3과목 - 3. 정형 데이터 마이닝

01.데이터마이닝

- 기업이 보유하고 있는 일일 거래 데이터, 고객 데이터, 상품 데이터 혹은 각종 마케팅 활동에 있어서의 고객 반응 데이터 등과 이외의 외부 데이터를 포함하는 **모든 사용가능한 원천 데이터를 기반으로 감춰진 지식, 기대하지 못했던 경향 또는 새로운 규칙 등을 발견하고 이를 실제 비즈니스 의사결정 등에 유용한 정보로 활용하는 일련의 작업**
- 데이터마이닝 5단계
  - 목적 정의 : 데이터 마이닝 도입 목적을 명확하게 함
  - **데이터 준비** : **데이터 정제(Cleaning)를 통해 데이터의 품질 확보까지 포함**

    > 필요시 데이터 양 충분하게 확보

  - **데이터 가공** : **목적 변수를 정의하고, 필요한 데이터를 데이터 마이닝 소프트웨어에 적용할 수 있게 가공 및 준비하는 단계**

    > 충분한 CPU와 메모리, 디스크 공간 등 개발환경 구축이 선행

  - 데이터 마이닝 기법 적용 : 모델을 목적에 맞게 선택하고 소프트웨어를 사용하는 데 필요한 값 지정
  - 검증 : 결과에 대한 검증 시행

```
잘못된 설명 찾기(18회 2문제)
- 선형 회귀에 대한 설명 : 종속변수 값을 예측하는 선형모형 추출 방법이다
- tv 광고와 매출액의 상관관계가 크면 당연히 인과관계도 있다고 할 수 있다. (NO!)
```

02.데이터마이닝 기법

- 분류(Classification)
  - 새롭게 나타난 현상을 검토하여 기존의 분류, 정의된 집합에 배정하는 것
  - 의사결정나무, memory-based reasoning 등 (질의에 따른 성격 분류)
- 추정(Estimation)
  - 주어진 입력 데이터를 사용하여 알려지지 않은 결과의 값을 추정하는 것
  - 연속된 변수의 값을 추정, 신경망 모형 (학습의 결과로 요인/관계 분석)
- **연관분석(Association Analysis)**
  - '같이 팔리는 물건 ' 같이 아이템의 연관성을 파악하는 분석
  - **카탈로그 배열 및 교차판매, 공격적 판촉행사 등의 마케팅 계획**
  - 조건 - 결과(if-then) 유형의 패턴을 발견하는데 사용하는 데이터마이닝 기법

- 예측(Prediction)
  - 미래에 대한 것을 예측, 추정하는 것을 제외하면 분류나 추정과 동일한 의미
  - 장바구니 분석, 의사결정나무, 신경망 모형
- 군집(Clustering)
  - **미리 정의된 기준이나 예시에 의해서가 아닌 레코드 자체가 가진 다른 레코드와의 유사성에 의해 그룹화되고 이질성에 의해 세분화 됨**
  - 데이터 마이닝이나 모델링의 준비단계로서 사용됨
- **기술(Description)**
  - **데이터가 가진 특징 및 의미를 단순하게 설명하는 것**, 데이터가 암시하는 바에 대해 설명 및 그에 대한 답을 찾아 낼 수 있어야 함
  - **사람, 상품에 관한 이해를 증가시키기 위한 것으로 데이터의 특징 및 의미를 표현 및 설명하는 기능**

```
분석 방법이 올바른 것 - 연관분석 (20회)
데이터마이닝의 목적 중 기술 찾기 (20회)
```

A.지도학습

- 로지스틱 회귀분석
- 신경망
- 의사결정나무

```
SOM과 함께 섞어 놓은 뒤 '기법 활용 분야'가 다른 것 찾기 (18회)
- SOM은 비지도학습!
신용카드 고객 파산여부를 예측하는 모형이 아닌 것 찾기 (20회)
- 선형회귀분석은 안됨 (로지스틱회귀분석, 의사결정나무, 앙상블모형됨)
- 파산은 범주형 데이터 유형
- 선형회귀분석은 독집변수, 종속변수 모두 연속형일 때 사용한다
```

03.로지스틱 회귀분석

- 종속변수가 성공 또는 실패인 **이항변수**로 되어 있을 때 종속변수와 독립변수 간의 관계식을 이용하여 두 집단 또는 그 이상의 **집단을 분류**하고자 할 때 사용되는 분석기법
- **반응변수가 범주형인 경우 적용되는 회귀분석 모형**
- 로지스틱 회귀모형에서 exp(x1)의 의미는 나머지 변수가 주어질 때 x1이 한 단위 증가할 때마다 성공(Y=1)의 **오즈(odds)**가 몇 배 증가하는지를 나타낸다`
- 오즈(odds)
  - = 성공률 / 실패율 = Pi / (1-Pi) 단, Pi 성공률
  - 일어날 가능성이 높은 경우는 1.0 보다 큰 값
  - 실패가 일어날 가능성이 높은 경우 1.0보다 작은 값을 갖음

오즈(odds) (16회)
로지스틱 회귀분석 찾는 문제 (20회, 22회)

04.로지스틱 회귀계수의 해석

```
> coef(b)
  (Intercept) Sepal.Length
  -27.831451    5.140336 (회귀계수)
# 로지스틱 회귀계수 값은 exp(5.140336)의 값이므로 약 170이 된다.
```

