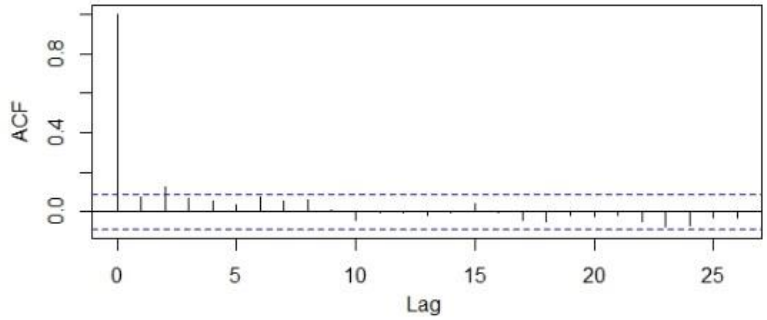
Senior Data Scientist 필기 예제 #3

1. 아래 그림은 시계열 y의 ACF(AutoCorrelation Function : 자기상관함수)를 나타낸다.

이 그래프를 가장 잘 설명한 것은?



① 서로 이웃한 시점 간의 자기 상관성이 없는 것으로 해석된다.

그러므로 시계열 y는 정상 시계열이다.

② 서로 이웃한 시점 간의 자기 상관성이 없는 것으로 해석된다.

그러므로 시계열 y는 비정상 시계열이다.

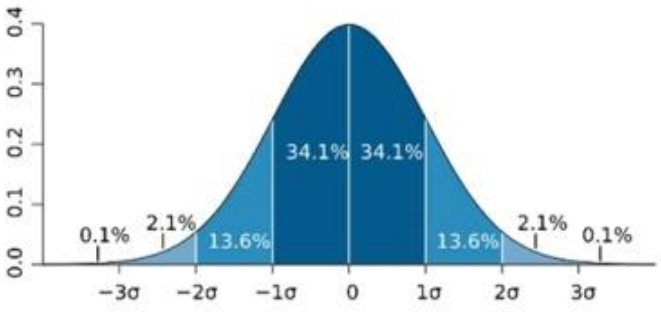
③ 서로 이웃한 시점 간의 자기 상관성이 있는 것으로 해석된다.

그러므로 시계열 y는 정상 시계열이다.

④ 서로 이웃한 시점 간의 자기 상관성이 있는 것으로 해석된다.

그러므로 시계열 y는 비정상 시계열이다.

2. 다음 중 정규분포에서 -1σ와 +1σ 사이의 곡선 아래의 면적은?



① 68.13%　② 47.72%　③ 27.18%　④ 15.87%

3. Volume, Variety, Velocity, Value, Veracity, Variability를 Big Data의 대표적인 특징을 나타내는 6V라고 부르기도 한다. (O, X)

4. 다음은 샘플링에 관한 설명들이다. 잘못된 것을 모두 고르시오.

(ㄱ) 데이터 집합이 클래스 불균형(Class Imbalanced)이더라도, Accuracy(정확도) 상황에 잘 맞는 척도(Measure)라고 볼 수 있다.

(ㄴ) 데이터 집합에 Negative 사례(Example)가 지나치게 많을 경우 Undersampling 방법을 사용하기도 한다.

(ㄷ) Oversampling과 Undersampling의 개념을 결합한 샘플링 기법은 알려진 방법이 거의 없다.

(ㄹ) 클래스 불균형(Class Imbalanced)인 데이터 집합으로 훈련(Training)한 분류기(Classifier)는, 다수(Majority) 클래스를 잘 맞추는 경향이 있다고 볼 수 있다.

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

5. 데이터에 주성분분석(Principal Component Analysis, PCA)를 적용한 뒤 새로운 모델을 생성하려고 한다. 이 때 새로운 모델이 사용하는 피처에 대한 설명으로 틀린 것은?

① 주성분을 각각이 설명하는 원래 데이터의 분산 양에 따라 정렬한 뒤, 정해진 기준값 이상을 설명할 수 있는 만큼의 주성분을 순서대로 골라서 사용한다.

② 일반적으로 원래의 데이터가 가진 피처의 갯수보다 적은 숫자의 주성분을 사용한다.

③ 원래의 데이터가 가진 분산(Variance) 중 일부는 포기한다.

④ 원래의 데이터가 가진 분산(Variance)을 모두 고려하면서도 피처 개수를 줄일 수 있다.

6. 선형 회귀에 아래와 같은 손실 함수(Loss Function)를 이용하려고 한다. 과적화(Overfitting)를 막기 위해 어떤 기법을 사용하고 있는가?

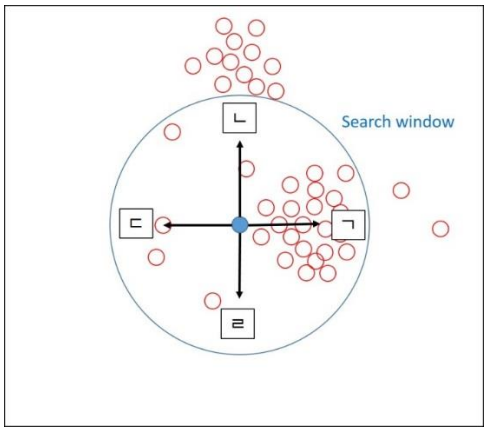
$$\min \sum_{i=1}^{n} \left(y_i - \sum x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

① Ridge 회귀　② Elastic Net 정규화　③ LASSO(Least Absolute Shrinkage And Selection) 정규화

④ 평균 제약 정규화 (Mean-Constrained Regularization)

7. 기계학습 알고리즘은 데이터에서 중요한 부분을 스스로 인식하므로 최대한 많은 종류의 피처값을 구해 입력으로 공급하는 것이 좋다. (O, X)

8. 다음 그림에서 Mean Shift Vector가 이동할 방향으로 옳은 것을 고르시오.



