3 과목 / 5 장
정형 데이터 마이닝

01. 다음 중 대용량 데이터 속에서 숨겨진 지식 또는 새로운 규칙을 추출해 내는 과정을 일컫는 것은? ④

① 지식경영

② 의사결정지원시스템

③ 데이터웨어하우징

④ 데이터마이닝

02. 다음 중 반응 변수가 범주형인 경우 예측모형의 주목적으로 가장 적절한 것은? ②

①연관분석 ②분류 ③시뮬레이션 ④최적화

16 회기출

03. 다음 데이터 마이닝의 대표적인 기능 중 이질적인 모집단을 세분화하는 기능으로 적절한 것은? ③

①분류분석 ②모수추정 ③군집분석 ④연관분석

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

① 시계열분석

② 랜덤포레스트

③ k-means 군집분석

④ 주성분분석

05. 데이터 마이닝 단계 중 모델링 목적에 따라 목적변수를 정리하고 필요한 데이터를 데이터 마이닝 소프트웨어에 적용할 수 있도록 준비하는 단계는? ③

① 데이터 마이닝 기법의 적용

② 목적 정의

③ 데이터 가공

④ 데이터 준비

06, 다음 중 기법의 활용 분야가 나머지와 다른 하나를 고르시오. ④

① 로지스틱 회귀 분석

② 인공신경망

③ 의사결정나무

④ SOM

07. 과대적합(overfitting)은 통계나 기계학습에서 모델에서 변수가 너무 많아 모델이 복잡하고 과대하게 학습될 때 주로 발생한다. 다음 중 과대 적합에 대한 설명으로 가장 부적절한 것은? ①

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.

08. 모형의 평가를 위해 관측치를 한번 이상 훈련용 자료로 사용하는 복원 추출법(sampling with replacement)에 기반하는 붓스트랩(bootstrap) 기법에서 일반적으로 훈련용 자료의 선정을 선번 반복할 때 하나의 관측치가 선정되지 않을 확률은 (1-1/d)e 이다. d 가 충분히 크다고 가정할 때 훈련용 집합으로 선정되지 않아 검증용 자료로 사용되는 관측치의 비율은? ④

① 20.5% ② 28.8% ③ 34.2% ④ 36.8%

09. Hitters 데이터셋은 메이저리그에서 활약하는 322 명의 선수에 대한 타자 기록으로 연봉을 비롯한 20 개의 변수를 포함하고 있다. 아래는 모형적합에 앞서 데이터를 train set 과 test set 으로 분할하는 과정이다. 다음 중 아래에 대한 설명으로 가장 부적절한 것은? ④

```
set.seed(1112)
train<-sample(l:nrow(Hitters), nrow(Hitters)/2)
Ytrain<-subset(Hitters[train,L select=Salary)
Xtrain<-subset(Hitters[train,], select=-Salary)
Ytest<-subset(Hitters[-trainJ, select=Salary)
Ytest<-subset(Hitters[-trainJ, select=-Salary)
```

① 50:50 으로 데이터를 분할하고 있다.

② 50%의 데이터(train set)를 사용하여 모형을 학습하고 나머지 50%의 데이터(test set)로 모형을 평가하기 위한 사전작업이다.

③ 모형 학습과 평가를 동일한 데이터셋에 진행하면 모형이 과적합 될 수 있다.

④ 일반적으로 test set 에 대한 모형평가 결과가 train set 에 대한 모형평가 결과보다 좋다.

10. 다음 중 기업이 보유하고 있는 거래데이터, 고객 데이터 등과 기타 외부 데이터를 포함하는 모든 데이터를 기반으로 새로운 규칙 등을 발견하고 이를 실제 비즈니스 의사결정 등에 유용한 정보로 활용하고자 하는 일련의 작업을 무엇이라고 하는가? ②

① 회귀분석 ② 데이터마이닝 ③ 데이터웨어하우징 ④ 의사결정지원시스템

11 . 귀납적 추론을 기반으로 하는 의사결정나무모형은 실무적으로 가장 많이 사용되는 모델 중 하나이다. 다음 중 의사 결정나무모형에 대한 설명으로 부적절한 것은? ②

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

② 의사결정나무모형에은 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 정지규칙과 가지치기 방법을 이용하여 트리를 조정하는 방법을 사용한다.

