실기단답_encoding_utf

title	answer
모델의 파라미터값을 측정하기 위해 알고리즘 구현 과정에서 사용, 주로 알고리즘 사용자에 의해 결정, 경험에 의해 결정 가능한 값이며 모델 성능 향상을 위해 조절해주는 값은?	하이퍼파라미터
신경망 모형의 활성함수 종류중 하나로 다음의 식으로 표현되는 활성함수는? y= 1/(1+exp(-z))	시그모이드함수
주어진 자료에서 단순랜덤 복원추출 방법을 활용하여 동일한 크기의 표본을 여러개 생성하는 샘플링 방법은?	부트스트랩
배깅의 개념과 속성의 임의선택을 결합한 앙상블 기법 § 예측변수들을 임의로 추출하고 추출된 변수 내에서 최적의 분할 을 만들어나가는 방법을 사용하는 방식의 모델은?	랜덤포레스트
L1 정규화를 위해 가중치의 절댓값과 동일한 페널티를 추가 제약조건으로 사용하는 임베디드 기법은?	라쏘
일련의 개체 또는 사건들 간의 규칙을 발견하기 위해 사용되는 대표적인 정형 데이터 마이닝 기법은?	연관성 분석
긍정으로 예측한 범주에서 실제로 긍정인 비율을 무엇이라고 하는가?	정밀도
인공신경망의 한 종류, 코호넨 네트워크에 근간을 두고있으며 차원축소와 군집화를 동시에 수행하는 기법은?	SOM
여러개의 분류모형에 의한 결과를 종합하여 분류의 정확도를 높이는 방법은?	앙상블
독립변수들 간에 강한 상관관계가 나타나서, 회귀분석의 전제가정인 독립변수들 간에 상관관계가 높으면 안된다는 조건을 위배하는 경우를 의미하는 용어는?	다중공선성
연관규칙의 측정지표중 하나로 품목 A가 포함된 거래 중에서 품목 A,B를 동시에 포함하는 거래일 확률은 어느정도인가를 나타내는 지표는?	신뢰도(confidence)
데이터를 저장하는 스토리지가 메인 메모리를 이용하는 방식의 데이터베이스 관리 시스템이며, 상대적으로 속도가 높은 성능을 보이는 것은?	인메모리 데이터 베이스
분해시계열 분석에서 고정된 주기를 가지고 자료가 변화하는 요인은?	계절요인
수집 대상데이터를 추출, 정제 후 데이터 웨어 하우스나 데이터 마트에 저장하는 기술을 의미하는 용어는?	ETL
기업 또는 기관의 전사 차원에서 식별된 다양한 분석과제를 대상으로 제한된 예산과 자원을 효과적으로 수행하기 위해 우선순위를 평가하고 평가 결과에 따른 단계별 군현 로드맵을 수립하는 실행 계획	분석마스터 플랜
개인과 집단들 간의 관계를 노드와 링크로서 모델링해 그것의 위상구조와 확산 및 진화과정을 계량저긍로 분석하는 방법론은?	사회연결망 분석(SNA)
M*N 차원의 행렬데이터에서 특이값을 추출하고 이를 통해 주어진 데이터 세트를 효과적으로 축약할 수 있는 차원 축소기법은?	특이값 분해
교차검증의 대표적인 기법으로 데이터를 k개의 하부 집합으로 나누고 k번째의 하부집합을 검증, 나머지 k-1개의 하부집합을 훈련용 자료로 사용하는 방법	k-fold 교차검증
의사결정나무 구조는 나무의 가지를 생성하는 가지분할의 과정과 생성된 가지를 잘라내어 모형을 단순화하는 ( ) 이 과정으로 구성되어있다.	가지치기(pruning)
서로 다른 분류에 속한 데이터 간에 간격이 최대가 되는 선을 찾아 이를 초평면이라는 기준으로 하고 데이터를 분류하는 모델?	svm
주어진 데이터에서 마진을 최대화 하는 초평면을 구하는 방법으로 학습하는 알고리즘은?	svm
자연어 처리 기술을 이용해 인간의 언어로 쓰인 비정형 텍스트에서 유용한 정보를 추출하거나 다른 데이터와의 연관성을 파악하기 위한 방법	텍스트마이닝
