다.

연습문제

20. 주성분 분석에서 주성분 수를 선택할 때 고려하지 않아도 되는 것은?

① Scree plot

② 개별 고윳값의 분해 가능 여부

③ 성분들이 설명하는 분산의 비율

④ 고윳값(Eigenvalue)

연습문제

21. 분해 시계열 분석에 대한 설명 중 옳지 않은 것은?

① 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법이다.

② 추세요인은 자료의 그림을 그렸을 때 그 형태가 오르거나 내리는 등 자료가 어떤 특정한 형태를 취할 때이다.

③ 계절요인은 고정된 주기에 따라 자료가 변화하는 경우이다.

④ 이동평균법은 최근 관측치에 더 높은 가중치를 부여하여 이동 평균을 계산하는 방법이다.

연습문제

22. 의사결정나무모형에 관한 내용으로 적절하지 않은 것은?

① 의사결정나무의 목적은 새로운 데이터를 분류(Classification)하거나 해당 범주의 값을 예측(Prediction)하는 것이다.

② 목표변수 유형에 따라 범주형 분류나무(Classification Tree)와 연속형 회귀나무(Regression Tree)로 분류된다.

③ 분리 변수의 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받지 않는다.

④ 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나간다.

연습문제

23. 다음 회귀분석에 결과에 대한 설명으로 틀린 것은?

```
Call:
lm(formula = wage ~ age + jobclass, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-107.234  -24.751   -6.311   16.308  197.278

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           76.6298     2.8320   27.06   <2e-16 ***
age                    0.6447     0.0638   10.11   <2e-16 ***
jobclass2. Information 15.9214     1.4732   10.81   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.16 on 2997 degrees of freedom
Multiple R-squared:  0.07435,   Adjusted R-squared:  0.07373
F-statistic: 120.4 on 2 and 2997 DF,  p-value: < 2.2e-16
```

① Age변수는 wage에 대해 유의하지 않다.

② 종속변수는 wage이다.

③ jobclass는 범주형 변수이다.

④ 데이터 개수가 3000개이다.

연습문제

24. 다음 Orange 나무에 대한 나이 및 둘레에 대한 분석 결과로 옳지 않은 것은?

```
> summary(Orange)
 Tree       age         circumference
 3:7   Min.   : 118.0   Min.   : 30.0
 1:7   1st Qu.: 484.0   1st Qu.: 65.5
 5:7   Median :1004.0   Median :115.0
 2:7   Mean   : 922.1   Mean   :115.9
 4:7   3rd Qu.:1372.0   3rd Qu.:161.5
       Max.   :1582.0   Max.   :214.0
```

① Circumference의 Median은 115이다.

