1회차

1. 다음 중 결측값에 대한 설명 중 적절하지 않은 것은?

* 정답1 결측값은 의미없는 값이므로 해당 관찰치의 정보는 일단 제거하는 것이 좋다. *내가 선택한 답*
* 2 결측값이 포함된 변수를 다른 변수로 설명하는 회귀모형을 이용하여 결측값을 예측할 수 있다.*내가 선택한 답*
* 3 결측값이 포함된 변수에서 결측값이 아닌 나머지 자료값들의 평균 등으로 대체할 수 있다.*내가 선택한 답*
* 4 결측값이 많은 자료의 경우 분석의 어려움이 따른다.*내가 선택한 답*

정답 : **결측값은 의미없는 값이므로 해당 관찰치의 정보는 일단 제거하는 것이 좋다.**

결측값도 의미있는 정보이며, 결측값이 포함된 관찰치의 다른 변수의 값들도 유용한 정보가 될 수 있으므로, 일방적으로 제거하는 것은 바람직하지 않다.

정답2. 다음 중 이상값으로 진단될 수 있는 경우는 무엇인가?

* 1 표준화 점수가 0인 자료*내가 선택한 답*
* 2 Q1-1.5×IQR보다 큰 자료*내가 선택한 답*
* 3 표준화 점수가 음수인 자료*내가 선택한 답*
* 정답4 Q3+1.5×IQR보다 큰 자료 *내가 선택한 답*

정답 : **Q3+1.5×IQR보다 큰 자료**

상자그림에 기반한 방식으로는 Q1-1.5×IQR보다 작거나, Q3+1.5×IQR보다 큰 자료는 이상값으로 진단될 수 있다. 표준화점수는 절대값이 2, 3등보다 크면 이상값으로 진단될 수 있다.

정답3. 다음 중 파생변수의 사례로 적절하지 않은 것은?

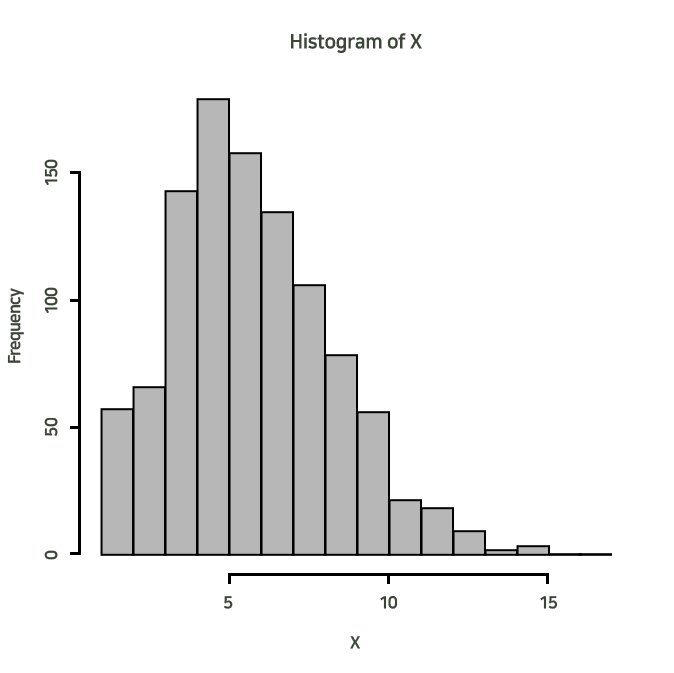
* 1 주요 활동지역 *내가 선택한 답*
* 2 선호 가격대*내가 선택한 답*
* 정답3 매장별 판매량*내가 선택한 답*
* 4 이벤트 민감도*내가 선택한 답*

정답 : **매장별 판매량**

나머지 문항은 사용자가 주관적으로 기준을 정의할 소지가 있는 사안이지만, 매장별 판매량은 수집된 정보를 객관적인 기준에 따라 요약한 변수에 해당한다.

2회차

1. 어느 연속형 변수 X에 관한 히스토그램이 다음과 같이 도출되었다. X의 분포가 종모양의 대칭형이 되도록 하려면 어떤 변환을 하는 것이 적절한가?

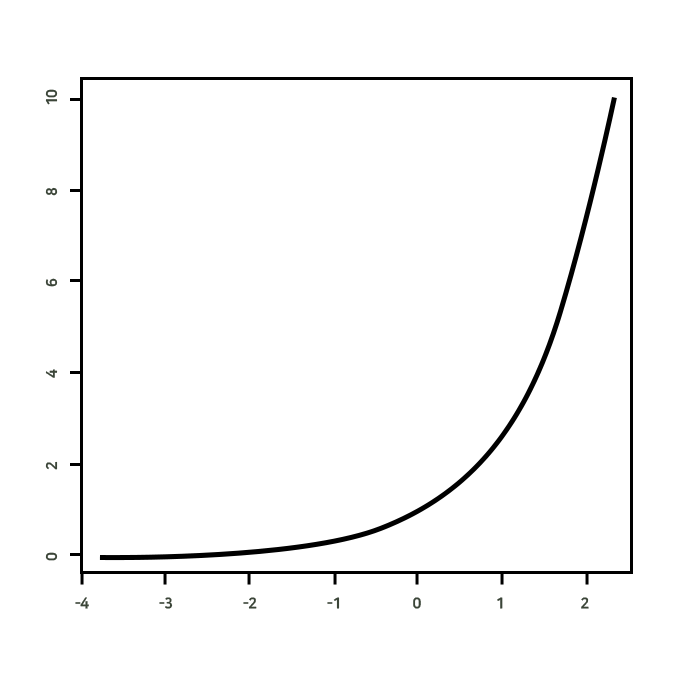


* 1 지수 변환*내가 선택한 답*
* 2 세제곱 변환*내가 선택한 답*
* 정답3 로그 변환 *내가 선택한 답*
* 4 제곱 변환*내가 선택한 답*

정답 : **로그 변환**

분포의 오른 꼬리가 긴 경우에 대칭화를 하기 위해서는 로그 또는 제곱근 유형의 변환을 해야 한다.

