

광주인공지능사관학교 교육생 평가지

Track 명	AI 모델링
과 정 명	사물지능
모 들 명	머신러닝
과 목 명	1과목 : 머신러닝 2과목 : ML 파이프라인
유 형	이론 / 실습
시험일자	2022년 8월 4일 (금요일) - 2시간

※ 제출 : 주피터노트북에 문제별 셀을 만들어서 작성 후 앨리스로 제출

출제자	황해도
검수자	강성관

1 과목	머신러닝
------	------

유형	이론	난이도	하	배점	6 점
----	----	-----	---	----	-----

문제 1	머신러닝으로 풀 수 있는 아래 예시에서 각각 지도 학습 예시와 비지도 학습 예시에 맞는 것을 모두 선택하시오.				
------	---	--	--	--	--

지도 학습	
비지도 학습	

〈예시〉

- (1) 의료 영상 이미지를 활용한 종양 유무 판단
- (2) 자율 주행차의 방향 결정
- (3) 다양한 악기 소리가 섞여 있는 소리데이터에서 악기 소리 구분
- (4) 블로그 글의 주제 구분
- (5) 의심되는 신용카드 거래 데이터를 이용한 불법 사용 판단
- (6) 고객들의 성향에 따라 고객 그룹핑
- (7) 바둑 프로그램에서 다음 수 결정
- (8) 손글씨 숫자 인식

유형	이론	난이도	하	배점	6 점
----	----	-----	---	----	-----

문제 2	다음은 통계학적 데이터 분석 방법과 머신러닝 기반 데이터 분석 방법에 대한 설명이다. 아래 보기 중 각 ()안에 알맞은 단어를 연결하시오.				
------	---	--	--	--	--

보기

- a. 통계학적
- b. 머신러닝 기반

- ㉠ ()데이터 분석 방법에서 강조되고 있는 영역은 추론과 검증이며, 주어진 데이터가 연구자의 가설과 이론에 얼마나 부합하는가 등을 설명하기 위해 다양한 방법론과 이론이 구축되어 있다.
- ㉡ ()데이터 분석 방법은 명시적인 알고리즘을 설계하기 어렵거나 프로그래밍하기 어려운 작업을 해결하기 위해 주로 사용된다.
- ㉢ ()데이터 분석 방법은 관측치로 부터 도출된 값이 실제 모집단의 모수를 얼마나 정확하게 추정하고 있는가를 설명한다.
- ㉣ ()데이터 분석 방법은 주어진 입력 데이터를 컴퓨터 프로그램이 학습하여 예측을 수행하고 스스로의 예측 성능을 향상시키는 과정을 말한다.

유형	이론	난이도	하	배점	6 점
문제 3	다음 문장이 설명하는 학습방법은 무엇인가?				
<div>- 설명변수(혹은 독립변수, 특성(Feature) 등으로 표현)와 목적변수(혹은 출력값, 반응변수, 종속변수, 목표변수 등으로 표현) 간의 관계성을 표현해내거나 미래 관측을 예측해 내는 것에 초점이 있으며 주로 인식, 분류, 진단, 예측 등의 문제 해결에 적합하다</div>					
답 :					

유형	이론	난이도	하	배점	6 점
문제 4	다음 문장이 설명하는 학습방법은 무엇인가?				
<div>목적변수(혹은 출력값, 반응변수, 종속변수, 목표변수 등으로 표현)에 대한 정보 없이 학습이 이루어지는 형태를 말하며, 예측의 문제보다는 주로 현상의 기술(Description)이나 특징 도출, 패턴 도출 등의 문제에 많이 활용된다.</div>					
답 :					

유형	이론	난이도	중	배점	9 점
문제 5	분류 문제 사례에 맞는 예시를 선택하시오.(답 3개)				
<div>(1) 스팸 메일 판단</div> <div>(2) 기업 고객 유형별 비율 예측</div> <div>(3) 고객 이탈 / 유지 예측</div> <div>(4) 고객 신용 점수 예측</div> <div>(5) 특정 질병 발생 여부 예측</div> <div>(6) 고객이 예상 구매액 판단</div>					

유형	이론	난이도	중	배점	9 점
문제 6	KNN 분류기에 대한 설명으로 맞는 것을 선택하시오.(답 4개)				
<div>(1) 데이터 세트의 확률분포 등에 대한 고려가 필요하다</div> <div>(2) 해당 데이터와 주변 데이터 세트 간의 거리 유사성을 측정한다</div> <div>(3) 알고리즘이 이해하기 쉽고 직관적이다</div> <div>(4) 일반적으로 이웃의 수는 짝수를 사용한다</div> <div>(5) 데이터의 수에 상관없이 연산량은 변하지 않는다</div> <div>(6) 이웃의 수가 많아지면 결정경계가 단순해진다</div> <div>(7) 이웃의 수가 작으면 과소적합이 발생하기 쉽다</div> <div>(8) 근접한 이웃의 수를 결정하는 것이 중요하다.</div>					

