Senior Data Scientist 필기 예제

1. (탐색적 데이터 분석과 확률분석) 일변량 분포의 왜도(skewness)에 대한 정보를 주지 않는 것은?

① 평균(mean) ② 중위수(median) ③ 상자그림(box plot) ④ 히스토그램(histogram)

2. (가설검정) 다음 중 ( )에 들어갈 알맞은 용어는?

(A)이란 주어진 x값에 대한 y 평균값의 구간 추정치를 말하고, (B)이란 주어진 x값에 대한 개별 y값의 구간 추정치를 말한다."

① A: 신뢰구간, B: 예측구간 ② A: 예측구간 B: 신뢰구간

③ A: 검정구간, B: 추정구간 ④ A: 추정구간, B: 검정구간

3. (가설검정) 정규 모집단의 분산에 대한 두 개의 독립적인 추정치 간의 비율에 기초한 분포로서 k개의 평균의 동일성을 검정하는 데 사용되는 것은?

① t분포 ② 정규분포 ③ 이항분포 ④ F분포

4. (NoSQL) Eventual consistency의 특징이 아닌 것을 고르시오.

① Replication이 사용될 때, 서로 다른 replica라도 write의 적용 순서는 동일

② 새로운 갱신 요청이 없으면, 모든 노드에 있는 데이터는 궁극적으로 일관성 유지

③ 어떤 작업은 stale data를 읽을 수 있음

④ 실시간 갱신 요청은 아주 적고, 대부분이 읽기 요청인 응용에 적합

5. (Spark) 다음은 Spark에서 Application, Job, Stage, Task에 관한 설명이다.

(ㄱ) 하나의 Application은 사용자 main function을 수행하는 Driver process를 포함하고 있으며, 여러 개의 Job을 생성할 수 있다.

(ㄴ) 한 Job은 Action이나 데이터 저장으로 끝난다.

(ㄷ) Shuffle을 야기시키는 Transformation들로는 repartition(), join(), count() 등이 있다.

(ㄹ) Task는 스케줄링의 가장 작은 단위다.

올바른 것을 모두 찾은 것을 고르시오.

① (ㄱ), (ㄴ), (ㄷ) ② (ㄱ), (ㄴ), (ㄹ) ③ (ㄱ), (ㄷ), (ㄹ) ④ (ㄴ), (ㄷ), (ㄹ)

6. (Data Cleaning) MySQL 등과 같은 관계데이터베이스(Relational DB)의 테이블에서는, 데이터 비일관성(inconsistency) 문제가 나타나는 일은 없다.

7. (Data Reduction) 전처리 과정에서 데이터 Reduction(축소)에 대한 설명으로 적합하지 않은 것을 고르시오.

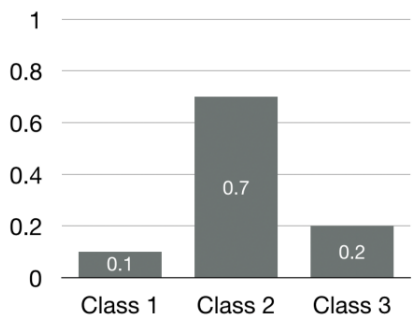
① 데이터 전체를 저장하는 대신, 데이터 분포를 가정하고 그 분포의 매개변수만을 저장하는 데이터 reduction 방법들도 있다.

② 데이터 큐브(Cube)는 미리 계산해 둔 집계(aggregated)값들을 관리하고 있다.

③ PCA에서는 분산이 작은 성분(component)을 우선적으로 선택한다.

④ 데이터가 특정 클래스에 과도하게 많거나 작을 경우, 인위적으로 특정 클래스의 객체들을 증가 또는 감소 시키기도 한다.

8. (Feature Selection) 분류(Classification)을 위해 사용한 의사결정트리(Decision Tree)의 한 노드 내부에서, 이 노드에 해당하는 데이터들 사이에 다음과 같은 클래스 분포가 존재한다. 이 노드의 엔트로피 값은? 계산 편의를 위해 주어진 log함수 근사값 표를 이용하시오.



x	log(x)
0.1	-3.3
0.2	-2.3
0.3	-1.7
0.4	-1.3
0.5	-1
0.6	-0.7
0.7	-0.5
0.8	-0.3
0.9	-0.1
1.0	0

① 1.14 ② -1.14 ③ 0.35 ④ -0.35

9. (Feature Selection) 공분산행렬(Covariance Matrix)의 고유벡터(Eigenvector)가 데이터의 주성분(Principal Component)과 방향이 같은 이유는 무엇인가?

① 임의의 벡터와 공분산행렬 고유벡터(Eigenvector)의 외적이 데이터 내에서 가장 큰 분산 방향을 가리키기 때문에

② 임의의 벡터에 공분산행렬을 계속 곱하면 데이터 내 가장 큰 분산 방향으로 벡터가 회전하는데 고유벡터는 이 회전의 수렴 방향이므로

③ 공분산행렬의 고유벡터는 분산이 가장 큰 피처에 해당하는 차원 방향이므로

④ 공분산행렬과 어떤 임의의 벡터를 곱해도 그 결과가 데이터의 주성분 방향을 가리키기 때문에

10. (Feature Selection) 주성분분석(PCA, Principal Component Analysis)에 대한 설명으로 옳지 않은 것은?

