32회 Final Test

데이터 준전문가

ADSP, Advanced Data Analytics semi-Professional

류영표 강사
ryp1662@gmail.com

연습문제

1과목. 데이터 이해

연습문제

1. 다음 SQL의 명령어 중 DML이 아닌 것은 무엇인가?

가 - SELECT , 나 – UPDATE, 다 – INSERT, 라-DELETE, 마 - CREATE

① 가, 나

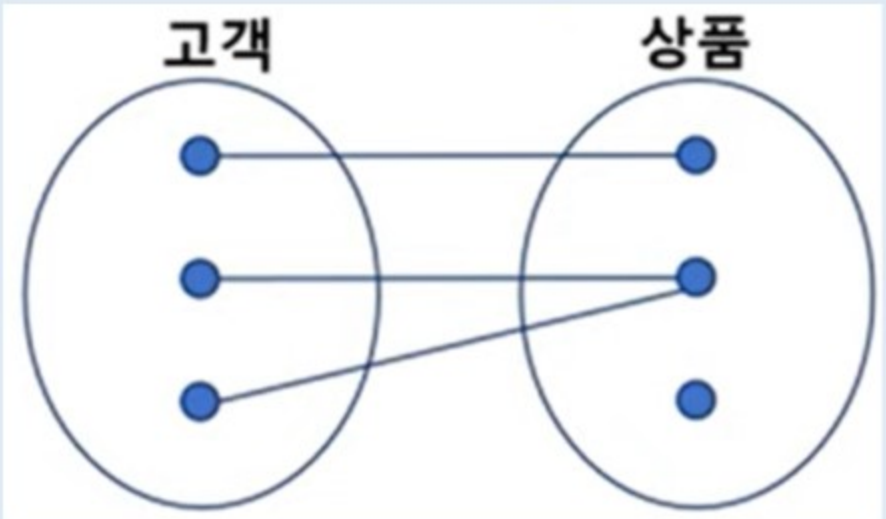
② 다

③ 라

④ 마

연습문제

2. 고객과 상품의 대응관계를 도식화 한 것이다. 대응비 관점에서 고객과 상품 간의 관계가 옳은 것은?



① 1:1

② N:1

③ N:M

④ 1:N

연습문제

3. 다음 어떤 기업내부 데이터베이스 솔루션에 대한 설명인가?

조직의 회계, 구매, 프로젝트 관리, 리스크 관리 규정 준수 및 공급망 운영 같은 일상적인 비즈니스 활용을 관리하는데 사용하는 소프트웨어 유형을 의미한다.
다양한 비즈니스 분야에서 생산, 구매, 재고, 주문, 공급자와의 거래, 고객서비스 제공 등 주요 프로세스 관리를 돕는 어플리케이션이다.

① ERP

② CRM

③ SCM

④ KMS

연습문제

4. 다음 중 딥러닝(Deep learning)과 가장 관련 없는 분석 기법은?

① LSTM

② Autoencoder

③ SVM

④ RNN

연습문제

5. 다음 중 머신러닝(Machine Learning) 학습 방법이 나머지와 다른 것은?

① 군집분석

② 로지스틱회귀분석

③ 인공신경망분석

④ 회귀분석

연습문제

6. 빅데이터 활용에 필요한 3요소는 무엇인가?

① 데이터, 기술, 인력

② 프로세스, 기술, 인력

③ 데이터, 프로세스, 인력

④ 인력, 데이터, 알고리즘

연습문제

7. 인문학 열풍의 외부 환경적인 측면 요소가 아닌 것은?

① 단순 세계화에서 복잡한 세계화로의 변화

② 비즈니스의 중심이 제품생산에서 서비스로 이동

③ 경제와 산업의 논리가 생산에서 시장창조로의 변화

④ 빅데이터 분석기법 이해와 분석 방법론의 확대

연습문제

8. 데이터 사이언티스트의 요구역량이 아닌 것은?

① 통찰력 있는 분석

② 설득력 있는 전달

③ 다분여간 협력

④ 알고리즘에 의해 부당하게 피해를 입는 인력 구제

연습문제

9. 다음은 어떤 기업 내부데이터 솔루션에 대한 설명인가?

물류, 유통업체 등 유통공급망에 참여하는 모든 업체들이 협력을 바탕으로 정보기술 (Information Technoloy)를 활용, 재고를 최적화하기 위한 솔루션이다.

연습문제

10. 인터넷으로 연결된 기계마다 통신 장치를 갖추고 있는 환경에서 사람 또는 기계끼리 자동으로 통신하는 기술로써 사물과 사람, 사물과 사물간의 정보를 상호 소통하는 방식을 무엇이라 하는가?

