35회 기출문제

데이터 준전문가

ADSP, Advanced Data Analytics semi-Professional

류영표 강사
ryp1662@gmail.com

연습문제

1과목. 데이터 이해

연습문제

1. 사용자와 데이터베이스 사이에서 사용자의 요구에 따라 정보를 처리해주고 데이터베이스를 관리해주는 소프트웨어는?

① DBMS

② Data Dicitionary

③ SQL

④ ERD

연습문제

2. 데이터 사이언티스트의 필요 역량 중 성격이 다른 것은?

① 통찰력이 있는 분석 능력

② 커뮤니케이션 능력

③ 네트워크 최적화 능력

④ 스토리텔링 능력

연습문제

3. 빅데이터의 위기 요인 중 사생활 침해 방지 기술에 해당하는 것은?

① 익명화

② 정규화

③ 일반화

④ 표준화

연습문제

4. 데이터 분석가, 인공지능 전문가 등이 만들어 낸 알고리즘으로 부당한 피해를 보는 사람을 방지하기 위해 생겨난 인력을 무엇이라 하는가?

① 애널리스트

② 데이터 관리자

③ 빅데이터 전문가

④ 알고리즈미스트

연습문제

5. 빅데이터가 미치는 영향에 대해 올바르지 않은 것은?

① 정치인은 선거 승리를 위해 사회관계망 분석을 통해 유세 지역을 선정하고 해당 지역의 유권자에게 영향을 줄 수 있는 내용을 선전해 효과적인 선거 활용을 펼친다.

② 인과관계 규명 없이 상관관계의 분석 결과만으로도 인사이트를 얻고 이를 바탕으로 수익을 창출할 기회가 늘어나고 있다.

③ 산업과 경제 구조 변화 동향 등 거시적인 흐름을 토대로 분석 기회를 도출한다.

④ 사물인터넷의 발달로 인해 사람이 최대로 개입하게 되었다.

연습문제

6. 빅데이터의 기술 활용에 대한 설명으로 옳지 않은 것은?

① 구글의 사용자의 로그 데이터를 활용하면서 기존의 페이지랭크(PageRank) 알고리즘을 혁신하였다.

② 정부의 실시간 교통정보 수집, 기후 정보 등을 실시간 모니터링을 한다.

③ 정부의 이익을 위해 개인의 정보를 활용한다.

④ 가수들은 팬들의 음악 청취 기록을 분석해 공연에 활용한다.

연습문제

7. 빅데이터 시대의 위기 요인이 아닌 것은?

① 사생활 침해

② 책임 원칙 훼손

③ 데이터 오용

④ 데이터 변화 관리

연습문제

8. 다음 중 데이터베이스 통신을 위해 고안된 언어로 가장 적절한 것은?

① Java

② R

③ Python

④ SQL

연습문제

9. “문자, 기호, 음성, 화상, 영상 등 상호 관련된 다수의 컨텐츠를 정보 정리 및 정보통신 기기에 의하여 체계적으로 수집, 축적하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보의 집합체“를 무엇이라 하는가?

연습문제

10. 다음 아래 보기에서 설명하는 빅데이터 분석 기법은 무엇인가?

- 존 홀랜드가 다윈 진화론의 적자생존에 기반하여 개발한 최적화 연산 방법이다.
- 최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가와 같은 최적화의 메커니즘을 찾아내는 분석기법이다.
- 어떤 미지의 함수 $y = f(x)$를 최적화하는 해 x를 찾기 위해, 진화를 모방한 탐색 알고리즘이다.

연습문제

2과목. 데이터 분석 기획

연습문제

1. 다음이 설명하는 데이터 거버넌스 체계 요소는 무엇인가?

- 데이터 정합성 및 활용의 효율성을 위하여 표준 데이터를 포함한 메타 데이터와 데이터 사전의 관리원칙을 수립한다.
- 수립된 원칙에 의거하여 항목별 상세한 프로세스를 만들고 관리와 운영을 위한 담당자 및 조직별 역할과 책임을 상세하게 준비한다.

① 데이터 표준화

② 데이터 관리 체계

③ 데이터 저장소 관리

④ 표준화 활동

연습문제

2. 분석조직 인력들을 협업부서로 직접 배치하여 신속한 업무 수행이 가능한 조직구조를 무엇이라 하는가?