대규모 분산 시스템 모니터링 및 agent와 collector 구성을 통해 하둡 분산 시스템으로 데이터를 수집하는 방식은?	척와
최소 지지도 보다 큰 집합만을 대상으로 높은 지지도를 갖는 품목집합을 찾는 연관분석 알고리즘으로 다음의 절차를 가지는 알고리즘의 이름은? 1. 최소 지지도 설정 2. 개별 품목 중에서 최소 지지도를 넘는 모든 품목찾음 3. 2에서 찾은 개별 품목만을 이용하여	apriori
( ) 은/는 필수 데이터가 입력이 안 되고 누락된 값을 의미하며 적재과정에서 임의 누락 또는 의도하지않은 결실등에 의해 발생한다	결측값
인터넷상에서 제공되는 다양한 웹 사이트로부터 소셜 네트워크 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠 수집 기술	크롤링
원하는 군집수만큼 초기값을 지정하고, 각 개체를 가까운 초기값에 할당하여 군집을 형성한 뒤, 각 군집의 평균을 재계산하여 초기값을 갱신하는 군집방법은?	k-평균 군집
네트워크를 통해 센서 데이터 및 오디오, 비디오 등의 미디어 데이터를 실시간으로 수집하는 기술	스트리밍
자연어 처리의 한 과정으로 접속사, 대명사 등을 제거해주고, 공통 어간을 가지는 단어를 묶기 위해 해주는 전처리는?	stemming
사후확률의계산시조건부독립을가정하여계산을단순화한방 법으로, 사후확률이 큰 집단으로 새로운 데이터를 분류하는 모델은?	나이브 베이즈 분류모형
다음 중 불균형 데이터 해결을 위한 샘플링 방법으로 적절하지 않은 것은?	홀드아웃
불순도의 측도로 출력변수가 범주형일 경우 지니지수를 이용, 연속형인 경우 분산을 이용한 이진분리를 사용하는 의사결정나무 알고리즘은?	CART
오버샘플링의 한 방식으로 데이터의 개수가 적은 클래스의 표본을 가져온 뒤 임의의 값을 추가하여 새로운 샘플을 만드는 데이터 불균형 해소 방법은?	SMOTE
붓스트랩 표본을 구성하는 재표본 과정 에서 각 자료에 동일한 확률을 부여하는 것이 아니라, 분류가 잘 못된 데이터에 더 큰 가중을 주어 표본을 추출하는 앙상블 모형은?	부스팅
문장에서 사용된 단어의 긍정과 부정여부에 따라 얼마나 긍정적인 단어가 많은지 소스를 부여해 긍정 문장인지 평가하는 데이터 마이닝의 한종류는?	감성분석
데이터 속성인 메타데이터를 가지며, 일반적으로 스토리지에 저장되는 데이터파일이며 HTML,XML,JSON등의 종류를 가지는 데이터 유형은?	반정형데이터
인터넷상에서 제공되는 다양한 웹 사이트로부터 소셜 네트워크 정보, 뉴스, 게시판 등의 웹 문서 및 콘텐츠 수집 기술	크롤링
비지도 신경망으로 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬하여 지도의 형태로 형상화하는 군집분석방법의 한종류는?	SOM
데이터 안에 관찰할 수 없는 잠재적인 변수(Latent Variable)가 존재한다고 가정하는 차원축소기법, 모형을 세운 뒤 관찰 가능한 데이터를 이용하여 해당 잠재 요인을 도출하고 데이터 안의 구조를 해석하는 기법은?	요인분석
커넥터(Connector)를 사용하여 관계형 데이터베이스(RDB)와 하둡 간 데이터 전송하는 기술	스쿱
GMM(Gaussian Mixture Model) 군집분석이 모수를 학습하는 방법은?	EM알고리즘
군집분석의 한 방법으로 가장 유사한 개체를 묶어 나가느 ㄴ과정을 반복하여 원하는 개수의 군집을 형성, 보통 덴드로그램의 형태로 결과가 주어지는 방식은?	계층적 군집
