|  |  |
| --- | --- |
| **설명** | **정답** |
| 크기가 같은 표본을 여러 번 단순임의 복원 추출하여 분류기를 생선한 후 앙상블하는 기법 / 모델의 안정성을 높이기 위하여 분석 데이터로부터 여러 개의 단순 복원 임의 추출하여 다수결을 통해 최종의 예측 모델을 도출하는 알고리즘 / 각각의 분류기가 모두 같은 유형의 알고리즘 기반이지만, 데이터 샘플링을 서로 다르게 가져가면서 학습을 수행해 보팅을 수행 하는 것 | **배깅** |
| 다층 신경망에서 은닉층이 많아 학습이 이루어지지 않는 문제는 무엇인가? | **기울기 소실 문제** |
| 출력값 z가 여러 개로 주어지고 목표치가 다범주인 경우 각 범주에 속할 사후 확률을 제공하는 함수는? | **소프트맥스 함수** |
| 군집내 거리와 군집간의 거리를 기준으로 군집 분할 성과를 측정하는 방식은? /  군집 내의 데이터 응집도와 군집 간 분리도를 계산하는 지표는? | **실루엣 계수** |
| 두 벡터 사이의 각도를 이용하여 벡터간의 유사 정도를 측정하는 방식은? | **코사인 유사도** |
| 계층적 군집분석에서 군집 내의 오차제곱합에 기초하여 거리를 측정하는 방법은? /  오차제 곱합의 합에 비해 증가한 정도가 작아지는 방향으로 군집하는 방법은? | **와드 연결법** |
| 연관규칙 분석에서 품목 간 상관관계를 기준으로 규칙의 예측력을 평가하는 지표는? /  A→B 의 연관 규칙에서 임의로 B가 구매되는 경우에 비해 A와의 관계가 고려되어 구매되는 경우의 비율이다. | **향상도** |
| 전체 항목 중 A와 B가 동시에 포함되는 항목의 비율은? | **지지도** |
| 정규화 방법 중 원 데이터의 분포를 유지하면서 정규화하는 방법은? /  모든 데이터를 0과 1 사이의 값으로 변환하는 기법은? | 최소-최대(Min-Max)정규화 |
| 일반 상용 서버로 구성된 클러스터에서 사용할 수 있는 분산 파일 시스템과,  대량의 자료를 처리하기 위한 분산 처리 시스템을 제공하는 오픈소스 프레임워크 | **하둡** |
| 데이터상의 주석 작업으로 딥러닝과 같은 학습 알고리즘이 무엇을 학습하여야 하는지 알려 주는 표식 작업 | **어노테이션(Annotation)** |
| 당사자의 동의 없는 개인정보 수집 및 활용하거나 제3자에게 제공하는 것을 금지하는 등  개인정보보호를 강화한 내용을 담아 제정한 법률 | 개인정보보호법 |
| 정보통신망의 개발과 보급 등 이용 촉진과 함께 통신망을 통해 활용되고 있는 정보보호에 관해 규정한 법률 | **정보통신망법** |
| 사생활 침해를 방지하기 위해 데이터에 포함된 개인정보를 삭제하거나 알아볼 수 없는 형태로 변환하는 방법 | **익명화** |
| 문제가 주어지고 이에 대한 해법을 찾기 위해 각 과정이 체계적으로 단계화되어 수행하는 분석 과제 발굴 방식 | **하향식 접근 방식** |
| 문제 정의 자체가 어려운 경우 데이터 기반으로 문제의 재정의 및 해결방안을 탐색하고 이를 지속적으로 개선하는 방식 | **상향식 접근 방식** |
| 개인 식별이 가능한 데이터를 직접적으로 식별할 수 없는 다른 값으로 대체하는 비식별화 방법 | **가명처리** |
| 통계 값을 적용하여 특정 개인을 식별할 수 없도록 하는 비식별화 방법 | **총계처리** |
| 특정 정보를 해당 그룹의 대푯값 또는 구간값으로 변환하는 비식별화 방법 | **데이터 범주화** |
| 데이터의 전부 또는 일부분을 대체값(공백, 노이즈 등)으로 변환하는 비식별화 방법 / 개인의 사생활 침해를 방지하고 통계 응답자의 비밀사항은 보호하면서 통계자료의 유용성을 최 대한 확보할 수 있는 데이터 변환 방법은? | 데이터 마스킹 |