05.선형회귀분석 vs 로지스틱 회귀분석

	일반 선형 회귀분석	로지스틱 회귀분석
종속변수	연속형 변수	이산형 변수
모형 탐색 방법	최소자승법(LSM, 최소제곱법)	최대우도법(MLE), 가중최소자승법
모형 검정	F-test, T-test	x2 test

최소자승법 : 단순회귀분석의 최소제곱추정량 (18회)

06.나이브 베이즈 분류(Naïve Bayes Classification)

- 베이즈 추론을 기반으로 한 방법론의 정확도는 일반적으로 머신러닝의 대표적인 방법인 랜덤 포레스트나 트리분류 방법보다도 높다고 평가받고 있다. 베이지안 추론을 활용한 대표적 분류 방법 알고리즘

주관식 문제 (16회)

07.클래스 불균형

- 분류모형에서 일부 범주형의 관측치가 현저히 부족하여 모형이 학습하기 힘든 문제

주관식 문제로 출제 (21회)

07.의사결정나무(Decision Tree)

- 의사결정나무의 종류(구분)

    - 목표변수가 이산형인 경우의 분류나무(classification tree)
    - 목표변수가 연속형인 경우의 회귀나무(regression tree)
    - 이때, 회귀나무는 분류기준으로 분산의감소량, F-통계량의 p-값 사용

- 의사결정나무의 목적은 새로운 데이터를 분류(classification)하거나 해당 범주의 값을 예측(Prediction)하는 것이다
- 분리 변수 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받는다
- 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나간다 (불순도가 감소한다)
- 최종마디가 너무 많으면 모형이 과대적합(Overfitting)된 상태로 현실 문제에 적용할 수 있는 적절한 규칙이 나오지 않게 되며, 이를 해결하기 위해 가지치기를 한다.
- 의사결정나무를 위한 알고리즘은 CHAID, CART, ID2, C5.0, C4.5가 있으며 **하향식 접근 방법**을 이용한다
- 장점
    - 구조가 단순하여 해석이 용이하다
    - 선형성, 정규성, 등분산성 등의 수학적 가정이 불필요한 비모수적 모형이다
    - 수치형 또는 범주형 변수를 모두 사용할 수 있다
- 단점
    - 분류기준값의 경계선 부근의 자료값에 대해서는 오차가 크다(비연속성)
    - 로지스틱회귀와 같이 각 예측변수의 효과를 파악하기 어렵다
    - 새로운 자료에 대한 예측이 불안정할 수 있다

08.의사결정나무 모형의 분리 기준

- **정지규칙**이란 더 이상 분리가 일어나지 않고 현재의 마디가 최종마디가 되도록 하는 여러 가지 규칙으로 **지니 지수, 엔트로피 지수, 카이제곱통계량** 등이 있다
- 지니지수
    - 불순도를 측정하는 지수이므로 값이 **작을수록 순수도가 높다**
    - **지니지수 구하기**
    - 1 - 합((각 도형별수/전체수)^2)

- 엔트로피 지수(Entropy measure)
    - 엔트로피 지수가 가장 작은 예측 변수와 이때의 최적 분리에 의해 자식 마디를 형성함
    - p=0.5일 때 이질성이 가장 크다
- 카이제곱 통계량의 유의 확률
    - 카이제곱 검정에 기반을 둔 것으로 **목표변수가 범주형일 때** Pearson의 카이제곱 통계량을 분리기준으로 사용함

- p-value 는 그 값이 **작을수록 자식 노드 내의 불확실성(이질성)이 크다**

09.의사결정나무의 결정규칙

- 분할규칙(Splitting rule)
  - 새 가지를 어디서부터 나오게 할까?
- **정지규칙(Stopping rule)**
  - 더 이상 분리가 일어나지 않고 현재의 마디가 최종마디가 되도록 하는 여러 가지 규칙으로 지니 지수, 엔트로피 지수, 카이제곱통계량 등이 있음
  - '불순도 감소량'이 아주 작을 때 정지함
- **가지치기 규칙(Pruning rule)**
  - 어느 가지를 쳐내야 예측력이 좋은 나무가 될까?
  - **최종노드가 너무 많으면 Overfitting 가능성이 커짐, 이를 해결하기 위해 사용**
  - 가지치기 규칙은 별도 규칙을 제공하거나 경험에 의해 실행할 수 있음
  - 가지치기의 비용함수(Cost Function)을 최소로 하는 분기를 찾아내도록 학습
  - Information Gain이란 어떤 속성을 선택함으로 인해 데이터를 더 잘 구분하게 되는 것을 의미함
- 알고리즘 별 분리기준변수 선택법
  - CART : 이산형 목표변수 - 지니지수, 연속형 목표변수 - 분산 감소량
  - C5.0 : 이산형 목표변수 - 엔트로피지수
  - CHAID : 이산형 목표변수 - Pearson의 카이제곱통계량, 연속형 목표변수 - ANOVA F-통계량의 p-값

```
지니지수 구하기 문제 (21회, 22회)
정지규칙 주관식 문제 (21회)
지니값, 카이제곱, 엔트로피 지수 설명으로 틀린 것 (23회)
- 지니지수는 작을 수록 순수도가 높다! (근데 떨어진다해서 틀림)


CART 찾기 문제 (20회)
의사결정나무 알고리즘 내용으로 틀린 것 찾기 (22회)
- 가지치기의 비용함수를 최대로 하는!! 이라고 해서 틀림
목표변수가 연속형인 경우 회귀나무의 분류기준 찾기 (23회)
```