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

1 오. 다음 중 의사결정 나무 모형에서 과대적합되어 현실 문제에 적응할 수 있는 적절한 규칙이 나오지 않는 현상을 방지하기 위해 사용되는 방법으로 가장 적절한 것은? ①

① 가지치기 (Pruning)

② 스테밍 (Stemming)

③ 정지규칙 (Stopping rule)

④ 랜덤포레스트(Random forest)

13. 다음 중 데이터를 무작위로 두 집단으로 분리하여 실험데이터와 평가데이터로 설정하고 검정을 실시하는 모형 평가방법으로 적절한 것은? ③

① k-fold 교차 검정

② ROC 그래프

⑤ 홀드아웃 방법

④ 이익도표

14. 소매점에서 물건을 배열하거나 카탈로그 및 교차판매 등에 적용하기 적합한 데이터마이닝 기법은 무엇인가? ③

① 분류 (classification)

② 예측 (prediction)

③ 연관분석 (association analysis)

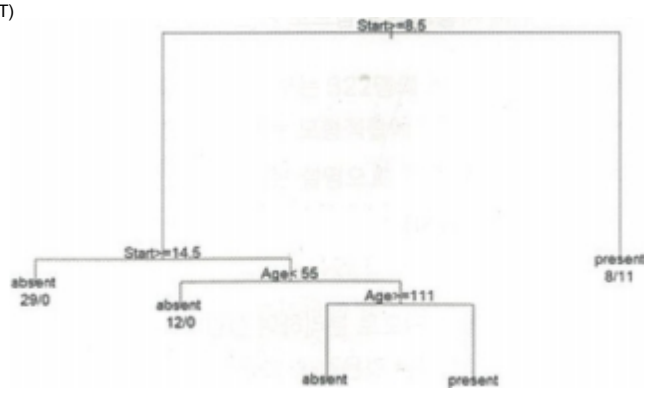
④ 군집 (clustering)