| 주어진 각 개체들의 유사성을 분석해서 높은 대상끼리 일반화된 그룹으로 분류하는 기법 | **군집분석** |
| 동일하거나 다른 학습 알고리즘을 사용해서 여러 모델을 학습하는 개념 / 주어진 자료로부터 여러 개의 예측모형들을 만든 후 예측모형들을 조합하여 하나의 최종 예측 모형을 만들 어 분류 정확성을 향상시키는 기법은? | 앙상블 기법 |
| 데이터의 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정과 기법 | **데이터 시각화** |
| 하나 이상의 변수에 대해서 변수 사이의 차이와 유사성 등을 표현하는 방법 | **비교 시각화** |
| 장소나 지역에 따른 데이터의 분포를 표현하는 것 | **공간 시각화** |
| 고정된 훈련 데이터 세트와 테스트 검증데이터 세트로 평가하여 반복적으로 튜닝할 시  테스트 데이터에 과적합되는 결과가 생기는 것을 방지하는 방법 | **교차 검증** |
| 반복을 통하여 점증적으로 개발하는 방법으로 처음 시도하는 프로젝트에 적용이 용이하지만, 반복에 대한 관리 체계를 효과적으로 갖추지 못한 경우 복잡도가 상승하여 프로젝트 진행이 어려울 수 있는 모델은 무엇인가? | **나선형 모델** |
| GMM(Gaussian Mixture Model) 군집분석이 모수를 학습하는 방법은? | **EM 알고리즘** |
| 빅데이터 저장 기술로 관계형 데이터베이스 관리 시스템으로 하나의 데이터베이스를  여러 개의 서버 상에 구축하는 시스템은? | **데이터베이스 클러스터** |
| 데이터를 분리하는 초평면 중에서 데이터들과 거리가 가장 먼 초평면을 선택하여 분리하는 지도 학습 기반의 이진 선형 분류 모델은? / 주어진 데이터에서 마진을 최대화하는 초평면 을 구하는 방법으로 학습하는 알고리즘은? | **SVM(Soft Vector Machine)** |
| 데이터 안에 관찰할 수 없는 잠재적인 변수가 존재한다고 가정하는 차원축소기법. 모형을 세운 뒤 관찰 가능한 데이터를 이용하여 해당 잠재 요인을 도출하고 데이터 안의 구조를 해석하는 기법은? | **요인분석** |
| 분석 대상 데이터 집합에서 준식별자 속성이 동일한 레코드가 적어도 K개 이상 존재하도록 제한하는 개인정보 보호 기법 | **k-익명성** |
| 대규모 분산 시스템 모니터링을 위해 에이전트와 컬렉터 구성을 통해 데이터를 수집하고  수집된 데이터를 하둡 파일 시스템(HDFS)에 저장하는 기능을 제공하는 데이터 수집 기술 | **Chukwa(척와)** |
| RDBMS와 하둡 사이의 데이터를 이동시켜주는 애플리케이션 | **Apache Sqoop (스쿱)** |
| 분산 환경에서 대량의 로그 데이터를 효과적으로 수집하여 합친 후 다른 곳으로 전송할 수 있는 신뢰할 수 lT는 서비스 | **Apache Flume (플럼)** |
| 빅데이터 저장 기술로 컴퓨터 네트워크를 통해 공유하는  여러 호스트 컴퓨터의 파일에 접근 할 수 있게 하는 파일 시스템은? | **분산 파일 시스템** |
| 실시간으로 기록 스트림을 게시, 구독, 저장 및 처리할 수 있는 분산 데이터 스트리밍 플랫폼 | Apache Kafka (카프카) |
| 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크. 간단하게 설명하자면, 한 명이 4주 작업할 일을 4명이 나누어 1주일에 끝내는 것 | MapReduce (맵리듀스) |
