Senior DS 예상문제 유형('19.2.25)

1. 다음과 같은 10개의 데이터가 있다고 하자. 여기서 Group 변수와 Gender 변수에 따라서 Values 변수의 평균값이 차이가 있는지 검정하고자 한다. 적절한 분석방법은 무엇인가? (중) 3

	Group	Gender	Values
1	A	M	2
2	A	M	3
3	A	F	4
4	A	F	5
5	B	M	4
6	B	M	3
7	B	F	2
8	B	F	2

① 이변량 T-검정 ② 일원분산분석 (Oneway ANOVA)

③ 이원분산분석 (Twoway ANOVA) ④ 회귀분석

2. 시계열분석에서 계절변동을 적용하는 방법으로서 수평성, 추세성, 계절성을 평활한 세 성분의 합으로 나타낸 시계열 방법은 다음중 무엇인가? (중) 4

① 선형지수평활법 ② 브라운의 계절지수평활법

③ 홀트의 계절지수평활법 ④ 윈터스의 계절지수평활법

3. Apache Spark에 관한 설명으로 맞는 것은 다음중 무엇인가? (중) 4

① Hadoop이 반드시 필요하다 ② Java를 사용해 개발되었다

③ 스파크는 스트리밍 데이터 처리에는 적합하지 않다

④ MapReduce의 대안으로 개발되었다.

4. HDFS (Hadoop Distributed File System)의 파일 읽기에 필요한 단계이다. 이를 순서대로 나열하시오. (중) 2

1. 어플리케이션이 클라이언트에게 파일 읽기를 요청
2. 클라이언트에게 요청된 블록을 전송
3. 메타데이터를 통해 파일이 저장된 블록 리스트를 반환
4. 요청된 파일이 어떤 블록에 저장되어 있는지를 네임노드에게 요청
5. 데이터 노드에 접근하여 블록 조회 요청
6. 클라이언트를 어플리케이션에 데이터를 전달

5 123456 ② 143526 ③ 135246 ④ 124356

5. 클래스 불균형 문제의 해결 방법 중 비용함수 기반 접근법의 설명이다. 설명이 옳으면 O를 그렇지 않으면 X를 표시하오. (중) X

모델 학습시, 소수 클래스 데이터에 대한 손실함수(Loss function)의 가중치를 다수 클래스 데이터에 대한 손실함수(Loss function)의 가중치보다 더 작게 준다.

6. Box-Cox 변환은 Log 변환을 포함하지 않는다. (중) X

7. 다음 PCA 방법의 아이겐벡터 결과물에서 제1주성분과 가장 관계성이 높다고 볼 수 있는 변수는 무엇인가? (상) 1

```
> winePCA$rotation
                       PC1          PC2         PC3         PC4
Alcohol        -0.144329395  0.483651548 -0.20738262  0.01785630
Malic           0.245187580  0.224930935  0.08901289 -0.53689028
Ash             0.002051061  0.316068814  0.62622390  0.21417556
Alcalinity      0.239320405 -0.010590502  0.61208035 -0.06085941
Magnesium      -0.141992042  0.299634003  0.13075693  0.35179658
Phenols        -0.394660845  0.065039512  0.14617896 -0.19806835
Flavanoids     -0.422934297 -0.003359812  0.15068190 -0.15229479
Nonflavanoid    0.298533103  0.028779488  0.17036816  0.20330102
Proanthocyanins -0.313429488  0.039301722  0.14945431 -0.39905653
Color           0.088616705  0.529995672 -0.13730621 -0.06592568
Hue            -0.296714564 -0.279235148  0.08522192  0.42777141
OD280          -0.376167411 -0.164496193  0.16600459 -0.18412074
Proline        -0.286752227  0.364902832 -0.12674592  0.23207086
```

① Flavanoids ② Nonflavanois ③ Ash ④ Phenols

8. Principle Component Analysis (PCA)와 Singular Value Decomposition (SVD)의 설명으로 적절하지 않은 것은? (중) 4

① PCA는 eigen value를 SVD는 singular value를 이용한다.

② 두 기법 모두 차원 축소를 위하여 사용한다.

③ PCA는 데이터의 공분산을, SVD는 데이터 행렬의 rank를 기반한다.

④ 두 기법 모두 지역 최적 해를 찾는 방법을 적용하여 차원을 축소한다.

9. Feature Selection은 크게 Forward Selection과 Backward Selection으로 나눌 수 있다. 이것에 대한 설명으로 적절한 것은? (중) 1

① Forward Selection은 변수를 추가하면서 진행하고, Backward Selection은 변수를 제거하면서 진행한다.