15. 아래는 kyphosis 라는 자료를 이용하여 의사결정나무 분석을 수행한 결과이다. 결과에 대한 해석으로 부적절한 것은? ①

```
> 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>=lll 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 present 와 absent 를 가장 잘 분리시킬 수 있다.

③ 위 결과의 단계에서 멈추지 않고 추가로 가지를 생성한다면, 새로운 자료에 대한 예측력은

떨어질 수도 있다.

④ 이 자료에서 Start 변수의 값이 14.5 이상인 관찰치는 Kyphosis 변수의 값이 모두 absent 였을 것이다.

16. 다음 중 의사결정 나무 모형의 학습 방법에 대한 설명으로 부족한 것은 무엇인가? ②

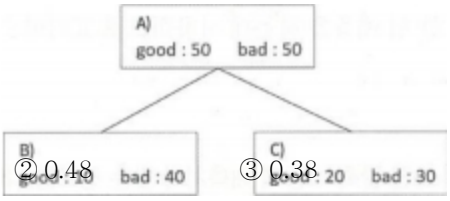
① 이익도표 또는 검정용 자료에 의한 교차타당성 등을 이용해 의사결정나무를 평가한다.

② 분리 변수의 p 차원 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.

17. 다음 중 아래 의사결정나무에서 B 의 지니지수를 계산한 결과로 적절한 것은? ④



① 0.5 ② 0.48 ③ 0.38 ④ 0.32

18. 이익도표(니 ft)를 작성함에 있어 평가도구 중 %Captured Response 를 표현한 계산식으로 올바른 것은? ①

① 해당집단에서 목표변수의 특정범주 빈도 / 전체 목표변수의 특정범주 빈도 x 100

② 해당집단에서 목표변수의 특정범주 빈도 / 해당■집단에서 전체 빈도 x 100

③ 전체에서 목표변수의 특정범주 빈도 / 전체 빈도 x 100

④ 해당집단의 %Response / BASE line Lift x 100

19. 다음 중 배깅(Bagging)에 대한 설명으로 가장 적절한 것은? ④

① 배깅은 데이터 간의 거리를 측정하여 군집화한다.

② 배깅은 트랜잭션 사이에 빈번하게 발행하는 규칙을 찾아낸다.

③ 배깅은 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화한다

④ 배깅은 반복추출 방법을 사용하기 때문에 같은 데이터가 한 표본에 여러 번 추출될 수 있고, 어떤 데이터는 추출되지 않을 수도 있다.

20. 다음 중 앙상블 기법이라고 할 수 없는 것은? ①

①시그모이드 ②부스팅 ⑤배깅 ④랜덤포레스트

21 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? ②

① 배깅은 주어진자료에서 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 Forrest)는 의사결정나무모형의 특징인 분산이 크다는 점을 고려하여 배깅보다 더 많은 무작위성을 추가한 방법으로 약한 학습기들을 생성하고 이를 선형결합해 최종 학습기를 만드는 방법이다.

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

22. 오분류표 중 실제 값이 True 인 관측치 중에 예측치가 맞는 정도를 나타내어 모형의 완전성(completeness)을 평가하는 지표를 무엇이라고 하는가? ①

①재현을 ②오분류을 ③정확도 ④특이도

23. 오분류표를 사용한 평가 지표 중 아래 설명이 나타내는 지표는 무엇인가? ①

정확도(precision)와 재현율(recall)은 한 지표의 값이 높아지면 다른 지표의 값이 낮아질 가능성이 높은 관계를 지니고 있어 이러한 효과를 보정하여 하나의 지표로 만들어 낸 지표

① F1 ②민감도 ③ 특이도 ④ 오즈비

24. 오분류표(effusion matrix)를 사용하여 계산할 수 있는 평가 지표 중 민감도와 동일하며 모형의 완전성(completeness)을 평가하는 지표는 ? ④

① F1 지표

② 정확도(precision)

③ 특이도 (specificity)

④ 재현율 (recall)

다음 중 아래 오분류표를 이용하여 구한 F1 값은 얼마인가? ③

		예측치		합계
		True	False	
실제값	True	40	60	100
	False	60	40	100
합계		100	100	200

① 0.15 ② 0.3 ③ 0.4 ④ 0.55

26.

아래와 같은 오분류표가 있을 때 민감도를 계산하는 방식으로 가장 적절한 것은? ④
분류모형의 성과 분석 중 ROC Curve 는 x 축에 FP Ratio, '/축에는 민감도를 나타낸다.

		예측치		합계
		True	False	
실제값	True	TP	FN	P
	False	FP	TN	N
합계		P,	N,	P+N

① (TP+ TN)+ (P+ N)

② TN+ N

③ TP + (TP+ FP)

④ TP+ P

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

교차표		확진결과	
		질병 유	질병 무
검사	양성	30	20
	음성	40	10

3

3

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 = $\frac{4}{5}$

28. R 에서 인공신경망의 학습 및 추론을 위해 대표적으로 사용되는 함수는 neuralnetO 이다.

다음 중 neuralnet 함수의 실행 결과로 도출되는 일반화 가중치(generalized weight)에 대한 설명으로 가장 적절한 것은? ②

① 각 자료점의 분산이 로그-오즈(log-odds)에 미치는 기여도를 나타낸다.
② 로지스틱 회귀모형에서의 회귀 계수와 유사하게 해석된다.
③ 로지스틱 회귀와 달리 일반화 가중치는 전역적인 기여도를 나타낸다.
④ 모든 자료에 대한 일반화 가중치의 분포는 가중치(weight)에 대한 신뢰구간을 나타낸다.

29. 단층신경망인 퍼셉트론(perceptron)에서 최종 목표값(Target value)은 활성함수에 의해 결정되는데 다양한 활성함수 중 출력값이 여러 개로 주어지고, 목표치가 다범주인 경우 각 범주에 속할 사후확률을 제공하는 함수는 무엇인가? ④

① Tanh 함수 ② Gauss 함수 ③ Sigmoid 함수 ④ Softmax 함수

30. 신경망 모형은 자신이 가진 데이터로부터 반복적인 학습과정을 거쳐 패턴을 찾아내고 이를 일반화하는 예측방법이다. 다음 중 신경망 모형에 대한 설명을 부적절한 것은 무엇인가? ②

① 피드포워드 신경망은 정보가 전방으로 전달되는 것으로 생물학적 신경계에서 나타나는 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.

31. 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 ? ②

① 은닉층(hidden layer)의 뉴런 수와 개수를 정하는 것은 신경망을 설계하는 사람의 직관과 경험에 의존한다. 뉴런수가 너무 많으면 과적합(overfittin)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 실제값의 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.

32. 다음 중 로지스틱 회귀모형에서 설명 변수가 한 개인 경우 해당 회귀 계수의 부호가 o 보다 작을 때 표현되는 그래프의 형태로 적절한 것은? ③

① s 자 그래프

② 양의 선형 그래프

③ 역 S 자 그래프

④ 음의 선형 그래프

33. 로지스틱 회귀모형은 독립변수이와 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? ③

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

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

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

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

34. College 데이터는 777 개의 미국 대학의 각종 통계치를 포함한다. 각 대학에 재학하는 비용이 졸업률(grade Rate)에 미치는 영향을 알아보기 우ㅣ해 사립학교 여부(Private), 고교성적 상위 10% 학생비율(ToplOperc), 등록금(Outstate), 기타지출(Expend)을 활용하기로 했다. 다음 중 아래의 결과물에 대한 설명으로 적절하지 않은 것은 무엇인가? ③

```
>summary(College)
  Private ToplOperc          Outstate        Expend        Grad.Rate
No :212 Min. : 1.00        Min. : 2340     Min. : 3186    Min. : 10.00
Yes:565 1st Qu.:15.00       1st Qu.: 7320   1st Qu.: 6751  1st Qu.: 53.00
          Median :23.00     Median : 9990   Median : 8377  Med ⊥an : 65.00
          Mean :27.56       Mean :10441     Mean : 9660    Mean : 65.46
          3rd Qu.:35.00     3rd Qu.:12925   3rd Qu.:10830  3rd Qu.: 78.00
          Max. :96.00       Max. :21700     Max. :56233    Max. :118.00


>summary(lm(Grad.Rate~.,data=College))