10.오분류표(confusion matrix)를 활용한 평가지표

Confusion matrix		예측값		
		TRUE	FALSE	
실제값	TRUE	40 (TP)	60 (FN) Type II Error	Sensitivity TP / (TP+FN)
	FALSE	60 (FP) Type I Error	40 (TN)	Specificity TN / (TN+FP)
		Precision TP / (TP+FP)	Negative Predictive Value TN / (TN + FN)	Accuracy (TP+TN) / (TP+TN+FP+FN)

confusion matrix		실제값	
		Y	N
예측값	Y	True Positive	False Positive
	N	False Negative	True Negative

11.오분류표를 활용한 평가지표

- Accuracy : (TP + TN) / (TP + FP + FN + TN)
  - 전체 예측에서 옳은 예측의 비율
- Error Rate : (FP + FN) / (TP + FP + FN + TN)
  - 전체 예측에서 틀린 예측의 비율
- 특이도(Specificity) : TN / (TN + FP)
  - 실제로 N 인 것들 중 예측이 N으로 된 경우의 비율
- FP Rate : FP / (FP + TN) : Y가 아닌데 Y로 예측된 비율
  - 1 - Specificity
  - 실제 N인것 중 예측이 Y로 된것
- Kappa : Accuracy - P(e) / (1 - P(e))
  - 두 평가자의 평가가 얼마나 일치하는지 평가하는 값
  - 0~1 사이의 값을 가짐
- 오분류표의 평가지표 중 F2(아래첨자)의 의미
  - 재현율에 정확도의 2배만큼 가중치를 부여하는 것

12.정분류율(Accuracy)

- 보기의 표를 보고 정분류율(Accuracy)를 구하시오

Confusion matrix		Predicted class	
		1	0
Actual class	1	a	b
	0	c	d

```
Accuracy = (a + d ) / (a + b + c + d)    (18회)
Error Rate, Misclassification rate = (b + c) / (a + b + c + d)
```

13.정확도, 재현율

- 정확도(Precision)
  - 예측값이 True인 것에 대해 실제값이 True인 지표
  - 식 :TP / (TP + F**P**)
- 재현율(Recall, Sensitivity)
  - 실제값이 True인 것에 대해 예측값이 True인 지표
  - 식 : TP / (TP + F**N**)
  - True Positive / (True Positive + False Negative)

```
재현율(Recall) 식 고르는 문항 (17회)
Recall을 고르는 문항 (19회)
Precision을 고르는 문항 (20회)
```

14.F1값 구하기

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

- F1은 데이터가 불균형 할 때 사용한다
- 오분류표 중 정확도와 재현율의 조화평균을 나타내며 정확도와 재현율에 같은 가중치를 부여하여 평균한 지표
- 2 * (Precision * Recall) / (Precision + Recall)

    - Precision : 정확도, Recall : 재현율
- Precision : True로 예측한 관측치 중 실제 True인 지표
- Recall : 실제 True인 것 중 예측이 True로 된 것

```
2 * (0.91 * 0.95) / (0.91 + 0.95)
```

```
0.9295698924731183
```

```
F1을 구하는 문항
= 2 * (Precision * Recall) / (Precision + Recall)   (16회)
Precision : TP / (TP + FP) = 100/110 = 0.91
Recall    : TN / (TN + FN) = 100/105 = 0.95
F1 = 2 * (0.91 * 0.95) / (0.91 + 0.95) = 0.93


F1을 고르는 문항 (17회)


F2의 의미 찾기 (21회)
kappa 단답형 쓰기 문제 (17회)
```

15.Confusion matrix 문제

Confusion matrix		예측값	
		TRUE	FALSE
실제값	TRUE	40	60
	FALSE	60	40

```
F1 구하기 (21회)
   Precision = TP/(TP + FP) = 0.4
   Recall = TN / (TN + FN) = 0.4
   F1 = 0.4
특이도 구하기 (22회)
   특이도 = TN / (TN + FP), 실제로 N 인 것들 중 예측이 N으로 된 경우의 비율
Error Rate(오분류율)  (23회)
   = (FP + FN) / (TP + FP + FN + TN)
   = 120 / 200 = 0.6
```

16.ROC, 향상도 곡선 VS 이익도표

- ROC(Receiver Operating Characteristic)
    - X축 **FR-Ratio(1 - Specificity(특이도))**, **Y축 민감도(Sensitivity)**를 나타내어 이 두 평면 값의 관계로 하는 모형 평가
    - 가장 **이상적인 X, Y축의 값은 0, 1** 이다
    - ROC 그래프의 밑 부분의 면적이 넓을수록 좋은 모형을 평가함

- 향상도 곡선(lift curve)
  - 랜덤 모델과 비교하여 해당 모델의 성과가 얼마나 향상되었는지를 각 등급별로 파악하는 그래프
- 이익도표
  - 분류 분석 모형을 사용하여 분류된 관측치가 각 등급별로 얼마나 포함되는지를 나타내는 도표