정답2. 어느 두 연속형 변수 X와 Y의 관계가 아래와 같은 비선형 관계라고 하자. 이를 선형 관계로 변환하려면, X 또는 Y에 어떤 변환을 적용하는 것이 적절한가?

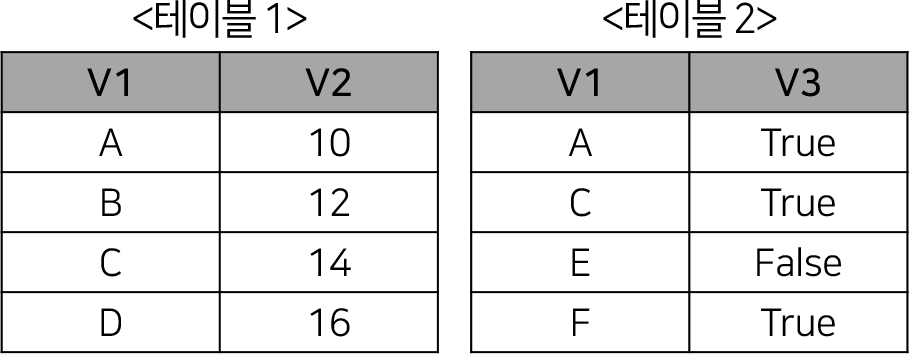


* 1 X는 변환하지 않음, Y는 제곱 변환*내가 선택한 답*
* 2 X는 변환하지 않음, Y는 지수 변환*내가 선택한 답*
* 정답3 X는 변환하지 않음, Y는 로그 변환 *내가 선택한 답*
* 4 X는 로그 변환, Y는 변환하지 않음*내가 선택한 답*

정답 : **X는 변환하지 않음, Y는 로그 변환**

기울기가 급격히 상승하는 형태이므로, X를 제곱 또는 지수변환을 하거나, Y를 제곱근 또는 로그변환을 하면 선형의 관계를 만들어 줄 수 있다.

정답3. 다음의 두 테이블을 결합한 결과 (V1='E', V2=NA(결측), V3=False)의 관찰값을 얻었다고 한다. 다음 중 어떤 유형의 결합을 실행한 것으로 볼 수 있나?



* 1 레프트 조인*내가 선택한 답*
* 정답2 라이트 조인*내가 선택한 답*
* 3 이너 조인*내가 선택한 답*
* 4 크로스 조인 *내가 선택한 답*

정답 : **라이트 조인**

V1의 값이 'E'인 경우는 오른쪽 테이블에 있으므로, 라이트 조인 또는 풀 아우터 조인을 수행한 경우만 얻을 수 있다.

3회차

1. 다음 중 지도학습(Supervised Learning)에 관한 설명으로 맞는 것은 무엇인가?

* 1 목표변수가 연속형인 경우는 분류모델이라고 한다.*내가 선택한 답*
* 2 목표변수 없이 특성변수 정보만으로 데이터를 탐색하는 것이 목적이다.*내가 선택한 답*
* 3 모델의 과적합 문제가 발생하지 않는다.*내가 선택한 답*
* 정답4 훈련 자료에서 여러 특성변수를 이용하여 목표변수인 라벨을 예측하는 것이 목적이다. *내가 선택한 답*

정답 : **훈련 자료에서 여러 특성변수를 이용하여 목표변수인 라벨을 예측하는 것이 목적이다.**

지도학습은 목표변수가 있는 훈련용 자료를 이용하여 목표변수를 예측하는 모델을 학습하는 것으로, 목표변수가 연속형인 경우는 회귀모델, 목표변수가 범주형인 경우는 분류모델이라고 한다.

정답2. 다음 중 자율학습(Unsupervised Learning) 방법에 해당하는 것은?

* 1 의사결정나무 모델 *내가 선택한 답*
* 정답2 K평균 군집분석*내가 선택한 답*
* 3 선형회귀 모델*내가 선택한 답*
* 4 로지스틱 회귀모델*내가 선택한 답*

정답 : **K평균 군집분석**

자율학습은 비지도 학습을 말하며, 군집분석은 대표적인 비지도 학습 알고리즘이다.

정답3. 이것은 비지도 학습법의 일종으로, 고차원의 특성변수로 구성된 데이터를 정보 손실을 최소로 하며 저차원의 데이터로 변환하는 방법을 말한다. 대표적인 알고리즘으로는 PCA, LDA, t-SNE 등의 있다. 이는 머신러닝 기법 중 무엇에 대한 설명인가?

* 1 분류 모델*내가 선택한 답*
* 2 군집화 기법*내가 선택한 답*
* 정답3 차원축소법 *내가 선택한 답*
* 4 추천시스템*내가 선택한 답*

정답 : **차원축소법**

고차원의 데이터를 저차원으로 변환하는 것은 차원축소법을 말하는 것으로 대표적인 알고리즘으로는 PCA, t-SNE 등이 있다.

4회차

1. 다음 중 과적합(overfitting) 문제를 막기위한 방법으로 올바른 것은?