① ㄱ

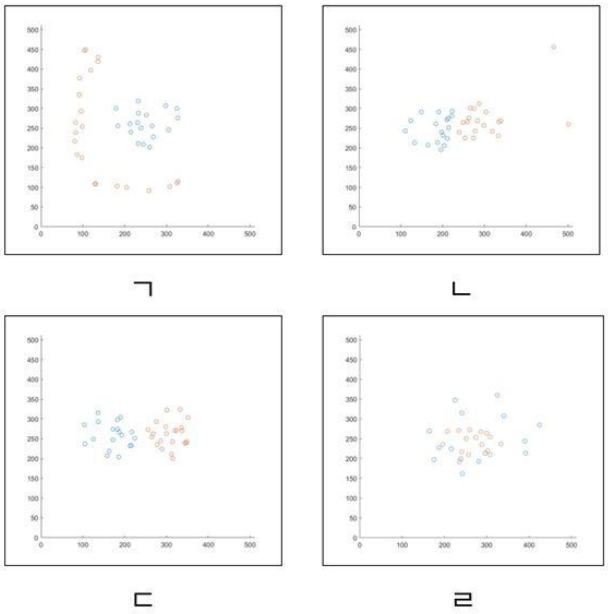
② ㄴ

③ ㄷ

④ ㄹ

9. 두 개의 집단 중 주어진 데이터를 K-nearest neighbor로 구분하려 할 때 k는 짝수여야 한다. (O, X)

10. 다음 중 k-means clustering이 제대로 작동할 수 있는 조건을 갖춘 데이터를 고르시오.



ㄱ ㄴ

ㄷ ㄹ

① ㄱ

② ㄴ

③ ㄷ

④ ㄹ

11. Logistic regression은 3개 이상의 class를 구분할 수 있다. (O, X)

12. Class y가 있고 Attributes x가 있을때 Likelihood는 클래스 y에서 x가 관측될 확률을 의미한다. (O, X)

13. Random Forest에 대한 설명으로 틀린 것을 고르시오.

① Training Data를 여러 Set으로 나눈 후 각각에 대해 Decision Tree를 구성하여, Random Forest로 활용한다

② Decision Tree를 기반으로 한 Ensemble Model의 일종이다

③ 여러 Decision Tree가 독립적으로 존재하여 Multi-Core 환경에서 고속으로 생성 가능하다

④ Classification을 진행할 수 있으나, Regression 지원은 불가능하다

14. Bagging에 대한 설명으로 틀린 것을 고르시오.

① Bootstrap Aggregation이라고도 불린다

② 기반 Classifier가 Unstable 한 경우 성능 향상이 우수하다

③ Over-Fitting 문제를 줄일 수 있다

④ 선택된 Data는 효율성을 위해 또다시 선택되지 않는다

15. Content-based Recommendation에 대한 설명으로 틀린 것을 고르시오

① Item의 특징에 대한 정보는 필요하나, User 특징에 대한 정보는 필요하지 않다

② 영화를 Item으로 생각할 때 영화의 장르, 주인공, 감독 등에 대한 정보를 활용하여 추천한다

③ Item과 관련된 여러 가지 Text 정보 등을 활용하여 추천을 진행할 수 있다

④ Content의 유사성을 구하기 위해 TF (Term Frequency)등을 활용할 수 있다

16. 다음 중 Unsupervised learning 기법에 통상 근거한 Neural Network이 아닌 것은 무엇인가?

① Generative Adversarial Network ② Recurrent Neural Network

③ Autoencoder ④ Restrict Boltzmann Machine

17. Least Square 방법론을 활용하여 데이터 (x,y)={(1,0),(1,2),(1,8),(1,10)}를 y=wx 로 fit하고자 한다. w값을 구하면 얼마인가?

① 400%　② 450%　③ 500%　④ 550%

18. 다음 중 EM(Expectation Maximization) 알고리즘에 대한 설명으로 옳지 않은 것은 무엇인가?

① Expectation 과정은 구하고자 하는 파라미터 값의 추정치를 요구한다.

② Maximization 과정은 구하고자 하는 파라미터 값의 추정치를 수정한다.

③ EM 알고리즘은 초기 파라미터 설정에 수렴 여부가 의존한다.

④ EM 알고리즘이 구한 파라미터 값이 꼭 최선의 값은 아니다.

19. 다음의 과업(Task)을 수행하기 위해 필요하지 않은 텍스트마이닝 기술은?

"Identify quiet hotels near the downtown of Barcelona, Spain, that serve high quality breakfast with a reasonable price."

① 자동요약　② 자동분류　③ 정보추출(Information Extraction)　④ 개체명인식

20. 순수 RNN 기반 언어모델이 갖는 "Vanishing Gradient" 문제를 극복하기 위한 모델이 아닌 것은?

① LSTM　② GRU　③ Neural Turing Machine　④ BERT

정답

1. ①	2. ①	3. O	4. ①	5. ④	6. ①	7. X	8. ①	9. X	10. ③
11. X	12. O	13. ④	14. ④	15. ①	16. ②	17. ③	18. ③	19. ①	20. ④