Call:  lm(formula = Grad.Rate ~ " data = College)
Residuals:
      Min IQ Median 3Q Max
 -47.317 -8.503 -0.245 7.741 58.760
Coefficients:
              Estimate    Std. Error    t value    Pr(>!t|)
(Intercept)   39.4130270  1.3579828     29.023     < 2e-16  **♦
PrivateYes    2.9131163   1.3431005     2.169      0.030391  *
ToplOperc     0.3209807   0.0379053     8.468      < 2e-16  ***
Outstate      0.0018820   0.0001988     9.467      < 2e-16  *♦*

Expend -0.0004723 0.0001423 -3.320                0.000943 ***


--- Signif. codes: 0 ***** 0.001 '*** 0.01 '** 0.05                    0.1 " 1

Residual standard error: 13.51 on 772 degrees of freedom
Multiple R-squared: 0.3843, Adjusted R-squared: 0.3811
F-statistic: 120.5 on 4 and 772 DF, p-value: < 2.2e-16
```

① Outstates 변수는 졸업률에 유의한 영향을 미치는 변수이다.

② 고교성적 상위 10% 학생의 비율이 높을수록 졸업률이 높다.

③ 다른 설명변수의 조건이 동일할 때 사립학교(Private Yes)의 경우 공립학교(Private No)에 비해 졸업률이 낮다.

④ 위의 모형은 유의수준 5% 하에서 유의하다.

35. Default 데이터셋은 10000 명의 신용카드 고객에 대한 카드대금 연체여부(default=Yes/No), 카드 대금납입 후 남은 평균 카드잔고(Balanced 연봉(Incomeh 학생여부(student=Yes/No)를 포함한다. 아래는 연체 가능성을 모형화하기 위한 로지스틱 회귀분석 결과이다. 다음 중 유의수준 0.05 하에서 아래에 대한 설명으로 가장 부적절한 것은? ②

```
> summary(Default)
 default      student       balance           income
 No :9667     No :7056   Min.   : 0.0    Min. :    772
 Yes: 333     Yes:2944   1st Qu. : : 481.7   1st Qu.:21340
                          Median  : 823.6    Median :34553
                          Mean               Mean : 33517
                                 : 835.4     3rd Qu.:43808
                           3rd Qu.:1166.3
                          Max.    :2654.3    Max.   :73554

> model<-glm(defauU~" data=Defa 니 lt, family="binomiaV')
> summary(model)

Call:
glm(formula = default - family = "binomial", data = Default)

Deviance Residuals :
      Min          IQ Median              3Q Max

-2.4 的 1 -0.1418 -0.0557 -0.0203              3.7383

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


(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

AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

① balance 는 default 를 설명하는 데 통계적으로 유의하다.

② income 는 default 를 설명하는 데 통계적으로 유의하다.

③ student 는 default 를 설명하는 데 통계적으로 유의하다.

④ balance 는 income[0] 동일할 때 학생일수록 default 가능성이 낮다.

36. 계층적 군집분석을 위해 거리 계산을 수행할 때 사용하는 dist 함수에서 지원하는 거리 측도로 부적절한 것은? ②

① minkowski ② cosine ③ binary ④ Canberra

37. 계층적 군집 방법은 n 개의 군집으로 시작해 점차 군집의 개수를 줄여나가는 방법이다. 다음 중 계층적 군집 분석 결과를 나타내는 도표로 가장 적절한 것은? ③

① 향상도곡선 ② ROC 그래프 ③ 덴드로그램 ④ 산점도

38. 150 개의 식물 개체를 4 개의 변수(꽃받침 길이, 꽃받침 폭, 꽃잎 길이, 꽃잎 폭)로 측정한 데이터를 사용하여 3 개의 식물 군으로 구분하려 한다. 이 때 사용 가능한 분석 방법으로 적절한 것은 무엇인가? ③

① 회귀분석 (Regression)
② 시계열분석 (Time series Analysis)
③ 군집분석 (Cluster Analysis)
④ 연관분석 (Association Analysis)

39. 계층적 군집분석 수행 시 두 군집을 병합하는 방법 가운데 병합된 군집의 오차제곱합이 병합 이전 군집의 오차제곱합의 합에 비해 증가한 정도가 작아지는 방향으로 군집을 형성하는 방법은? ③

①단일연결법 ②중심연결법 ③와드연결법 ④완전연결법

40.

