PART 3. 데이터 분석 - 5장. 정형 데이터 마이닝

데이터 준전문가

ADSP, Advanced Data Analytics semi-Professional

류영표 강사
ryp1662@gmail.com

강사 소개

류영표

Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 AI tech 1~4기 멘토

前 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 ㈜셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 공공데이터 청년 인턴 / SW공개개발자대회 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 이젠 종로 아카데미(파이썬, ADSP 강사) / 강남 : ADSP
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원
- 보건 · 바이오 AI활용 S/W개발 및 응용전문가 양성과정 강사
- Upstage AI X KT 융합기술원 기업교육 모델최적화 담당 조교

주요 프로젝트 및 기타사항

- 개인 맞춤형 당뇨병 예방·관리 인공지능 시스템 개발 및 고도화(안정화)
- 폐플라스틱 이미지 객체 검출 경진대회 3위
- 인공지능(AI)기반 데이터 사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

연습문제

1. 다음 데이터마이닝의 대표적인 기능 중 이질적인 모집단을 세분화하는 기능으로 적절한 것은?

① 분류 분석

② 모수 추정

③ 군집분석

④ 연관분석

연습문제

2. 한 보험회사에서 자사 고객의 보험갱신 여부를 고객의 인구통계학적 특성, 보험가입 채널, 상품 종류 등의 정보를 사용하여 예측하려고 한다. 다음 중 가장 적절한 분석 기법은 무엇인가?

① 시계열 분석

② 랜덤포레스트

③ K-menas 군집분석

④ 주성분분석

연습문제

3. 다음 중 기법의 활용 분야가 나머지와 다른 하나를 고르시오.

① 로지스틱 회귀 분석

② 인공신경망

③ 의사결정나무

④ SOM

연습문제

4. 과대적합(Overfitting)은 통계나 기계학습에서 모델에서 변수가 너무 많아 모델이 복잡하고 과대하게 학습할 때 주로 발생한다. 다음 중 과대 적합에 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 종료하고 업데이트 하는 과정을 반복해 과대적합의 방지할 수 있다.

연습문제

5. 귀납적 추론을 기반으로 하는 의사결정나무모형은 실무적으로 가장 많이 사용되는 모델 중 하나이다.
다음 중 의사결정나무모형에 대한 설명으로 부적절한 것은?

① 대표적인 적용 사례는 대출신용평가, 환자 증상 유추, 채무 불이행 가능성 예측 등이 있다.

② 의사결정나무에는 ID3, C4.5, CART 등 여러 가지 알고리즘이 있는데, 핵심적인 공통 개념은 상향식 의사결정 흐름과 해시 탐색(Hash Search)기반의 구조를 가지고 있다는 것이다.

③ 과적합(Overfitting)의 문제를 해결하기 위해 가지치기 방법을 이용하여 트리를 조정하는 방법을 사용한다.

④ 불순도 측도인 엔트로피 개념은 정보이론의 개념을 기반으로 하며, 그 의미는 여러 가지 임의의 사건이 모여 있는 집합의 순수성(Purity) 또는 단일성(homogeneity) 관점의 특성을 정량화해서 표현한 것이다.

연습문제

6. 다음 중 의사결정 나무 모형에서 과대적합되어 현실 문제에 적응할 수 있는 적절한 규칙이 나오지 않는 현상을 방지하기 위해 사용되는 방법으로 가장 적절한 것은?

① 가지치기(Pruning)

② 스테밍(Stemming)

③ 정지규칙(Stopping rule)

④ 랜덤포레스트(Random forest)

연습문제

7. 다음 중 데이터를 무작위로 두 집단으로 분리하여 실험데이터와 평가데이터로 설정하고 검정을 실시하는 모형 평가방법으로 적절한 것은?

① K-fold 교차 검정

② ROC 그래프

③ 홀드아웃 방법

④ 이익 도표

연습문제

8. 데이터마이닝의 목적 중 사람, 상품에 관한 이해를 증가시키기 위한 것으로 데이터의 특징 및 의미를 표현 및 설명하는 기능을 무엇이라고 하는가?

① 기술(Desciption)

② 예측(Forecast)

③ 추정(Estimate)

④ 분류(Classification)