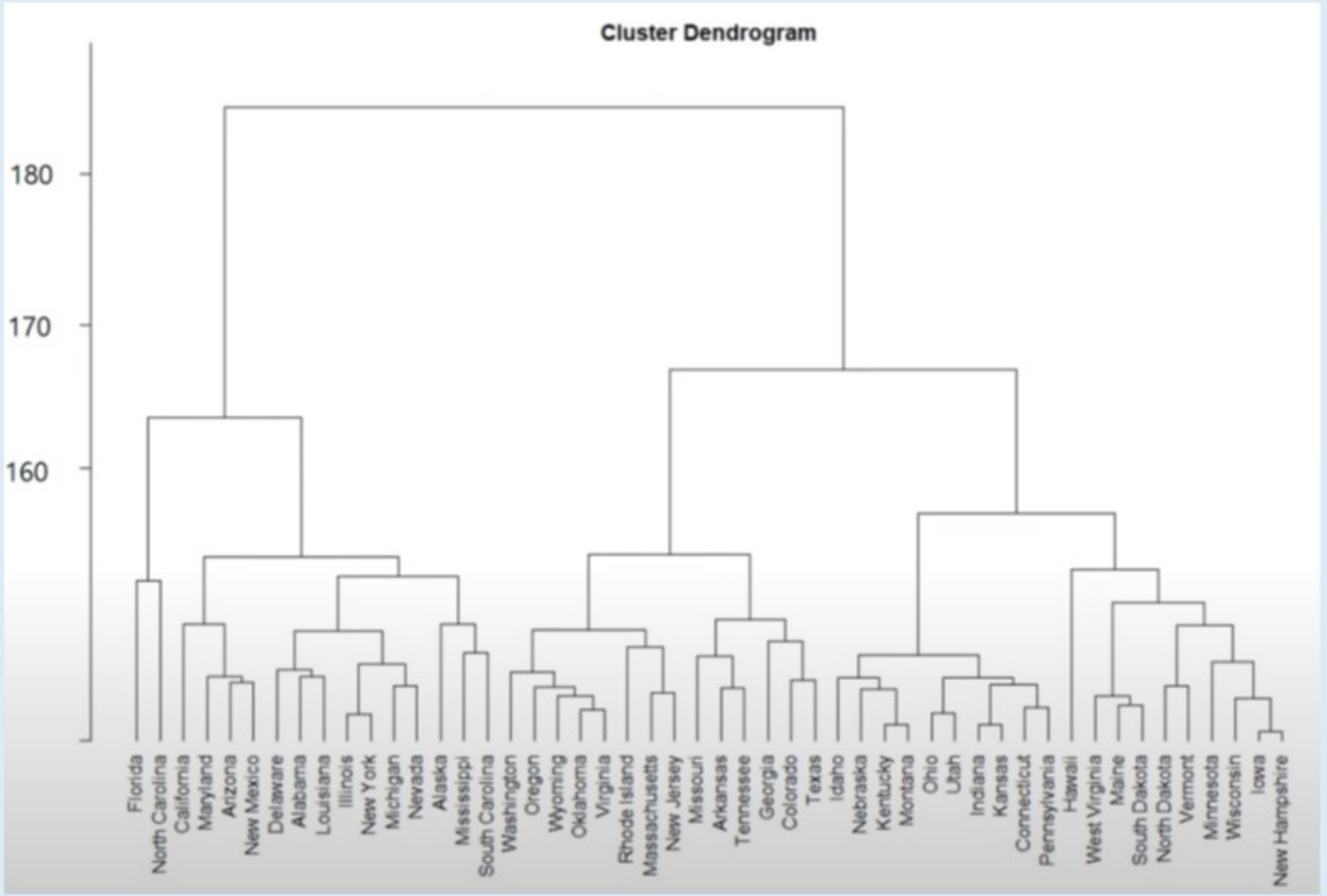
② 나무 age와 circumference가 유의한 관계를 가진다.

③ age의 IQR은 888이다.

④ Tree의 종류는 5가지이며 각 종류당 7개의 sample이 존재한다.

연습문제

25. 다음 덴드로그램에서 height가 160일 때의 군집 개수는?



연습문제

26. 귀무가설이 실제로 사실이어서 채택하여야 함에도 불구하고 이를 기각하는 오류를 무엇이라 하는가?

27. 신경망 모형에서 표준화 지수함수로 불리며, 출력 값 z가 여러 개로 주어지고, 목표치가 다 범주인 경우 각 범주에 속할 사후확률을 제공하여 출력 노드에 주로 사용되는 함수는 무엇인가?