② 두 방법은 입력 변수들의 가능한 모든 조합을 검사하여 최적을 선택한다.

③ Forward Selection에서 추가된 변수는 다시 제거될 수 있고, Backward Selection 제거된 변수는 다시 추가될 수 있다.

④ 두 방법은 예측 성능의 향상을 보장한다.

10. 다음 중 Bagging 방법의 성공요인이 아닌 것은 무엇인가? (상) 3

① 데이터의 다양성 ② 나무모형 생성의 다양성

③ 분류방법의 안정성 ④ 여러 모형의 통합성

11. Gradient Boosting에 대한 설명으로 틀린 것은? (상) 4

① Weaker Learner를 결합하여 Strong Leaner를 만드는 방식이다

② 사용자가 Loss Function에 대한 정의를 해 줘야한다.

③ Loss Function이 미분 가능해야 한다.

④ 여러 개의 Weaker Learner를 사용하므로써, 어떤 Loss Function을 사용하더라도 Outlier에 강한 특징이 있다.

12. One Class Classification에 대한 설명으로 적절하지 않은 것은? (중) 1

① 주어진 데이터가 여러 클래스 중 어디에 속하는지를 판별하는 문제를 말한다.

② Outlier Detection, Novelty Detection, Concept Learning이라고도 불린다.

③ 대부분의 경우 Positive Training Data만 주어진다.

④ Support Vector Data Description (SVDD)가 대표적인 방법 중의 하나이다.

13. 다음은 Soft Margin을 갖는 Linear SVM의 Loss Function이다. 이에 대한 설명으로 적절한 것은? w는 모델이며, ϵ_k는 k번째 데이터에 대한 분류 모델의 오차에 해당하는 슬랙변수(slack variable)이다. R은 학습 데이터의 개수이다. (상) 3

$$\min_{\mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{k=1}^{R} \epsilon_k \right)$$

① C=0일 때 학습데이터에 대한 정확도는 최소가 된다.

② C가 작아질수록 모델은 underfitting된다.

③ 모델의 예측 정확도는 C가 증가하면서 증가하다가, C가 어느 값이 넘어가면 감소할 하는 경향을 보일 것이다.

④ C가 증가할 때 모델의 복잡도는 최대가 된다.

14. K-Nearest Neighbor Classifier는 일반적으로 모든 학습데이터를 저장한 뒤, 주어진 테스트 데이터에 대해서 k개의 Nearest Neighbor를 찾아 그것들을 기반으로 클래스를 추론한다. 이러한 과정에서 생기는 문제에 대응하기 위한 기법 중 성격이 다른 것은 무엇인가? (상) 4

① Sampling ② Clustering ③ Dimension Reduction ④ Locality Sensitive Hashing

15. 모형 기반 협업필터링에 관한 내용으로 관계없는 것은 다음 중 무엇인가?(상) 4

①ALS(Alternating lease square) ② SVD 근사

③ Slope one ④ 코사인 거리 (cosine distance)

16. 추천에 사용되는 Matrix Factorization 기법의 목적함수는 일반적으로 아래와 같다. 이에 대한 설명으로 적절하지 않은 것은? 수식에서 r_ui는 사용자 u의 상품 i에 대한 선호도이다. (상) 2

$$Loss = \min_{q,p} \sum_{u,i} \left(r_{ui} - q_i^T p_u \right)^2 + \lambda (\|q_i\|^2 + \|p_u\|^2)$$

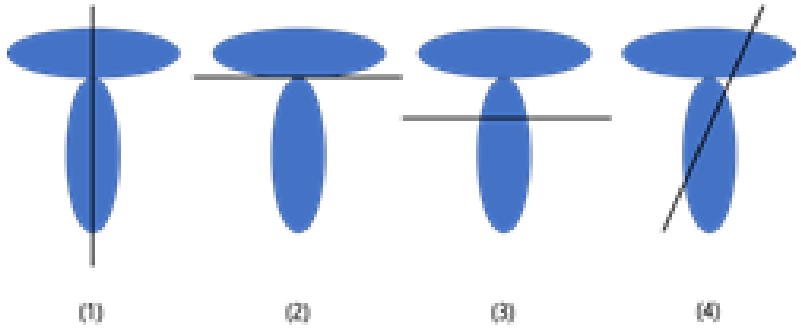
① λ=0인 경우에는 Loss함수를 0으로 하는 q와 p를 찾을 수 있다.

② λ의 값을 크게 주면 예측은 overfitting될 수 있다.

③ Alternating Least Square 기법을 이용하여 해결 할 수 있다.