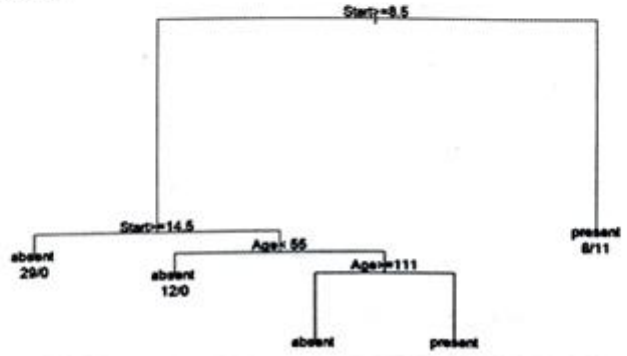
연습문제

9. 아래는 kphosis라는 자료를 이용하여 의사결정나무 분석을 수행한 결과이다. 결과에 대한 해석으로 부적절한 것은?

```
> a<-rpart(Kyphosis~Age +Number+Start, data=kyphosis)
n= 81

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 81 17 absent (0.79012346 0.20987654)
   2) Start>=8.5 62  6 absent (0.90322581 0.09677419)
     4) Start>=14.5 29  0 absent (1.00000000 0.00000000) *
     5) Start< 14.5 33  6 absent (0.81818182 0.18181818)
      10) Age< 55 12  0 absent (1.00000000 0.00000000) *
      11) Age>=55 21  6 absent (0.71428571 0.28571429)
        22) Age>=111 14  2 absent (0.85714286 0.14285714) *
        23) Age< 111 7  3 present (0.42857143 0.57142857) *
   3) Start< 8.5 19  8 present (0.42105263 0.57894737) *
> plot(a)
> text(a, use.n=T)
```



① 뿌리마디에서 아래로 내려갈수록 각 마디에서의 불순도는 점차 증가한다.

② 뿌리마디의 자료는 Start 변수를 이용하여 분리했을 때 Present와 absent를 가장 잘 분리시킬 수 있다.

③ 위 결과의 단계에서 멈추지 않고 가지를 생성한다면, 새로운 자료에 대한 예측력은 떨어질 수 있다.

④ 이 자료에서 Start 변수의 값이 14.5 이상인 관찰치는 kyphosis 변수의 값이 모두 absent였을 것이다.

연습문제

10. 다음 중 의사결정 나무 모형의 학습 방법에 대한 설명으로 부족한 것은 무엇인가?

① 이익도표 또는 검정용 자료에 의한 교차타당성 등을 이용해 의사결정나무를 평가한다.

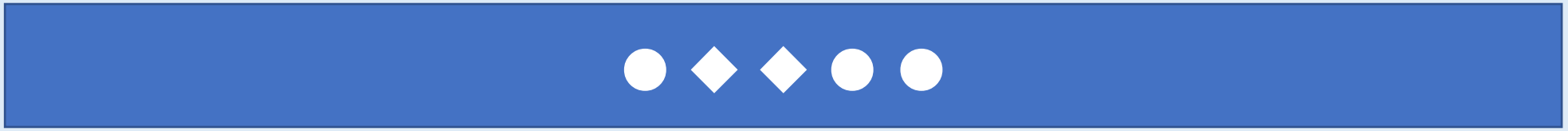
② 분리 변수의 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받지 않고 이루어지며, 공간을 분할하는 모든 직사각형들이 가능한 순수하게 되도록 만든다.

③ 각 마디에서의 최적 분리규칙은 분리변수의 선택과 분리기준에 의해 결정된다.

④ 가지치기는 분류 오류를 크게 할 위험이 높거나 부적절한 규칙을 가지고 있는 가지를 제거하는 작업이다.

연습문제

11. 아래 집단에 대해 지니지수(gini index)는 얼마인가?



① 1

② 2

③ 1/2

④ 12/25

연습문제

12. 다음 중 앙상블 기법이라고 할 수 없는 것은?

① 시그모이드

② 부스팅

③ 배깅

④ 랜덤포레스트

연습문제

13. 앙상블모형(Ensemble)이란 주어진 자료로부터 여러 개의 예측모형을 만든 후 이러한 예측모형들을 결합하여 하나의 최종 예측모형을 만드는 방법을 말한다. 다음 중 앙상블모형에 대한 설명으로 적절하지 않은 것은?

① 배깅은 주어진 자료에서 여러 개의 붓스트랩(Bootstrap) 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 방법이다.

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

④ 앙상블모형은 훈련을 한 뒤 예측을 하는데 사용하므로 교사학습법(Supervised learning)이다.

연습문제

14. 다음 중 오분류표의 평가지표 중 True로 예측한 관측치 중 실제 True인 지표를 무엇이라 하는가?

① Precision

② Specificity

③ Recall

④ Sensitivity

연습문제

15. 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

16. 다음 중 아래 오분류표에 대한 F1은 얼마인가?

		예측치		
		True	False	합계
실제값	True	30	70	100
	False	60	40	100
합계		90	110	200

① 4/10

② 18/57

③ 6/19

④ 7/11

연습문제

17. 아래는 오분표를 나타낸 것이다. 다음 중 특이도(Specificity)는 얼마인가?