* 1 평가(test) 자료 수를 늘린다.*내가 선택한 답*
* 정답2 훈련(training) 자료 수를 늘린다.*내가 선택한 답*
* 3 목적함수에 규제(regularization)를 없앤다. *내가 선택한 답*
* 4 모수(parameter)가 더 많이 포함된 모델을 선택한다.*내가 선택한 답*

정답 : **훈련(training) 자료 수를 늘린다.**

머신러닝 모델에서의 과적합을 방지하기 위해서는 모델 파라미터에 규제를 적용하거나 특성변수를 줄여 모수의 개수를 줄이는 등의 모델 복잡도를 낮추거나, 모델이 잘 적합되도록 훈련 데이터의 수를 늘이는 방법이 있다. 평가데이터는 모델을 학습하는데 활용되지 않으므로, 평가데이터를 늘인다고 과적합을 줄일 수는 없다.

정답2. 자료에 비하여 너무 복잡한 모델을 적용하는 경우에 발생할 수 있는 현상과 거리가 먼 것은?

* 정답1 편향(bias)이 증가한다.*내가 선택한 답*
* 2 분산(variance)이 증가한다. *내가 선택한 답*
* 3 과적합(overfitting)의 문제가 발생할 수 있다.*내가 선택한 답*
* 4 훈련 자료는 잘 예측하지만 평가자료에 대한 예측력은 떨어지는 현상을 보일 수 있다.*내가 선택한 답*

정답 : **편향(bias)이 증가한다.**

모델 복잡도가 증가하면, 편향은 줄어들지만, 분산이 늘어나 일반화 오차가 커지는 현상을 보인다. 또 훈련 데이터는 잘 예측하지만, 평가 데이터의 예측력은 그에 미치지 못하는 과적합 현상을 보이게 된다.

정답3. 모델을 최적의 성능으로 튜닝하기 위해 사용하는 데이터로, 모델에 필요한 하이퍼 파라미터를 조정하거나, 특성변수의 선택 등의 목적으로 활용되는 자료를 무엇이라고 하는가?

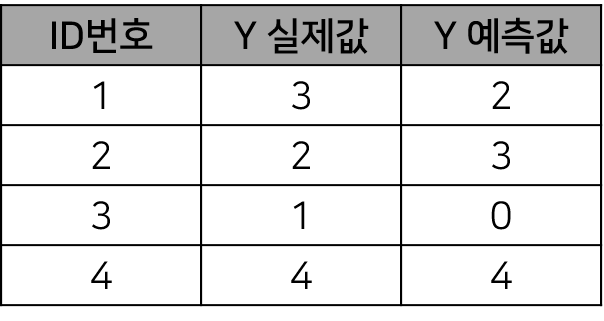
* 1 훈련(training) 데이터 *내가 선택한 답*
* 2 평가(test) 데이터*내가 선택한 답*
* 정답3 검증(validation) 데이터*내가 선택한 답*
* 4 학습(learning) 데이터*내가 선택한 답*

정답 : **검증(validation) 데이터**

검증 데이터는 훈련 데이터의 일부로, 하이퍼파라미터 및 변수 선택 등의 모델 튜닝을 위해 활용된다.

5회차

1. 다음은 어느 자료의 목표변수 Y에 대한 실제값과 예측값을 기록한 것이다. 이 결과에 대한 RMSE는 얼마인가?



* 1 0.5*내가 선택한 답*
* 2 0.75 *내가 선택한 답*
* 정답3 0.866*내가 선택한 답*
* 4 1*내가 선택한 답*

정답 : **0.866**

MSE는 (1+1+1+0)/4=0.75이며, RMSE는 MSE의 제곱근으로 0.866이 된다.

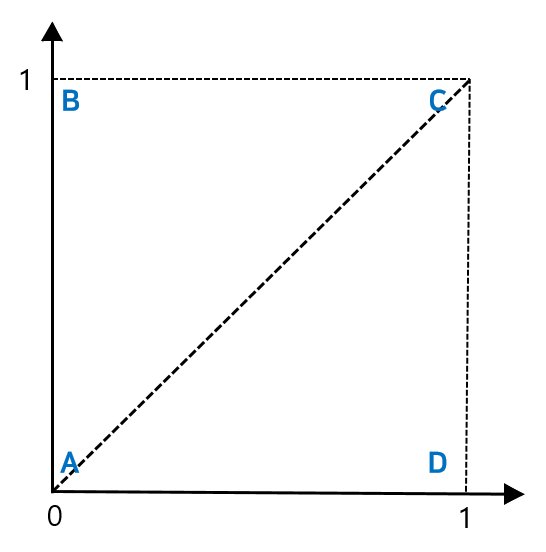
정답2. 다음 중 정밀도(precision)에 대한 설명으로 올바른 것은 무엇인가?

* 1 정밀도(precision)가 높으면 재현율(recall)은 함께 높은 경향이 있다. *내가 선택한 답*
* 2 정밀도(precision)가 높으면 정확도(accuracy)도 함께 높은 경향이 있다.*내가 선택한 답*
* 3 분류 성능을 평가하기 위한 지표이다.*내가 선택한 답*
* 정답4 Positive로 예측한 것 중 실제로 Positive인 자료의 비율로 정의된다*내가 선택한 답*

정답 : **Positive로 예측한 것 중 실제로 Positive인 자료의 비율로 정의된다**

정밀도가 높다고 재현율이나 정확도가 높은 것은 아니고, 정밀도와 재현율은 실제로 반대로 움직이는 경우가 많다.

정답3. 다음 그림에 ROC 커브를 그리고자 한다. 이에 관한 설명으로 옳지 않은 것은?