ROC 주관식 문제로 출제 (17회)
이상적 X,Y축 값 찾기 (22회)
향상도 곡선 주관식 문제로 출제됨 (19회) 향상도곡선 주관식 문제 (22회)

17.인공 신경망

- 인공 신경망의 특징
  - 딥러닝은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계학습 알고리즘이다.
  - 이와 관련한 딥러닝 기법의 기반이 되는 모형이 '인공신경망'이다
  - 신경망은 입력층, 은닉층, 출력층 3개의 층으로 구성되어 있음
  - 각 층에 뉴런(노드)이 몇 개씩 포함되어 있음
  - 분석가의 주관과 경험에 따름
  - **풀고자 하는 문제 종류에 따라 활성화 함수의 선택이 달라짐**
  - 역전파 알고리즘은 동일 입력층에 대해 원하는 값이 출력되도록 개개의 weight를 조정하는 방법으로 사용됨
- 신경망 모형의 장점
  - 변수의 수가 많거나 입,출력변수 간에 복잡한 비선형 관계에 유용
  - **이상치 잡음에 대해서도 민감하게 반응하지 않음**
  - 입력변수와 결과변수가 연속형이나 이산형인 경우 모두 처리 가능
- 신경망 모형의 단점
  - 결과에 대한 해석이 쉽지 않음
  - 최적의 모형을 도출하는 것이 상대적으로 어려움
  - 데이터 정규화를 하지 않으면, 지역해(local minimum)에 빠질 위험 있음

인공신경망 특징 중 틀린 것 (19회)
- 활성화 함수 선택을 입력변수의 속성에 따라 달라진다해서 틀림
인공신경망 설명 써주고 단어 찾기 (22회)

18.Min-Max Normalization

- 데이터 전처리 방법 중 데이터를 일정범위로 Feature scaling 범위 0~1 사이로 적용해주고 원 데이터의 분포를 유지하는 정규화 방법

```
min-max normalization 찾기 (21회)
```

19.신경망 활성화 함수(activation function)

- 결과값을 내보낼 때 사용하는 함수로, 가중치 값을 학습할 때 에러가 적게 나도록 도움
- **풀고자 하는 문제 종류에 따라 활성화 함수의 선택이 달라짐**
- 문제 결과가 직선을 따르는 경향이 있으면 '선형함수'를 사용
- 목표 정확도와 학습시간을 고려하여 선택하고 혼합 사용도 함
- 활성화 함수의 종류
  - 계단함수 : 0 또는 1의 결과
  - 부호함수 : -1 또는 1의 결과
  - **sigmoid 함수** : 연속형 0~1, Logistic 함수라 불리기도 함

    > **선형적인 Multi-Perceptron에서 비선형 값을 얻기 위해 사용**

  - **softmax 함수**

    > 1. **sigmoid 함수의 일반화된 형태로 목표치가 다 범주인 경우 각 범주에 속할 사후 확률을(Posterior probability) 제공하는 활성화 함수**
    > 2. 주로 3개 이상 분류 시 사용함

- 포화 상태
  - 신경망에서 일반적으로 가중치 초기화는 -1.0 ~ 1.0 사이의 임의값으로 설정하지만 가중치를 지나치게 큰 값으로 초기화하면 활성화 함수를 편향시키게 되며 활성화 함수가 과적합 되는 상태

```
softmax 주관식 문제로 출제 (17회)
시그모이드 함수의 범위 찾기 (20회)
시그모이드 함수 찾기 (21회)
포화 상태 설명 (21회)
```

20.신경망 은닉 층, 은닉노드 수

- 신경망 은닉 층 및 은닉 노드 수
  - 다층신경망은 단층신경망에 비해 훈련이 어려움
  - 은닉층 수와 은닉 노드 수의 결정은 '분석가가 분석 경험에 의해 설정'함

- **은닉층 노드가 너무 적으면**
  - **네트워크가 복잡한 의사결정 경계를 만들 수 없다**
  - Underfitting 문제 발생
- **은닉층 노드가 너무 많으면**
  - 복잡성을 잡아낼 수 있지만, 일반화가 어렵다
  - **레이어가 많아지면 기울기 소실 문제가 발생**할 수 있다
  - **과적합(Overfitting) 문제 발생**
- **기울기 소실 문제(Vanishing Gradient Problem)**
  - 다층신경망에서 은닉층이 많아 인공신경망 기울기 값을 베이스로 하는 역전파 알고리즘으로 학습시키려고 할 때 발생하는 문제
  - 인공신경망에서 역전파 알고리즘은 출력층으로부터 하나씩 앞으로 되돌아가며 각 층의 가중치를 수정하는 방법이다. 은닉층이 늘어나면서 기울기가 중간에 0이 되어 버리는 문제를 말한다.
- 역전파 알고리즘(Backpropagation Algoritm)
  - 출력층에서 제시한 값에 대해, 실제 원하는 값으로 학습하는 방법으로 사용
  - **동일 입력층에 대해 원하는 값이 출력되도록 개개의 weight를 조정하는 방법으로 사용됨**

```
틀린 것 찾기 (17회)
- 은닉층, 은닉노드수가 자동 설정이라 해서 틀림
기울기 소실 문제 용어 - 객관식 (22회)
은닉노드가 너무 적으면 발생하는 문제는? (23회)
기울기 소실 문제 - 주관식으로 (23회)
```