		예측치		
		True	False	합계
실제값	True	30	70	100
	False	60	40	100
합계		90	110	200

① 4/10

② 18/57

③ 6/19

④ 7/11

연습문제

18. ROC 커브는 민감도와 1-특이도로 그려지는 커브이다. 아래 오분류표에서 민감도와 특이도는?

교차표		확진결과	
		질병유	질병무
검사	양성	30	20
	음성	40	10

① 민감도 = $\frac{3}{7}$, 특이도 = $\frac{1}{3}$

② 민감도 = $\frac{3}{5}$, 특이도 = $\frac{1}{3}$

③ 민감도 = $\frac{4}{7}$, 특이도 = $\frac{2}{3}$

④ 민감도 = $\frac{2}{5}$, 특이도 = $\frac{4}{5}$

연습문제

19. 다음 중 아래 ( ) 에서 설명하는 활성함수로 가장 적절한 것은?

> 입력층이 직접 출력층에 연결되는 단층신경망(single-layer neural network)에서 활성함수를 ( )로 사용하면 로지스틱 회귀 모형과 작동원리가 유사해진다.

① 계단(Step) 함수

② Tanh함수

③ ReLU 함수

④ 시그모이드(sigmoid) 함수

연습문제

20. 신경망 모형은 자신이 가진 데이터로부터 반복적인 학습과정을 거쳐 패턴을 찾아내고, 이를 일반화하는 예측방법이다. 다음 중 신경망 모형에 대한 설명을 부적절한 것은 무엇인가?

① 피드포워드 신경망은 정보가 전방으로 전달되는 것으로 생물학적 신경계에서 나타는 형태이며 딥러닝에서 가장 핵심적인 구조 개념이다.

② 은닉층의 뉴런 수와 개수는 신경망 모형에서 자동으로 설정된다.

③ 일반적으로 인공신경망은 다층퍼셉트론을 의미한다. 다층 퍼셉트론에서 정보의 흐름은 입력층에서 시작하여 은닉층을 거쳐 출력층으로 진행된다.

④ 역전파 알고리즘은 연결강도를 갱신하기 위해 예측된 결과와 실제값의 차이인 에러의 역전파를 통해 가중치를 구하는데서 시작되었다.

연습문제

21. 신경망 모형은 동물의 뇌신경계를 모방하여 분류를 위해 만들어진 모형이다. 신경망의 학습 및 기억 특성들은 인간의 학습과 기억 특성을 닮았고 특정 사건으로부터 일반화하는 능력도 갖고 있다. 다음 중 신경망 모형에 대한 설명으로 부적절한 것은?

① 은닉층(hidden layer)의 뉴런 수와 개수를 정하는 것은 신경망을 설계하는 사람의 직관과 경험에 의존한다. 뉴런수가 너무 많으면 과적합(overfitting)이 발생하고 뉴런 수가 너무 적으면 입력 데이터를 충분히 표현하지 못하는 경우가 발생한다.

② 신경망 모형에서 뉴런의 주요 기능은 입력과 입력 강도의 가중합을 구한 다음 활성화 함수에 의해 출력을 내보내는 것이다. 따라서 입력 변수의 속성에 따라 활성화 함수를 선택하는 방법이 달라지게 된다.

③ 역전파(back propagation) 알고리즘은 신경망 모형의 목적함수를 최적화하기 위해 사용된다. 연결강도를 갱신하기 위해서 예측된 결과와 실제 값의 차이인 에러(Error)를 통해 가중치를 조정하는 방법이다.

④ 신경망 모형은 변수의 수가 많거나 입출력 변수 간에 복잡한 비선형관계가 존재할 때 유용하며, 잡음에 대해서도 민감하게 반응하지 않는다는 장점을 가지고 있다.

연습문제

22. 로지스틱 회귀모형은 독립변수(x)와 종속변수(y) 사이의 관계를 설명하는 모형으로서 종속변수가 범주형(y=0 또는 y=1)값을 갖는 경우에 사용하는 방법이다. 다음 중 로지스틱 회귀모형에 대한 설명으로 가장 부적절한 것은?

① 이러한 데이터에 대해 선형회귀모형을 적용하는 것이 기술적으로 가능하지만, 선형회귀의 문제점은 0이하의 값이나 1 이상의 값을 예측값으로 줄 수 있다는 것이며 따라서 이를 확률값으로 직접 해석할 수 있다.

② 로지스틱 회귀모형은 클래스가 알려진 데이터에서 설명변수들의 관점에서 각 클래스 내의 관측치들에 대한 유사성을 찾는데 사용할 수 있다.

③ 종속변수 y대신 로짓(logit)이라 불리는 상수를 사용하여 로짓을 설명변수들의 선형함수로 모형화하기 때문에 이 모형을 로지스틱 회귀모형이라고 한다.

④ Odds(오즈)란 클래스 0에 속할 확률(1-p)이 클래스 1에 속할 확률 p의 비로 나타낸다. 즉, odds = p/(1-p)로 나타낸다.