① 공분산행렬(Covariance Matrix)의 고유벡터(Eigenvector)를 구해야 한다.

② 각각의 주성분은 데이터에서 분산(Variance)이 큰 방향을 가리킨다.

③ 피처 추출(Feture Extraction)에 사용할 수 있다.

④ 데이터에 존재하는 주요 분산(Variance)의 방향은 구할 수 있지만, 해당 방향으로 데이터가 얼마나 실제로 분산되어 있는지는 알 수 없다.

11. (Tree Model) Decision Tree를 구성하는 방식 중에 Depth First (DF) 또는 Breadth First (BF) 방식에 대한 설명으로 틀린 것을 고르시오

① DF 방식은 Recursive 방식으로 Training Data를 분할한다

② DF 방식에서는 모든 데이터를 Memory에 저장할 필요가 없어서, 메모리 측면에서 효율적이다

③ DF 방식은 적은 Training Data에 적합하다

④ BF 방식은 Tree Level 별로 데이터를 처리하면서 진행할 수 있다

12. (Tree Model) Ensemble Classifier를 사용하는 목적은 무엇인가?

① Stable Classifier 의 성능을 극대화 한다

② Unstable Classifier의 성능을 향상 시킨다

③ 대규모 Training Data를 생성시킬 수 있다

④ Classifier 의 동작 속도를 향상 시킨다

13. (Tree Model) Boosting에 대한 설명으로 틀린 것을 고르시오

① 분류가 틀린 Data에 대한 가중치를 높인다

② 여러 개의 Classifier 들을 만든 후 그로부터 Voting으로 최종 결정을 한다

③ 분류가 맞는 Data에 대한 가중치를 증가시켜 정확도를 더 높인다

④ 대표적인 예로 AdaBoost 기법이 있다

14. (Recommendation) Collaboration Filtering (CF)에 대한 설명으로 틀린 것을 고르시오

① CF의 Objective Function에 일반성(Generalization)을 높이기 위해 Regularization Term을 추가한다

② CF은 Explicit Feedback 뿐만 아니라 Implicit Feedback도 처리할 수 있다

③ CF Objective Function을 최적화하기 위한 방식으로 ALS (Alternating Least Squares)을 실행할 수 없다

④ 사용자의 평가 Bias를 고려하도록 확장 가능하다

15. (텍스트분석) 자연어처리를 통해 하나의 문장을 분석하는 경우 여러 단계에서의 분석이 가능하다. 그 중 문장의 구조를 분석하는 것을 (가) 분석이라고 하고 문장의 의도를 파악하는 것을 (나) 분석이라고 한다. (가)와 (나) 괄호에 들어갈 적절한 용어는?

① 어휘(Lexical), 의미 (Semantic)　　② 구문(Syntax), 화용 (Pragmatic)

③ 담화(Discourse), 의미 (Semantic)　　④ 의미 (Semantic), 담화 (discourse)

16. (텍스트분석) 품사태깅(POS Tagging)은 파싱(Parsing)에 반드시 필요한 과정으로 파싱과정에 사전을 참조해서 각 단어의 품사를 결정하는 것이 일반적이다.

17. (텍스트분석) 언어모델(Language Modeling)을 통해 다양한 언어처리 문제를 해결할 수 있다. 다음 중 확률기반 언어모델 적용이 가장 적절치 못한 과업은?

① 컴파일러 구축　　② 질의응답　　③ 정보검색　　④ 기계번역

18. (텍스트분석) 워드임베딩(Word Embedding)은 약 300차원의 벡터 공간에 각 단어를 위치시키는 방법으로 각 차원이 어떤 특성이나 내용을 가지고 있는지 해석이 가능하므로 단어 간의 관계에 대한 연산(예: "King" – "Man" + "Woman" = "Queen")이 가능하다.

19. (텍스트분석) 개체명 인식(NER)을 위해 주위 단어 뿐만 아니라 단어와 관련된 다양한 자질이나 단어의 맥락을 표현하는 자질들을 사용하는 경우 더욱 좋은 성능을 기대할 수 있다. 다음 중 이렇게 다양한 종류의 서로 중첩되는 자질들을 사용하는 것이 어려운 sequence labeling 모델은?

① Decision Tree

② Maximum Entropy Markov Model

③ Hidden Markov Model

④ Conditional Random Fields Model

20.(텍스트분석) LDA는 관찰 가능한 문서의 단어열로부터 모든 잠재변수(Latent Variables)를 추론하기 위해 샘플링 기술을 사용한다. 이와 관련하여 맞지 않는 것은?

① Gibbs Sampling방법에서는 주어진 단어에 대한 특정 토픽 확률을 전체 컬렉션의 나머지 단어가 가지는 토픽에 의거하여 결정한다.

② 한 단어가 갖는 토픽 별 확률 값은 초기에 동일한 값으로 설정된다.

③ 한 단어가 갖는 토픽 별 확률 값은 문서 별 토픽 분포와 토픽 별 단어 분포 데이터를 사용하여 점진적으로 갱신된다.

④ Gibbs Sampling 은 MCMC (Markov Chain Monte Carlo) 알고리즘의 하나이다.

정답

1. ① 2. ① 3. ④ 4. ① 5. ② 6. X 7. ③ 8. ① 9. ② 10. ④

11. ② 12. ② 13. ③ 14. ③ 15. ② 16. X 17. ① 18. X 19. ③ 20. ②