|  |  |
| --- | --- |
| **설명** | **정답** |
| 큰 용량과 복잡성으로 기존 애플리케이션이나 툴로는 다루기 어려운 데이터셋의 집합 | 빅데이터 |
| 추론과 추정의 근거를 이루는 객관적 사실 / 현실에서 관찰하거나 측정하여 수집한 사실 | 데이터 |
| 지식을 도출할 때 사용하는 데이터 / 데이터의 가공 및 데이터 간 관계를 통해 패턴을 인식 하는 것 | 정보 |
| 데이터를 통해 도출된 다양한 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험 을 결합해 내재화한 것 | 지식 |
| 지식의 축적과 아이디어가 결합된 창의적 산물 | 지혜 |
| 빅데이터의 특징인 3V는? | Volume(크기), Variety(다양성), Velocity(속도) |
| 고객의 대규모 거래데이터로부터 함께 구매가 발생하는 규칙으로 연관성 높은 상품을 추천하는 분석  어떤 변인 간에 주목할 만한 상관관계가 있는지를 찾아내는 방법 | 장바구니 분석 |
| 새로운 사건이 속할 범주를 찾아내는 일 | 유형 분석 |
| 최적화가 필요한 문제의 해결책을 자연 선택, 돌연변이 등과 같은 매커니즘을 통해 점진적으로 진화시키는 방법 | 유전 알고리즘 |
| 이것은 인공지능의 한 분야로 간주된다. 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술 을 개발하는 분야이다. /  학습 데이터로 학습한 알려진 특성을 활용해 ‘예측’하는 방법 | 머신러닝 |
| 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계 학습의 한 분야 | 딥러닝 |
| 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정해 내는 분석 방법이다. /  독립변수를 조작하면서 종속변수가 어떻게 변하는지를 보며 두 변인의 관계를 파악하는 것 | 회귀분석 |
| 특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석 | 감정분석 |
| 오피니언 리더, 즉 영향력 있는 사람을 찾아낼 수 있으며, 고객 간 소셜 관계를 파악 | 소셜 네트워크 분석 |
| 여러 사람이 공유하여 사용할 목적으로 체계화해 통합, 관리하는 데이터의 집합 /  데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합 | 데이터베이스 |
| 데이터베이스를 관리하며 응용 프로그램들이 데이터베이스를 공유하며 사용할 수 있는 환경 을 제공하는 소프트웨어 | DBMS(DataBase Management System) |
| 사용자의 의사결정에 도움을 주기 위해 수집된 대량의 비즈니스 데이터베이스 /  기업의 의사결정 과정을 지원하기 위한 주제 중심적으로 통합적이며 시간성을 가지는 비휘발성 데이터의 집합 | 데이터 웨어하우스(Data Warehouse, DW) |
| 고객과의 관계를 지속적으로 강화하기 위한 정보시스템 | CRM(Customer Relationship Management) |
| 기업이 시간과 비용을 최적화 시키기 위해 외부 공급업체와 연계하여 통합한 정보시스템 | SCM(Supply Chain Management) |
| 재무, 제조, 소매유통, 공급망, 인사 관리, 운영 전반의 비즈니스 프로세스를 자동화하고 관리하는 시스템 | ERP(Enterprise Resource Planning) |
| 기업의 목표를 달성하기 위한 성과 지표 | KPI(Key Performance Indicator) |
| 데이터의 원래 소유자인 개인이 데이터에 대한 권리를 소유하고 행사할 수 있어야 한다는 개념 | 마이 데이터 |
| 직무에 특정한 구체적인 기술이며, 실제로 업무를 수행하는 데 필요한 기술 | 하드 스킬 |
