데이터마이닝 일반 / 앙상블 모형

문제 1. 다음 중 데이터마이닝에서 구축용(training), 검정용(validation), 시험용(test) 데이터로 분리하는 이유로 가장 타당한 것은?

① 과잉 또는 과소맞춤에 대한 미세조정 절차를 수행하기 위해 데이터를 준비한다.

② 모형이 잘못된 가설을 가정하여 발생되는 2종 오류의 발생을 사전에 방지한다.

③ 주어진 데이터에서만 높은 성과를 보이고 새로운 데이터에는 성과가 낮은 현상을 방지한다.

④ 모델을 구축하고 평가하는데 소요되는 시간을 단축한다.

문제 2. 데이터 양이 충분하지 않을 때 모델의 성능측정의 신뢰도를 높이기 위해 사용하는 방법으로 샘플을 k개의 집합으로 나눈 후 k-1개의 집합으로 학습 후 나머지 1개의 집합으로 성능을 측정하는 과정을 반복하는 검증방법을 무엇이라 하는가?

① 층화추출법(Stratiifed sampling)

② 배깅(bagging)

③ 교차검증(cross validation)

④ 부트스트래핑(bootstraping)

문제 3. 아래의 데이터 마이닝 분석 예제 중 비지도(unsupervised) 분석을 수행해야 하는 예제는?

가. 우편물에 인쇄된 우편번호 판별 분석을 통해 우편물을 자동으로 분류

나. 고객의 과거 거래 구매 패턴을 분석하여 고객이 구매하지 않는 상품을 추천

다. 동일 차종의 수리 보고서 데이터를 분석하여 차량 수리에 소요되는 시간을 예측

라. 상품을 구매할 때 그와 유사한 상품을 구매한 고객들의 구매 데이터를 분석하여 쿠폰을 발행

① 나, 다

② 가, 라

③ 가, 다

④ 나, 라

문제 4. 다음 중 이상값 검색을 활용한 응용시스템으로 가장 적절한 것은?

① 장바구니분석 시스템

② 부정사용방지 시스템

③ 데이터 마트

④ 교차판매 시스템

문제 5. 다음 중 그 성격이 나머지 하나와 다른 것은?

① 연관분석

② K-평균

③ 의사결정나무

④ 분포추정

문제 6. 다음 중 그 성격이 나머지 하나와 다른 것은?

① 의사결정나무

② 서포트 벡터 머신

③ 로지스틱 회귀

④ 계층적 군집

문제 7. 다음 중 분류를 위해 사용되는 데이터 마이닝 기법은?

① Association Rule

② K-means

③ Collaborative filtering

④ K-Nearest Neighbors

문제 8. 다음 중 과적합(overfitting)에 대한 설명으로 가장 부적절한 것은?

① 과대적합이 발생할 것으로 예상되면 학습을 종료하고 업데이트 하는 과정을 반복해 과대적합을 방지할 수 있다.

② 과대적합은 분석 변수가 너무 많이 존재하고 분석 모델이 복잡할 때 발생한다.

③ 분석 데이터가 모집단을 특성을 설명하지 못하면 발생한다.

④ 생성된 모델은 분석 데이터에 최적화되었기 때문에 훈련 데이터의 작은 변화에 민감하게 반응하는 경우는 발생하지 않는다.

문제 9. 관측표본 n개 개체를 n번 복원추출하면, 수학적으로 특정 개체가 복원추출표본에 있지 않을 확률은 $(1-(1/n))^n$ 임이 알려져 있다. n이 충분히 크다면 이 확률의 값은 약 36,8%로 수렴된다. 즉, 복원추출 과정에서 전체 데이터 중 36.8%가 선택되지 않는 반면, 63.2%가 선택되어진다고 할 수 있는데, 선택된 63.2% 데이터로 훈련된 모델이 예측한 값과 선택되지 않은 36.8% 데이터 간의 차이를 계산하여 모델의 유효함을 검증하는데 사용하는 오차를 무슨 오차라고 하는가?

문제 10. 모형의 평가를 위해 관측치를 한 번 이상 훈련용 자료로 사용하는 복원 추추법에 기반하는 부트스트랩 기법에서 일반적으로 훈련용 자료의 선정을 d번 반복할 때 하나의 관측치사 선정되지 않을 확률은 $(1-(1/d))^d$ 이다. d가 충분히 크다고 가정할 때 훈련용 집합으로 선정되지 않아 검증용 자료로 사용되는 관측되의 비율은 얼마인가?