| 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작한다. 자율 학습의 일종으로, 레이블이 달려 있지 않은 입력 데이터에 레이블을 달아주는 역할을 수행한다. | **k-평균 군집화 알고리즘** |
| 데이터베이스의 테이블이 어떻게 구성되는지, 어떤 정보를 담고 있는지에 대한 기본적인 구조를 정의하는 것 | **스키마** |
| 지역별 매출액, 영업이익률, 판매량과 같이 수치로 명확하게 표현되는 데이터로,  그 양이 크게 증가하더라도 이를 DBMS에 저장, 검색, 분석하여 활용하기가 용이하다. | **정량적 데이터** |
| 번호를 부여한 샘플을 나열하여 k개씩 n개의 구간을 나누고 첫 구간에서 하나를 임의로 선택한 후에  k개씩 띄어서 표본을 선택하고 매번 k번째 항목을 추출하는 표본 추출 방법 | **계통추출방법** |
| 시계열 분석의 기본이 되는 중요한 개념으로 시계열의 평균과 분산이 일정하고  일정한 추세가 없는 것을 무엇이라 하는가? | **정상 시계열** |
| 베이즈 정리와 특징에 대한 조건부 독립을 가설로 하는 알고리즘으로 클래스에 대한 사전 정보와 데이터로부터 추출된 정보를 결합하고 베이즈 정리를 이용하여 어떤 데이터가 특정 클래스에 속하는지를 분류하는 알고리즘 | **나이브 베이즈 분류** |
| 이것은 데이터 안의 두 변수 간의 관계를 알아보기 위해 사용하는 값이다. 두 변수간의 공 분산으로는 음과 양의 관계를 파악할 수 있으나 관계 정도를 확인하기는 힘들다. 그래서 각 변수의 표준편차를 곱하여 공분산을 나누어 -1에서 1사이의 값으로 표준화하여 두 변수 간 의 관계 정도를 확인 할 수 있도록 수치화 한 이것을 활용한다. 이것은 무엇인가? | **상관 계수** |
| 풀어야 할 문제에 대한 상세한 설명 및 해당 문제를 해결했을 때 발생하는 효과를 명시함으로써  향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용하도록 하는 것은 무엇인가? | **분석 유스 케이스** |
| 의사결정나무 중 연속형 타깃변수(또는 목표변수)를 예측하는 의사결정나무를 무엇이라고 하 는가? | **회귀나무** |
| 이것은 비즈니스 측면에서 일반적으로 ‘**공동 활용의 목적으로 구축된 유무형의 구조물**’을 의미한다. 수집된 데이터를 가공, 처리, 저장해두고 이 데이터에 접근할 수 있도록 API를 공개 한다. 그러면 다양한 서드파티 사업자들이 비즈니스에 필요한 정보를 추출해 활용하게 되고 빅데이터는 그 자체로 이 역할을 수행하게 된다. | **플랫폼** |
| 로지스틱 회귀분석에서 어떠한 일이 일어날 확률을 일어나지 않을 확률로 나누어 log를 취하고  이를 0~1의 값이 아닌 (-무한대, 무한대) 범위에서 선형함수를 시그모이드 함수로 변환 하는 방법은 무엇인가? | **로짓 변환** |
| 변수들의 자기상관성을 기반으로 한 시계열 모형으로 현**시점의 자료를 p시점 전의 과거 자료를 통해 설명할 수 있는 모델**이다. 자기 자신의 과거 값이 이후 자신의 값에 영향을 준다. / 현시점의 자료가 k 시점 이전의 유한 개의 과거 자료로 설명할 수 있는 모형 / 자기 자신 의 과거 값이 이후 자신의 값에 영향을 주기 때문에 이름이 붙음 | **자기회귀모형(AutoRegressive,AR모형)** |
| **현재 데이터가 과거 백색잡음의 선형 가중합으로 구성된다는 모형** / 시간이 갈수록 관측치의 평균값이 지속해서 증가하거나 감소하는 시계열모형 / 백색잡음 과정은 서로 독립이고 평균이 0인 확률변수이므로 항상 정상성을 만족함 | **이동평균모형(Moving Average,MA모형)** |