| 개인이 보유하고 있는 고유한 속성, 성격 특성 및 의사소통 역량 | 소프트 스킬 |
| 전사 차원의 모든 데이터에 대해 정책 및 지침, 표준화, 운영 조직 및 책임 등의  표준화된 관리 체계를 수립하고 운영을 위한 프레임워크 및 저장소를 구축하는 것 | 데이터 거버넌스 |
| 2020년 1월에 국회에서 통과된 이른바 데이터 3법이라 불리는 3가지? | 개인정보보호법, 정보통신망법, 신용정보법 |
| 전산시스템을 필요로 하는 곳으로부터 하청을 받아 시스템의 기획 개발, 유지보수, 운영 등 을 대신해주는 업종 | SI(System Integration, 시스템 구축) |
| 정답이 있는 데이터를 활용하여 분석 모델을 학습시키는 것 / 레이블이 범주형인 분류와 연속형인 회귀로 나뉜다. | 지도 학습 |
| 정답을 알려주지 않고 학습하는 것 /  정답 레이블이 없는 데이터를 비슷한 특징을 가진 데이터끼리 군집화하여 새로운 데이터에 대한 결과를 예측 | 비지도 학습 |
| 웹을 운영하는 주체가 누구나 사용할 수 있게 공개한 데이터를 개발자나 사용자가 수집해 사용하는 기술을 의미 | Open API(Application Programming Interface) |
| 고정된 구조로 정해진 필드에 저장된 데이터.  엑셀 스프레드시트, RDBMS(관계형 데이터베 이스), CSV 파일 형태가 대표적 | 정형 데이터 |
| 고정된 필드에 저장되어 있지는 않지만, 데이터와 메타데이터, 스키마 등을 포함하는 데이터 (XML, HTML, JSON 등). 규칙을 가지고 있어 필요 시 정형 데이터로 변형 가능 | 반정형 데이터 |
| 정해진 구조가 없고 고정된 필드에 저장되어 있지 않은 데이터 | 비정형 데이터 |
| 데이터에 개인을 식별할 수 있는 정보가 있는 경우 일부 또는 전체를 삭제하거나  일부를 대체 처리함으로써 개인을 식별할 수 없게 하는 것 | 데이터 비식별화 |
| 여러 변수의 변량을 서로 상관성이 높은 변수들의 선형 조합으로 만든 새로운 변수로 요약 및 축소하는 기법 | PCA(Principal Components Analysis) |
| 전체 데이터 중 분석에 필요한 데이터만 선택적으로 이용하는 것 | 샘플링 |
| 관계형 데이터베이스를 SQL을 사용해 CRUD(Create, Read, Update, Delete)를 수행하고 관리할 수 있는 소프트웨어 | RDBMS |
| Not Only SQL의 약자로 SQL을 사용하는 전통적인 관계형 데이터베이스 시스템보다 상대적으로 제한이 덜한 데이터 모델을 기반에 둔 분산 데이터베이스 기술 / 데이터 저장을 위 한 스키마가 필요 없으며 조인 연산을 지원하지 않는다. | NoSQL |
| 데이터 원천으로부터 데이터를 추출 및 변환하여 데이터 웨어하우스 등에 데이터를 적재하는 작업 | ETL(Extraction, Transformation and Load) |
| 동일한 데이터셋에서 일반적인 데이터 값의 범위를 벗어난 값 | 이상값 |
| 모든 독립변수 가운데 기준 통계치에 가장 많은 영향을 줄 것으로 판단되는  변수(p값, AIC 가 낮은 유의한 변수)부터 하나씩 추가하면서 모형을 선택하는 방법 | 전진 선택법 |
| 최적방정식을 선택 하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수를 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하는 방법은 무엇인가? | 후진 제거법 |
| 회귀분석에서 사용된 모형의 일부 설명 변수가 다른 설명 변수와 상관 정도가 높아  데이터 분석 시 부정적 영향을 미치는 것(회귀분석의 기본 가정인 독립성에 위배) | 다중공선성 |
| 변수를 연속적으로 추가 혹은 제거하면서 AIC가 낮아지는 모델을 찾는 방법 | 단계적 방법 |