* 1 커브가 B에 더욱 근접할 수록 좋은 성능을 가진다고 말할 수 있다.*내가 선택한 답*
* 2 가로축은 FPR이다. *내가 선택한 답*
* 3 ROC 커브의 곡선아래 면적이 클수록 좋은 성능을 나타낸다.*내가 선택한 답*
* 정답4 세로축은 특이도(sensitivity)를 나타낸다.*내가 선택한 답*

정답 : **세로축은 특이도(sensitivity)를 나타낸다.**

ROC 곡선 그림의 세로축은 민감도(sensitivity) 또는 TPR에 해당한다.

6회차

1. 다음 중 특징 선택(feature selection)과 관련이 없는 설명은 무엇인가?

* 정답1 기존의 특성변수를 결합하여 새로운 변수를 생성한다. *내가 선택한 답*
* 2 전체 특성변수들 가운데 가장 좋은 특성변수의 조합을 찾고자 한다.*내가 선택한 답*
* 3 특성변수들 가운데 중복된 정보를 제거하는 효과가 있다.*내가 선택한 답*
* 4 특성변수의 조합을 찾기 위해 순차탐색법을 사용하기도 한다.*내가 선택한 답*

정답 : **기존의 특성변수를 결합하여 새로운 변수를 생성한다.**

기존의 특성변수를 결합하여 새로운 변수를 생성하는 방식은 특징 추출(feature extraction)에 해당한다.

정답2. 특징 선택(feature selection) 기법 중 Filter 방식의 특징으로 볼 수 없는 것은?

* 1 선택된 특성변수는 목표변수의 예측에 중요한 변수로 볼 수 있다.*내가 선택한 답*
* 정답2 선택된 입력변수 간에는 거의 상관관계가 존재하지 않게 된다. *내가 선택한 답*
* 3 계산 비용이 적고, 속도가 빠른 방식이다.*내가 선택한 답*
* 4 대부분의 기법들이 특성변수와 출력변수 사이의 선형성을 기반으로 선택한다.*내가 선택한 답*

정답 : **선택된 입력변수 간에는 거의 상관관계가 존재하지 않게 된다.**

Filter 방식은 특성변수와 목표변수 간의 일대일 관계로 중요성 여부를 판단하므로, 선택된 특성변수들 간의 상관성은 고려하지 않는다. 따라서 서로 상관관계가 높은 중복된 특성변수들이 함께 선택될 수 있다는 것이 단점이다.

정답3. 이는 특징 선택 방법의 하나로, 주어진 특성 변수들 중 최적의 조합을 찾기 위하여 다양한 특성변수의 조합에 대하여 매번 목표변수를 예측하기 위한 모델을 훈련하는 방식을 말한다. 따라서 여러 차례 모델을 학습해야 해서 계산비용이 크지만, 특성 변수 간 상호작용을 고려한 변수선택이 가능하다는 것이 장점이다. 이는 어떤 특징 선택 방법에 대한 설명인가?

* 1 Filter 방식*내가 선택한 답*
* 정답2 Wrapper 방식 *내가 선택한 답*
* 3 Embedded 방식*내가 선택한 답*
* 4 Extraction 방식*내가 선택한 답*

정답 : **Wrapper 방식**

Wrapper 방식은 특성변수 조합별로 모델을 적합한 뒤 예측 성능을 비교하는 방식으로 최적의 특성 변수의 조합을 찾아내는 방식을 말하며, 순차 탐색법 등이 포함된다

7회차

1. 다음 중 특징 추출(feature extraction)과 관련이 없는 설명은 무엇인가?

* 1 기존에 없던 새로운 특성변수가 생성된다. *내가 선택한 답*
* 2 특성공간의 차원을 축소할 수 있다.*내가 선택한 답*
* 3 주성분분석, SVD 등의 방법이 고려될 수 있다.*내가 선택한 답*
* 정답4 중요한 특성변수부터 순차적으로 선택하는 방법을 이용할 수 있다.*내가 선택한 답*

정답 : **중요한 특성변수부터 순차적으로 선택하는 방법을 이용할 수 있다.**

특성변수를 순차 탐색하는 방법은 특징선택법의 일종이다.

정답2. 다음은 주성분분석(PCA)에 관한 설명이다. 적절하지 않은 것은 무엇인가?

* 1 특성변수의 차원축소법으로 활용될 수 있다.*내가 선택한 답*
* 2 변환된 특성변수의 공분산행렬은 대각행렬이 된다.*내가 선택한 답*
* 정답3 가장 평균이 커지는 축으로 특성공간을 회전변환하는 방법이다.*내가 선택한 답*
* 4 자료가 유클리디안 공간에 존재하는 것으로 가정한다. *내가 선택한 답*

정답 : **가장 평균이 커지는 축으로 특성공간을 회전변환하는 방법이다.**

주성분분석은 평균이 아니라 분산이 가장 커지는 방향으로 특성공간의 축을 회전변환하는 방법이다.

정답3. 주성분분석(PCA)와 SVD에 관한 공통된 사항으로 잘못된 것은 무엇인가?

* 정답1 지도학습 알고리즘으로 볼 수 있다. *내가 선택한 답*
* 2 자료 행렬에 대한 특성값 분해와 밀접한 연관이 있다.*내가 선택한 답*
* 3 차원 축소법으로 활용될 수 있다.*내가 선택한 답*
* 4 특성 추출법의 일종이다.*내가 선택한 답*

정답 : **지도학습 알고리즘으로 볼 수 있다.**

주성분분석과 SVE는 모두 목표변수 없이 특성변수들의 정보를 이용하여 차원을 축소하는 머신러닝 알고리즘으로, 비지도학습 알고리즘에 해당한다.