	A	B
키	185	180
앉은키	70	75

아래 데이터 셋에서 A, B 의 유클리드 거리(Euclidean distance)를 계산하시오. ④

① 0 ② $\sqrt{10}$ ③ $\sqrt{25}$ ④ $\sqrt{50}$

41. 아래는 학생들의 키와몸무게를 정규화한 데이터이다. 최단연결법을 통해 학생들을 3 개의 군집

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

으로 나누고자 한다. (유클리디안 거리 사용) 다음 중 가장 적절한 것은? ④

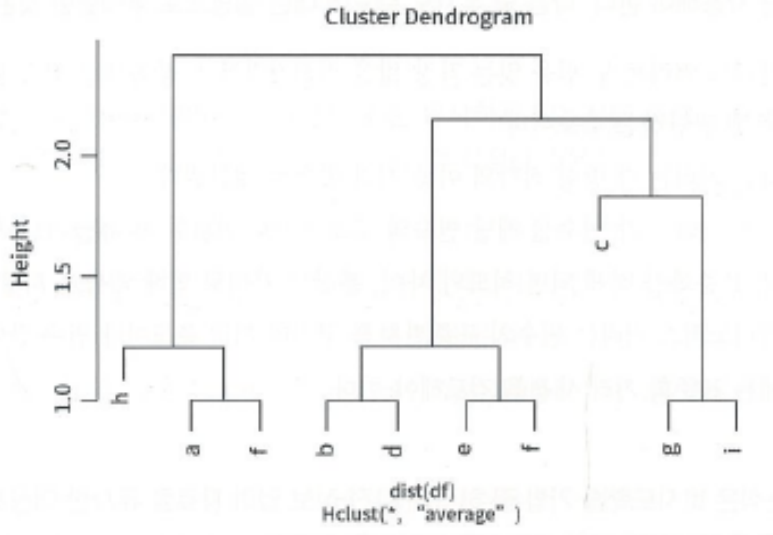
① (A,C), (B), (D.E)

② (A,D), (B), (C,E)

③ (A,E), (C), (B,D)

④ (A,B), (C), (D,E)

42. 아래 그림은 평균연결법을 통한 계층적 군집화 예제이다. 데이터 분석 목적 상 Height 값을 1.5 을 기준으로 하위 군집을 구성할 때 다음 중 생성된 하위 군집을 가장 잘 나타낸 것은? ④



① {i, a, f}, {b, d}, {e,j}, {c }, {g,i}

② {h,a,f}, (b,d), {e,j}, {c,g,i}

③ {h,a,f}, {b,d,e,j}, {c,g,i}

④ {h,a,f}, {b,d,e,j}, {c }, {g,i}

43. 계층적 군집방법은 두 개체(또는 군집) 간의 거리(또는 비유사성)에 기반하여 군집을 형성해 나가므로 거리에 대한 정의가 필요한데, 다음 중 변수의 표준화와 변수 간의 상관성을 동시에 고려한 통계적 거리로 적절한 것은? ③

① 표준화 거리(Standardized distance)
② 민코우스키 거리 (Minkowski distance)
③ 마할라노비스 거리 (MahalanoMs distance)
④ 자카드 계수(Jaccard coefficient)

44. 다음 중 k-means 군집의 단점으로 가장 부적절한 것은? ④

① 볼록한 형태가 아닌 군집이 존재하면 성능이 떨어진다.
② 사전에 주어진 목적이 없으므로 결과 해석이 어렵다.
③ 잡음이나 이상값에 영향을 많이 받는다.
④ 한번 군집이 형성되면 군집내 객체들은 다른 군집으로 이동 할 수 없다.

45. 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? ④

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 측도이나 변수 간에 상관성이 있는 경우에는 표준화 거리 사용을 검토해야 한다.

46. 군집분석은 비지도학습 기법 중 하나로 사전 정보 없이 자료를 유사한 대상끼리 묶는 방법이다. 다음 중 군집분석에 대한 설명으로 부적절한 것은? ①

① 군집분석에서는 군집의 개수나 구조에 대한 가정없이 다변량 데이터로부터 거리 기준에 의한 자발적인 군집화를 유도하지 않는다.
② 군집 결과에 대한 안정성을 검토하는 방법은 교차타당성을 이용하는 방법을 생각할 수 있다. 데이터를 두 집단으로 나누어 각 집단에서 군집분석을 한 후 합쳐서 군집분석한 결과와 비교하여 비슷하면 결과에 대한 안전성이 있다고 할 수 있다.
③ 군집의 분리가 논리적인가를 살펴보기 위해서는 군집 간 변동의 크기 차이를 검토한다.

④ 개체를 분류하기 위한 명확한 기준이 존재하지 않거나 기준이 밝혀지지 않은 상태에서 유용하게 이용할 수 있다.

47. K-means 군집분석에 대한 설명으로 틀린 것은? ④

① K-means 군집분석은 원하는 군집의 개수를 초기에 정하고 seed 중심으로 군집을 형성한다.
② K-means 군집분석은 각 개체를 가장 가까운 seed 가 있는 군집으로 분류한다.
③ 군집으로 분류된 개체들의 정보를 활용하여 새로운 seed 를 계산하면서 개체의 적용에 따른 seed 의 변화를 관찰한다.
④ 95% 이상의 개체가 seed 에 할당 되면 seed 의 조정을 멈춘다.