| 데이터 학습을 위해 차원이 증가하면서 학습데이터 수가 차원의 수보다 적어져 성능이 저하 되는 현상 /  차원이 증가함에 따라(=변수의 수 증가) 모델의 성능이 하락하는 현상 | 차원의 저주 |
| 고차원에 존재하는 데이터 간의 거리를 최대한 보존하면서 데이터 간의 관계를 저차원으로 축소해 시각화하는 방법 | t-SNE(t-분포 확률적 임베딩) |
| 행렬의 크기가 다른 M\*N 행렬에 대해 세 행렬의 곱으로 분해하는 것으로 데이터 압축 등의 많은 분야에서 활용 /  M\*N 차원의 행렬 데이터에서 특잇값을 추출하여 데이터 세트를 효과적으로 축약할 수 있는 차원 축소 기법은? | 특잇값 분해(Singular Value Decomposition,SVD) |
| 기존 변수들을 조합하여 새롭게 만들어진 변수 | 파생변수 |
| 데이터를 이해하고 의미 있는 관계를 찾아내기 위해 데이터의 통곗값과 분포 등을 시각화하고 분석하는 것 | EDA(Exploratory Data Analysis) |
| 데이터의 최댓값에서 최솟값을 뺀 것으로 순서 통계량의 산포를 의미 | 범위 |
| 데이터 분포의 비대칭성을 나타내는 지표 | 왜도 |
| 데이터가 분포의 중심에 어느 정도 몰려 있는가를 측정할 때 사용하는 척도 | 첨도 |
| 다양한 문서 자료 내 비정형 텍스트 데이터에 자연어 처리(NLP) 기술 및 문서 처리 기술을 활용해  인사이트를 도출하는 기술 | 텍스트 마이닝 |
| 자연어 분석 작업의 대상이 되는 대량의 텍스트 문서들을 모아놓은 집합 | 말뭉치 |
| 구조화되어 있지 않은 문서를 단어로 나누는 과정 | 토큰화 |
| 말뭉치에서 자주 등장하지만, 분석에 있어 기여하는 바가 없는 단어 | 불용어 |
| 관심을 갖고 있는 모집단의 특성을 나타내는 대푯값 | 모수 |
| 표본을 조사하여 얻은 데이터로 표본의 특징을 수치화한 값 /  모수를 추정하기 위해 구하는 표본 값들에대한 용어 | 통계량 |
| 모집단을 어떤 특성에 따라 서로 겹치지 않는 여러 개의 층으로 분할한 후  각 층에서 표본 을 단순 무작위 추출법에 따라 추출하는 방법(계층 간 이질성, 계층 내 동질성) | 층화 표본 추출 |
| 모집단의 분포를 모르더라도 표본의 크기가 충분히 크면 표본 평균들의 분포가 정규분포에 근사하는 것 | 중심극한정리 |
| 추정량의 기대값이 모수와 같아진다면, 이 추정량을 무엇이라고 하는가? | 불편추정량 |
| 실험, 연구를 통해 기각하고자 하는 어떤 가설 | 귀무가설 |
| 실험, 연구를 통해 증명하고자 하는 새로운 아이디어 혹은 가설 | 대립가설 |
| 가설 검정에서 사용된 샘플 데이터로부터 계산된 표본 통계량.  이것으로 P-value를 계산하 며 귀무가설을 기각할 것인지 판별 | 검정통계량 |
| 통계적인 가설검정에서 사용되는 기준값으로 α로 표시 | 유의수준 |
| 귀무가설 분포에서 검정통계량보다 극단적인 값이 관측될 확률값.  이 값이 작을수록 검정통 계량이 귀무가설의 내용에 적합하지 않음을 나타낸다. | 유의 확률(P-value) |
| 귀무가설이 참일 때 귀무가설을 기각하는 오류 | 1종 오류(알파 오류) |
| 귀무가설이 거짓일 때 귀무가설을 채택하는 오류 | 2종 오류(베타 오류) |
| 대립가설이 참일 때 귀무가설을 기각하고 대립가설을 채택할 확률 | 검정력 |
| 가설검정은 귀무가설과 대립가설 중에서 하나의 가설을 양자택일한다. 그래서 1-α는 귀무가 설을 채택시키므로, 1-α의 영역을 “(1)”이라고 부르고, 반대로 α는 귀무가설을 기각(탈락)시 키므로, α의 영역을 “(2)”이라고 부른다. | (1)채택역 (2)기각역 |
| 수집한 데이터를 요약, 묘사, 설명하는 통계 기법 | 기술 통계 |
| 수집한 데이터를 바탕으로 모수에 대하여 추론 또는 예측하는 통계 기법 | 추론(추측) 통계 |