연습문제

2과목. 데이터 분석 기획

연습문제

1. 하향식 접근방법의 문제탐색 관련한 거시적 관점의 요인이 아닌 것은?

① 사회(Social)

② 기술(Technological)

③ 환경(Environmental)

④ 채널(Channel)

연습문제

2. 다음 중 빅데이터 특징 중 비즈니스 효과에 해당되는 것은?

① Volume

② Variety

③ Velocity

④ Value

연습문제

3. 다음 중 데이터 거버넌스 중 무엇에 관한 설명인가?

데이터 표준용어 설정, 명명규칙 수립, 메타 데이터 구축, 데이터 사전 구축

① 데이터 표준화

② 표준화 활동

③ 데이터 저장 관리

④ 데이터 관리 체계

연습문제

4. 전사 차원의 모든 데이터에 대하여 정책 및 지침, 표준화, 운영조직 및 책임 등의 표준화된 관리체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축을 무엇이라 하는가?

① 데이터 거버넌스

② 데이터웨어하우스

③ 데이터베이스관리시스템

④ 데이터베이스

연습문제

5. 빅데이터 분석 방법론의 분석기획 단계에서 발생하는 산출물로 프로젝트에 참여하는 관계자들 이해를 일치시키기 위한 결과물을 무엇이라 하는가?

① 데이터 스토어

② SOW(Statement of Work)

③ 상세 알고리즘

④ WBS(Work Breakdown Structure)

연습문제

6. 빅데이터 분석 방법 프로세스 순서로 올바른 것은?

① 분석 기획 -> 데이터 준비 -> 데이터 분석 -> 시스템 구현 -> 평가 및 전개

② 데이터 준비 -> 분석 기획 -> 데이터 분석 -> 시스템 구현 -> 평가 및 전개

③ 데이터 준비 -> 분석 기획 -> 데이터 분석 -> 평가 및 전개 -> 시스템 구현

④ 분석 기획 -> 데이터 준비 -> 시스템구현-> 평가 및 전개 -> 데이터 분석

연습문제

7. 분석과제 발굴에 대한 설명 중 적절하지 않은 것은?

① 분석 유즈 케이스는 향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용하도록 한다.

② 상향식 접근 방법은 원천 데이터를 대상으로 분석을 수행하여 가치 있는 문제를 도출하는 일련의 과정이다.

③ 하향식 접근법은 특정 주제별로 새로운 문제를 탐색하여 분석과제를 발굴한다.

④ 하향식 접근방식은 문제가 주어지고 이에 대한 해법을 찾기 위하여 각 과정이 체계적으로 단계화되어 수행하는 방식이다.

연습문제

8. 기업의 데이터 분석 도입의 수준을 명확하게 파악하기 위한 방법으로 분석준비도(Readiness)를 진단할 수 있다. 다음 중 분석준비도를 측정하기 위한 요소로 가장 부적절한 것은?

① 분석업무파악

② 인력 및 조직

③ 분석기법

④ 분석성과

연습문제

9. 동일한 사안이라고 해도 제시되는 방법에 따라 그에 관한 해석이나 의사결정이 달라지는 왜곡 현상을 무엇이라 하는가?

연습문제

10. 분석 수준 진단 방법 중 조직의 분석 및 활용을 위한 역량 수준을 파악하기 위해 도입 -〉( ) -〉 확산 -〉 최적화의 분석 성숙도 단계 포지셔닝을 파악하게 된다. 빈칸에 알맞은 용어는?

연습문제

3과목. 데이터 분석

연습문제

1. 다음 중 시계열 데이터에 대한 설명 중 올바르지 않는 것은?

① 시계열 데이터의 모델링은 다른 분석모형과 같이 탐색 목적과 예측 목적으로 나눌 수 있다.

② 짧은 기간 동안의 주기적인 패턴을 계절변동이라 한다.

③ 잡음은 무작위적인 변동이지만, 일반적으로 원인은 알려져 있다.

④ 시계열분석의 주목적은 외부인자와 관련해 계절적인 패턴 추세와 같은 요소를 설명할 수 있는 모델을 결정하는 것이다.

연습문제

2. 귀무가설이 실제로 사실임에도 불구하고, 귀무가설이 기각하는 확률은?

① 검정력

② 제2종 오류

③ 유의수준

④ 유의확률

연습문제

4. TRUE로 예측한 관측치 중 실제값이 TRUE인 정도를 나타내는 분류 모형 평가지표를 무엇이라 하는가?

① Precision

② Accuracy

③ Recall

④ Sensitivity

연습문제

5. 시계열 데이터의 정상성(Stationary)에 해당되지 않는 것은?