① 집중구조

② 기능구조

③ 분산구조

④ 확산구조

연습문제

3. 분석 마스터 과제 우선순위 결정과 관련한 설명 중 가장 적절하지 않은 것은?

① 빅데이터의 특징을 고려한 ROI 요소에서 가치(Value)는 투자 비용(Investment) 요소이다.

② 전략적 중요도, 실행 용이성, ROI 등은 분석과제 우선순위 결정에 고려할 사항이다.

③ 시급성의 판단기준은 전략적 중요도가 핵심이다.

④ 적용기술의 안정성 검증은 기술 용이성이 평가요소이다.

연습문제

4. 분석과제 우선순위 선정 매트릭스에서 일반적으로 데이터 분석과제를 가장 먼저 추진해야 하는 영역은?

① 난이도 : 쉬움, 시급성 : 현재

② 난이도 : 어려움, 시급성 : 미래

③ 난이도 : 쉬움, 시급성 : 미래

④ 난이도 : 어려움, 시급성 : 현재

연습문제

5. 다음 중 빅데이터 분석방법론의 분석 기획 단계에서 프로젝트 위험 대응 계획을 수립할 때 예상되는 위험을 대응하는 방법으로 가장 부적절한 것은?

① 회피(Avoid)

② 관리(Management)

③ 완화(Mitigate)

④ 수용(Accept)

연습문제

6. 기업의 “도입형” 분석 수준 진단 결과 항목이 아닌 것은?

① 조직 및 인력

② 분석업무

③ 분석기법

④ 관리 목표

연습문제

7. 분석과제 접근방법 중 상향식 접근 프로세스 절차로 옳은 것은?

① 프로세스 분류 -> 프로세스 흐름 분석 -> 분석 요건 식별 -> 분석 요건 정의

② 프로세스 흐름 분석 -> 프로세스 분류 -> -> 분석 요건 식별 -> 분석 요건 정의

③ 프로세스 흐름 분석 -> 프로세스 분류 -> -> 분석 요건 정의 -> 분석 요건 식별

④ 프로세스 분류 -> 프로세스 흐름 분석 -> 분석 요건 정의 -> 분석 요건 식별

연습문제

8. 다음 중 데이터 거버넌스의 구성요소가 아닌 것은?

① 원칙(Principle)

② 조직(Organization)

③ 프로세스(Process)

④ 분석 방법(Method)

연습문제

9. 문제를 주고 이에 대한 해법을 찾기 위하여 각 과정이 체계적으로 단계화되어 수행하는 분석과제 접근방식을 무엇이라 하는가?

연습문제

10. 다음이 설명하는 조직구조는 무엇인가?

- 전사 분석업무를 별도의 분석 전담 조직에서 담당
- 전략적 중요도에 따라 분석조직이 우선순위를 정해서 진행 가능
- 현업 업무부서의 분석업무와 이중화 및 이원화 가능성이 큼.

연습문제

3과목. 데이터 분석

연습문제

1. 다음 중 회귀분석에서의 결정계수에 대한 설명으로 적절하지 않은 것은?

① 결정계수의 값은 0에서부터 1사이의 값을 갖는다.

② 결정계수의 값이 클수록 추정된 회귀식의 설명력이 높다.

③ 종속변수와 독립변수 상의 표본 상관계수와 같다.

④ 총 변동 중에서 회귀선에 의해 설명이 되는 변동이 차지하는 비율을 의미한다.

연습문제

2. 다음 중 이산형 확률변수 x에 대한 설명 중 옳은 것은?

$$P(X=1)=\frac{1}{3}, P(X=2)=\frac{1}{6}, P(X=3)=\frac{1}{2}$$

① 확률변수 X의 기댓값은 $\frac{13}{6}$이다.

② 확률변수 X가 1 또는 2일 때, 확률은 $\frac{1}{6}$ 이다.

③ 확률변수 X가 4일 확률은 0보다 크다.

④ 확률변수 X의 확률의 합은 1보다 작다.

연습문제

3. 다음 중 목표변수가 연속형인 회귀 나무의 분류 변수와 분류 기준 값의 선택 방법은 무엇인가?

① 카이제곱 통계량, 지니 지수

② 지니 지수, F-통계량

③ F-통계량의 p값, 분산 감소량

④ 분산 감소량, 엔트로피 지수

연습문제

4번. 다음 중 오분류표에서 특이도(Specificity)를 구하시오.