| 대규모로 저장된 데이터 속에서 분석을 통해 유의미한 패턴과 규칙을 찾아내는 과정 | 데이터 마이닝 |
| 모델이 학습 데이터를 과하게 학습하여 훈련용 데이터에 대한 성능은 높게 나오지만,  테스트 데이터에 대한 성능은 낮게 나오는 것 | 과(대)적합 |
| 모델이 너무 단순해서 학습 데이터조차 제대로 예측하지 못하는 경우 | 과소적합 |
| 재표본추출 방법의 일종으로 중복추출을 허용하는 방법 / 주어진 자료에서 단순 랜덤 복원 추출 방법을 활용하여 동일한 크기의 표본을 여러개 생성하는 샘플링 방법 | 부트스트랩 |
| 회귀분석의 5가지 기본 가정 | 선형성, 독립성, 등분 산성, 정규성, 비상관성 |
| 우리는 모집단의 실제값과 회귀선과의 차이인 (1)을 알아낼 수 없기에  표본에서 나온 관측값과 회귀선의 차이인 (2)를 이용해 분석을 수행한다. | (1)오차 (2)잔차 |
| 전체 변동 중 회귀모형에 의해 설명되는 변동의 비율로, 표본에 의해 추정된 회귀식이 자료를 잘 설명하는지를  보여주는 값 / 주어진 데이터에 회귀선이 적합 정도를 평가하는 척도이자 독립변수들이 종속변수를 얼마나 잘 설명하는지 보여 주는 지표 | 결정계수 |
| 인공신경망은 노드에 입력되는 값을 바로 다음 노드로 전달하지 않고 비선형 함수에 통과시킨 후  전달하는데 이때 사용되는 비선형 함수를 무엇이라고 하는가? | 활성화 함수 |
| 코호넨 맵이라고도 불리며, 인공신경망을 기반으로 차원축소와 군집화를 동시에 수행할 수 있는 알고리즘은? / 비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화하는 클러스터링 방법은 무엇인가? | 자기조직화지도 (SOM, Self Organizing Maps) |
| 이진 분류에서 모형이 예측한 값과 실제 값의 조합을 교차표 형태로 정리한 행렬 | 혼동행렬 |
| 전체 데이터에서 올바르게 분류한 데이터의 비율 | 정확도 |
| Positive로 예측한 것 중에서 실제 값이 Positive인 비율 | 정밀도 |
| 실제 Positive인 값 중 Positive로 분류한 비율 | 재현율, 민감도, 참 긍정률 |
| 실제 Negative인 값 중 Negative로 분류한 비율 | 특이도, 참 부정률 |
| 실제 Negative인 값 중 Positive로 잘못 분류한 비율 | 거짓 긍정률 |
| 정밀도와 재현율의 조화평균으로, 정밀도와 재현율 중 한쪽만 클 때보다 두 값이 골고르 클 때 큰 값이 된다. | F1-스코어 |
| 가장 단순한 종류의 교차검증 방법으로 데이터를 랜덤으로 추출해 학습 데이터와 테스트 데이터로 나누는 것 /  모형 평가방법 중 주어진 원천 데이터를 랜덤하게 두 분류로 분리하여 교차 검정을 실시하는 방법으로,  하나는 모형의 학습 및 구축을 위한 훈련용 자료로 다른 하나는 평가를 위한 검증용 자료로 사용하는 방법은 무엇인가? | 홀드아웃 |
| 빅데이터 분석에 경제성을 제공해준 기술은? / 인터넷상의 서버에서 다른 물리적인 위치에 존재하는 컴퓨팅 자원을 가상화 기술을 통해 서비스를 제공하는 컴퓨팅 기술은? | 클라우드 컴퓨팅 |
| 데이터로부터 의미 있는 정보를 추출해 내는 학문은? | 데이터 사이언스 |
| 데이터에 관한 구조화된 데이터 / 어떤 목적을 가지고 만들어진 데이터 / 데이터 그 자체가 아니라, 자료의 속성, 구조 등을 설명하는 데이터 | 메타데이터 |
| 합리적 의사결정을 방해하는 요소로 표현 방식 및 발표자에 따라 판단을 달리하는 것은? | 프레이밍 효과 |
| 분석용 데이터를 이용한 가설 설정을 통해 통계모델을 만들거나 기계학습을 이용한 모델을 만드는 과정 | 모델링 |
| 상향식 접근 방식의 발산단계와 하향식 접근 방식의 수렴단계를 반복하는 과제 발굴 방법 | 디자인 사고 (Design Thinking) |