① 평균이 일정하다.

② 분산이 시점에 의존하지 않는다.

③ 공분산은 단지 시차에만 의존하고 시점 자체에는 의존하지 않는다.

④ 시계열 자료는 독립성을 충족해야 한다.

연습문제

6. 이질적인 모집단을 동질성을 지닌 그룹별로 세분화하는 데이터 마이닝 기법을 무엇이라 하는가?

① 연관분석

② 인공신경망

③ 군집분석

④ 로지스틱회귀분석

연습문제

7. 아래 거래 데이터에서 연관규칙 커피 -> 우유의 향상도는?

품목	거래건수
커피	100
우유	100
맥주	100
커피, 우유, 맥주	50
우유, 맥주	200
커피, 우유	250
커피, 맥주	200

① 30.5%

② 50.0%

③ 83.3%

④ 93.3%

연습문제

8. 아래 거래 데이터를 활용하여 연관성 측정 지표 중 빵 -〉 우유의 신뢰도를 구하시오

장바구니	구입품목
1	(빵, 맥주)
2	(빵, 우유, 계란)
3	(맥주, 우유)
4	(빵, 맥주, 계란)
5	(빵, 맥주,우유, 계란)

① 50%

② 55%

③ 60%

④ 65%

연습문제

9. 군집분석 중 모형기반(Model-Based)의 군집방법으로 데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단의 모형으로 나왔다는 가정하에서 모수와 함께 가중치를 자료로부터 추정하는 방법은?

① 밀도기반군집

② 혼합분포군집

③ 비계층적군집

④ 격자기반군집

연습문제

10. 데이터 마이닝 기법 중 아래 보기와 같은 분석 기법을 무엇이라 하는가?

- 물건 배치계획, 카탈로그 배치 및 교차판매
- 카탈로그의 공격적 판촉행사 등의 마켓팅 계획

① 회귀분석

② 주성분분석

③ 군집분석

④ 연관분석

연습문제

11. 아래의 오분류표의 민감도(Sensitivity) 값은?

		예측값		합계
		True	False	합계
실제값	True	40	60	100
	False	60	40	100
합계		100	100	2000

① 0.2

② 0.4

③ 0.6

④ 0.8

연습문제

12. ROC커브는 X축에는 1-특이도, Y축에는 민감도를 나타내 두 평가 값의 관계로 모형을 평가한다. 아래 혼동행렬에서 특이도(Specificity)는?

		예측값		합계
		True	False	합계
실제값	True	TP	FN	P
	False	FP	TN	N
합계		P	N	P+N

① TP/(TP+FP)

② TP/N

③ TN/N

④ TP/P

연습문제

13. 비계층적 군집분석인 k-means 군집분석의 수행 순서로 올바른 것은?

가) 초기 군집의 중심으로 K개의 객체를 임의로 선택한다.
나) 각 자료를 가장 가까운 군집 중심에 할당한다.
다) 각 군집 내의 자료들의 평균을 계산하여 군집의 중심을 갱신한다.
라) 군집 중심의 변화가 거의 없을 때까지 단계 2와 단계3을 반복한다.

① 가 → 나 → 다 → 라

② 나 → 가 → 다 → 라

③ 다 → 나 → 가 → 라

④ 라 → 가 → 나 → 다

연습문제

14. 계층적 군집분석의 거리에 대한 설명 중 적절하지 않은 것은?

① 코사인 유사도는 벡터간의 코사인 각도를 이용하여 서로간에 얼마나 유사한지를 산정한다.

② 맨해튼 거리의 특징은, 두 점 사이의 도로가 모두 x축 또는 y축에 평행한 경우라면, 두 점 사이의 최단거리는 항상 맨해튼 거리와 일치하게 된다는 점이다.

③ 유클리디언 거리가 각 속성들 간의 차이를 모두 고려한 거리라면, 민코스키 거리는 가장 큰 차이만을 가지고 거리를 이야기한다. 계산값이 0에 가까울수록 유사하다.

④ 마할라노비스 거리는 표준화의 상관성을 고려하지 않는 거리로 상관성 분석을 위해서는 표준화 거리를 사용해야 한다.

연습문제

15. 일정한 시간동안 수집 된 일련의 순차적으로 정해진 데이터 셋의 집합을 무엇이라 하는가?

