



# Day37; 20221028

📅 날짜	@2022년 10월 28일
👤 유형	@2022년 10월 28일
☰ 태그	

**GitHub - u8yes/Python**

You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

<https://github.com/u8yes/Python>

**u8yes/Python**



1 Contributor 0 Issues 1 Star 0 Forks

ADsP)

- 승산비(odds ratio) = 관심있는 사건이 발생할 상대 비율,  $x=1$ 일 때,  $y=1$ 이 되는 상대적 비율
  - $\text{odds\_a} / \text{odds\_b} = \exp(\text{coef}) = \exp(5.140336) = 170.7731385$
  - 로지스틱 회귀에서  $\exp(x_1)$ 의 의미 (단,  $x_1$  : 회귀계수)
  - 나머지 변수가 주어질 때  $x_1$ 이 한 단위 증가할 때마다 성공( $Y=1$ )의 odds가 몇 배 증가하는지를 나타냄

```

2 data(iris)
3 a <- subset(iris, Species=='setosa' | Species=='versicolor')
4 a$Species<-factor(a$Species)
5 b <- glm(Species~Sepal.Length, data=a, family = binomial)

> coef(b)
(Intercept) Sepal.Length
-27.831451      5.140336
> exp(coef(b))['Sepal.Length']
Sepal.Length
170.7732
      ▪ Y=1은 versicolor 일 경우, Sepal.Length가 한 단위
      증가하면 Versicolor일 odds가 170배 증가를 의미함
  
```

- ☞ 1. 로지스틱 회귀모형에서  $\exp(x_1)$ 의 의미는 나머지 변수가 주어질 때  $x_1$ 이 한 단위 증가할 때마다 성공( $Y=1$ )의 ( ) 가 몇 배 증가하는지를 나타낸다

오즈(odds)

- ☞ 2. 종속변수가 성공 또는 실패인 이항변수로 되어 있을 때 종속변수와 독립변수 간의 관계식을 이용하여 두 집단 또는 그 이상의 집단을 분류하고자 할 때 사용되는 분석기법을 무엇이라 하는가?

- 1 로지스틱 회귀분석    2 다중 회귀분석    3 의사결정나무    4 양상블 모형

로지스틱 회귀분석: 종속변수가 성공 또는 실패인 이항변수로 되어 있을 때 종속변수와 독립변수 간의 관계식을 이용하여 두 집단 또는 그 이상의 집단을 분류하고자 할 때 사용되는 분석기법

### 3-82. 의사 결정 나무(Decision Tree) 모형

특징

- 목적은 새로운 데이터를 분류하거나 값을 예측하는 것이다
- 분리 변수 P차원 공간에 대한 현재 분할은 이전 분할에 영향을 받는다
- 부모마디보다 자식마디의 순수도가 증가하도록 분류나무를 형성해 나간다 (불순도 감소)

독립변수	종속변수
<ul style="list-style-type: none"> <li>▪ 설명변수</li> <li>▪ 예측변수</li> <li>▪ Feature</li> </ul>	<ul style="list-style-type: none"> <li>▪ 목표변수</li> <li>▪ 반응변수</li> <li>▪ Label</li> </ul>

목표변수(= 종속변수)

## 3-82. 불순도 측정 지표

### ▣ 목표변수가 범주 형일 때 사용하는 지표 (분류에서 사용)

지니 지수	<ul style="list-style-type: none"> <li>불순도 측정 지표, 값이 작을수록 순수도가 높음(분류 잘됨)</li> <li>가장 작은 값을 갖는 예측 변수와 이때의 최적 분리에 의해 자식 마디 형성</li> <li><math>Gini(T) = 1 - \sum_{i=1}^k (각 범주별수/전체수)^2 = 1 - \sum_{i=1}^k P_i^2</math></li> </ul> <p style="text-align: center;"><input checked="" type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/></p>	
엔트로피 지수 Entropy measure	<ul style="list-style-type: none"> <li>불순도 측정 지표, 가장 작은 값을 갖는 방법 선택</li> <li><math>Entropy(T) = - \sum_{i=1}^k P_i \log_2 P_i</math></li> </ul>	
카이제곱 통계량의 유의 확률(p-value)	<ul style="list-style-type: none"> <li>가장 작은 값을 갖는 방법 선택</li> </ul>	
	$1 - ((3/4)^2 + (1/4)^2) = 1 - 5/8 = 3/8$	$1 - ((2/4)^2 + (2/4)^2) = 1 - 4/8 = 4/8$
원본 데이터	색으로 구분	모양으로 구분

### ▣ 4. 다음 중 의사결정나무(Decision Tree)에 대한 설명 중 틀린 것은?

- 1 정지규칙이란 더 이상 분리가 일어나지 않고 현재의 마디가 최종마디가 되도록 하는 여러 가지 규칙으로 지니 지수, 엔트로피 지수, 카이제곱통계량 등이 있다
- 2 최종마디가 너무 많으면 모형이 과대적합된 상태로 현실 문제에 적용할 수 있는 적절한 규칙이 나오지 않게 되며, 이를 해결하기 위해 가지치기를 한다.
- 3 의사결정나무를 위한 알고리즘은 CHAID, CART, ID2, C5.0, C4.5가 있으며 상향식 접근 방법을 이용한다
- 4 의사결정나무는 목표변수가 이산형인 경우의 분류나무(classification tree)와 목표변수가 연속형인 경우의 회귀나무(regression tree)로 구분된다.