		예측값		합계
		True	False	합계
실제값	True	200	400	600
	False	300	100	400
합계		500	500	1000

① 0.5

② 0.4

③ 0.2

④ 0.3

연습문제

5번. 두 좌표 간의 맨해튼 거리(Manhanttan distance)를 구하시오.

	A	B
키	175	180
몸무게	70	65

① 0

② 10

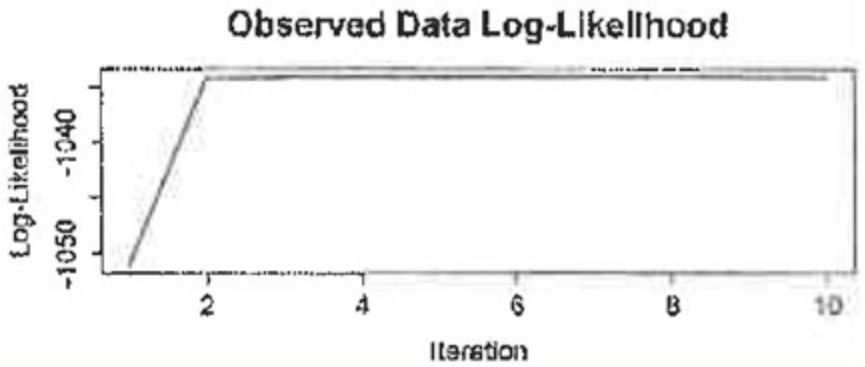
③ $\sqrt{10}$

④ $\sqrt{50}$

연습문제

6. EM 알고리즘을 사용하여 혼합분포 모형을 추정하고자 한다. 아래와 그래프가 도출되었을 때 가장 적절한 해석은?

```
>wait1<-normalmixEM(waiting,lambda
=.5,mu=c(55,80),sigma=5)
>number of iterations= 9
>plot(wait1,density=TRUE,cex.axis=1.4,cex.
lab=1.4,cex.main=1.8)
```



① 반복횟수 2회만에 로그-가능도 함수가 최대가 됨을 알 수 있다.

② 모수의 추정을 위해 10회 이상이 반복했다.

③ 최종 로그가능도는 -10400이다.

④ Lambda는 성분들의 표준편차 시작 값이다.

연습문제

7. 이산형 확률변수 x의 기댓값을 구하시오.

x	1	2	3
p(x)	0.5	0.3	0.2

① 1

② 1.7

③ 2.5

④ 10

연습문제

8. 다음 중 데이터 마이닝 프로세스 순서로 올바른 것은?

> ㉮ 목적 정의
> ㉯ 데이터 준비
> ㉰ 데이터 가공
> ㉱ 데이터 마이닝 기법의 적용
> ㉲ 검증

① ㉮ → ㉯ → ㉰ → ㉱ → ㉲

② ㉰ → ㉱ → ㉲ → ㉮ → ㉯

③ ㉯ → ㉰ → ㉱ → ㉲ → ㉮

④ ㉰ → ㉱ → ㉲ → ㉯ → ㉮

연습문제

09. 라쏘(Lasso) 회귀모형에 대한 설명으로 적절하지 않은 것은?

① 회귀계수들의 절댓값이 클수록 패널티(Penalty)를 부여하는 방법이다.

② L_2-norm을 통해 제약을 주는 방법이다.

③ 선형 회귀 계수에 대한 제약 조건을 추가하여 모델이 과도하게 최적하게 현상(과적합, overfitting)을 막는 방법이다.

④ 라쏘 회귀모형은 변수 선택이 가능하다.

연습문제

10. 군집분석에 대한 설명으로 가장 적절하지 않은 것은?

① 비 계층적 군집분석 기법의 경우 분석자가 사전지식 없이 군집의 수를 설정하기 때문에 군집결과가 잘 나오지 않을 수 있다.

② 논리성과 안정성을 검증하기가 지도학습에 비교하여 어렵다.

③ 안정성 검증 방법으로 지도학습과 동일한 교차검증을 이용한다.

④ 군집분석은 이상치 자료에 민감하다.

연습문제

11. SOM 정의로 적절하지 않은 것은?

① 입력층과 출력층으로 구성되어 있다.

② 한 개의 입력층과 한 개의 출력층을 가진다.

③ 입력층과 출력층이 부분 연결되어 있다.

④ 뉴런들은 승자 뉴런이 되기 위해 경쟁하고 오직 승자만이 학습한다.