48. 다음 중 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬화하여 지도의 형태로 형성화하는 클러스터링 방법으로 적절한 것은? ④

① 의사결정나무(Decision Tree)
② 연관규칙 (Association Rule)
③ 랜덤포레스트 (Random Forest)
④ 자기조직화지도(self-organizing Map)

49. 군집분석에서는 관측값들이 얼마나 유사한지 또는 유사하지 않은지를 측정할 수 있는 측도가 필요하다. 다음 중 유사도 측도에 대한 설명으로 가장 부적절한 것은? ④

① 유클리드 거리는 공통으로 점수를 매긴 항목의 크기를 통해 판단하는 측도이다.
② 코사인 거리는 두 단위 벡터의 내적을 이용하여, 단위 벡터의 내각의 크기로 유사도를 측정한다
③ 자카드는 Boolean 속성으로 이루어진 두 객체 간의 유사도 측정에 사용된다.
④ 피어슨 상관계수는 각 객체의 데이터 집합이 직선으로 표현되는 정도를 측정한다.

50. som 은 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도 형태로 형상화하는 방법이다. 다음 중 SOM 방법에 대한 설명으로 부적절한 것은? ②

① SOM 은 입력변수의 위치 관계를 그대로 부존한다는 특징이 있다. 이러한 SOM 의 특징으로 인해 입력 변수의 정보와 그들의 관계가 지도상에 그대로 나타난다.
② SOM 을 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 위치 관계를 그대로 보존하기 때문에 실제 데이터가 유사하면 지도상 가깝게 표현된다.

④ SOM 은 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.

51. 비계층적 군집 방법의 기법인 k-means Clustering 의 경우 이상값(Outlier)에 민감하여 군집 경계의 설정이 어렵다는 단점이 존재한다. 이러한 단점을 극복하기 위해 등장한 비계층적 군집 방법으로 가장 적절한 것은? ①

① PAM(Partitioning Around Medoids)

② 혼합 분포 군집 (mixture distribution clustering)

③ Density based Clustering

④ Fuzzy Clustering

52. 아래는 22 개의 미국 전투기에 대한 4 개의 변수 값을 사용한 군집분석의 결ㄱ ㅏ이다. 이에 대한 설명 중 부적절한 것은? ④

```
> kmeansCjetUJU)
K-means clustering with 3 clusters of sizes 7, 6, 9

Cluster means:
    SPR  RGF   PLF   SLF
1   6.9   5.1  0.20   2.7
2   1.6   4.6  0.16   3.1
3   1.8   4.1  0.15   1.4

Clustering vector:
 FM-1 FJ-1 F-86A      F9F-2     F-94A F30-1   F-89A   XF10 F-I     F9F-6   F100-A   F4D-1
    3     3      2        3   3           3       3      3     3        3     2          2
 F11F-1     F-101A    F3H-2   F-102A     F-8A    F-104B   F-105B     YF-107A   F-106A F-4B
    2 1                   2       2         3         1       1               1 1 1

 F-111A
      1
Within cluster sum       of squares by      cluster :
[1] 11.1 6.4 13.9
(between_SS / total_SS = 79.2 %)

Available components:


[1] "cluster"         "centers"          "totss"          "withinss" "tot.withinss" "betweenss"
[7] "size"            "iter"             "ifault"
```

① 비계층적 군집분석의 결과이다.

② 위의 방법을 사용할 때 군집 개수를 사전에 결정해야 한다.

③ 각 군집은 7 개, 6 개, 9 개의 전투기를 포함한다.

④ 각 군집의 중심에 대한 정보가 포함되어 있다.

53. k-평균 군집으로 대표되는 비계층적 군집 방법에서는 군집의 개수인 k 를 미리 정해주어야 한다. 다음 중 군집수를 정하는 데 활용할 수 있는 그래프로 가장 적절한 것은 무엇인가? ②

① ROC 그래프
② 집단 내 제곱합 그래프
③ 덴드로그램
④ 향상도 곡선

54. 다음 중 k 평균군집에 대한 설명으로 부적절한 것은? ①

① 한번 군집이 형성되면 군집에 속하는 개체들은 다른 군집으로 이동할 수 없다.
② 초기 군집의 중심을 임의로 선택해야한다.
③ 군집의 개수를 미리 선택해야 한다.
④ 이상점에 영향을 많이 받는다.

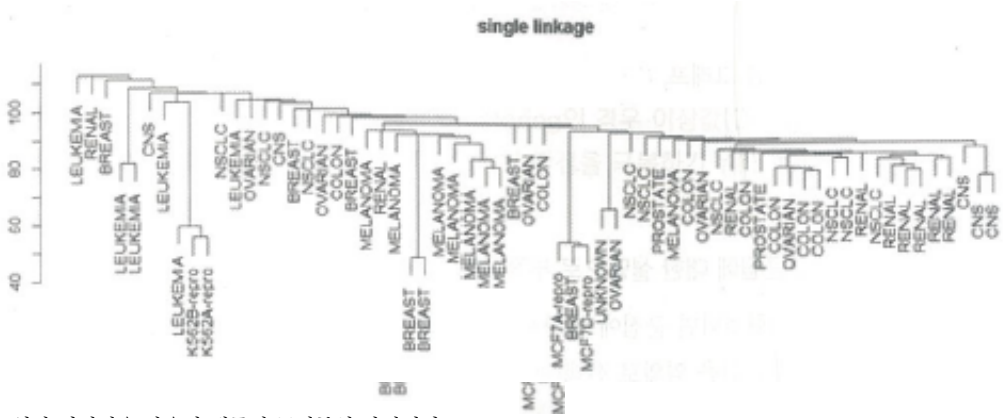
55. 다음 군집화 방법 중 DBSCAN, DENCLUE 기법 등 임의적인(arbitrary) 모양의 군집 탐색에 가장 효과적인 방법은? ①