- CHAID, CART, ID2, C5.0, C4.5 등은 하향식 접근 방법을 이용한다

### ▣ 10. 의사결정나무에서 이산형 목표변수는 지니지수, 연속형 목표변수는 분산 감소량을 사용하는 알고리즘은 무엇인가?

1 CHAID

2 CART

3 C4.5

4 C5.0

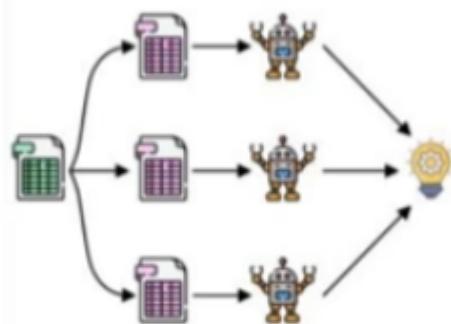
알고리즘	이산형 목표변수	연속형 목표변수
CART	지니지수	분산 감소량
C5.0	엔트로피지수	
CHAID	Pearson의 카이제곱 통계량	ANOVA F-통계량 p-값

☞ 11. 목표변수가 연속형인 경우 회귀나무의 경우 사용하는 분류기준은 무엇인가?

- 1 카이제곱통계량, 지니지수
- 2 지니지수, 엔트로피지수
- 3 엔트로피지수, 분산감소량
- 4 분산감소량, F-통계량의 p-값

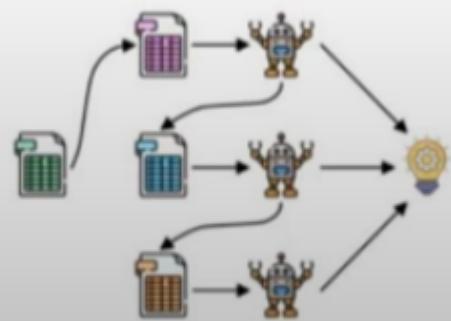
알고리즘	이산형 목표변수	연속형 목표변수
CART	지니지수	분산 감소량
C5.0	엔트로피지수	
CHAID	Pearson의 카이제곱 통계량	ANOVA F-통계량 p-값

## Bagging



Parallel

## Boosting



Sequential

☞ 9. 부스팅(boosting) 알고리즘 중 Leaf-wise-node 방법을 사용하는 알고리즘을 무엇이라 하는가?

1 AdaBoost

2 GBM

3 Xgboost

4 Light GBM

▪ Light GBM : Boosting 알고리즘, Leaf-wise-node 방법을 사용

☞ 1. lazy learning(게으른 학습)이 사용되는 지도학습 알고리즘을 무엇이라 하는가?

1 kNN

2 SVM

3 로지스틱회귀

4 의사결정나무

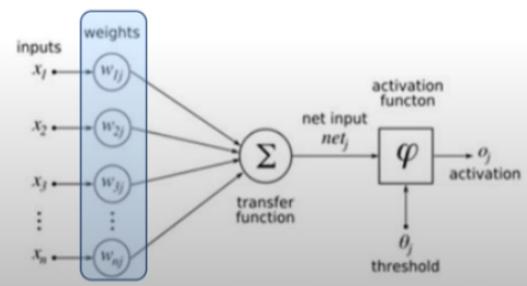
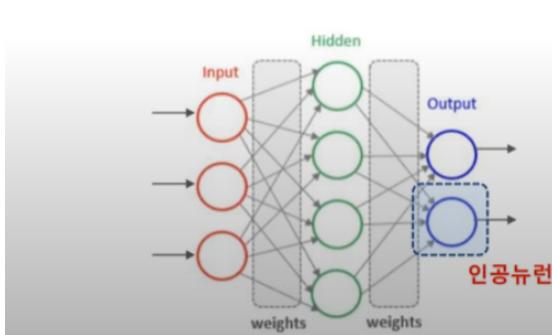
- 새로운 데이터에 대해 주어진 이웃의 개수( $k$ ) 만큼 가까운 멤버들과 비교하여 결과를 판단하는 방법
- 모형을 미리 만들지 않고, 새로운 데이터가 들어오면 그때부터 계산을 시작하는 lazy learning(게으른 학습)이 사용되는 지도학습 알고리즘

## 3-86. 인공 신경망(ANN) 모형

19

2

- 인공신경망을 이용하면 분류 및 군집을 할 수 있음
- 인공신경망은 입력층, 은닉층, 출력층 3개의 층으로 구성되어 있음
- 각 층에 뉴런(노드)이 여러 개 포함되어 있음
- 학습: 입력에 대한 올바른 출력이 나오도록 가중치(weights)를 조절하는 것
- 가중치 초기화는 -1.0~1.0 사이의 임의 값으로 설정하며, 가중치를 지나치게 큰 값으로 초기화하면 활성화 함수를 편향 시키게 되며, 활성화 함수가 과적합 되는 상태를 포화상태라고 함