| **데이터가 비정상성이 아닌 증거를 나타내는 경우에 적용되며, 초기 차분 단계(모델의 "통합된" 부분에 해당)를 한 번 이상 적용하여 비정상성을 제거할 수 있다. / 분기, 반기, 연간 단위로 다음 지표를 예측하거나 주간, 월간 단위로 지표를 리뷰하여 경향을 분석하는 기법** | **자기회귀누적 이동평균모형(AutoRegressive Integrated Moving Average, ARIMA모형)** |
| 대용량의 정형 및 비정형 데이터를 저장하고 손쉽게 접근할 수 있게 하는 대규모 저장소 | **데이터 호수 (Data Lake)** |
| K-익명성의 동질설 문제나 배경지식을 이용하는 문제를 해결하기 위하여 익명성을 향상시키는 방법 | **L-다양성** |
| 동질 집합에서 민감정보의 분포와 전체 데이터 집합에서의 민감정보 분포가 유사한 차이를 보이게 만드는 기법 | **T-접근성** |
| 데이터의 결측값(치)을 채우거나 이상값을 제거하여 데이터 품질을 높이는 과정 | **데이터 정제** |
| 데이터의 결측값을 처리하는 방법 중 이것은 보통 m번 대체를 수행하고  그에 따른 m개의 자료가 생성되면 이를 각각 분석하는 방법이다. | **다중 대체법** |
| 포본평균들의 표준편차 | **표준 오차** |
| 통계적 추정을 할 때 표본자료 중 모집단에 대한 정보를 주는 독립적인 자료의 수 | **자유도** |
| 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법 /  분석목적에 따라 특정 요인만 분리 분석하거나 제거하는 작업을 함 | **분해시계열** |
| 성능이 약한 학습기를 여러 개 연결하여 순차적으로 학습하여, 정답을 맞히지 못한 부분에 가중치를 부여함으로써 강한 학습기를 생성하는 앙상블 기법 / 모델의 정확성을 높이기 위해 오분류된 개체들에 가중치를 부여함으로써 새로운 분류규칙을 생성 및 반복하여 약한 분류 모델을 강한 분류모델로 변형하는 알고리즘 | **부스팅** |
| 적중확률(Y축,True Positive Rate, Sensitivity) 대 오경보확률(X축, False Positive Rate, 1- Specificity)의 그래프 /  민감도와 특이도를 이용하여 분류 모델의 수준을 면적으로 표현 하여, 모델 평가를 가시화한 도구 | **ROC 커브** |
| 중요 정보를 하나의 그래픽으로 표현하여 정보를 쉽게 이해할 수 있도록 만드는 시각화 기법 | **인포그래픽** |
| 다양한 데이터를 통합적으로 분석하여 기업 의사결정권자가 합리적인 의사결정이 가능하도록 지원하는 일련의 활동 | **BI(Business Intelligence)** |
| 누구나 열람할 수 있는 디지털 장부에 거래 내역을 투명하게 기록하고,  여러 대의 컴퓨터에 이를 복제해 저장하는 분산형 데이터 저장기술 | **블록체인** |
| 새로운 모델을 만들 시, 기존의 만들어진 모델을 사용하여 학습을 빠르게 하며, 모델 성능을 높이는 방법은? | **전이학습** |
| 인간의 간섭을 가능한 최소한으로 하여 금융 서비스나 투자 관리를 온라인으로 제공하는 투자 자문역할의 일종이다. 수리적 규칙이나 알고리즘에 기반한 디지털 금융 서비스를 제공한다. | 로보 어드바이저(Robo Advisor) |
| 정규 분포의 평균을 측정할 때 주로 사용되는 분포로 모집단의 분산(혹은 표준편차)이  알려져 있지 않은 경우에 정규분포 대신 이용하는 확률분포는? | t-분포 |
| 총 평균과 각 집단의 평균 차이에 의해 생긴 집단 간 분산 비율을 나타내는 분포 | F-분포 |