| 작은 규모의 데이터 웨어하우스는? / 데이터 웨어하우스에서 추출한 데이터를 특정 주제영역으로 분석 후 그 결과를 조직이나 팀에서 활용하도록 제공한 데이터 | 데이터 마트 (Data Mart) |
| 평균으로부터 K\*표준편차만큼 떨어져 있는 값들을 이상값으로 판단하는 방법 | ESD(Extreme Studentized Deviation) |
| 주어진 시간 또는 영역에서 어떤 사건의 발생 횟수를 나타내는 확률 분포는? | 포아송분포 |
| 비선형적인 관계도 파악할 수 있는 상관계수는? | 스피어만 상관계수 |
| 회귀모형의 계수를 추정하는 방법으로써 잔차제곱합을 최소화하는 계수를 찾는 방법은? | 최소제곱법 |
| 회귀모형의 잔차항이 정규분포를 이뤄야 한다는 가정은? /  미래는 확률적으로 과거와 동일 하다는 것을 의미하는 시계열 용어는? | 정상성 |
| 현 시점의 자료값에서 전 시점의 자료를 빼는 방법은? | 차분 |
| 동일한 상대적 거리를 가진 실수 공간의 점들로 대상들을 배치시키는 방법은? / 개체들 사 이의 유사성, 비유사성을 측정하여 2차원 또는 3차원 공간상에 점으로 표현하여 개체들 사 이의 집단화를 시각적으로 표현하는 분석 방법 | 다차원 척도법 |
| 의사결정 나무에서 더 이상 분기되지 않도록 하는 규칙은? | 정지규칙 |
| 여러 개의 결정 트리 분류기(같은 알고리즘이 여러 개)가 전체 데이터에서 배깅 방식으로 각자의 데이터를 샘플링해 개별적으로 학습을 한 뒤 최종적으로 모든 분류기가 보팅을 통해 예측을 결정하는 앙상블 알고리즘은? | 랜덤 포레스트 |
| 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 |
| 모델의 파라미터값을 측정하기 위해 알고리즘 구현 과정에서 사용, 주로 알고리즘 사용자에 의해 결정, 경험에 의해 결정 가능한 값이며 모델 성능 향상을 위해 사용자가 직접 설정해 주는 조절해주는 값은? | 하이퍼 파라미터 |
| 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) |
| 인터넷상에서 제공되는 다양한 웹 사이트로부터 소셜 네트워크 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠 수집 기술 | 크롤링 |
| 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 |
| 우리는 모집단을 조사하기 위해 추출한 모집단의 일부 원소를 이용한다.  통계자료의 획득 방법 중 모집단을 조사하기 위해 추출한 집단을 무엇이라 하는가? | 표본 집단, 샘플 |
| 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 |
| 각종 사물에 센서와 통신 기능을 내장하여 인터넷에 연결하는 기술.  즉, 무선 통신을 통해 각종 사물을 연결하는 기술을 의미 | 사물인터넷(Internet of Things, IoT) |
| 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 현시점의 자료를 p시점 전의 과거 자료를 통해 설명할 수 있는 모델이다. 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, 그래프 |
| 과대 적합을 방지하기 위해 인공지능 학습 과정에서 일부 신경망 일부만 동작하고 일부는 동작하지 않도록 하는 방법 | 드롭아웃 |
| 중요 정보를 하나의 그래픽으로 표현하여 정보를 쉽게 이해할 수 있도록 만드는 시각화 기법 | 인포그래픽 |
| 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법 | 워드 클라우드 (Word Cloud) |
| 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 |