① 밀도기반 군집
② 모형기반 군집
③ 격자기반 군집
④ 커널기반 군집

56. SOM (Self Organizing Maps) 알고리즘은 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런(Neuron)으로 정렬하여 지도(Map)의 형태로 형상화하는 방법이다. 다음 중 SOM 방법의 설명으로 적절하지 않은 것은 무엇인가? ④

① 입력 벡터와 가장 비슷한 연결강도 벡터를 가진 경쟁층의 뉴런이 승자이며, 승자와 그 주변의 경쟁층 뉴런에 대해서만 연결강도를 수행하는 학습 방법이다.
② 고차원의 표현을 1 차원으로 표현할 수 있는 장점이 있다.
③ 지도 형태의 형상화는 입력변수의 위치 관계를 그대로 보존한다는 특징이 있다.
④ 자율적인(Unsupervised) 신경망 모델로서 역전파(Back Propagation) 알고리즘처럼 여러 단계의 피드백을 처리하면서 전방 패스(Feed-Forward Flow)를 사용하는 방법이다.

57. nci.data 는 64 개 암세포주에 대한 6830 개 유전자 microarray 데이터이다. 아래는 이 자료를 이용한 군집분석 결과이다. 다음 중 아래 결과에 대한 설명으로 가장 부적절한 것은? ④



① 최단 연결법을 사용한 계층적 군집분석 방법이다.

② 두 군집 사이의 거리를 각 군집에서 하나의 관측값을 뽑았을 때 나타날 수 있는 거리의 최소값으로 측정한다.

③ 사슬모양의 군집이 생길 수 있다.

④ 평균연결법에 비해 계산량이 많다.

58. 다음 중 연관성 분석에 대한 설명으로 부적절한 것은? ④

① Apriori 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 A 와 품목 B 의 구매가 상호 관련이 없다면 향상도는 1 이 된다.

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

59. 다음 중 이상값 자료에 민감한 k 평균 군집의 단점을 보완하기 위해 평균 대신 사용되는 것으로 적절한 것은? ①

①중앙값 ②최대값 ③조화평균 ④가중평균

60. 다음 중 R 에서 연관성 분석을 위해 apriori 함수를 활용하여 연관 규칙을 생성하였다. 다음 중 생성된 연관 규칙을 보기 위해 사용되는 함수로 가장 적절한 것은? ③

① sort() ② arule() ③ inspectO ④ aprioriO

61. 다음 중 아래 거래 전표에서 연관 규칙 "빵-->우유"의 향상도를 구한 것으로 알맞은 것은? ③

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

①30% ②50% ③83% ④100%

62. 아래 거래 전표에서 연관 규칙, "A→B"의 향상도는 얼마인가? ③ (소수점 첫째자리에서 반올림)

	거래건수
{A}	100
{B, C}	100
{C}	100
{A, B, C, D}	50
{B, 0}	200
{A, B, D}	250
{A, C}	200

①30% ②50% ③ 83% ④100%

63. 아래는 쇼핑몰의 거래 내역이다. 다음 중 규칙 "사과 -->딸기"에 대한 향상도(lift)는 얼마인가?

항목	거래수
사과	40
딸기	20
포도	30
사과, 딸기	20
사과, 포도	40
딸기, 포도	10
사과, 딸기, 포도	40
전체거래 수	200

① O.3/(O.6 x 0.45)

② 0.4/(0.7 x 0.45)

③ 03/(07 x 0.45)

④ 0.4/(0.6 x 0.45)

64. 분류분석의 모형평가 방법으로 랜덤모델과 비교하여 해당 모델의 성과가 얼마나 향상되었는지를 각 등급별로 파악하는 그래프는 무엇인가? (향상도 곡선)

65. 오분류표(Confusion Matrix)를 활용하여 모형을 평가하는 지표 중 범주 불균형(Class Imbalance Problem 信 가지고 있는 데이터에 대한 중요한 범주만을 다루기 위해 사용되는 지표로 실제값이 False 인 관측치 중 예측치가 적중한 정도를 나타내는 지표는 무엇인가? (Specificity(특이도))

66. 베이즈 정리(Bayes Theory)와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전 정보와 데이터로부터 추출된 정보를 결합하고 베이즈 정리를 이용하여 어떤 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘은 무엇인가? (나이프 베이지안 분류)