시계열모형중 과거시점의 관측자료와 과거시점의 백색잡음의 선형결합으로 현시점의 자료를 표현하는모형은?	자기회귀 이동평균모형
변수간의 상호 작용을 감지할 수 있도록 변수 일부만을 모델링에 사용한 후, 그결과를 평가하는 작업을 반복수행하며 변수를 증가, 감소 할지 결정하는 변수 선택기법은?	래퍼 기법
대규모 분산 시스템 모니터링을 위해 에이전트(Agent)와 컬렉터(Colector) 구성을 통해 데이터를 수집하고, 수집된 데이터를 하둡 파일 시스템(HDFS)에 저장하는 기능을 제공하는 데이터 수집기술	척와
연관규칙의 측정지표중 하나로 전체 거래중에서 두 품목이 동시에 포함되는 거래의 비율을 나타내는말은?	지지도(support)
혼합분포에서의 모수 추정은 단일 분포의 경우와는 달리 가능도함수에 기초한 최대가능도 추정이 쉽지않다 혼합분포에 대한 최대 가능도 추정을 위해 사용 되는 알고리즘은?	EM 알고리즘
사회 연결망 분석 기법 중 중심성을 측정하는 방법중의 하나로 간접적으로 연결된 모든 노드 간의 거리를 합산해 중심성을 측정하는 방법은?	근접 중심성
시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법. 회귀분석적인 방법을 주로 사용하는 것을 의미 하는것은?	분해 시계열
정지규칙과 가지치기로 모델을 학습하며 다수의 독립변수 중에서 종속변수에 큰 영향을 미치는 변수를 탐색하는 기장 기본적인 모델은?	의사결정나무
데이터 정제 기술중 하나로 데이터를 정제 규칙적용하기 위해 최소 단위로 분할하는 작업을 의미하는 용어는?	파싱
데이터를 분리하는 초평면(Hyperlane) 중에서 데이터들과 거리가 가장 먼 초평면을 선택하여 분리하는 지도 학습 기반의 이진 선형 분류 모델은?	svm
여러 개의 모형을 결합하여 개별 모형보다 좋은 예측성능을 얻는 분석 기법을 표현한 말은?	앙상블
적정 수준의 학습이 부족하여 실제 성능이 떨어지는 현상이며 데이터 수집 시 단편화된 방법으로 인한 학습 부족 현상을 나타내는 용어는?	과소적합
모집단을 성격에 따라 집단으로 분류하고 각 집 단 내에서 원하는 크기의 표본을 무작위 추출	층화 추출
정상성을 만족하는 요인으로 평균값과 분산값은 시간 t에 상관없이 일정해야하며, 공분산은 시간에 의존하지 않고 오직 ( )에만 의존한다. 괄호에 들어갈 단어는?	, 시차
컴퓨터 단위를 나타낸다. 괄호안에 들어갈 단어는? 페타바이트 < 엑사바이트 < 제타바이트 < ()바이트	요타
다중공선성을 측정하는 지표로 알려져있으며, 독립 변수간 상관 관계가 있는지 정량적으로 나타내는 용어는?	분산팽창요인
실제값이 False인 관측치 중 예측치가 적중한 정도를 나타내는 분류모델의 평가방식은?	특이도(specificity)
모든 변수가 연속형인 경우에 데이터 좌표간 거리를 구하는 방식중 하나로 좌표간 거리차이의 절댓값의 총합을 나타내는 표현은?	맨하튼거리
인터넷상의 서버에서 데이터 저장, 처리, 네트워크, 콘텐츠 사용 등 서로 다른 물리적인 위치에 존재하는 컴퓨팅 자원을 가상화 기술을 통해 IT 관련 서비스를 한 번에 제공하는 혁신적인 컴퓨팅 기술을 의미하는 단어는?	클라우드 컴퓨팅
표본 추출 방법 중 질적인 원소들로 구성된 모집단에서 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법(유사한 원소끼리 층으로 나누어 각 층에서 랜덤 추출)	층화추출법