연습문제

12. 시계열 자료의 정상성에 대한 설명 중 옳지 않은 것은?

① 평균이 일정하다.

② 분산이 시점에 의존하지 않는다.

③ 자료가 추세를 보일 때에는 변환을 통해 비정상 시계열을 정상 시계열로 변환할 수 있다.

④ 공분산은 단지 시차에만 의존하고 시점 자체에는 의존하지 않는다.

연습문제

13. 다음 중 인공신경망에서 입력신호의 총합을 출력신호로 변환하는 함수로 로지스틱 회귀모형의 작동 원리가 유사한 활성화 함수를 무엇이라 하는가?

① 계단함수

② Tanh 함수

③ ReLU 함수

④ 시그모이드 함수.

연습문제

14. 다음 중 군집의 수를 미리 지정하지 않아도 탐색적 기법에 적합한 군집 방법은?

① SOM

② K- 평균 군집

③ 계층적 군집

④ 혼합분포 군집

연습문제

15. 다음 중 혼합분포 군집의 특징으로 가장 적절하지 않은 것은?

① 확률분포를 도입하여 군집을 수행하는 모형-기반의 군집 방법이다.

② 군집을 몇 개의 모수로 표현할 수 있다.

③ EM 알고리즘을 이용한 모수 추정에서 데이터가 커지면 수렴하는 데 시간이 걸릴 수 있다.

④ 군집의 크기가 작을수록 추정의 정도가 쉬워진다.

연습문제

16. 다음 중 과적합(Overfitting)을 방지하기 위한 방법이 아닌 것은?

① 홀드 아웃(Hold- Out)

② K-fold

③ 붓스트랩

④ 의사결정나무

연습문제

17. 다음 중 분류모형에 대한 설명으로 옳은 것은?

① 현상을 이해하기 위해 어떤 사실들을 분류하고, 범주화하고, 등급으로 나눈다.

② 같은 아이템의 연관성을 파악하는 분석이다.

③ 이질적인 모집단을 동질성을 지닌 그룹별로 세분화하는 것을 의미한다.

④ 장바구니 분석이라고 불린다.

연습문제

18. 다음 중 로지스틱 회귀에 대한 특징으로 가장 적절한 것은?

① 반응변수가 범주형일 때 적용되는 회귀분석이다.

② 종속변수가 비율척도일 때 적합하다.

③ 최소제곱법으로 분석모형을 추정한다.

④ 로지스틱 회귀계수는 해당 변수가 1 증가함에 따른 오즈로 해석이 불가하다.

연습문제

19. 다음 중 연관분석의 특징으로 적절하지 않은 것은?

① 품목 수의 증가와 관계없이 계산량은 동일하다.

② 너무 세분화된 품목이 연관규칙 분석에 적합하다.

③ 거래량이 적은 품목이 연관규칙 분석에 적합하다.

④ 조건 반응(if-then)으로 표현되는 연관분석의 결과를 이해하기 쉽다.

연습문제

20. 다중회귀분석에서 모형이 통계적으로 유의성을 검증하는 통계량은?