1

67. 신경망 모형에서 아래의 식으로 계산되는 함수로서 표준화 지수 함수로 불리며, 출력값 고가 여러개로 주어지고, 목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공하여 출력노드에 주로 사용되는 함수는? (softmax 함수)

$$y_i = \frac{\exp(z_j)}{\sum_{i=1}^{L}\exp(z_i)}, \quad j = 1, \cdots, L$$

68. 두 개체 간의 거리에 기반하여 군집을 형성해가는 계층적 군집방법에서 사용되는 측도 중 두 개체의 벡터 내적을 기반하여 아래의 수식으로 계산할 수 있는 유사성 측도는 무엇인가? (코사인 유사도(cosine similarity))

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

69. 아래는 오분류표를 나타낸 것이다. F1 값을 구하시오 (0.2)

		예측값		합계
		True	False	
실제값	True	30	70	100
	False	170	40	210
합계		300	110	310

70. 혼합분포군집(mixture disribution clustering)은 모형 기반의 군집 방법으로서 데이터가 k 개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정 하에서 분석을 하는 방법이다. k 개의 각 모형은 군집을 의미하며 이 혼합모형의 모수와 가중치의 최대가능도(Maximum Likelihood)추정에 사용되는 알고리즘은 무엇인가?

(Expectation-Maximization)알고리즘)

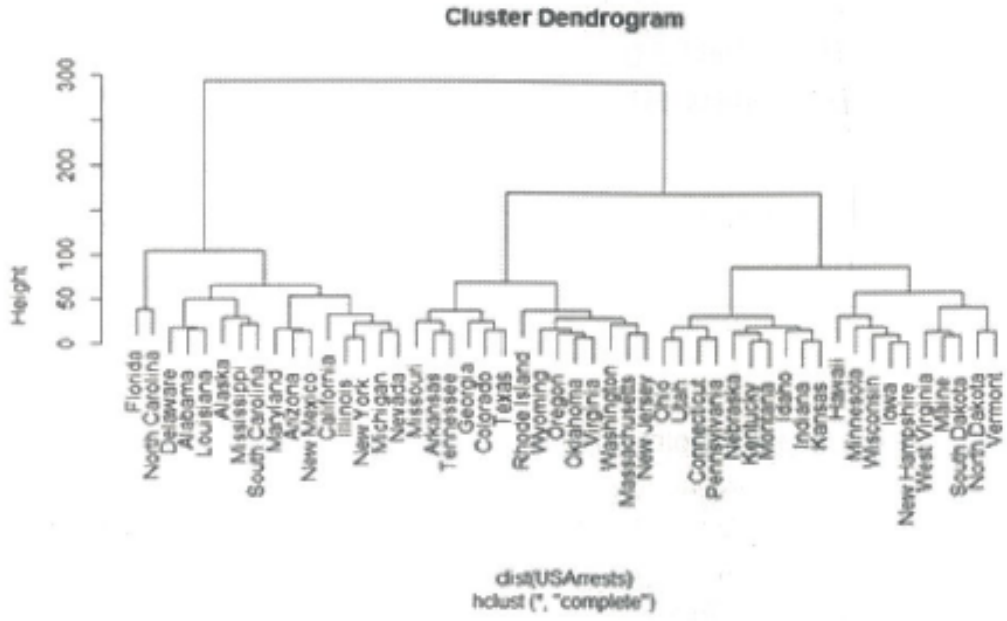
71. 군집분석의 품질을 정량적으로 평가하는 대표적인 지표로 군집 내의 데이터 응집도(cohesion)와 군집간 분리도(separation)를 계산하여 군집 내의 데이터의 거리가 짧을수록, 군집 간 거리가 멀수록 값이 커지며 완벽한 분리일 경우 1 의 값을 가지는 지표는?

(실루엣(shilouette))

72. SOM(Self- Organizing Maps)에서는 각 학습 단계마다 입력층의 데이터 집합으로부터 하나의 표본 벡터를 임의로 선택하고 경쟁층의 프로토타입 벡터와의 거리를 계산하고 가장 가까운 프로토타입 벡터를 선택하는데 이 때 선택된 프로토타입 벡터를 나타내는 용어는 무엇인가?

(BMU(Best-Matching neighbors))

73. 아래는 미국 50 개 주의 범죄 유형으로 군집분석을 한 결과이다. height=150 에서 아래의 덴드로그램을 통해 군집 결과를 도출하면 총 군집의 수는 몇 개인가? (3 개)



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

75. 아래는 학생들의 키와 몸무게를 정규화한 데이터이다. 맨하탄 거리를 이용하여 군집분석을 고자 한다. 맨하튼 거리를 이용하여 A와 B의 거리를 구하시오. (|2-1| + |4-5| = 1+ 1 = 2)

사람	(키, 몸무게)
A	(1,5)
B	(2,4)