랜덤 모델과 비교하여 해당 모델의 성과가 얼마나 향상되었는지를 각 등급별로 파악하는 그래프는?	향상도곡선(lift curve)
수집 대상 데이터를 추출, 가공하여 데이터 웨어하우스 및 데이터마트에 저장하는 기술을 나타내는 단어는?	ETL
합리적의사방해 요소로서 문제의 표현방식에 따라 동일한 사건이나 상황임에도 불구하고 개인의 판단이나 선택이 달라질 수 있는 현상은?	프레이밍 효과
x의 편차와 y의 편차의 곱의 기댓값이며 두 변수 간의 선형관계에 대한 방향을 나타내는 용어는?	, 공분산
다차원의 데이터를 대화식으로 분석하기 위한 소프트웨어를 나타내는 용어는?	OLAP
빅데이터 저장 기술로 RDBMS와 다른 DBMS를 지칭하기 위한 용어로, 고정된 테이블 스키마가 필요하지 않으며, JOIN 연산을 사용할 수 없는 수평적 확장가능한 DBMS는?	NoSQL
원 데이터 집합으로 부터 크기가 같은 표본을 여러번 단순 임의 복원추출하여 각 표본에 대해 분류기를 생성한 후 그 결과를 앙상블하는 방법은?	배깅
단순분류, 확인 목적으로 숫자를 부여. 숫자 자체로서 가지는 의미는 없는 변수를 나타내는 용어는?	명목척도
귀무가설이 참인데 잘못하여 이를 기각하게 되는 오류는?	제 종 오류
변수의 차원을 감소하는 다변량분석 방법중 하나로 크로스칼의 스트레스 값을 통해 검증하며 개체들 사이의 유사성/비유사성을 측정하여 2차원 또는 3차원 공 간상에 표현하는방법은?	다차원척도법
ROC그래프의 x축에는 FP ratio(1-특이도)를 나타내며 y축에는 ( ) 를 나타낸다. 빈칸에 들어갈 단어는?	민감도
빅데이터 저장 기술로 컴퓨터 네트워크를 통해 공유하는 여러 호스트 컴퓨터의 파일에 접근할 수 있게 하는 파일 시스템은?	분산파일시스템
사회 연결망 분석 기법 중 중심성을 측정하는 방법중의 하나로 한노드에 직접적으로 연결된 노드들의 합으로 연산하는방법은?	연결정도 중심성
이상값 검출방식중 하나로 오름차순으로 정렬된 데이터에서 범위에 대한 관측치 간 차이의 비율을 이용하여 이상값으로 판단하는 방법은?	딕슨의 Q검정
개체들 사이의 유사성, 비유사성을 측정하여 2차원 또는 3차원 공간상에 점으로 표현하여 개체들 사이의 집단화를 시각적으로 표현하는 분석 방법	다차원 척도법
최소 지지도보다 큰 집합만을 대상으로 높은 지지도를 갖는 품목 집합을 찾는 연관규칙을 찾는 대표적인 방법은?	apriori
원격지 시스템 간에 파일을 공유하기 위한 서버-클라이언트 모델로 TCP/IP 기반으로 파일을 송-수신하는 응용계층 통신 프로토콜	FTP
1-특이도로 계산되며, 실제로 부정인 범주에서 긍정으로 잘못예측한 비율을 무엇이라고 하는가?	FPR
표준편차를 평균으로 나눈 것으로, 자료의 측정치 단 위가 서로 다르거나 관찰점 수가 다른 경우 표준편차 값을 서로 비 교하기 어려워지는 문제를 해결하기 위해 사용. 단위가 다른 두 집 단 간의 산포를 비교할 때 가장 적합한 측도는 ?	변동계수
코호넨 네트워크라고도 불리는 인공신경망 일종의 비지도학습 방법이며, 클러스터링, 고차원시각화, 데이터압축, 특성추출 등의 용도로 사용되는 기법은?	자기 조직화지도 SOM
잔차가 정규분포를 잘 따르고 있는지를 확인하는 그래프를 나타내는 용어로 잔차들이 그래프 선상에 있어야 이상적임을 나타낸다. 이는 무엇인가?	q-q
스트리밍 데이터 흐름(Data Flow)을 비동기 방식으로 처리하는 분산형 로그 수집 기술	플럼