21.데이터 분할

1. 훈련 데이터에 대한 학습만을 바탕으로 모델의 설정(Hyperparameter)를 튜닝하게되면 과대적합(overfitting)이 일어날 가능성이 매우 크다
2. 모델이 너무 간단하여 정확도가 낮은 모델을 과소적합(underfitting) 되었다고 말한다
3. 과대적합이나 과소적합의 문제를 최소화하고 모델의 정확도를 높이는 가장 좋은 방법은 더 많고 다양한 데이터를 확보하고, 확보한 데이터로부터 더 다양한 특징(feature)들을 찾아 학습에 사용하는 것이다
4. training set결과가 일반적으로 test set 결과보다 좋다

```
4번 traning set, test set 을 바꿔서 틀리게 한 것 선택하는 답 (16회)
```

```
상향식 접근 방법이라해서 틀린 예가 있었음 (16회)
P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받지 않는다해서 틀림 (17회)
끝 노드로 갈수록 불순도가 상승한다로 해서 틀림 (18회)
```

가지치기 주관식 (19회)
과대적합 주관식 (23회)

22.홀드아웃, 교차검증, 붓스트랩

- **홀드아웃(Hold Out)**
  - 과적합(overfitting) 발생 여부를 확인하기 위해서 **주어진 데이터의 일정 부분을 모델을 만드는 훈련 데이터로 사용하고, 나머지 데이터를 사용해 모델을 평가**한다. 이렇게 데이터를 훈련, 테스트 데이터로 분리하여 검증하는 방법
  - **주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차검점을 실시하는 방법**으로 하나의 모형 학습 및 구축을 위한 훈련용 자료로 하나는 성과평가를 위한 검증용 자료로 사용하는 방법
- 교차검증(Cross Validation)
  - 주어진 데이터를 가지고 반복적으로 성과를 측정하여 그 결과를 평균한 것으로 분류분석 모형을 평가하는 방법
  - K-fold 교차검증 : K개로 데이터를 분할 하여 K번째의 하부 집합을 검증용 자료로, 나머지 K-1개는 훈련용 자료로 사용, 이를 K번 반복 측정하고 그 결과를 평균 낸 값을 최종 평가로 사용함
- **붓스트랩(Bootstrap)**
  - 평가를 반복하는 측면에서 교차검증과 유사하지만, 훈련용 자료를 반복 재선정한다는 점에서 차이가 있음
  - 붓스트랩은 관측치를 한 번 이상 훈련용 자료로 사용하는 복원추출법에 기반함 전체 데이터 양이 크지 않을 경우의 모형 평가에 가장 적합
  - 일반적인 **훈련 데이터의 양은 63.2% 임**

홀드아웃을 찾는 문항 (17회)
홀드아웃 주관식 문제 (20회, 22회)
붓스트랩의 훈련 데이터 양 63.2% 찾기(16회)

23.앙상블 모형

- 하나의 모델만을 학습시켜 사용하지 않고 여러 모델을 학습시켜 결합하는 방식으로 문제 처리
- 약하게 학습되 여러 모델들을 결합하여 사용하는 것을 앙상블 학습이라 할 수 있음
- 성능을 분산시키기 때문에 과적합(overfitting) 감소 효과가 있음
- 상호 **연관성이 높으면 분류가 쉽지 않음**
- 배깅(bagging)

    - bootstrap aggregating의 줄임말
    - 원 데이터로 집합으로부터 **크기가 같은 표본의 중복을 허용**하고 복원추출하여 각 표본에 대해 분류기(classifiers)를 생성하는 기법
- 랜덤 포레스트(random forest)
    - **보험사에서 해지할 예상 고객을 예측시 사용할 수 있는 적절한 기법**
    - 매번 분할을 수행할 때마다 설명변수의 일부부만을 고려함으로 성능을 높이는 방법
- 부스팅(boosting)
    - 재표본 과정에서 각 자료에 동일한 확률을 부여하지 않고, **분류가 잘못된 데이터에 더 가중**을 주어 표본을 추출하는 분석 기법
    - Leaf-wise-node 방법을 사용하는 알고리즘 : Light GBM

```
매우 자주 출제되는 앙상블!  하나씩 모두 이해해두자
앙상블 모형이 아닌 것 찾기 (16회) - 시그모이드가 답이었지
부스팅을 주관식 문제 (18회)
보험사 해지 예상 고객에 대한 적절한 기법으로 랜덤포레스트 찾기 (18회)
배깅의 설명써주고 배깅 찾기 (19회, 20회)
앙상블 특징으로 틀린 것 (21회)
- 상호 연관성이 높으면 분류가 쉽지 않음~!
랜덤포레스트 찾는 문제 (21회)
Light GBM을 찾는 문제 (23회)
배깅(Bagging) 주관식 문제 (22회, 23회)
```

B.비지도학습(Unsupervised Learning)

- 군집분석
- 연관분석
- SOM(Self-Organizing Maps, 자기조직화지도)

24.군집분석(Cluster analysis)