인공 뉴런 (=퍼셉트론)

### 3-86. 경사하강법(Gradient descent)

- 함수 기울기를 낮은 쪽으로 계속 이동시켜 극값에 이를 때까지 반복시키는 것
- 제시된 함수의 기울기의 최소값을 찾아내는 머신러닝 알고리즘
- 비용함수(cost function)을 최소화 하기 위해 parameter를 반복적으로 조정하는 과정

#### ▣ 경사하강법 과정

1. 임의의 parameter값으로 시작
2. Cost Function 계산, cost function - 모델을 구성하는 가중치 w의 함수, 시작점에서 곡선의 기울기 계산
3. parameter 값 갱신 :  $W = W - \text{learning rate} * \text{기울기미분값}$
4. n 번의 iteration, 최소값을 향해 수렴함, learning rate가 적절해야 함

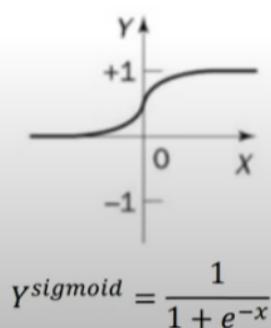


#### 신경망 모형의 장점

- 변수의 수가 많거나 입, 출력변수 간에 복잡한 비선형 관계에 유용
- 이상치 잡음에 대해서도 민감하게 반응하지 않음
- 입력변수와 결과변수가 연속형이나 이산형인 경우 모두 처리 가능

#### softmax 함수

- 모든 logits의 합이 1이 되도록 output을 정규화
- sigmoid 함수의 일반화된 형태로 결과가 다 범주인 경우 각 범주에 속할 사후 확률을(posterior probability) 제공하는 활성화
- 주로 3개 이상 분류 시 사용함



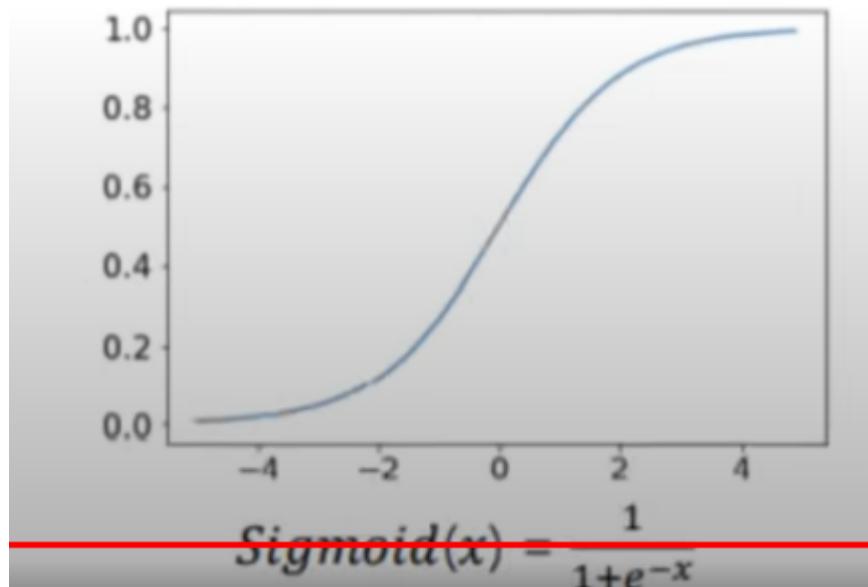
$$\begin{bmatrix} 1.2 \\ 0.9 \\ 0.4 \end{bmatrix} \xrightarrow{\text{Softmax}} \begin{bmatrix} 0.46 \\ 0.34 \\ 0.20 \end{bmatrix}$$
$$\text{softmax}(x)_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

☞ 1. 인공신경망의 특징으로 부적절한 것은?

- 1 분석가의 주관과 경험에 따른다.
  - 2 입력변수의 속성에 따라 활성화 함수의 선택이 달라진다.
  - 3 역전파 알고리즘이 동일 입력층에 대해 원하는 값이 출력되도록 개개의 weight를 조정하는 방법으로 사용된다.
  - 4 이상치 잡음에 민감하지 않다.
- ↳
- 풀고자 하는 문제 종류에 따라 활성화 함수의 선택이 달라진다

## sigmoid 함수

- 연속형 0~1, Logistic 함수라 불리기도
- 선형적인 Multi-Perceptron에서 비선



5. 신경망에서 가중치 초기화는 -1.0~1.0 사이의 임의 값으로 설정하지만, 가중치를 지나치게 큰 값으로 초기화하면 활성화 함수를 편향시키게 되며 활성화 함수가 과적합 되는 상태를 무엇이라 하는가?

1 포화상태

역전파 알고리즘 Backpropagation Algorithm	<ul style="list-style-type: none"><li>▪ 출력층에서 제시한 값에 대해, 실제 원하는 값으로 학습하는 방법으로 사용</li><li>▪ 동일 입력층에 대해 원하는 값이 출력되도록 개개의 weight를 조정하는 방법으로 사용됨</li></ul>
기울기 소실 문제 Vanishing Gradient Problem	다층신경망에서 <b>은닉층이 많아</b> 인공신경망 기울기 값을 베이스로 하는 역전파 알고리즘으로 학습시키려고 할 때 발생하는 문제

1. 다층 신경망에서 은닉층이 많아 인공신경망 기울기 값을 베이스로 하는 역전파 알고리즘으로 학습시키려고 할 때 발생하는 어려움을 무엇이라 하는가?