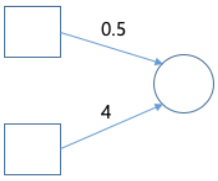
④ 사용자가 상품을 선택하는 과정에, 관찰하기 어려운 요소(Latent Factor)가 있다는 가정에 기반 둔 방법이라 할 수 있다.

17. 다음 그림에서 타원은 데이터에 존재하는 군집(Cluster)이며, 직선은 GMM (Gaussian Mixture Model) 혹은 k-Means를 적용했을 때 얻어진 군집 간의 경계이다. 이 중에서 GMM은 가능하나, k-Means로는 불가능한 것은 무엇인가? (상) 2



18. "아래 좌측과 같은 단층신경망에서 x1=-2, x2=1일때, 신경망을 통한 출력값은 어떻게 산출되는지 계산하시오. 활성함수는 아래 우측과 같은 시그모이드 함수를 이용하라. (중) 3

힌트 $e^3 \approx 20$ 이다"



$$\frac{e^x}{1+e^x}$$

① 3 ② 20 ③ 0.95 ④ 0.05

19. 하나의 Perceptron으로 Classification문제를 해결하고자 한다. Classification 문제는 두 개의 입력(x1,x2)과 하나의 출력(y)를 가지며 입출력 변수는 참(1)과 거짓(0) 값만 갖는다. 다음 중 하나의 Perceptron으로 해결할 수 없는 것은 무엇인가? 단, 아래에서 and, or, not, xor는 논리연산로 각각 논리곱, 논리합, 부정, 배타적 논리합을 의미한다. 단, activation fucntion은 다음과 같다. (상) 4

$$f(x) = \begin{cases} 0 & x \le 0 \\ 1 & x > 0 \end{cases}$$

① y = x1 and x2 ② y = x1 or x2 ③ y = x1 or (not x2) ④ y = x1 xor x2

20. 신경망의 단점에 대한 내용으로 해당되지 않은 것은 다음 중 무엇인가? (중) 4

① 구조가 복잡하면 훈련과정에 시간이 많이 소용될 수 있다

② 초기값에 따라 전역해(global optimum)가 아닌 지역해(local optimum)로 수렴할 수 있다

③ 은닉층의 수와 은닉노드 수의 결정이 어렵다

④ 노이즈(Noise)에 민감하게 반응한다

21. 딥러닝(deep learning)에 관한 설명으로 옳지 않은 것은 다음 중 무엇인가? (중) 2

① 신경망모형으로부터 비롯된 기계학습 방법이다

② 입력층이 많이 쌓여 가면서 복잡하고 깊은 구조를 가졌다

③ 선형적합과 비선형변환을 통해 복잡한 문제를 모형화한다

④ 사진인식, 음성인식, 자연어처리 등의 영역에서 효과를 나타내고 있다

22. Bag-of-Words (BOW)모델에서 n-gram을 사용할 때, 적절하지 않은 설명은? (중) 4

① bigram, trigram을 사용하면 단어의 순서를 반영한 문서 모델을 생성할 수 있다.

② n이 커질수록 문서 모델의 sparsity는 높아진다.

③ 문서 내에서 단어와 단어가 서로 독립이라면 bigram을 사용하는 것이 낫다.

④ n이 커질수록 더 긴 단어의 순서를 고려할 수 있으므로 가급적 큰 값을 사용해야 한다

23. "다음과 같은 내용을 가진 문서 d1, d2, d3가 있다. TF는 문서 안에서 단어의 총 출현 횟수로 정의하고, IDF는 DF의 역수의 상용log라고 정의할 때, d3에서의 my의 TF-IDF 값은? (중) 3

d1: I love dogs.

d2: I hate dogs and knitting.

d3: Knitting is my hobby and my passion.

① 0.18 ② 0.48 ③ 0.95 ④ 앞의 3개 모두 답이 아님

24. LDA(Latent Dirichlet Allocation)의 설명으로 적절하지 않은 것은? (중) 3

① 문서 내 단어의 순서를 고려되지 않는다.

② 사전에 토픽의 개수를 결정해야한다.

③ 문서의 토픽과 문서 내의 단어들의 토픽은 일치한다.

④ 일반적으로 EM 기법으로 최적화 된다.

25. 문장 "cat sat on mat"에서 얻을 수 있는 모든 bigram은? (중) 1

① cat sat, sat on, on mat

② cat sat, cat on, cat mat, sat on, sat mat, on mat

③ cat, sat, on, mat

④ cat, sat, on, mat, cat sat, sat on, on mat