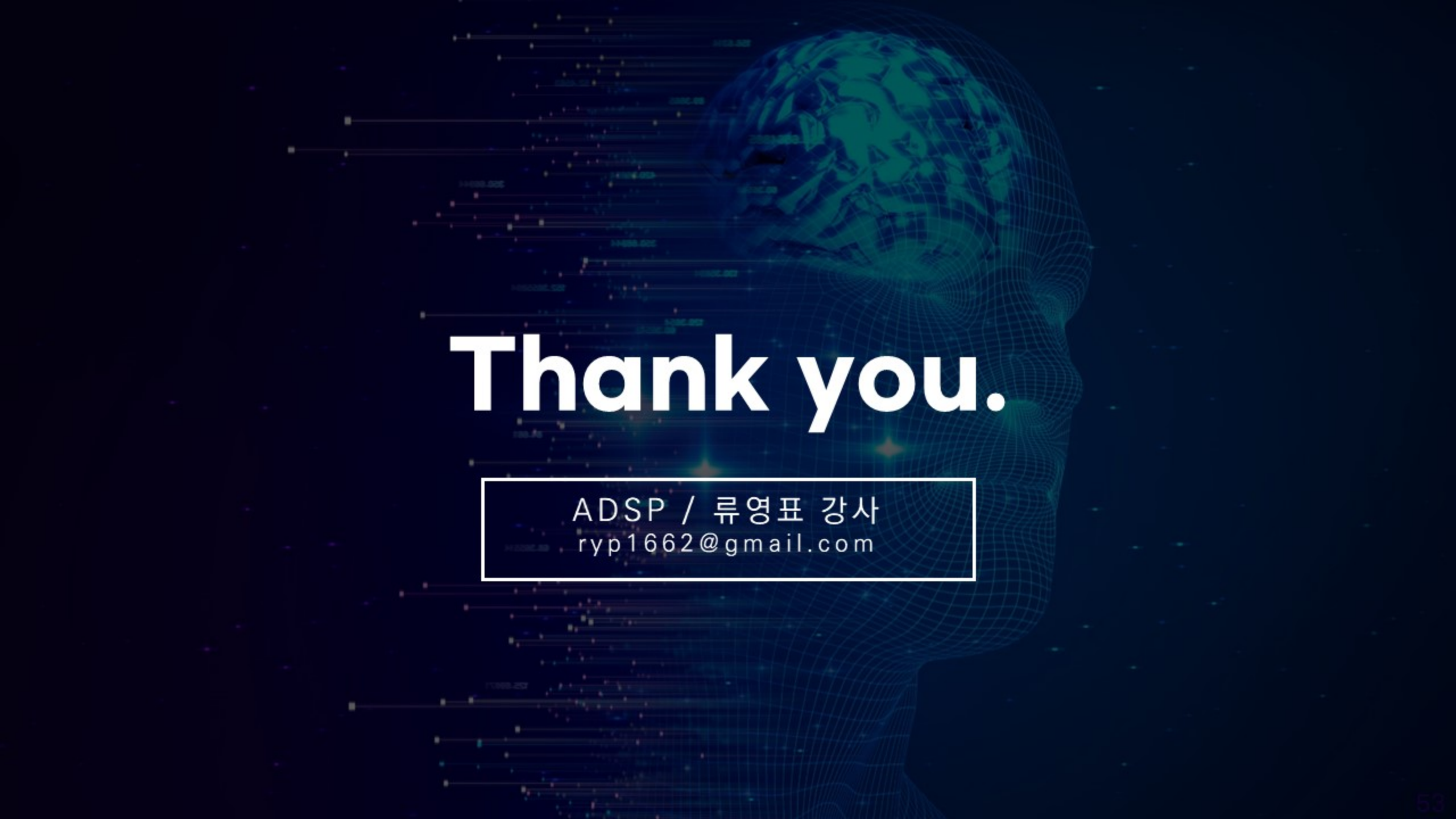
연습문제

28. 오분류표 용어 중 실제로 False일 때 예측이 적중하는 경우를 무엇이라고 하는가?

29. 은닉층이 다층인 신경망을 학습하다 보면 역전파 과정에서 초기 부분의 입력층이 갈수록 기울기 변화가 점차적으로 작아지는 현상은?

연습문제

30. 로지스틱 회귀모형에서 exp(x1)의 의미는 나머지 변수가 주어질 때 x1이 한 단위 증가할 때마다 성공(Y=1) ( ) 가 몇 배 증가 하는지를 나타낸다. ( ) 에 들어가는 용어는?


Thank you.
ADSP / 류영표 강사
ryp1662@gmail.com