빅데이터 저장 기술중 하나로 RDBMS와 구분된 DBMS를 지칭하기 위한 용어로, 고정된 테이블 스키마가 필요하지 않으며, JOIN 연산을 사용할 수 없는 수평적 확장가능한 DBMS를 뜻하는 단어는?	NoSQL
비선형관계의 상관정도나 이산형 변수에 관한 상관정도를 표현하는 상관계수는?	스피어판
모델 검증방법중 하나로 일반적으로 전체 데이터 중 70%의 데이터는 훈련용 자료로 사용하고 나머지는 검증용 자료로 사용하는 방식은?	홀드아웃(hold-out)
회귀모형 변수 선택시 개선도가 높은 변수를 차례로 추가는 방법은?	전진선택법
지식의 피라미드 단계중 지식의 축적과 아이디어가 결합된 창의적산물을 지칭하는 단계는?	지혜
모형기반의 군집 방법으로 데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정하에서 모수와 함께 가중치를 자료로부터 추정하는 방법을 사용하는 군집방식은?	혼합분포군집
센서로 부터 수집 및 생성된 데이터를 네트워크를 통해 수집 및 활용	센싱
임계값을 기준으로 활성화되거나 혹은 비활성화되는 형태의 활성화 함수로0또는 1로 출력는 함수의 이름은 무엇인가?	계단함수
언어 분석이 가능한 텍스트 데이터, 형태와 구조가 복잡한 이미지 동영상 같은 멀티미디어 데이터를 의미하는 데이터 유형의 종류는?	비정형데이터
회귀 모형의 평가 방법중 하나로 회귀제곱합을 총제곱합의 값으로 나눈 것은?	결정계수
복원 랜덤샘플링 방식을 통해 표본을 추출하여 모델을 학습시키고 집계하는 앙상블 방식은 무엇인가	배깅
독립변수들 간에 강한 상관관계가 나타나서, 회귀분석의 전제가정인 독립변수들 간에 상관관계가 높으면 안된다는 조건을 위배하는 경우를 의미하는 용어는?	다중공선성
실제값이 True인 관측치 중 예측치가 적중한 정도를 나타내는 분류모델의 평가방식은?	민감도(sensitivity)
의사결정나무분석을 위한 알고리즘중 CHAID 방식의 이산형 목표변수 선택법은 무엇인가?	카이제곱통계량
텍스트마이닝의 전처리 과정 중에서 변형된 단어형태에서 접사 등을 제거하고 그 단어의 원형 또는 어간을 찾아내는 것	스태밍
통계의 표본 분포의 표준 편차를 의미하는 단어는?	표준 오차
로지스틱회귀모형의 해석방법중 하나로 특정변수가 한단위 증가할때 성공하는 확률을 표현하는 단위는?	오즈
빅데이터 저장 기술로 관계형 데이터베이스 관리 시스템으로 하나의 데이터베이스를 여러 개의 서버상에 구축하는 시스템은?	데이터베이스 클러스터
결측치의 대체를 하는 방법중의 하나로 회귀 분석을 실시한 결과로 얻은 추정치를 결측값의 대체값으로 사용하는 방법은?	회귀 대체법
계측정 군집 수행시 거리측정 방법중 하나로, 모든 항목에 대한 거리 평균을 구하면서 군집화를 하는 방식은?	평균연결법
구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작, 발표한 소프트웨어 프레임워크이며 대용량 데이터를 신뢰도가 낮은 컴퓨터로 구성된 클러스터 환경에서 병렬 처리를 지원하기 위해서 개발된 것은?	맵 리듀스
이것은 데이터베이스의 구조와 제약조건에 관한 전반적인 명세를 의미하는 것으로서, 데이터베이스를 구성하는 데이터 개체(Entity), 속성(Attribute), 관계(Relationship) 및 데이터 조작 시 데이터 값들이 갖는 제약 조건 등에 관해 전반적으로 정의한다	스키마
실시간에 근접하게 대량 로그 데이터를 수집하고 처리하는 기술은?	플럼