8회차

1. 다음 중 군집분석(clustering)에 대한 설명에 해당하는 것은 무엇인가?

* 1 지도학습 알고리즘으로 볼 수 있다.*내가 선택한 답*
* 2 주요 기법으로 KNN(K-nearest neighbor)를 들 수 있다. *내가 선택한 답*
* 정답3 유사한 관찰치들을 묶어서 그룹을 형성한다.*내가 선택한 답*
* 4 유사한 특성변수를 묶어 그룹을 형성한다.*내가 선택한 답*

정답 : **유사한 관찰치들을 묶어서 그룹을 형성한다.**

군집분석은 비지도학습으로 유사한 관찰치를 묶는 방식이다.

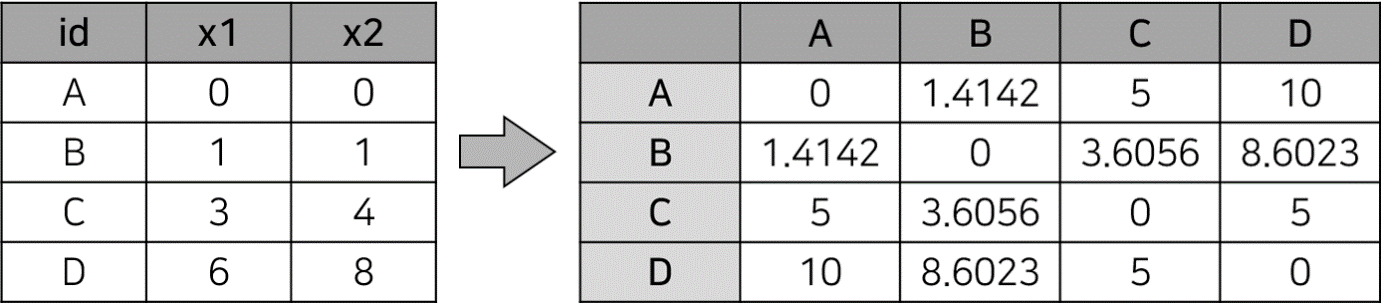
정답2. 다음은 계층적 군집분석(hierarchical clustering)에서 군집 간 거리를 정의하는 여러 기법이다. 두 개의 군집 내에 속한 모든 자료간 거리를 구한 뒤 그 평균을 두 군집간 거리로 정의하는 기법을 무엇이라고 하는가?

* 1 완전연결법(complete linkage)*내가 선택한 답*
* 2 와드연결법(ward linkage)*내가 선택한 답*
* 정답3 평균연결법(average linkage) *내가 선택한 답*
* 4 중심연결법(centroid linkage)*내가 선택한 답*

정답 : **평균연결법(average linkage)**

제시된 설명은 평균연결법에 대한 것이며, 중심연결법은 두 군집 중심간의 거리를 이용하는 방식이다.

정답3. 다음은 두 개의 변수 x1, x2에 대한 4개의 관찰값과 이 자료를 토대로 거리행렬을 도출한 것이다. 이 자료에 대해 단일연결법(single linkage)를 적용하여 두 개의 군집을 발견하고자 한다면, 각 군집은 어떻게 구성되겠는가?



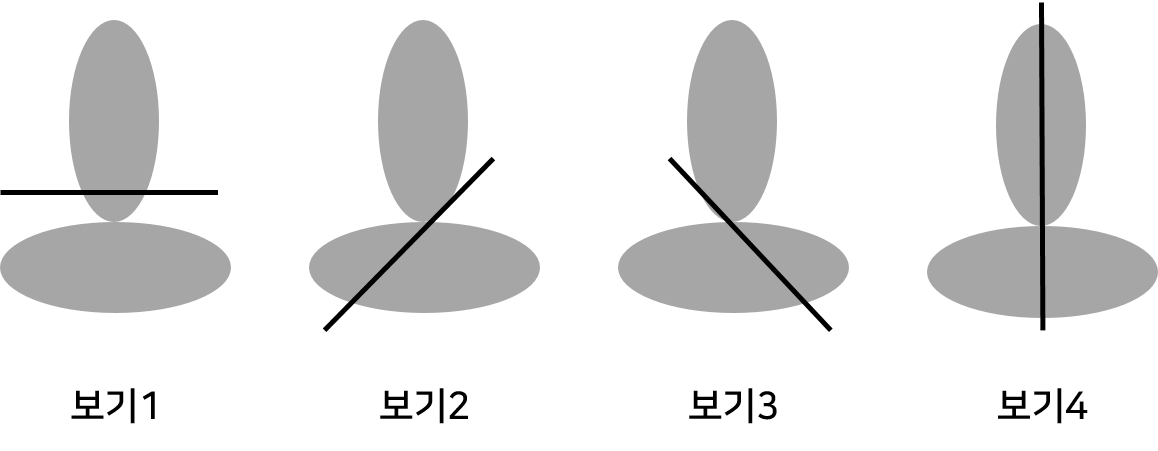
* 1 A, BCD*내가 선택한 답*
* 2 AB, CD*내가 선택한 답*
* 3 AC, BD *내가 선택한 답*
* 정답4 ABC, D*내가 선택한 답*

정답 : **ABC, D**

단일연결법은 군집간 거리를 군집 내 자료 간의 거리 중 최소값으로 정의한다. 첫 단계에서는 주어진 거리행렬에서 가장 최소거리를 가지는 AB가 먼저 군집을 이룬다. 그 다음 단계에서는 (AB), C, D 간의 거리를 먼저 구해보아야 하는데, (AB)와 C의 거리는 d((AB), C)=min(d(A,C), d(B,C))=3.6056이고, (AB)와 D의 거리는 d((AB), D)=min(d(A,D), d(B,D))=8.6023이며, C와 D의 거리는 5가 된다. 따라서 (AB)와 C를 묶은 ABC와 D가 두번째 단계의 군집이 된다. 최종적으로 D까지 포함된 ABCD가 하나의 군집이 된다. 따라서 두 개의 군집을 만들고자 한다면, (ABC)와 D로 구성될 것이다.