① t - 통계량

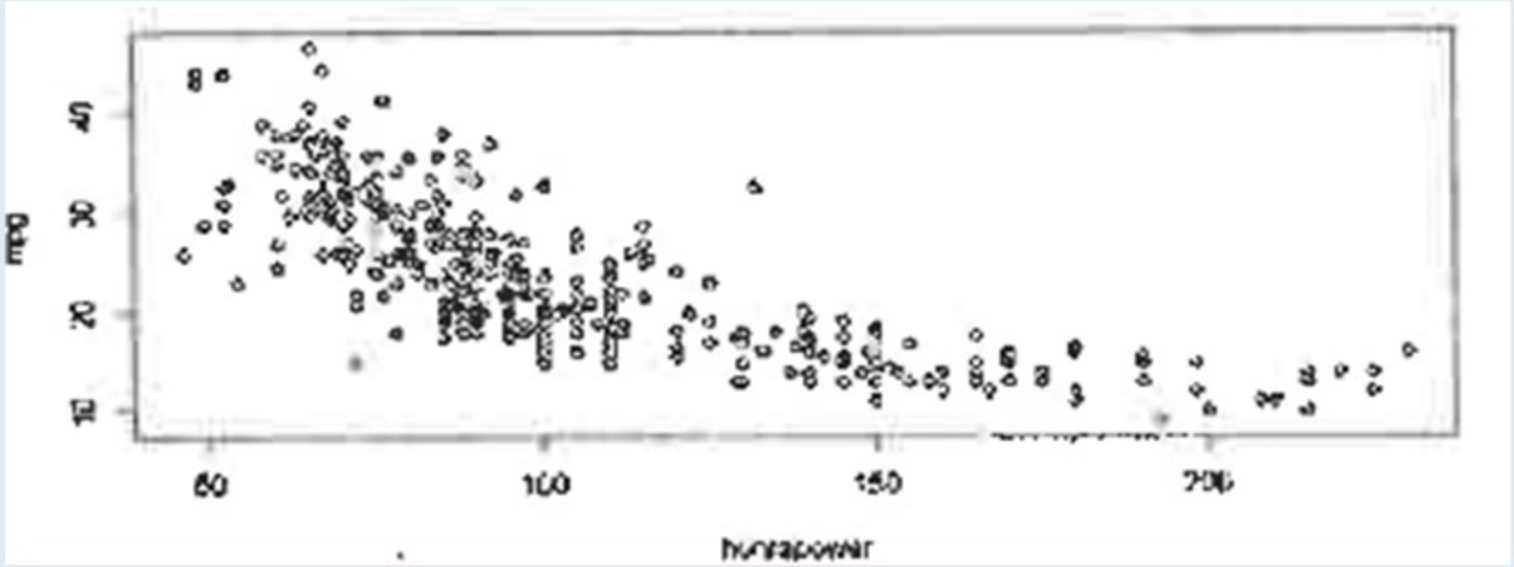
② F - 통계량

③ 카이제곱 통계량

④ R-square

연습문제

21. 아래 산점도는 392대의 연비(mpg)와 마력(housepower)에 관한 그래프 설명으로 가장 적절하지 않은 것은?



① 마력이 향상할 때 연비가 감소하는 경향이 있다.

② 연비와 마력의 상관관계는 피어슨 상관계수로 분석이 가능하지 않다.

③ 연비와 마력은 음의 상관관계이다.

④ 연비와 마력 간의 영향력으로 단순 선형회귀모형 추정도 가능하다.

연습문제

22. 다음 중 로지스틱 회귀분석 결과에 대한 해석으로 옳지 않은 것은?

```
summary(glm(default~.,data=Default,family=
"binomial"))
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
```

① Balance는 default(연체 여부)에 통계적으로 유의미한 영향을 주는 변수이다.

② Income은 default에 통계적으로 유의미한 영향을 주는 변수이다.

③ 로지스틱 회귀분석은 지도학습 방법이다.

④ studentYes는 default를 설명하는데 통계적으로 유의한 변수이다.

연습문제

23. 다음 중 k-평균 군집에 대한 설명 중 옳은 것은?

① K-평균 군집은 오목 군집이 존재할 때 성능이 우수하다.

② 잡음이나 이상값에 영향을 받지 않는다.

③ K-개의 초기 중심값에 따라 군집 결과가 크게 달라질 수 있다.

④ 군집의 수는 자동으로 결정된다.

연습문제

24. 연관규칙 학습 중 A->B 구입의 지지도(Support) 공식은 무엇인가?

① A와 B가 동시에 포함된 거래 수/전체 거래 수

② A와 B가 동시에 포함된 거래 수/A가 포함된 거래 수

③ A와 B가 동시에 포함된 거래 수/B가 포함된 거래 수

④ A 거래 수/전체 거래 수

연습문제

25. 여러 대상 간의 거리가 주어졌을 때, 대상들을 동일한 상대적 거리를 가진 실수공간의 점들로 배치시키는 차원 축소 기법을 무엇이라 하는가?

26. 독립변수 후보 모두를 포함한 모형에서 출발해 제곱 합의 기준으로 가장 적은 주는 변수부터 하나씩 제거하면서 다시는 유의하지 않은 변수가 없을 때까지 설명변수들을 제거하고 모형을 선택하는 변수선택 기법을 무엇이라 하는가?

연습문제

27. 데이터를 랜덤하게 두 분류로 분리하고 교차검증을 시행하는 방법으로 하나는 모형의 학습 및 구축을 위한 훈련용 자료로 다른 하나는 성과 평가를 위한 검증용 자료로 사용하는 검증 기법은 무엇인가?