연습문제

23. 계층적 군집분석을 위해 거리 계산을 수행할 때 사용하는 dist 함수에서 지원하는 거리 측도로 부적절한 것은?

① minkowski

② cosine

③ binary

④ canberra

연습문제

24. 계층적 군집분석 수행시 두 군집을 병합하는 방법 가운데 병합된 군집의 오차제곱합이 병합 이전 군집의 오차제곱합의 합에 비해 증가한 정도가 작아지는 방향으로 군집을 형성하는 방법은?

① 단일연결법

② 중심연결법

③ 와드연결법

④ 완전연결법

연습문제

25. 아래 데이터셋 A,B 간의 유사성을 맨하튼(Manhatten)거리로 계산하면 얼마인가?

	키	몸무게
A	165	65
B	170	70

① 25

② 20

③ 15

④ 10

연습문제

26. 아래는 학생들의 키와 몸무게를 정규화한 데이터이다. 최단연결법을 통해 학생들을 3개의 군집으로 나누고자 한다.(유클리디안 거리 사용) 다음 중 가장 적절한 것은?

사람	(키,몸무게)
A	(1,5)
B	(2,4)
C	(4,6)
D	(4,3)
E	(5,3)

① (A,C), (B), (D,E)

② (A,D), (B), (C,E)

③ (A,E), (C), (B,D)

④ (A,B), (C), (D,E)

연습문제

27. 계층적 군집방법은 두 개체(또는 군집) 간의 거리(또는 비유사성)에 기반하여 군집을 형성해 나가므로 거리에 대한 정의가 필요한데, 다음 중 변수 간의 상관성을 동시에 고려한 통계적 거리로 적절한 것은?

① 표준화 거리(Standardized distance)

② 민코우스키 거리(Minkowski distance)

③ 마할라노비스 거리(Mahalanobis distance)

④ 자카드 계수(Jaccard coefficient)

연습문제

28. 거리를 이용하여 데이터 간 유사도를 측정할 수 있는 척도는 데이터의 속성과 구조에 따라 적합한 것을 사용해야 한다. 다음 중 유사도 측도에 대한 설명으로 부적절한 것은?

① 유클리드 거리는 두 점을 잇는 가장 짧은 직선거리이다. 공통으로 점수를 매긴 항목의 거리를 통해 판단하는 측도이다.

② 맨하튼 거리는 각 방향 직각의 이동 거리 합으로 계산된다.

③ 표준화 거리는 각 변수를 해당 변수의 표준편차로 변환한 후 유클리드 거리를 계산한 거리이다. 표준화를 하게 되면 척도의 차이, 분산의 차이로 인해 왜곡을 피할 수 있다.

④ 마할라노비스 거리는 변수의 표준편차를 고려한 거리 측도나 변수 간에 상관성이 있는 경우에는 표준화 거리 사용을 검토해야 한다.

연습문제

29. 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 입력 변수의 정보와 그들의 관계가 지도상에 그대로 나타난다.

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

③ SOM 알고리즘은 고차원의 데이터를 저차원의 지도 형태로 형상화하기 때문에 시각적으로 이해하기 쉬울 뿐 아니라, 변수의 위치관계를 그대로 보존하기 때문에 실제 데이터가 유사하면 지도상 가깝게 표현된다.

④ SOM은 경쟁 학습으로 각각의 뉴런이 입력 벡터와 얼마나 가까운가를 계산하여 연결강도를 반복적으로 재조정하여 학습한다. 이와 같은 과정을 거치면서 연결강도는 입력 패턴과 가장 유사한 경쟁층 뉴런이 승자가 된다.

연습문제

30. 아래는 K-평균군집을 수행하는 절차를 단계별로 기술한 것이다. 다음 중 K-평균군집 수행 절차로 가장 올바른 것은?

가. 각 자료를 가장 가까운 군집 중심에 할당한다.

나. 군집 중심의 변화가 거의 없을 때(또는 최대 반복 수)까지 단계2와 단계3을 반복한다.

다. 초기 (군집의) 중심으로 K개의 객체를 임의로 선택한다.

라. 각 군집 내의 자료들의 평균을 계산하여 군집의 중심을 업데이트한다.

① 다 →라 →가 →나

② 가→ 다→ 라→ 나

③ 가→ 라→ 다 →나

④ 다 →가→ 라 →나

연습문제

31. 다음은 군집화 방법 중 DBSCAN, DENCLUE 기법 중 임의적인(arbitrarity) 모양의 군집 탐색에 가장 효과적인 방법은?

① 밀도기반 군집

② 모형기반 군집

③ 격자기반 군집

④ 커널기반 군집

연습문제

32. 다음 중 자기조직화지도(Self-Organizing Maps, SOM)에 대한 것으로 옳은 것은?

① 군집 분할을 위해 역전파 알고리즘을 사용한다.