1 기울기 소실문제(Vanishing Gradient Problem)

1. 다음 중 신용카드 고객 파산여부를 예측하는 모형이 아닌 것은?

1 로지스틱회귀분석

2 선형회귀분석

3 의사결정나무

4 앙상블 모형

- 신용카드 고객 파산 여부는 범주형 데이터로 범주에 대한 '분류'에 해당하며, 선형 회귀 분석은 연속형 데이터를 대상으로 한다

3. 븁스트랩 방식을 이용하였을 때 일반적인 훈련 데이터의 양은?

1 63.20%

2 10.20%

3 23.80%

4 36.80%

- 전체 데이터 양이 크지 않을 경우의 모형 평가에 가장 적합
- 훈련 데이터를 63.2% 사용하는 0.632 븁스트랩이 있음

6. 분류모형에서 일부 범주형의 관측치가 현저히 부족하여 모형이 학습하기 힘든 문제를 무엇이라 하는가?

1 클래스 불균형

2 집중 샘플링

3 ROSE 샘플링

4 차원 축소

- class의 비율이 한쪽에 치우쳐 있는 클래스 불균형 상태라면 다음 기법 사용을 고려한다
  - under sampling : 적은 class의 수에 맞추는 것
  - over sampling : 많은 class의 수에 맞추는 것

## 3-91. 오분류표를 활용한 평가 지표

Confusion matrix		예측값		Sensitivity TP / (TP+FN)	실Sen, 예Pre																						
		TRUE	FALSE																								
실제값	TRUE	40 (TP)	60 (FN) Type II Error	Sensitivity TP / (TP+FN)	실Sen, 예Pre																						
	FALSE	60 (FP) Type I Error	40 (TN)																								
		Precision TP / (TP+FP)	Negative Predictive Value TN / (TN + FN)	Accuracy (TP+TN) / (TP+TN+FP+FN)																							
<table border="1"> <tr> <td>T/F</td> <td>P/N</td> </tr> <tr> <td>실제 == 예측 : True 실제 != 예측 : False</td> <td>True 예측 : Positive False 예측 : Negative</td> </tr> <tr> <td>T P</td> <td>T N</td> </tr> <tr> <td>F P</td> <td>F N</td> </tr> </table>		T/F	P/N	실제 == 예측 : True 실제 != 예측 : False	True 예측 : Positive False 예측 : Negative	T P	T N	F P	F N	<table border="1"> <tr> <th colspan="2">confusion matrix</th> <th colspan="2">실제값</th> </tr> <tr> <th colspan="2"></th> <th>Y</th> <th>N</th> </tr> <tr> <th rowspan="2">예측값</th> <th>Y</th> <td>True Positive</td> <td>False Positive</td> </tr> <tr> <th>N</th> <td>False Negative</td> <td>True Negative</td> </tr> </table>			confusion matrix		실제값				Y	N	예측값	Y	True Positive	False Positive	N	False Negative	True Negative
T/F	P/N																										
실제 == 예측 : True 실제 != 예측 : False	True 예측 : Positive False 예측 : Negative																										
T P	T N																										
F P	F N																										
confusion matrix		실제값																									
		Y	N																								
예측값	Y	True Positive	False Positive																								
	N	False Negative	True Negative																								

정밀도 Precision	<ul style="list-style-type: none"> <li>예측값이 True인 것에 대해 실제값이 True인 지표</li> <li>식 : <math>TP / (TP + FP)</math></li> </ul>	실Sen, 예Pre
재현율, 민감도 Recall, Sensitivity	<ul style="list-style-type: none"> <li>실제값이 True인 것에 대해 예측값이 True인 지표</li> <li>식 : <math>TP / (TP + FN)</math></li> </ul>	
F1	<ul style="list-style-type: none"> <li>F1은 데이터가 불균형 할 때 사용한다</li> <li>오분류표 중 정밀도와 재현율의 조화평균을 나타내며 정밀도와 재현율에 같은 가중치를 부여하여 평균한 지표</li> <li><math>2 * (Precision * Recall) / (Precision + Recall)</math></li> </ul>	

### ■ 재현률(Recall, Sensitivity):

9. 두 평가자의 평가가 얼마나 일치하는지 평가하는 값으로 0~1 사이의 값을 가진다. P(e)는 두 평가자의 평가가 우연히 일치할 확률을 뜻하는 모델 평가 메트릭을 무엇이라 하는가?

kappa

↳

- $(\text{Accuracy} - P(e)) / (1 - P(e))$
  - 두 평가자의 평가가 얼마나 일치하는지 평가하는 값으로 0~1 사이의 값을 가짐