- 군집 분석은 신뢰성과 타당성을 점검하기 어려움
- 군집 결과에 대한 안정성을 검토하는 방법으로 '군집타당성지표(clustering validity index)'를 사용함
    - 지도학습과 다름!
- 이질적인 모집단을 세분화시키기 위한 방법
- 밀도기반 기법(density-based methos) - DBSCAN
    - 어느 점을 기준으로 반경 x내에 점이 n개 이상 있으면 하나의 군집으로 인식하는 방식을 의미, 임의적 모양의 군집분석

```
틀린 것 찾기 (17회)
```

- 지도학습과 동일한 교차타당성.. 이라고 해서 틀림

- 이질적 모집단 세분화시키는 방법으로 '군집분석'을 찾는 문항 (16회)
- 군집분석 설명 중 옳은 것 찾는 문항 (16회)
- DBSCAN 찾는 문항 (21회)

25.계층적 군집(Hierarchical)

- 계층적 군집분석
  - 계층적 군집은 두 개체 간의 거리에 기반하므로 거리 측정에 대한 정의가 필요함
  - 유클리드, 맨해튼, 마할라노비스, 민코프스키 등
  - 이상치에 민감함
  - 사전에 군집 수 k를 설정할 필요가 없는 탐색적 모형
  - 병합적 방법에서 한 번 군집이 형성되면 군집에 속한 개체는 다른 군집으로 이동할 수 없음
- 계층적 군집 중 병합법(agglomerative)
  - 가까운 개체들끼리 묶어 감으로써 군집을 만들어 나가는 방법으로 우선 가장 가까운 2개의 개체를 묶어서 하나의 군집을 만들고 나머지 [N-2] 개의 개체는 각각 하나의 군집을 만듦
  - 이와 같은 방법으로 [N-1] 단계를 반복하면 결국 N개의 개체가 모두 묶여서 하나의 군집을 만들게 되는 군집 방법
- 최단연결법 (병합법 종류)
  - 두 군집 사이의 거리를 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최소값을 측정하며, 고립된 군집을 찾는데 중점을 둔 방식
- 와드연결법 (병합법 종류)
  - 계층적 군집내의 오차제곱합(within group sum of squares)에 기초하여 군집을 수행하는 군집 방법

병합법(agglomerative) - 주관식으로 출제됨 (20회)
와드연결법 설명 써놓고 와드연결법 찾기 (19회)
계층적 군집분석 찾기 (21회)
- 군집수 k를 설정할 필요 없음
와드연결법 주관식 (21회)
- 군집간의 거리에 따라 데이터를 연결하기 보다는 군집내 편차들의 제곱합에 근거를 두고 군집들을 병합

26.거리를 활용한 측도

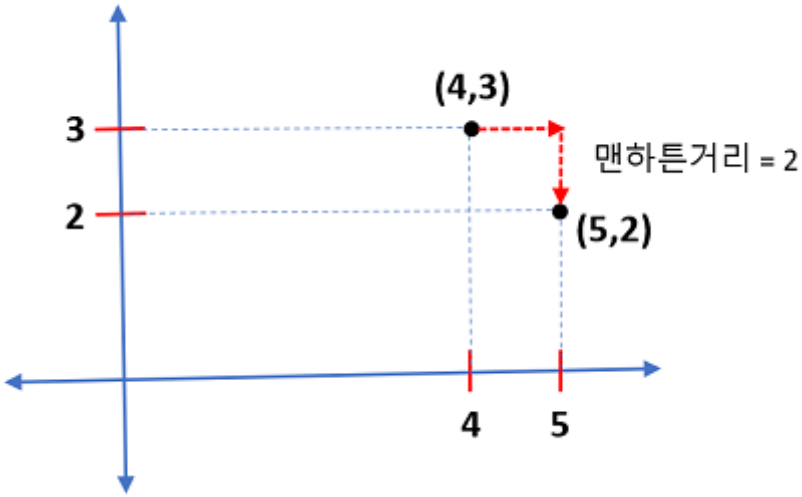
- 유클리드 : 두점 사이의 거리로 가장 직관적이고 일반적인 거리의 개념
- 맨해튼 거리 : 두 점의 좌표 간의 절대값 차이
- 마할라노비스 : 변수간의 상관성을 고려함 (18회)
- 표준화, 마할라노비스 거리는 통계적 거리의 개념

- dist() 함수에 의해 지원되는 거리 : 유클리드, 맨해튼, 민코프스키, maximum, canberra, binary 등이 있음

마할라노비스가 변수간 상관성을 고려하지 않는다 해서 틀림 (18회)
dist() 함수에 의해 지원되는 거리 아닌 것 찾기 (19회)

27.맨해튼 거리

- 맨해튼 거리 : x좌표 차이 + y좌표 차이
- A, B 두 점의 맨해튼 거리를 구하시오.
  - A(5, 2) B(4, 3)
  -



맨하튼 거리 : x좌표 차이 + y좌표 차이 (18회)

28.맨해튼 거리 계산

구분	a	b
Score	90	80
Time	60	75

맨해튼 거리는 차이의 절대값의 합이므로 10 + 15 = 25가 됨 (22회)

29.코사인 유사도

- 거리에 대한 개념으로 두 벡터 사이의 사잇각을 계산하여 유사한 정도를 구하는 것

코사인 유사도 찾기 (21회)

30.비계층적 군집(k- 평균 군집)

- 비계층적 군집분석 기법의 경우 사용자가 사전 지식 없이 그룹의 수를 정해주는 일이 많기 때문에 결과가 잘 나오지 않을 수 있음
- 잡음이나 이상값에 영향을 받기 쉽다
- 평균 대신 중앙값을 사용하는 k-medoids(중앙값) 군집을 사용할 수 있다.
- 또한, k-mean 분석 전에 이상값을 제거하는 것도 좋은 방법이다.
- k-mean 군집은 사전에 군집의 수를 정해 주어야 한다
- 만일 군집수 k가 원데이터 구조에 적합하지 않으면 좋은 결과를 얻을 수 없다.
- 따라서 Nbclust 패키지를 통해 군집의 수에 대한 정보를 참고해야 한다.