① 주성분 데이터

② 금융 데이터

③ 시계열 데이터

④ 군집 데이터

연습문제

16. 상자 그림(box plot) 중앙에 선이 한 줄 그어져 있는데, 이 중간선의 의미는?

① Median

② Mean

③ Standard deviation

④ variance

연습문제

17. 상자그림에서 제3사분위수에서 1사분위수를 뺀 값으로 전체 자료의 중간에 있는 절반의 자료들이 지니는 값의 범위를 무엇이라 하는가?

① 사분위수 범위

② 사분위수

③ 변동계수

④ 공분산

연습문제

18. Boxplot에서 상한(최댓값)과 하한(최솟값)은 얼마인가?

Q1(1사분위수) = 4, Q3(34사분위수) = 12

① 하한 = −8, 상한 =24

② 하한 = −6, 상 한=22

③ 하한 = −4, 상 한=20

④ 하한 = −2, 상 한=19

연습문제

19. 다음 중 비모수적 검정(Non-parametric test)에 해당하지 않는 것은?

① Run test

② Wilcoxon signed rank test

③ Sign test

④ Chi- squared test

연습문제

20. 다음 중 파생변수에 대한 설명 중 올바르지 않은 것은?

① 파생변수는 기존 변수에 특정 조건 혹은 함수등을 사용하여 새롭게 재정의한 변수를 의미한다.

② 파생변수는 재활용성이 높고, 다른 많은 모델을 공통으로 사용할 수 있는 장점이 있다.

③ 파생변수는 논리성과 대표성을 나타나게 할 필요가 있다.

④ 일반적으로 1차 분석마트의 개별 변수에 대한 이해 및 탐색을 통해 고려하여 파생변수를 생성한다.

연습문제

21. 차원축소 기법 중 하나로, 원 데이터의 분포를 최대한 보전하면서 고차원 공간의 데이터들을 저차원 공간으로 변환하는 분석 기법을 무엇이라 하는가?

① 랜덤포레스트

② 앙상블모형

③ 주성분분석

④ 인공신경망

연습문제

22. 다음 중 K-fold 교차검증에 대한 설명 중 적절하지 않은 것은?

① 교차검증은 주어진 데이터를 가지고 반복적으로 성과를 측정하여 그 결과를 평균한 것으로 분류분석 모형을 평가하는 방법이다.

② 대표적인 k-fold 교차검증은 일반적으로 10-fold 교차검증이 사용된다.

③ 전체데이터 N개에서 2개의 샘플을 선택하여 그것을 평가 데이터 셋으로 모델 검증에 사용하고 나머지는 N-2개는 모델을 학습시키는 교차검증을 LOOCV라 한다.

④ 교차검증을 하는 이유는 과적합을 피하고 일반화된 모델을 생성할 수 있다.

연습문제

23. 표본조사에 대한 설명이 부적절한 것은?

① 조사과정에서 발생하는 오류는 표본추출 오류와 비표본추출 오류로 분류할 수 있다.

② 표본편의(Sampling Bias)는 표본추출방법에서 기인하는 오차를 의미한다.

③ 표본편의는 확률화(Randomization)에 의해 최소화하거나 없앨 수 있다.

④ 표본오차와 비표본오차 모두 표본크기가 증가함에 따라 감소한다.

연습문제

24. 다음 중 주성분 분석에 설명 중 적절하지 않은 것은?

① 제1주성분이라함은 데이터들의 분산이 가장 작은 축을 의미한다.

② 주성분분석은 상관관계가 있는 변수들을 결합해 상관관계가 없는 변수로 분산을 극대화하는 변수로 선형결합을 해 변수를 축약하는데 사용하는 방법이다.

③ 공분산행렬은 변수의 측정단위를 그대로를 반영한 것으로 상관행렬은 모든 변수의 측정단위를 표준화한 것이다.

④ 공분산행렬을 이용한 분석의 경우 변수들의 측정단위에 민감하다.

연습문제

25. 시계열에 영향을 주는 일반적인 요인을 시계열에서 분석하는 방법을 무엇이라 하는가?

26. 아래 설명하는 시계열모형은 무엇인가?

과거 시점의 관측자료와 과거시점의 백색잡음의 선형결합으로 현 시점의 자료를 표현하는 모형

연습문제

27. 아래 빈칸에 알맞은 말은?