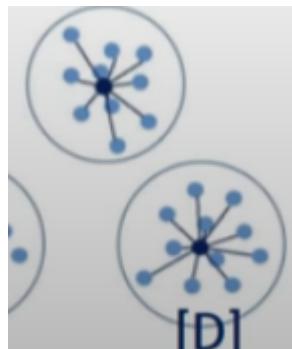
4

재현율에 정확도에 2배만큼의 가중치를 부여

- F2 : 재현율에 정밀도의 2배 만큼 가중치를 부여하는 것

## 와드 연결법

- 계층적 군집내의 오차제곱합에 기초하여 군집을 수행하는 군집 방법
- 크기가 비슷한 군집끼리 병합하는 경향이 있음



## 와드 연결법

누적의 개념으로 사용하는 것이 count += 1 (보통 입금할 때나 출금할 때 가진 것에서 누적, 인출해서 관리하기 위해 보는 개념)

## 마할라노비스

- 변수의 표준화와 함께 변수 간의 상관성을 동시에 고려한 통계적 거리

### dist 함수

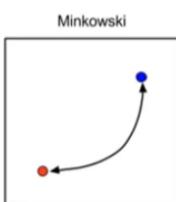
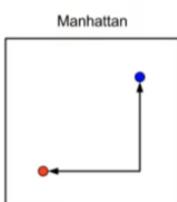
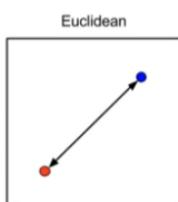
- 거리측정에 사용하는 함수로 사용가능한 거리 개념으로 유클리드, 맨해튼, 민코프스키, maximum, canberra, binary 등이 있음

### 코사인(cosine)거리

- 두 벡터 사이의 사잇각을 계산해서 유사한 정도를 구하는 것
- 값이 1인 경우 유사도가 크며, -1인 경우 유사도가 매우 작음을 의미함

## 3-94. 계층적 군집의 거리

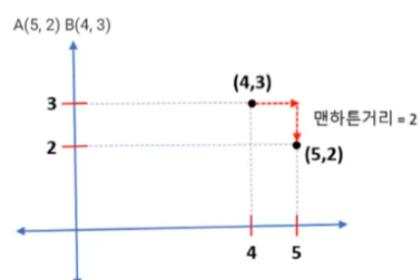
### ▣ 수학적 거리개념의 종류



[https://subscription.packtpub.com/book/big\\_data\\_and\\_business\\_intelligence/9781785882104/6/ch06lvl1sec40/measuring-distance-or-similarity](https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781785882104/6/ch06lvl1sec40/measuring-distance-or-similarity)

$$\text{Euclidean, } d(I, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}$$

$$\text{Manhattan, } d(I, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$



비계층적 군집 - 분할적 군집 방법	k-중심 군집
<p><b>k-means</b></p> <p>Nbclust 패키지를 통해 군집 수에 대한 정보 참고</p>	<ul style="list-style-type: none"> <li>▪ k-mean 방법은 사전에 군집의 수 <math>k</math>를 정해 주어야 함 (<math>k</math> : hyperparameter)</li> <li>▪ 군집수 <math>k</math>가 원데이터 구조에 적합하지 않으면 좋은 결과를 얻을 수 없음</li> <li>▪ 알고리즘이 단순하며 빠르게 수행되어 계층적 군집보다 많은 양의 자료를 처리</li> <li>▪ k-means 군집은 잡음이나 이상값에 영향을 받기 쉬움</li> <li>▪ k-means 분석 전에 이상값을 제거하는 것도 좋은 방법</li> <li>▪ 평균 대신 중앙값을 사용하는 k-medoids 군집을 사용할 수 있음</li> </ul>
<p><b>k-means 절차</b></p> <ol style="list-style-type: none"> <li>1. 초기 군집의 중심으로 <math>k</math>개의 객체를 임의로 선택한다</li> <li>2. 각 자료를 가장 가까운 군집의 중심에 할당한다</li> <li>3. 각 군집 내의 자료들의 평균을 계산하여 군집의 중심을 갱신한다</li> <li>4. 군집 중심의 변화가 거의 없을 때까지 2, 3을 반복한다</li> </ol>	

1. K-평균 군집화와 달리 군집 수  $k$ 를 설정할 필요가 없는 탐색적 모형을 무엇이라 하는가?

- |   |  |
|---|--|
| <span style="border: 1px solid #ccc; border-radius: 50%; padding: 5px;">1</span> 계층적 군집분석 | <span style="border: 1px solid #ccc; border-radius: 50%; padding: 5px;">2</span> 비계층적 군집분석 |
| <span style="border: 1px solid #ccc; border-radius: 50%; padding: 5px;">3</span> 판별분석     | <span style="border: 1px solid #ccc; border-radius: 50%; padding: 5px;">4</span> 밀도기반 군집분석 |

- I
- 계층적 군집은 두 개체 간의 거리에 기반하므로 거리 측정에 대한 정의가 필요함
  - 사전에 군집 수  $k$ 를 설정할 필요가 없는 탐색적 모형
  - 병합적 방법에서 한 번 군집이 형성되면 군집에 속한 개체는 다른 군집으로 이동할 수 없음

7. 어느 점을 기준으로 반경  $x$ 내에 점이  $n$ 개 이상 있으면 하나의 군집으로 인식하는 방식을 의미하며, 임의적 모양의 군집분석을 무엇이라 하는가?