② 지도(map)형태로 형상화가 이루어지지만 입력 변수의 위치 관계를 보존하지는 않는다.

③ 학습횟수(epochs)와 군집 내 거리는 반비례한다.

④ 승자 독점의 학습 규칙에 따라 입력 패턴과 가장 유사한 경쟁층 뉴런이 승자가 된다.

연습문제

33. 다음 중 연관성 분석에 대한 설명으로 부적절한 것은?

① Aprori 알고리즘은 최소지지도보다 큰 빈발항목집합에서 높은 측도(신뢰도, 향상도) 값을 갖는 연관규칙을 구하는 방법이다.

② 연관성 분석은 하나 이상의 제품이나 서비스를 포함하는 거래 내역을 이용하여 동시에 구매되는 제품별 거래 빈도표를 통해 규칙을 찾는데서 시작했다.

③ 품목 A와 품목 B의 구매가 상호 관련이 없다면 향상도는 1이 된다.

④ 사건들이 어떤 순서로 일어나고 이 사건들 사이에 연관성을 알아내는 것이 시차 연관분석이지만 원인과 결과 형태로 해석되지는 않는다.

연습문제

34. 아래는 쇼핑몰의 거래내역이다. 연관 규칙 “우유 → 커피” 에 대한 지지도(Support)는 얼마인가?

품 목	거래건수
우유	10
커피	20
{우유, 커피}	30
{커피, 초코렛}	40
전체거래수	100

① 0.1

② 0.2

③ 0.3

④ 0.4

연습문제

35. 아래 거래 전표에서 연관 규칙 “A → B”의 향상도는 얼마인가(소수점 첫째자리에서 반올림)

품 목	거래건수
{A}	100
{B,D}	100
{C}	100
{A,B,C,D}	50
{B,C}	200
{A,B,D}	250
{A, C}	200

① 30%

② 50%

③ 83%

④ 100%

연습문제

36. 아래는 쇼핑몰의 거래 내역이다. 다음 중 규칙 “사과–〉딸기”에 대한 향상도(lift)는 얼마인가?

항 목	거래수
사과	40
딸기	20
포도	30
사과, 딸기	20
사과, 포도	40
딸기, 포도	10
사과, 딸기, 포도	40
전체 거래 수	200

① 0.3 / (0.6 X 0.45)

② 0.4 / (0.7 X 0.45)

③ 0.3 / (0.7 X 0.45)

④ 0.4 / (0.6 X 0.45)

연습문제

37. 모형 평가방법 중 주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차 검정을 실시하는 방법으로 하나는 모형의 학습 및 구축을 위한 훈련용 자료로, 다른 하나는 성과 평가를 위한 검증용 자료로 사용하는 방법은 무엇인가?

연습문제

38. 분류문제를 예측하기 위한 모형을 개발하여 그 결과를 분석하고자 할 때, 특이도를 산출하는 방식을 (a) ~ (d) 로 나타내시오.