① 20.5%

② 28.8%

③ 34.2%

④ 36.8%

문제 11. 데이터 속성(feature)의 수가 증가하면 (같은 비율의 공간을 채우기 위해 변수 1개당) 분석에 필요한 데이터의 양이 기하급수적으로 증가하게 되고 충분한 데이터가 없다면 적은 데이터로 설명해야 하기 때문에 결국 과적합(Overfitting)이 발생하여 모델의 성능이 떨어지게 되는데 이러한 현상을 의미하는 용어는?

문제 12. 데이터셋의 분포가 불균형할때 부족한 데이터의 수를 늘려 불균형을 극복하는 방법으로 단순임의복원추출을 통해 데이터의 확률분포를 알아낼 때 사용하며, 기계학습에서 훈련데이터를 무작위복원추출하여 여러 개의 모델을 만들어 과적합을 방지하는 역할도 할 수 있는 있는 방법은 무엇인가?

① boosting(부스팅)

② bootstrapping(부트스트래핑)

③ stacking(스태킹)

④ ensemble(앙상블)

문제 13. 부트스트래핑을 통해 생성된 조금씩 다른 훈련 데이터들을 학습하여 각각의 모델을 형성하고 최종적으로 학습된 모델들의 예측변수를 집계하여 그 결과로 모델을 생상하는 앙상블 기법을 무엇이라 하는가?

① boosting(부스팅)

② stacking(스태킹)

③ bagging(배깅)

④ LSTM

문제 14. 다음 중 랜덤포레스트 알고리즘과 가장 관련이 깊은 것은 무엇인가?

① boosting(부스팅)

② stacking(스태킹)

③ bagging(배깅)

④ LSTM

문제 15. 잘못 분류된 객체들에게 높은 가중치를 부여하고 반대로 올바르게 분류된 객체들에게는 낮은 가중치를 부여하여 오분류된 객체들이 더 잘 분류되도록 함으로써 예측모형의 정확도를 향상시키기 위한 방법은?

① boosting(부스팅)

② stacking(스태킹)

③ bagging(배깅)

④ bootstrapping(부트스트래핑)

문제 16. 다음은 어느 한 앙상블 기법의 절차를 기술한 내용이다. 이 앙상블 기법은 무엇인가?

가. 전체 데이터셋에서 동일한 크기의 데이터를 무작위복원추출 후 서로 다른 데이터세트들을 만들다.

나. 각 훈련데이터셋을 이용해 각각의 모델을 생성한다.

다. 모델의 예측값이 연속형일 경우에는 평균값을, 범주형일 경우에는 투표를 통해 최종 결과를 산출한다.

① boosting(부스팅)

② stacking(스태킹)

③ bagging(배깅)

④ bootstrapping(부트스트래핑)

문제 17. 약한 검출기들을 여러 개 모아 강한 검출기를 생성하는 방법으로 순차적으로 이전 학습 분류기의 결과를 토대로 다음 학습 데이터의 샘플 가중치를 조정하면서 학습을 진행하는 앙상블 기법은?

① boosting(부스팅)

② stacking(스태킹)

③ bagging(배깅)

④ bootstrapping(부트스트래핑)

문제 18. 서로 다른 모델들을 조합해서 최고의 성능을 내는 모델을 생성하는 앙상블 기법으로 다양한 알고리즘의 조합을 통해 강점은 취하고 약점은 보완하는 장점이 있으나, 필요한 연산량이 상당하다는 단점이 있는 것은?

① boosting(부스팅)

② stacking(스태킹)

③ bagging(배깅)

④ bootstrapping(부트스트래핑)

문제 19. 다음은 어느 한 앙상블 기법의 절차를 기술한 내용이다. 이 앙상블 기법은 무엇인가?

> 가. 학습 데이터로 모델 1을 생성한다.
>
> 나. 생성된 모델 1이 잘못 예측한 데이터의 가중치를 높인다.
>
> 다. 수정된 데이터로 새로운 모델 2를 생성한다.
>
> 라. 상기 "나"~"다" 의 과정을 반복한다.

> ① boosting(부스팅)
>
> ② stacking(스태킹)
>
> ③ bagging(배깅)
>
> ④ bootstrapping(부트스트래핑)

답안

1. 4
2. 3
3. 4
4. 2
5. 3
6. 4
7. 4
8. 4
9. OOB오차(Out of bag Error)
10. 4
11. 차원의 저주
12. 2
13. 3
14. 3
15. 1
16. 3
17. 1
18. 2
19. 1