- |  |   |
|--|---|
| <span style="border: 1px solid #ccc; border-radius: 50%; padding: 5px;">1</span> kNN     | <span style="border: 1px solid #ccc; border-radius: 50%; padding: 5px;">2</span> PAM    |
| <span style="border: 1px solid #ccc; border-radius: 50%; padding: 5px;">3</span> K-means | <span style="border: 1px solid #ccc; border-radius: 50%; padding: 5px;">4</span> DBSCAN |

- DBSCAN
- I
- 밀도 기반 클러스터링으로 점이 세밀하게 몰려 있어 밀도가 높은 부분을 클러스터링 함
  - 어느 점을 기준으로 반경 내에 점이  $n$ 개 이상 있으면 하나의 군집으로 인식하는 방식
  - Gaussian 분포가 아닌 임의적 모양의 군집분석에 적합함

## 8. 아래 데이터 세트(dataset)에서 a, b 간의 맨해튼 거리는 얼마인가?

구분	a	b
Score	90	80
Time	60	75

1 25

2 20

3 15

4 10

- 맨해튼 거리는 절대값의 합으로 구한다
- $|90 - 80| + |60 - 75| = 25$

## 9. 거리를 활용한 측도에 대한 설명으로 틀린 것은?

- 1 유클리드는 두 점 사이의 거리로, 가장 직관적이고 일반적인 거리의 개념이다
- 2 맨해튼 거리는 두 점의 좌표 간의 절대값 차이를 구하는 것이다
- 3 마할라노비스는 변수의 표준화를 고려하고, 변수 간의 상관성을 고려하지 않는다
- 4 표준화, 마할라노비스 거리는 통계적 거리의 개념이다

### 마할라노비스

- 변수의 표준화와 함께 변수 간의 상관성을 동시에 고려한 통계적 거리

민코프스키 : 거리 차수를 사용하며 거리 차수가 1인 경우 맨해튼, 2인 경우 유클리드 거리로 사용됨  
표준화 거리 : 각 변수를 해당 변수의 표준편차로 척도 변환한 후에 유클리드 거리를 계산한 것으로  
통계적 거리(Statistical distance)라고도 함

## 3-93-96. 군집분석(Clustering Analysis) - 문제11



11. 계층적 군집은 두 개체 간의 거리에 기반하므로 거리측정에 대한 정의가 필요하다. 다음 중 dist() 함수에서 지원하지 않은 거리는?

- 1 유클리드
- 2 맨하튼
- 3 민코프스키
- 4 cosine

▪ dist는 유클리드, 맨해튼, 민코프스키, maximum, canberra, binary 등의 거리 개념을 지원함

### 13. k 평균 군집에 대한 설명 중 적절하지 않은 것은?

- 1 초기값 선택이 최종 군집 선택에 영향을 미친다.
- 2 초기 군집수를 결정하기 어렵다.
- 3 한 개체가 속해 있던 군집에서 다른 군집으로 이동해 재배치가 가능하지 않다.
- 4 각 군집내의 자료들의 평균을 계산하여 군집의 중심을 갱신한다.

▪ 한 개체가 속해 있던 군집에서 다른 군집으로 이동해 재배치가 가능하다

### 14. 다음 군집분석(Cluster analysis) 관한 설명 중 올바르지 않은 것은?

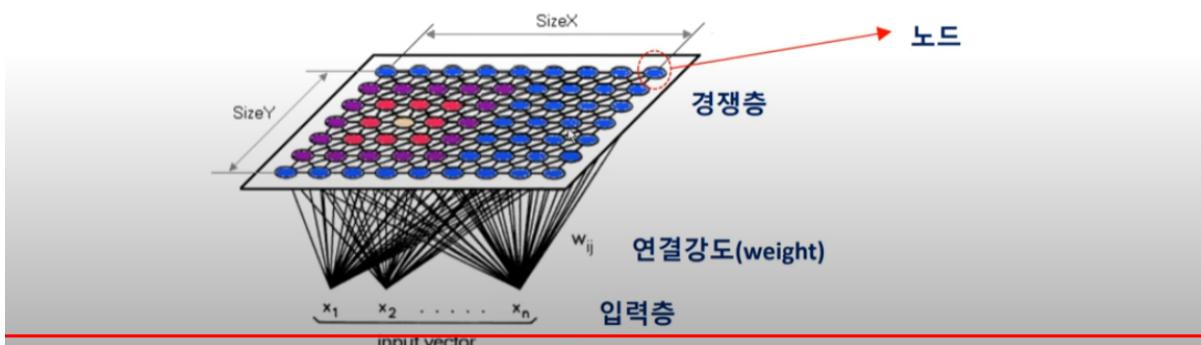
- 1 비계층적 군집분석 기법의 경우 사용자가 사전 지식 없이 그룹의 수를 정해주는 일이 많기 때문에 결과가 잘 나오지 않을 수 있다.
- 2 군집분석은 신뢰성과 타당성을 점검하기 어렵다
- 3 군집 결과에 대한 안정성을 검토하는 방법으로 지도학습과 동일한 교차타당성을 이용한다.
- 4 계층적 군집분석은 이상치에 민감하다.