		예측값	
		True	False
실제값	True	a	b
	False	c	d

연습문제

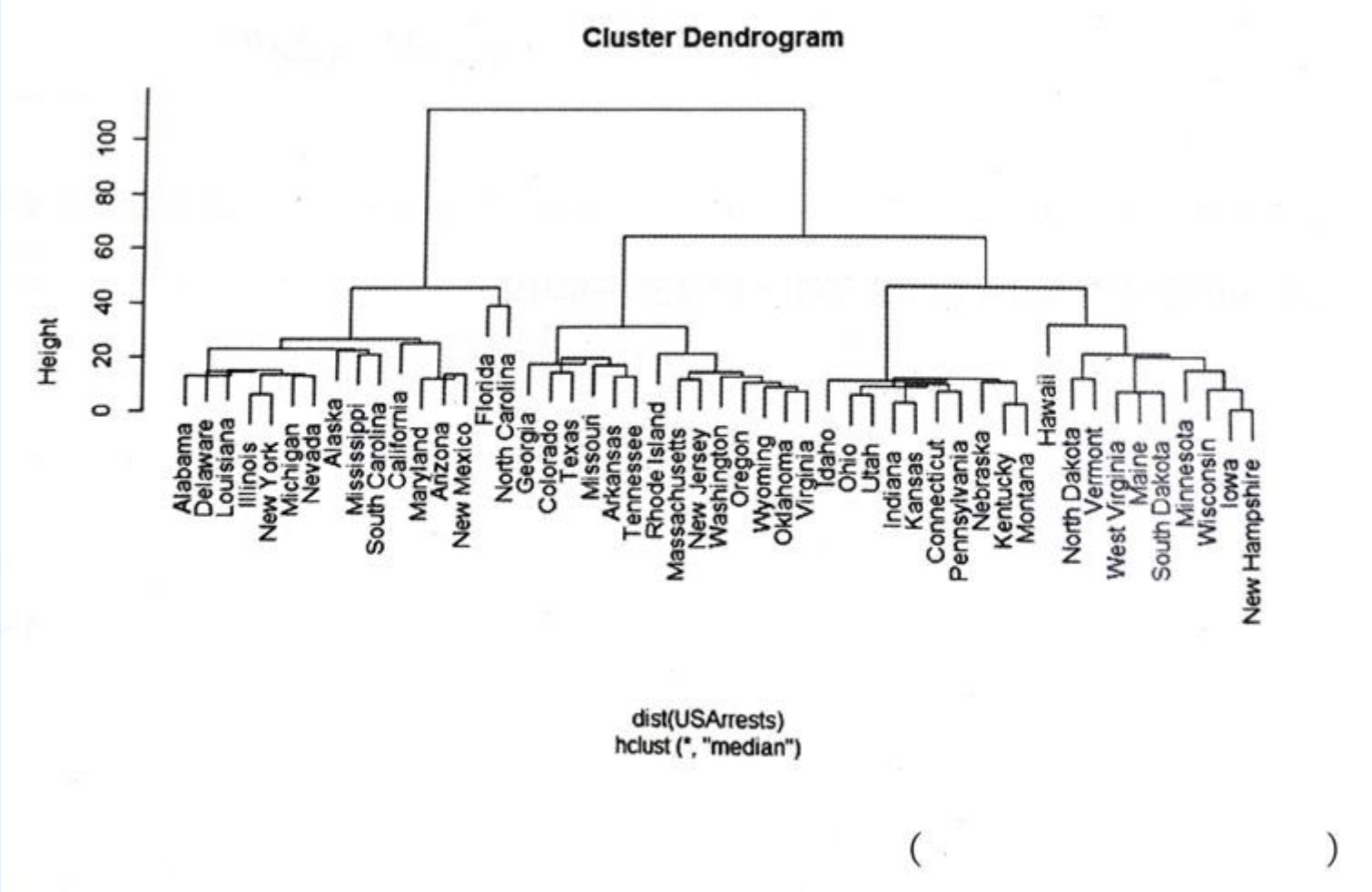
39. 베이즈 정리(Bayes Theory)와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전정보와 데이터로부터 추출된 정보를 결합하고 베이즈 정리를 이용하여 어떤 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘은 무엇인가?

연습문제

40. 군집분석의 품질을 정량적으로 평가하는 대표적인 지표로 군집 내 응집도(Cohesion)와 군집간 분리도(Separation)를 계산하여 군집 내의 데이터의 거리가 짧을수록, 군집 간 거리가 멀수록 값이 커지며 완벽한 분리일 경우 1의 값을 가지는 지표는?

연습문제

41. 아래는 미국 50개 주의 범죄유형으로 군집분석을 한 결과이다. Height=60에서 아래의 덴드로그램을 통해 군집 결과를 도출하면 총 군집의 수는 몇 개인가?



( )

연습문제

42. 랜덤 모델과 비교하여 해당 모델의 성과가 얼마나 좋아졌을지를 각 등급별로 파악하는 그래프로 상위등급에서 매우 크고 하위 등급으로 갈수록 감소하게 되면 일반적으로 모형의 예측력이 적절하다고 판단하게 된다. 모형 평가에 사용되는 이 그래프는 무엇인가?

연습문제

복원 문제에 관련된 것들이다.

연습문제

1. 모형의 성능을 평가할 때, 사용되는 방법론 중 사후확률과 각 분류기준값에 의해 오분류 행렬을 만든 다음, 민감도(Sensitivity)와 특이도(Specificity)를 산출하여 도표에 도식화하여 평가하는 방식은 무엇인가?

① ROC(receive operating characteristics)

② 이익도표(Lift)

③ AUROC

④ 예측률(Prediction rate)

연습문제

2. K-means 군집분석과 계층적 군집분석의 차이를 잘못 설명한 것은?

① K-means 군집분석은 계층적 군집분석과는 달리 한 개체가 처음 속한 군집에서 다른 군집으로 이동해 재배치 될 수 있다.

② K-means 군집분석은 초기값에 대한 의존이 커서 초기값을 어떻게 하느냐에 따라 군집이 달라질 수 있다.

③ K-means 군집분석은 동일한 거리계산법을 적용하면 몇 번을 시행해도 동일한 결과가 나온다.

④ 계층적 군집분석은 동일한 거리계산법을 적용하면 몇 번을 시행해도 동일한 결과가 나온다.

연습문제

3. 아래 거래 전표에서 연관성 규칙 A->B 때의 지지도는?

물품	거래건수
{A}	10
{B}	5
{C}	25
{A, B, C}	5
{B, C}	20
{A, B}	20
{A, C}	15

① 15%

② 20%

③ 25%

④ 30%

연습문제

4. 비계층적 군집분석의 군집분석의 장점에 대한 설명이 잘못된 것은?