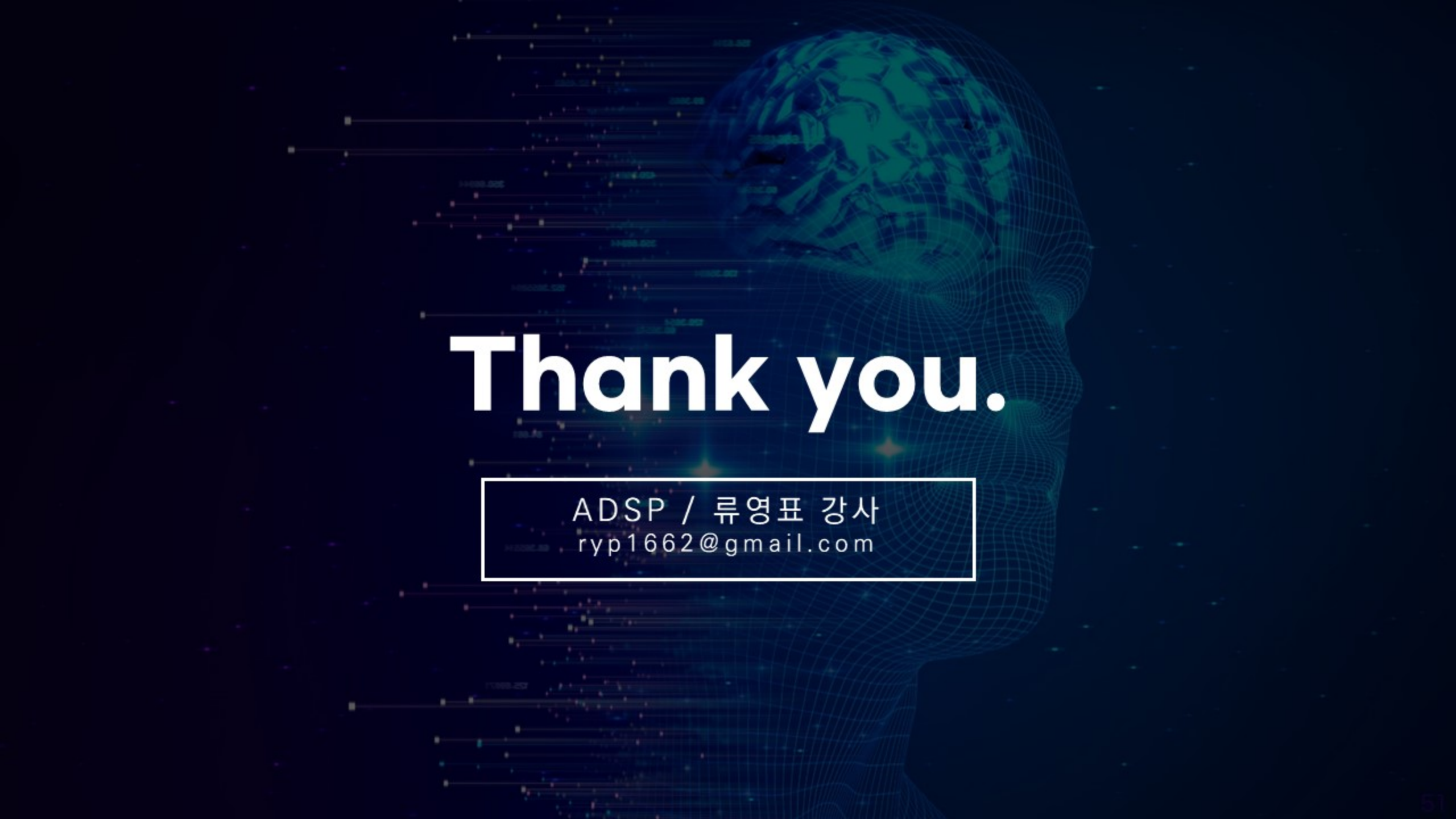
> 인공신경망의 노드가 많을수록 변수의 복잡성을 학습하기 쉽지만 ( ) 문제가 발생한다. 훈련용 데이터에서는 높은 성능을 보여주지만, 일반화 시키기는 어렵다.

28. 인간의 뉴런 구조를 본떠 만든 기계학습 모델을 무엇이라 하는가?

연습문제

29. 이익(Gain)는 목표범주에 속하는 개체들이 각 등급에 얼마나 분포하고 있는지 나타내는 값을 의미한다. 분류된 관측치가 각 등급별 얼마나 포함되는지 나타내는 평가를 무엇이라 하는가?

30. 전사적으로 구축된 데이터 웨어하우스로부터 특정 주체, 부서 중심으로 구축된 소규모 단일 주제의 데이터 웨어하우스로 재무, 생산, 운영과 같이 특정 조직의 특정 업무 분야에 초점을 두고 사용하는 저장소를 무엇이라 하는가?


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