▪ 군집 결과에 대한 안정성 검토는 실루엣, Dunn Index를 사용함

## 3-97. SOM(Self-Organizing Maps)

### SOM 이란?

- 자기조직화지도
- 인공신경망의 한 종류로, 차원축소와 군집화를 동시에 수행하는 기법
- 비지도 학습(Unsupervised Learning)의 한 가지 방법
- 고차원으로 표현된 데이터를 저차원으로 변환해서 보는데 유용함
- 입력층과 2차원의 격자 형태의 경쟁층(= 출력층)으로 이루어져 있음(2개의 층으로 구성)



### 1. SOM에 대한 설명으로 가장 적절한 것은?

- 1 지도학습이다
- 2 인공신경망과 같은 역전파 알고리즘을 이용한다.
- 3 다수의 입력층과 다수의 출력층으로 구성이 되어 있다.
- 4 출력 뉴런들은 승자 뉴런이 되기 위해 경쟁하고 오직 승자만이 학습한다.

- SOM은 비지도 학습(Unsupervised Learning)
- 신경망은 역전파 알고리즘이지만, SOM은 전방패스를 사용해 속도가 매우 빠름

### 3. 기법 활용 분야가 다른 것은?

1 SOM

2 로지스틱 회귀분석

3 신경망

4 의사결정 나무

- SOM은 군집분석
- 로지스틱 회귀분석, 신경망, 의사결정 나무는 분류 분석에 해당한다

### 3-98. 연관분석(Association Analysis)

#### 연관분석

- 연관규칙(Association rule) : 항목들 간의 '조건-결과' 식으로 표현되는 유용한 패턴
- 이러한 패턴 규칙을 발견해내는 것을 연관분석이라 함
- 장바구니 분석이라고 함(미국 마트에서 기저귀를 사는 고객은 맥주를 동시에 구매한다는 연관규칙을 알아낸 것에 기인함)

#### Apriori 알고리즘

- 연관규칙의 대표적 알고리즘으로 현재도 많이 사용됨
- 데이터들에 대한 **발생 빈도를 기반**으로 각 데이터 간의 연관관계를 밝히는 방법
- 데이터셋이 큰 경우 모든 후보 itemset에 대해 하나하나 검사하는 것이 비효율적임

#### FP Growth

- Apriori 단점을 보완하기 위해 FP-tree와 node, link라는 특별한 자료 구조를 사용

#### 장점

- **조건반응(if-then)**으로 표현되는 연관 분석의 결과를 이해하기 쉬움
- 강력한 비목적성 분석 기법이며, 분석 계산이 간편함

#### 단점

- 분석 품목 수가 증가하면 분석 계산이 **기하급수적으로** 증가함
- 너무 세분화된 품목을 가지고 연관규칙을 찾으려면 의미 없는 분석 결과가 도출됨
- 상대적 거래량이 적으면 규칙 발견 시 제외되기 쉬움

규칙표기 :  $A \rightarrow B$

- if A then B  $\Rightarrow$  A가 팔리면 B가 같이 팔린다

지지도  
Support

- 전체 거래항목 중 상품 A와 상품 B를 동시에 포함하여 거래하는 비율
- 전체 거래 중 차지하는 비율을 통해 해당 연관 규칙이 얼마나 의미가 있는 것인지를 확인함
- 지지도 =  $P(A \cap B)$  : A와 B가 동시에 포함된 거래 수 / 전체 거래 수

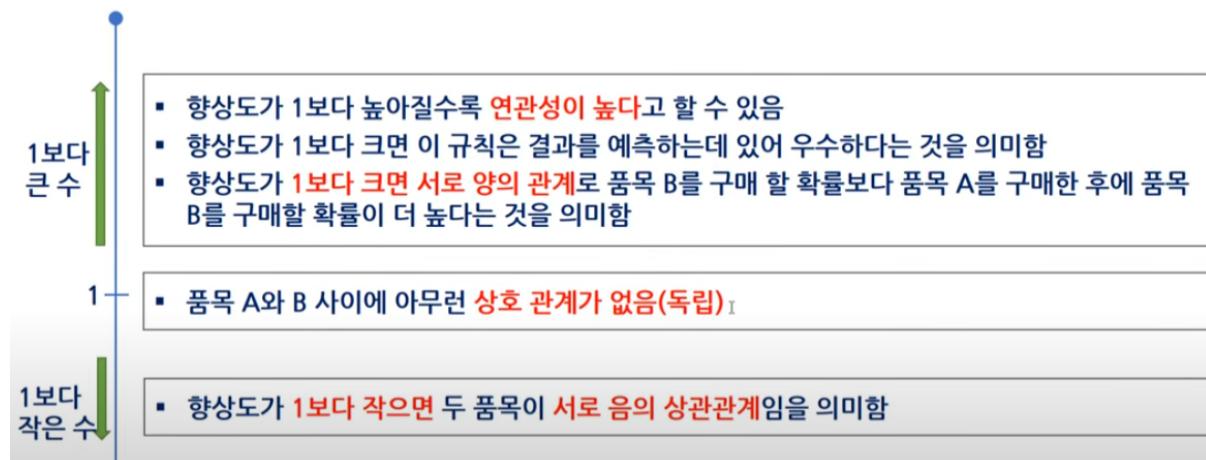
신뢰도  
Confidence

- 상품 A를 포함하는 거래 중 A와 B가 동시에 거래되는 비율
- 상품 A를 구매했을 때 상품 B를 구매할 확률이 어느 정도 되는지를 확인
- 신뢰도 =  $P(B|A) = P(A \cap B) / P(A)$  : A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수