9회차

1. 다음과 같은 형태로 분포하는 자료에 K=2인 K-평균 군집분석을 수행한다고 하자. 다음 4가지 보기 중에서 이 군집분석에서 초기 군집으로 사용하기에 가장 좋은 것은 무엇인가?



* 정답1 이미지1*내가 선택한 답*
* 2 이미지2*내가 선택한 답*
* 3 이미지3*내가 선택한 답*
* 4 이미지4 *내가 선택한 답*

정답 : **이미지1**

K-평균 군집분석은 초기 군집을 어떻게 설정하는가에 따라 최종 결과가 영향을 많이 받게 되는데, 가능 한 초기군집 중심 간의 거리가 먼 것이 좋다.

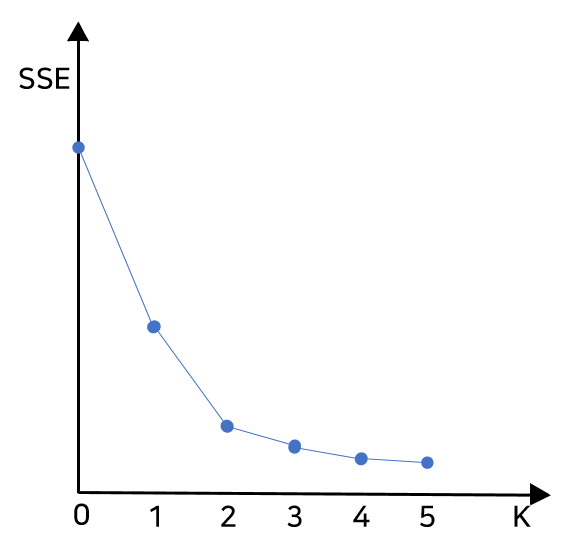
정답2. K-평균 군집분석에 관한 설명 중 적절하지 않은 것은?

* 1 초기군집을 어떻게 설정하는지에 따라 결과의 영향이 크기 때문에 주의해야 한다.*내가 선택한 답*
* 2 계층적 군집분석에 비해 수행시간이 빠르다.*내가 선택한 답*
* 정답3 계층적 군집분석에 비해 이상치나 잡음의 영향을 덜 받는다. *내가 선택한 답*
* 4 군집의 수를 사전에 결정해야 한다.*내가 선택한 답*

정답 : **계층적 군집분석에 비해 이상치나 잡음의 영향을 덜 받는다.**

K-평균 군집분석에서 이상치나 잡음이 있으면 각 군집 중심점이 영향을 받기 때문에, 군집 분석의 결과가 이상치의 영향이 큰 편이다.

정답3. K-평균 군집분석에서 적절한 군집 수를 선택하기 위하여 각 군집의 개수별로 오차제곱합(SSE)을 구해 보았더니 아래와 같은 결과가 도출되었다. 이 경우 가장 적절한 군집의 수 K는 얼마가 적절한가?



* 정답1 2개*내가 선택한 답*
* 2 3개 *내가 선택한 답*
* 3 4개*내가 선택한 답*
* 4 5개*내가 선택한 답*

정답 : **2개**

군집 수의 결정에 오차제곱합을 이용하는 경우, K의 증가에 따라 오차제곱합의 크기가 급격히 감소하다가 완만해지는 야상을 보이게 되는데, 이 때 완만해지기 시작하는 시점을 적절한 군집의 수로 선택한다.

10회차

1. 다음은 선형회귀모형에 관한 것이다. 올바른 설명은 무엇인가?

* 1 다중회귀모형에서는 종속변수가 여러 개다.*내가 선택한 답*
* 2 단순회귀모형에서는 독립변수가 여러 개다.*내가 선택한 답*
* 3 다중회귀모형에서의 종속변수는 범주형이다.*내가 선택한 답*
* 정답4 단순회귀모형은 하나의 독립변수와 하나의 종속변수에 관한 모형이다. *내가 선택한 답*

정답 : **단순회귀모형은 하나의 독립변수와 하나의 종속변수에 관한 모형이다.**

다중회귀모델은 독립변수가 여러 개, 종속변수가 하나이며, 회귀모델의 종속변수는 숫자형이어야 한다.

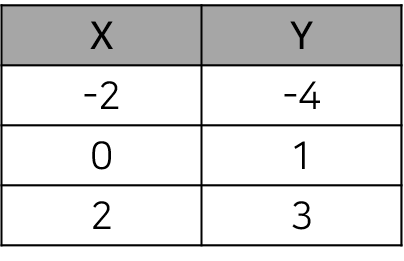
정답2. 단순 선형회귀모형에서 오차항에 관한 가정으로 적절하지 않은 것은?