① 주어진 데이터의 내부 구조에 대한 사전 정보가 없어도 의미 있는 결과를 얻을 수 있다.

② 다양한 형태의 데이터의 적용이 가능하다.

③ 분석방법의 적용이 용이하다.

④ 사전에 주어진 목적이 없으므로 결과 해석이 쉽다.

연습문제

5. 아래의 데이터 마이닝 분석 예제 중 비지도(unsupervised learning) 분석을 수행해야 하는 예제는?

가. 우편물에 인쇄된 우편번호 판별 분석을 통해 우편물을 자동으로 분류
나. 고객의 과거 거래 구매 패턴을 분석하여 고객이 구매하지 않은 상품을 추천
다. 동일 차종의 수리 보고서 데이터를 분석하여 차량 수리에 소요되는 시간을 예측
라. 상품을 구매할 때 그와 유사한 상품을 구매한 고객들의 구매 데이터를 분석하여 쿠폰을 발행

① 나, 다

② 가, 라

③ 가, 다

④ 나, 라

연습문제

6. 다음 중 연관분석에서 ‘항목 A를 포함한 거래 중에서 항목 A와 항목 B가 같이 포함될 확률은 어느 정도인가를 나타내 주는 연관성의 정도 ’ 로 정의하는 측도로 가장 적절한 것은?

① 지지도

② 신뢰도

③ 특이도

④ 민감도

연습문제

7. 다음 분류 분석 모형 중 훈련용 데이터 집합으로부터 미리 모형을 학습하는 것이 아니라 새로운 자료에 대한 예측 및 분류를 수행할 때 모형을 구성하는 lazy learning 기법을 사용하는 것은 무엇인가?

① 유전자 알고리즘(genetic algorithm)

② 최근접 이웃(nearest neighbor) 모형

③ 신경망(artificial neural network) 모형

④ 서포트 벡터 기계(support vector machine)

연습문제

8. 앙상블(ensemble) 모형은 여러 모형의 결과를 결합함으로써 단일 모형으로 분석했을 때보다 신뢰성 높은 예측값을 얻을 수 있다. 다음 중 앙상블 모형의 특징으로 옳지 않은 것은?

① 이상값(outlier)에 대한 대응력이 높아진다.

② 전체적인 예측값의 분산을 감소시켜 정확도를 높일 수 있다.

③ 모형의 투명성이 떨어져 원인 분석에는 적합하지 않다.

④ 각 모형의 상호 연관성이 높을수록 정확도가 향상된다.

연습문제

9. 다음 중 의사결정나무를 앙상블(ensemble)하는 방법 중 전체 변수 집합에서 부분 변수 집합을 선택하여 각각의 데이터 집합에 대해 모형을 생성한 후 결합하는 방식은?

① 부스팅(boosting)

② 배깅(bagging)

③ 랜덤포레스트(random forest)

④ 붓스트랩(boostrap)

연습문제

10. 다음 중 아래에서 설명하는 문제를 나타내는 용어로 적절한 것은?

분류 모형을 구성하는 경우 예측 실패의 비용이 큰 분류 분석의 대상에 대한 관측치가 현저히 부족하여 모형이 제대로 학습되지 않는 문제가 발생한다.

① 과대적합 문제(overfitting problem)

② 과소적합 문제(underfitting problem)

③ 범주 불균형 문제(case imbalance problem)

④ 정보과부하 문제(information overload problem)

연습문제

11. 군집화 기법 중 특정 공간에서 가까이 있는 데이터가 많은 지역을 중심으로 클러스터를 구성하며 비교적 비어 있는 지역을 경계로 하는 군집 기법으로 임의적인 모양의 군집 탐색에 효과적인 기법은 무엇인가?

① 계층적 군집 기법

② 분리 군집 기법

③ 밀도 기반 군집 기법

④ 격자 기반 군집 기법

연습문제

12. 군집 모형 평가 기준 중 하나이며 군집의 밀집정도를 계산하는 방법으로 군집 내의 거리와 군집간의 거리를 기준으로 군집 분할의 성과를 평가하는 것은 다음 중 무엇인가?

① 피어슨 상관 계수(Pearson Correlation Coefficient)

② ARI(Adjusted Rand Index)

③ NMI(Normalized Mutual Information)

④ 실루엣 계수(Silhouette Coefficient)

연습문제

13. 다음 군집 모형 중 군집의 개수를 미리 지정하지 않아도 되는 장점으로 탐색적 분석으로 사용하는 모형은 무엇인가?

① K-평균군집 모형

② SOM(Self-Organizing Maps) 모형

③ 계층적 군집

④ 혼합분포군집 모형

연습문제

14. 다음 중 자기조직화지도(Self-Organizing Maps, SOM)에 대한 것으로 옳은 것은?