31.k-평균 군집 (k-Means) - 비계층적 군집

1. 알고리즘이 **단순**하며, 빠르게 수행되며 **계층적 군집보다 많은 양의 자료**를 다룰 수 있다
2. 이상값 자료에 민감한 k-평균 군집의 **단점을 보완**하기 위해 군집을 형성하는데 매단계 마다 평균 대신 중앙값을 사용하는 **k-중앙값 군집을 사용**한다
3. 초기값 선택이 최종 군집 선택에 영향을 미친다
4. 초기 군집수를 결정하기 어렵다
5. 각 군집내의 자료들이 평균을 계산하여 군집의 중심을 갱신한다
6. 한 개체가 속해있던 군집에서 **다른 군집으로 이동해 재배치가 가능**하다
7. 분석 절차 순서
    - a. 초기 군집중심으로 k개의 객체로 임의 선택
    - b. 각 자료를 가장 가까운 군집 중심에 할당
    - c. 각 군집내의 자료들의 평균을 계산하여 군집의 중심 갱신
    - d. 군집 중심의 변화가 없을 때까지 b, c 반복

옳은 것을 찾아라로 1번에 대한 것 찾기 (16회)
단점 해결 방법으로 2번 찾기 (16회)
절절하지 않은 것 찾기 6번 찾기 (17회)
k평균군집분석의 분석 절차 찾기 (20회, 22회)

32.실루엣(silhouette)

- 군집내 거리와 군집간의 거리를 기준으로 **군집 분할 성과를 측정하는 방식**
    - 군집화의 성능 평가!
- 클러스터 안의 데이터들이 다른 클러스터와 비교해서 얼마나 비슷한가를 나타내는 군집 평가
- 군집분석에서 중요한 지표로서, **거리가 가까울수록 높고 멀수록 낮은 지표**이자 완전히 분리된 경우 1이 되는 지표
    - 지표가 **0.5 보다 크면 군집결과가 타당한 것으로 평가**
    - 지표가 1에 가까울수록 군집화가 잘 되었다고 판단

주관식 문제 (16회, 23회)
객관식으로 단어 찾기 (21회)

33. 연관분석(Association analysis)

1. 품목 수가 증가하면 분석에 필요한 계산은 기하급수적으로 늘어난다
2. 너무 세분화된 품목을 가지고 연관규칙을 찾으려고 하면 의미 없는 분석 결과가 나올 수도 있다
    - 세분화 분석 품목은 필요함!
3. 향상도가 **1이면** 두 품목간에 연관성이 없는 서로 **독립적인** 관계이고, 1보다 작으면 음의 관계로 품목 간에 연관성이 없다
4. **시차 연관분석은 순서와 연관된** 분석이 가능하다
5. 교차판매/물건배치 등에 이용되는 분석 기법
6. 조건반응(if then)으로 표현되는 연관규칙의 결과를 이해하기 쉽다
7. 비목적성 분석 기법, 분석 계산이 간편하다
8. 고객의 과거 거래 구매 패턴을 분석하여 고객이 구매하지 않은 상품 추천
9. 상품을 구매할 때 유사한 상품을 구매한 고객들의 구매 데이터를 분석하여 쿠폰 발행
10. 기저귀를 사는 고객은 맥주를 동시에 구매한다는 연관규칙을 알아냄

4번 시차 연관분석을 인과관계라해서 틀린거 찾기로 출제 (16회)
5번을 적어주고 어떤 분석법인지 찾기 (18회)
2번을 세분화 분석 품목 없이 연관 규칙을 찾을 수 있다해서 틀림 (19회)
6번을 설명하고 '연관규칙' 찾기 (20회)
연관성 설명으로 틀린 것 (23회)
비지도 학습의 예 찾기 (23회) (8번 9번)

34. 연관규칙의 지지도, 향상도, 신뢰도

- 지지도(support) : A와 B가 동시에 포함된 거래수/전체 거래수

> A->B라고 하는 규칙이 전체 거래 중 차지하는 비율을 통해 해당 연관 규칙이 얼마나 의미가 있는 규칙인지 확인함 A,B의 교집합

- 향상도
    - 품목B를 구매한 고객 대비 품목 A를 구매한 후 품목 B를 구매하는 고객에 대한 확률을 의미

    1. 향상도가 1보다 크면 이 규칙은 결과를 예측하는 데 있어 우수하다는 것을 의미함
    2. 향상도가 1이면 두 품목은 독립적임
    3. 향상도가 1보다 작으면 두 품목은 서로 음의 상관관계를 의미하며 연관성이 없음
- 신뢰도 : 품목 A가 포함된 거래 중에서 품목 A, B를 동시에 포함하는 거래일 확률 (A를 구매한 사람이 B도 구매하는 확률)