* 1 오차는 서로 독립임.*내가 선택한 답*
* 2 오차는 정규분포를 따름.*내가 선택한 답*
* 정답3 오차의 분산이 알려짐. *내가 선택한 답*
* 4 오차의 기대값은 0임.*내가 선택한 답*

정답 : **오차의 분산이 알려짐.**

오차는 정규, 등분산, 독립 가정이 필요하다. 오차의 분산은 일반적으로 알려져 있지 않은 것으로 가정된다.

정답3. 두 숫자형 변수 X와 Y에 관한 다음 3개의 관찰값이 주어져 있다고 하자. 최소제곱법에 의해 단순선형회귀모형을 적합한다면, 기울기계수는 얼마로 추정되는가?





정답 : **1.75**

X와 Y의 평균이 모두 0이므로, 최소제곱법에 의한 기울기 계수 베타의 추정치는 ((-2)\*(-4)+0\*1+2\*3)/((-2)\*(-2)+0\*0+2\*2)=14/8=1.75로 구해진다.

11회차

1. 어느 기업의 매출액 Y을 광고비 X1, 제품가격 X2, 원료가격 X3를 이용한 다중선형회귀모형을 적합하고자 한다. 광고비 X1에 대한 회귀계수가 0.3으로 추정된 경우 이를 적절하게 해석한 것은?

* 1 광고비가 1단위 증가하면 매출액의 평균은 0.3 감소할 것으로 추정된다. *내가 선택한 답*
* 2 광고비가 1단위 증가하면 매출액의 평균은 0.3 증가할 것으로 추정된다.*내가 선택한 답*
* 3 다른 독립변수는 고정된 상태에서 광고비가 1단위 증가하면 매출액의 평균은 0.3 감소할 것으로 추정된다.*내가 선택한 답*
* 정답4 다른 독립변수는 고정된 상태에서 광고비가 1단위 증가하면 매출액의 평균은 0.3 증가할 것으로 추정된다.*내가 선택한 답*

정답 : **다른 독립변수는 고정된 상태에서 광고비가 1단위 증가하면 매출액의 평균은 0.3 증가할 것으로 추정된다.**

다중선형회귀모형의 회귀계수는, 해당 변수 이외 나머지 변수는 상수로 고정된 상태에서 그 변수가 1단위 증가할 때 종속변수가 평균적으로 어느만큼 변하는가를 나타낸다.

정답2. 어느 데이터에 대하여 다중선형회귀모형을 적합한 결과 총제곱합(SST)는 60, 오차제곱합(SSE)는 15였다고 하자. 이 적합된 회귀모형의 결정계수(r-square)는 얼마인가?



정답 : **0.75**

결정계수는 1-SSE/SST로 정의되므로, 1-15/60=0.75가 된다.

정답3. 근로자의 임금을 학력으로 설명하는 회귀분석을 수행하고자 한다. 근로자의 학력은 중졸, 고졸, 대졸이상의 세 범주로 표현된 범주형 자료라고 할 때, 다음 설명 중 적절하지 않은 것은 무엇인가?

* 1 독립변수가 범주형인 경우 더미변수를 생성하여 분석해야 한다.*내가 선택한 답*
* 2 독립변수가 가지는 범주의 개수가 3개이므로 더미변수는 2개가 필요하다.*내가 선택한 답*
* 정답3 대졸이상인지 중졸인지에 따라 회귀선의 기울기가 달라진다.*내가 선택한 답*
* 4 대졸이상인지 중졸인지에 따라 회귀선의 절편이 달라진다. *내가 선택한 답*

정답 : **대졸이상인지 중졸인지에 따라 회귀선의 기울기가 달라진다.**

독립변수가 범주형인 경우 (범주의 개수-1)만큼의 더미변수를 생성하여 분석을 수행한다. 중졸인 경우와 대졸이상인 경우 더미변수값이 1 또는 0이 되며, 이는 회귀선의 기울기가 아니라 절편에만 영향을 준다.

12회차

1. 회귀모형에서 변수선택법 가운데 후진제거법(backward elimination)에 관한 설명으로 적절하지 않은 것은?

* 1 초기 모형은 모든 독립변수가 포함된 모형이다.*내가 선택한 답*
* 정답2 한번 제거된 변수도 경우에 따라 다시 선택될 수 있다. *내가 선택한 답*
* 3 독립변수는 한번에 하나씩 제거된다.*내가 선택한 답*
* 4 독립변수가 제거될 때마다 모델을 다시 적합한다.*내가 선택한 답*

정답 : **한번 제거된 변수도 경우에 따라 다시 선택될 수 있다.**

한번 제거된 변수가 다시 선택되는지를 고려하는 방법은 단계선택법에 해당한다.

정답2. 절편만 포함된 모형에서 출발하여, 종속변수를 설명하기에 가장 중요한 독립변수부터 하나씩 모형에 추가하되, 새로운 변수가 모형에 추가되는 매 단계마다 모형에 포함되어 있던 변수 중 더 이상 중요하지 않은 변수가 있는지 파악하여 이를 제거하는 변수선택법을 무엇이라고 하는가?

* 1 전진선택법*내가 선택한 답*
* 2 후진제거법*내가 선택한 답*
* 정답3 단계선택법 *내가 선택한 답*
* 4 모든 가능한 조합을 탐색하는 방법*내가 선택한 답*

정답 : **단계선택법**

제시된 설명은 단계선택법에 관한 것으로, 모든 가능한 조합을 탐색하지 않고도 독립변수들의 상호작용이 고려된 최적의 조합을 빠르게 찾을 수 있다는 점에서 가장 선호되는 변수선택법이다.

정답3. 다음은 회귀모형의 다양한 적합도 지표에 관한 설명이다. 잘못된 것은 무엇인가?