① 군집 분할을 위해 역전파 알고리즘을 사용한다.

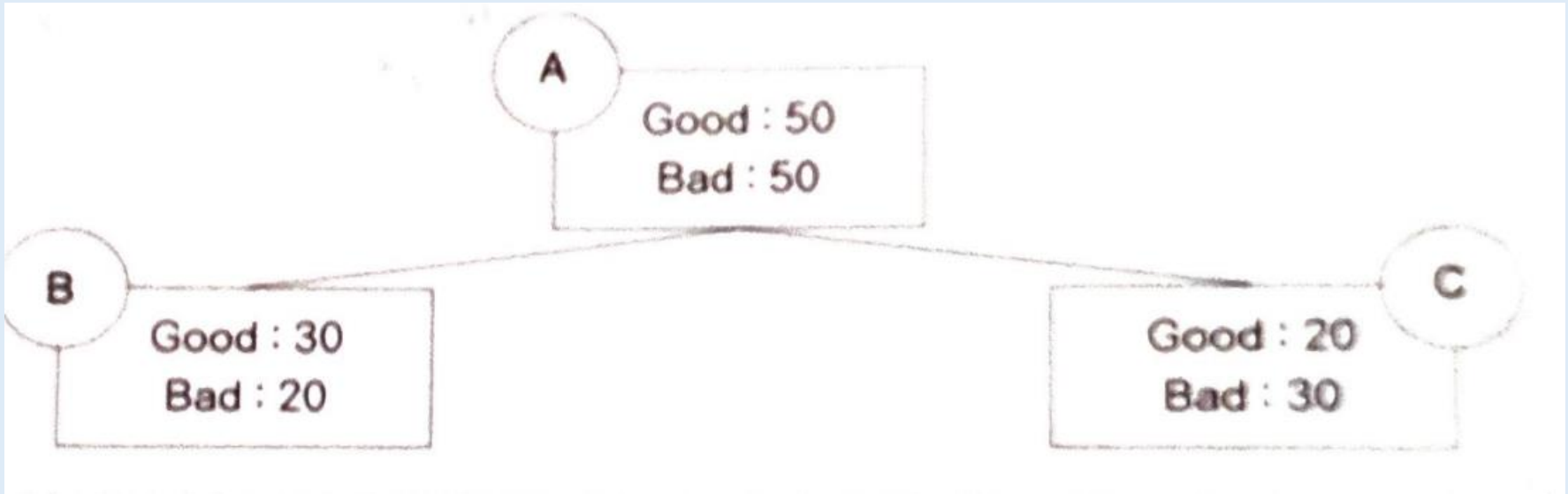
② 지도(map) 형태로 형상화가 이루어지지만 입력 변수의 위치관계를 보존하지는 않는다.

③ 학습횟수(epochs)와 군집 내 거리는 반비례한다.

④ 승자 독점의 학습 규칙에 따라 입력 패턴과 가장 유사한 경쟁층 뉴런이 승자가 된다.

연습문제

15. 아래는 의사결정나무를 나타낸 것이다. C의 지니 지수(gini index)는 얼마인가?



① 0.2

② 0.48

③ 0.4

④ 0.32

연습문제

16. 계층적군집을 수행할 때, 두 군집간의 거리를 측정하는 방법 중 아래에서 설명하는 방법은?

17. 의사결정 나무에서 더 이상 분기가 되지 않고 현재의 마디가 끝마디(leaf node, terminal node)가 되도록 하는 규칙을 나타내는 용어는 무엇인가?

18. 연관규측의 측정 지표 중 도출된 규칙의 우수성을 평가하는 기준으로 두 품목의 상관관계를 기준으로 도출된 규칙의 예측력의 평가하는 지표는 무엇인가?

연습문제

19. 다층 신경망 모형에서 은닉층(hidden layer)의 개수를 너무 많이 설정하게 되면 역전파 과정에서 앞쪽 은닉층의 가중치 조정이 이루어지지 않아 신경망의 학습이 제대로 이루어지지 않는다. 이러한 현상을 나타내는 용어는?

① 기울기 소실 문제

② 지역 최적화 문제

③ XOR 문제

④ 과적합 문제

연습문제

20. ROC(Receiver Operating Characteristic) 그래프에서 이상적으로 완벽히 분류한 모형의 x축과 y축 값으로 옳은 것은? (x축,y축)

① (0, 0)

② (0, 1)

③ (1, 0)

④ (1, 1)

연습문제

21. 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 사의의 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 거리라면, 민코스키 거리는 가장 큰 차이만을 가지고 거리를 이야기한다. 계산값이 0에 가까울수록 유사하다.

④ 마할라노비스 거리는 표준화와 상관성을 고려하지 않는 거리로 상관성 분석을 위해서는 표준화 거리를 사용해야한다.


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