> A, B의 조건부 확률 : P(B|A)

지지도 설명 찾기 (19회)
지지도 설명 보고 지지도 찾기 (23회)
향상도 이해 문제 출제됨 (16회)
향상도 설명이 틀린 것은 (18회)
연관규칙지표의 설명으로 옳은 것 찾기 (19회)
- 향상도 1보다 큰 것

35.지지도 구하기

거래 번호	판매 상품
1	소주, 콜라, 맥주
2	소주, 콜라, 와인
3	소주, 주스
4	콜라, 맥주
5	주스

콜라 -> 맥주의 지지도는? (x회, 21회)
= (콜라, 맥주가 함께 구입됨) / 전체 거래수
= 2 / 5 = 0.4

36.향상도 계산 - 1

- 아래의 거래 데이터에서 추출된 연관규칙 중 하나인 **"사과 --> 딸기"에 대한 향상도**는?

거래 번호	판매 상품
1	배, 사과, 딸기
2	배, 사과, 포도
3	배, 자몽
4	사과, 딸기
5	배, 사과, 딸기, 포도
6	자몽

- 향상도= P(A|B) / (P(A) * P(B))

향상도 구하기 주관식 (16회)
향상도 = (사과 딸기가 함께 판매되는 확률) / (사과 판매 확률) * (딸기 판매 확률)
= (3/6) / ((4/6) * (3/6))
= 1/3

37.향상도 계산 - 2

항목	거래수
딸기	100
사과	100
배	50
[딸기, 사과]	500
[딸기, 배]	300
[사과, 배]	200
[딸기,사과, 배]	100
전체 거래건수	1450

향상도 구하기 객관식 (19회)
- 딸기, 사과를 동시 구입할 확률 / ((딸기 구입확률 * 사과 구입확률))
= (600/1450) / ((1000/1450) * (900/1450))
= 0.96

38.신뢰도 구하기

장바구니	item
1	빵, 맥주, 우유
2	빵, 우유, 계란
3	맥주, 우유
4	빵, 맥주, 계란
5	빵, 맥주, 우유, 계란

빵 -> 우유의 신뢰도는?
= (빵, 우유가 함께 구입됨) / 빵 구매
= 3 / 4 = 0.75

39.Apriori 알고리즘

- 연관규칙의 대표적인 알고리즘으로 현재도 많이 사용됨
- 기본 개념은 데이터들에 대한 발생 빈도를 기반으로 각 게이터 간의 연관관계를 밝히기 위한 방법임

40.FP-Growth

- 연관분석의 대표적 알고리즘 apriori 단점을 보완하기 위해 트리와 노드링크라는 특별한 자료구조를 사용하는 알고리즘

FP-Growth를 찾는 문제 (17회)

41.SOM (Self-Organizing Maps, 자기조직화지도)

- 차원축소와 군집화를 동시에 수행하는 기법
- 입력 벡터를 훈련집합에서 match 되도록 가중치가 조정되는 인공신경세포 격자에 기초한 Unsupervised Learning의 방법
- 주요 기능 중 데이터의 특징을 파악하여 유사 데이터를 클러스터링 함
- 대표적 **비지도학습**
- 한개의 입력과 한개의 출력층
- 입력층과 출력층이 완전 연결
- 출력 뉴런들은 승자 뉴런이 되기 위해 경쟁하고 오직 승자만 학습한다

SOM의 설명으로 틀린 것 (18회)
- 역전파 알고리즘 사용한다해서 틀림 (역전파 알고리즘은 AI의 인공신경망꺼임)

42.SOM Process

1. SOM 맵의 노드에 대한 연결강도로 초기화한다
2. 입력 벡터와 경쟁층 노드 간의 유크리드 거리를 계산하여 입력벡터와 가장 짧은 노드를 선택한다
3. 선택된 노드와 이웃 노드의 가중치(연결강도)를 수정한다.
4. 2번 단계로 가서 반복하면서 연결강도는 입력 패턴과 가장 유사한 경쟁층 뉴런이 승자가 된다. 결국 승자 독식 구조로 인해 경쟁층에는 승자 뉴런만이 나타난다.

- 신경망은 역전파 알고리즘이지만, SOM은 전방패스를 사용하여 속도가 매우 빠르다.

43.BMU (Best Matching Unit)

- SOM Process에서 입력 벡터와 경쟁층 노드 간의 유크리드 거리를 계산하여 그 중에서 제일 가까운 Neuron을 무엇이라 하는가?

주관식 문제로 출제 (17회)

44.SOM vs 신경망 모형

1. 신경망 모형은 연속적인 layer로 구성, SOM은 2차원 그리드(격자) 구성 (입력층, 경쟁층)
2. 신경망 모형은 에러를 수정하는 학습, SOM은 경쟁 학습 실시
3. SOM은 비지도학습

45.순차 패턴 분석

- 연관규칙 분석과 유사한 아이디어에서 출발하지만 시간 또는 순서에 따른 사건의 규칙을 찾는다는 점에서 다른 분석

순차 패턴 분석의 의미를 적어 주고 순차 패턴 분석 찾기 (22회)