센서 데이터, HTTP 트랜잭션, 알람 등과 같이 네트워크를 통해서 실시간으로 전송되는 데이터를 칭하는 말	스트림 데이터
사용자의 요구에따라정보를 처리해주고 데이터베이스를 관리해주는 소프트웨어는?	DBMS
다수의 서버로부터 실시간으로 스트리밍되는 로그 데이터를 수집하여 분산 시스템에 데이터를 저장하는 대용량 실시간 로그 수집 기술	스크라이브(scribe)
회귀모형 변수 선택시 도움이 되지 않는 변수들을 하나씩 제거하는 방법은?	후진제거법
항목들 간의 '조건-결과'식으로 표현되는 유용한 패턴, 규칙을 발견해내는 분석종류의 방법으로 장바구니 분석이라고도 부르는 이 방법은?	연관분석
사회 연결망 분석 기법 중 중심성을 측정하는 방법중의 하나로 네트 워크 내에서 한 노드가 담당하는 매개자 혹은 중재자 역할의 정도로 중심성을 측정하는 방법은?	매개중심성
시계열 분석에서 시점 t와 바로 전 시점인 t-1간의 상관관계를 의미하는 것은?	자기 상관
데이터 마이닝의 절차 중 데이터의 정제, 통합, 선택, 변환의 과정을 거친 구조화된 단계 로서 더 이상 추가적 절차 없이 데이터 마이닝 알고리즘 실험에서 활용가능한 상태를 나태내는 말은?	corpus
동시에 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합을 의미하는 용어는?	데이터베이스
기존의 변수를 조합해 만든 새로운 변수로 특정 의미를 갖는 작위적 의미의 변수는?	파생변수
대용량 실시간 로그 처리를 위해 기존 메시징 시스템과 유사하게 레코드 스트림을 발행, 구독하는 방식의 분산 스트리밍 플랫폼 기술	아파치 카프카
x축을 1- 특이도 y축을 민감도로 그려지는 분류모형 평가용 그래프는?1-	roc
데이터에 포함된 개인 식별 정보를 삭제하거나 알아볼 수 없는 형태로 변환하는 것을 의미하는 단어는?	데이터 익명화
참긍정률이라고하며, 실제로 긍정인 범주에서 긍정으로 예측한 비율을 무엇이라고 하나.	민감도
부트스트랩 표본을 구성하는 재표본 과정에서 각 자료에 동일한 확률을 부여하는 것이 아니라, 분류가 잘못된 데이터에 더 큰가중을 주어 표본을 추출하는 방법은?	부스팅
데이터를 크기가 있는 값들로 이뤄진 자료를 순서대로 나열했을 때 백분율로 나타낸 특정 위치의 값을 이르는 척도에 따라 구분하고 극단적인 이상치를 제거하는 방법은?	사분위수 기법
자연계의 유전자 진화 과정을 활용하여 선택, 교배,변이,평가를 반복하여 최적화 문제를 해결하는 기법은?	유전자 알고리즘
문장에서 사용된 단어의 긍정과 부정여부에 따라 얼마나 긍정적인 단어가 많은지 소스를 부여해 긍정 문장인지 평가하는 데이터 마이닝의 한종류는?	감성분석
입력받는 값을 출력으로 0~1사이의 값으로 모두 정규화하여 출력값의 총합은 항상 1이되는 특성을 가진 활성화 함수는?	소프트 맥스 함수
변수를 공분산 행렬이나 상관행렬을 이용하여 원래 데이터 특징을 잘 설명해주는 성분을 추출하기 위하여 고차원 공간의 표본들을 선형 연관성이 없는 저차원의 공간으로 변환하는 기법	PCA
임의개의 소집단의 중심좌표를 이용하여 각 객체와 중심좌표간의 거리를 산출하고, 가장 근접한 소집단에 배정한 후 해당 소집단의 중심좌표를 업데이트하는 방식으로 군집화하는 방식	k-평균 군집화
RDBMS의 고정된 필드에 저장될 수 있으며, 데이터 스키마 지원을 하는 데이터 유형은?	정형데이터
프라이버시보호모델 중 특정인임을 추론할 수 있는지 여부를 검토, 일정 확률수준 이상 비식별되도록 하는 기법은?	k-익명성