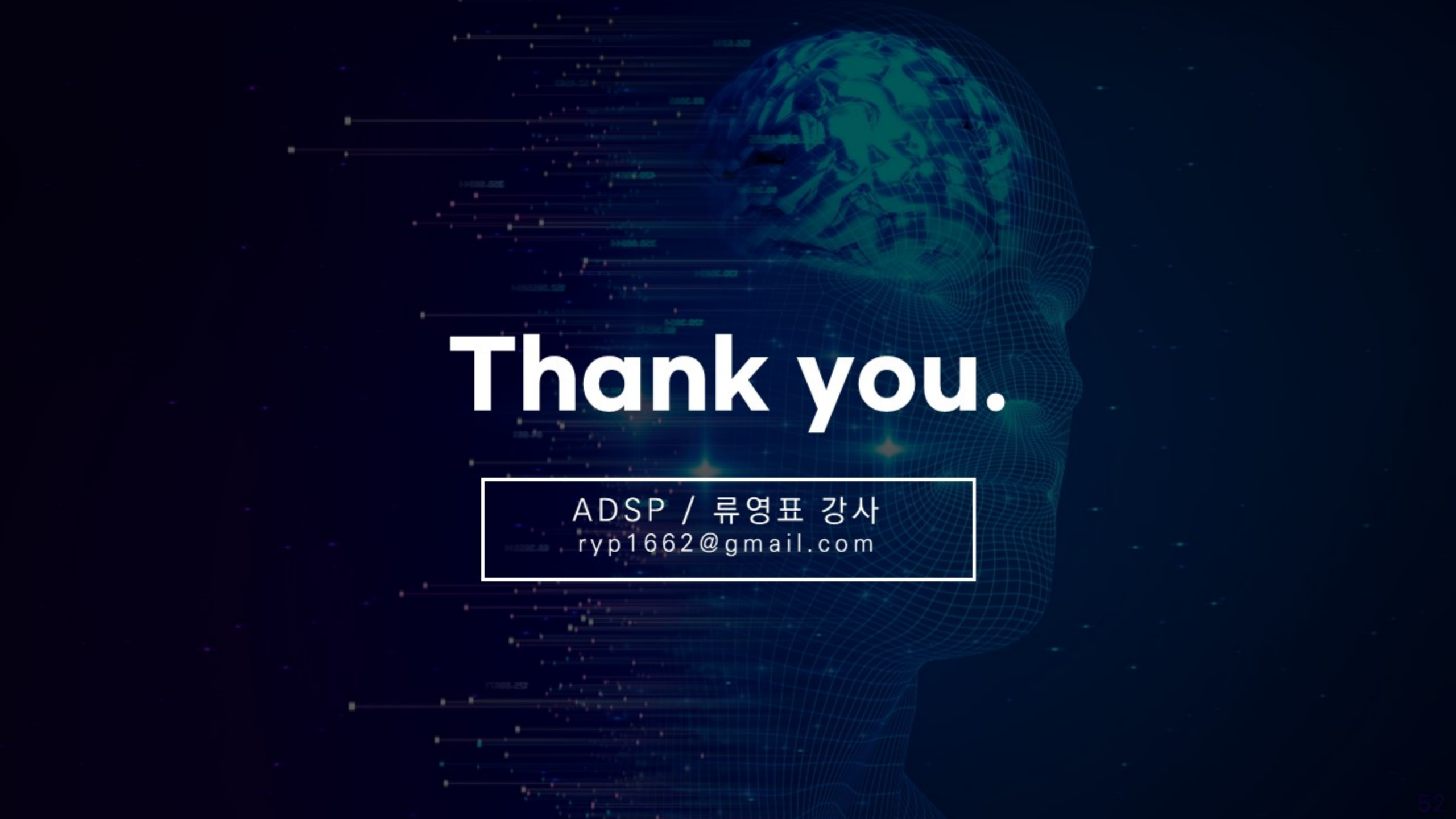
28. P(A) = 0.3 , P(B) = 0.4, A,B는 서로 독립이다. 이 때, P(B|A)의 확률?

연습문제

29. 단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현하는 이산 확률 분포를 무엇이라 하는가?

30. 다음 빈칸에 알맞은 용어는?

( )은 두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최솟값으로 측정한다. 최단 거리를 사용할 때 사슬 모양으로 생길 수 있으며, 고립된 군집을 찾는 데, 중점을 둔 방법이다.


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