향상도  
Lift

- A가 주어지지 않았을 때 B의 확률 대비 A가 주어졌을 때 B의 확률 증가 비율
- 품목B를 구매한 고객 대비 품목 A를 구매한 후 품목 B를 구매하는 고객에 대한 확률
- 향상도 =  $P(B|A)/P(B) = P(A \cap B) / (P(A)*P(B))$
- = 상품 A의 거래 중 상품 B가 포함된 거래의 비율 / 전체 상품 거래 중 상품 B가 거래된 비율
- = A와 B가 동시에 일어난 확률 / A, B가 독립된 사건일 때 A, B가 동시에 일어날 확률

### ▣ 향상도 해석



자바스크립트 함수: 이름이 없는 익명함수, 이름이 있는 선언적 함수가 있었다.

익명함수는 변수(참조변수)를 담아서 가리킬 수 있었다.

R 함수도 자바스크립트처럼 function()

8비트는 3개씩 쪼개서, 16비트는 4개씩 쪼갬.

매개변수(parameter) vs 인자(argument)

Script snippet #1 ×

```

1 function sum(number1, number2) {
2     return number1 + number2; 매개변수(Parameter)
3 }
4
5 sum(4, 3); 인자(Argument)

```

매개변수(Parameter)와 인자(Argument)의 정의

Oracle 공식 홈페이지에서는 매개변수와 인자를 다음과 같이 정의한 내용을 확인할 수 있습니다.

---

**Note:** Parameters refers to the list of variables in a method declaration. Arguments are the actual values that are passed in when the method is invoked. When you invoke a method, the arguments used must match the declaration's parameters in type and order.

---

Parameter and Argument 정의 - 오라클 공식 홈페이지

위의 내용을 해석하면 다음과 같습니다. "매개변수는 메서드 선언의 변수 목록을 나타냅니다. 인수는 메서드가 호출될 때 전달되는 실제 값입니다." 모든 프로그래밍 언어에서 이와 같이 매개변수(Parameter)와 인자(Argument)가 정의된다고 단정 지을 수는 없지만 프로그래밍 전반에 있어서는 큰 무리 없이 사용할 수 있는 정의라고 생각합니다.

매개변수와 인자는 흔히 쓰이는 단어이지만 정확한 구분 없이 사용하는 경우가 생각보다 많습니다. 사소한 부분이지만 아는 만큼 보인다고 프로그래밍을 할 때 매개변수와 인자에 대한 차이를 인식한다면 더 재미있게 개발을 할 수 있습니다.

## ▶ 2. 조건-결과(if-then) 유형의 패턴을 발견하는데 사용하는 데이터마이닝 기법은?

1 SOM

2 연관규칙

3 다차원척도

4 의사결정나무

- 연관규칙(Association rule) : 항목들 간의 '조건-결과' 식으로 표현되는 유용한 패턴
- 이러한 패턴 규칙을 발견해내는 것을 연관분석이라 함

3. 연관분석의 대표적 알고리즘 apriori의 단점을 보완하기 위해 트리와 노드 링크라는 특별한 자료 구조를 사용하는 알고리즘은?

1 FP-Growth

2 arules

3 kohonen

4 spade

#### FP-Growth

- Apriori의 경우 데이터셋이 큰 경우 모든 후보 itemset에 대해 하나하나 검사하는 것이 비효율적임
- 이러한 Apriori 단점을 보완하기 위해 FP-tree와 node, link라는 특별한 자료 구조를 사용

신뢰도

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

1 0.75

2 0.65

3 0.6

4 0.7

I

- 신뢰도 : 상품 A를 구매했을 때 상품 B를 구매할 확률이 어느 정도 되는지를 확인
- $= P(A \cap B) / P(A)$  : A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수
- $= 3 / 4 = 0.75$

6. A 제품구입  $\rightarrow$  B 제품 구입의 연관규칙 측정지표 중 지지도(support)란?

1 A와 B가 동시에 포함된 거래수/전체 거래수

## arbitrary

미국식['ɑ:rbitrəri] 영국식['a:bɪtrəri; 'a:bɪtri]

형용사

### 1 임의적인, 제멋대로인

The choice of players for the team seemed completely arbitrary.

그 팀의 선수 선발은 완전히 제멋대로인 것처럼 보였다.

### 2 전횡을 일삼는, 독단적인

the arbitrary powers of officials

공무원들의 독단적인 권력

영어사전 다른 뜻 1

## # - 가변 매개 변수(Arbitrary Argument List)

```
def merge_string(*text_list):
    print(type(text_list)) # <class 'tuple'>

    result = ""
    for s in text_list:
        result += s

    return result

print(merge_string('세상을 이처럼 ', '사랑하사 ', '독생자 예수를 주셨느니라 (요3:16)'))
```

<class 'tuple'>  
세상을 이처럼 사랑하사 독생자 예수를 주셨느니라 (요3:16)