* 1 결정계수는 종속변수의 총 변동 중 적합된 회귀모형으로 설명되는 변동의 비율을 의미한다.*내가 선택한 답*
* 정답2 종속변수를 잘 설명하지 못하는 독립변수가 추가되는 경우에도 수정결정계수는 항상 증가한다.*내가 선택한 답*
* 3 AIC 값이 작을수록 적합도가 높은 것이다.*내가 선택한 답*
* 4 동일한 모형과 자료에 대하여 결정계수보다 수정결정계수의 값이 항상 더 작다. *내가 선택한 답*

정답 : **종속변수를 잘 설명하지 못하는 독립변수가 추가되는 경우에도 수정결정계수는 항상 증가한다.**

종속변수를 잘 설명하지 못하는 독립변수가 추가되는 경우에 결정계수는 항상 증가하지만, 수정결정계수의 경우는 그렇지 않다. 수정결정계수는 새로운 독립변수의 추가로 SSE/(n-k-1)가 감소하는 경우에만 증가한다. 또한 결정계수는 1-(SSE/SST)이고, 수정결정계수는 1-(n-1)/(n-k-1)\*(SSE/SST)이므로 동일한 조건에서 결정계수보다 항상 조금 더 작은 값이 된다.

13회차

1. 선형회귀모형에서 오차에 대한 어떤 가정이 위반된 경우에 가중최소제곱(WLS) 방법을 고려하는가?

* 1 오차의 정규성 *내가 선택한 답*
* 정답2 오차의 등분산성*내가 선택한 답*
* 3 오차의 기대값은 0*내가 선택한 답*
* 4 오차의 독립성*내가 선택한 답*

정답 : **오차의 등분산성**

가중최소제곱은 오차의 등분산성이 위반된 경우의 해결방안이다. 독립성이 위반된 경우는 시계열 분석을 적용하며, 정규성이 위반된 경우는 변수변환을 적용한다.

정답2. 다중선형회귀모형에서 다중공선성(multicollinearity)에 대한 설명으로 잘못된 것은?

* 1 주성분분석(PCA)를 이용하여 해결할 수 있다.*내가 선택한 답*
* 정답2 적절한 다중회귀모형은 다중공선성이 존재한다. *내가 선택한 답*
* 3 변수선택이 해결방안이 될 수 있다.*내가 선택한 답*
* 4 다중공선성이 심할수록 회귀계수 추정의 정확도가 낮아진다.*내가 선택한 답*

정답 : **적절한 다중회귀모형은 다중공선성이 존재한다.**

다중회귀모형은 독립변수 간 상관관계가 어느정도 존재하는 경우가 많지만, 그 상관관계가 너무 강한 상태를 다중공선성이라고 하며, 다중공선성이 진단되는 모형은 적절한 모형으로 볼 수 없다.

정답3. 다중선형회귀모형에서 다중공선성 문제를 진단하기 위한 지표로 적절한 것은?

* 정답1 분산팽창요인(VIF) *내가 선택한 답*
* 2 회귀계수 추정값의 표준오차*내가 선택한 답*
* 3 회귀계수 추정값*내가 선택한 답*
* 4 결정계수*내가 선택한 답*

정답 : **분산팽창요인(VIF)**

어느 독립변수에 대한 분산팽창요인(VIF)는 해당 독립변수를 제외한 나머지 독립변수로 주어진 독립변수를 얼마나 설명할 수 있는지를 반영하는 지표로, 이 값이 5나 10 등으로 큰 값이 나오면 다중공선성이 있는 것으로 진단한다.

14회차

1. 선형회귀모형에서 모형의 파라미터에 대한 제약을 가하는 규제(regularization)가 필요한 이유는 무엇인가?

* 1 모형의 편향을 줄이기 위해서 *내가 선택한 답*
* 정답2 모형의 과적합을 방지하기 위해서*내가 선택한 답*
* 3 훈련자료가 부족해서*내가 선택한 답*
* 4 모형의 변동성을 키우기 위해서*내가 선택한 답*

정답 : **모형의 과적합을 방지하기 위해서**

선형회귀모형에 파라미터에 대한 규제를 가하는 경우 편향이 늘어나지만 변동성을 줄여줘 과적합을 방지할 수 있다.

정답2. 선형회귀모델에 대한 규제 방법에 대한 다음의 설명 중 잘못된 것은?

* 1 릿지회귀는 모델 파라미터들의 제곱합이 일정 수준을 넘지 않도록 제한한다.*내가 선택한 답*
* 2 라쏘회귀는 모델 파라미터들의 절대값의 합이 일정 수준을 넘지 않도록 제한한다. *내가 선택한 답*
* 정답3 릿지회귀를 적용하면 일부 파라미터 추정치가 0이 되어, 변수선택의 효과가 있다.*내가 선택한 답*
* 4 엘라스틱넷은 릿지회귀와 라쏘회귀에서의 제약이 모두 반영된다.*내가 선택한 답*

정답 : **릿지회귀를 적용하면 일부 파라미터 추정치가 0이 되어, 변수선택의 효과가 있다.**

일부 파라미터를 0으로 추정하여 변수선택의 효과가 있는 규제방식은 라쏘회귀이다.

정답3. 규제가 있는 선형회귀모형(릿지, 라쏘, 엘라스틱넷 등)은 모형의 복잡도를 증가시켜서 예측력을 높이고자 하는 방법이다. (O, X)

* O*내가 선택한 답*
* 정답X *내가 선택한 답*

정답 : **X**

규제는 모형의 복잡도를 낮춰서 과대적합을 줄이기 위한 기법이다.