유형	이론	난이도	중	배점	9 점
----	----	-----	---	----	-----

문제 7	Decision Tree 분류기에 대한 설명으로 맞는 것을 선택하시오.(답 4개)
<div><div>(1) 분류분석에만 활용 가능한 모델이다.</div><div>(2) 데이터의 이상치에 영향을 받지 않는다</div><div>(3) 특성 중요도를 계산할 수 있어 다른 모델의 분석에서 특성선택방법으로 활용할 수 있다</div><div>(4) 트리가 깊을수록 모델의 성능이 향상된다.</div><div>(5) 목표변수와 가장 연관성이 높은 변수의 순서대로 불순도나 엔트로피 등이 낮아지는 방향으로 나무 형태로 가지를 분할하면서 분류 규칙을 만들어 내는 기법</div><div>(6) 높은 수학적 지식이 필요한 모델이고 트리 모형에 대한 이해가 어렵다</div><div>(7) 분류결과에 대한 Rule기반의 해석 가능하여 분류결과 이유를 설명해야 할 경우 유용하다</div><div>(8) 과소적합이 발생하기 쉬운 모델이다.</div></div>	

유형	이론	난이도	중	배점	9 점
----	----	-----	---	----	-----

문제 8	Decision Tree는 모델이 복잡해지는 것을 방지하기 위해 사전가지치기를 수행한다. 사전가지치기의 설명에 맞은 파라미터를 기술하시오.
<div>(1) 트리의 최대 깊이를 설정</div> <div>(2) 리프 노드의 최대 개수를 설정</div> <div>(3) 리프 노드로 분리가능한 최소 샘플의 개수를 설정</div> <div>(4) 리프 노드를 구성하는 최소 샘플의 개수를 설정</div>	

유형	이론	난이도	상	배점	12 점
----	----	-----	---	----	------

문제 9	선형회귀 모델에 대한 설명으로 맞는 것을 선택하시오.(답 3개)
<p>(1) 이상치나 데이터의 변화에 강인하다.</p> <p>(2) 결측치와 불완전 자료가 있을 때도 잘 작동한다.</p> <p>(3) 관측된 사건들을 정량화해서 독립변수와 종속변수의 관계를 함수식으로 설명하는 방법</p> <p>(4) 독립변수와 목표변수 간 관계에 대한 설명과 해석이 용이하며, 각 독립변수의 영향력을 파악하기가 쉽다.</p> <p>(5) 다른 모델에 비해 훈련시간이 느리다.</p> <p>(6) 해당 함수식이 모수에 대해 선형일 경우 선형회귀분석이라고 하며, 독립변수가 한개인 경우 단순 선형회귀, 여러 개일 경우 다중선형회귀분석을 적용한다.</p> <p>(7) 복잡한 데이터를 예측할 때는 정확도가 향상된다.</p>	

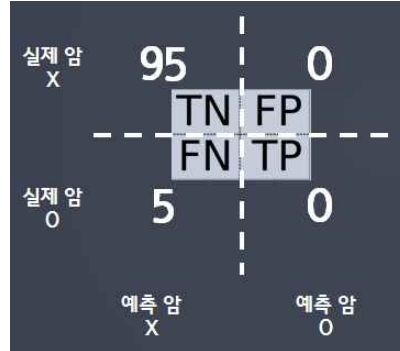
유형	이론	난이도	상	배점	12 점
----	----	-----	---	----	------

문제 10	선형회귀 모델을 규제하기 위한 방법과 그 특징에 대한 설명 중 옳은 것을 선택하시오.(답 2개)
<p>(1) L1 규제와 L2 규제로 구분된다.</p> <p>(2) L1 규제는 손실함수에 규제항으로 가중치 제곱의 총합을 사용한다.</p> <p>(3) L1 규제는 0이 되는 가중치가 발생할 수 있다.</p> <p>(4) L1 규제는 특성간의 가중치의 차이를 줄여준다.</p> <p>(5) L2 규제는 손실함수에 규제항으로 가중치 절대값의 총합을 사용한다.</p> <p>(6) L2 교제는 특성선택시 사용하기도 한다.</p> <p>(7) Ridge는 L1 규제를 Lasso 모델은 L2 규제를 사용한다.</p>	

유형	이론	난이도	상	배점	16 점
----	----	-----	---	----	------

다음 아래의 문제를 읽고 답을 서술하시오.

아래 그림은 100명의 암환자에 대해 예측한 한 모델의 Confusion_matrix이다.



문제 11-1	분류평가 지표 중 정확도만으로는 평가가 안 되는 경우가 있어서 정밀도와 재현율을 보는 경우가 있다. 항목에 맞게 사례를 선택하시오.
---------	---

정밀도가 중요한 사례	
재현율이 중요한 사례	

- (1) 암 진단
- (2) 스팸 메일 판단
- (3) 유해 동영상 판단
- (4) 범죄자 판단

문제 11-1	위 모델의 정확도(Accuracy), 재현율(Recall), 정밀도(Precision)를 계산하시오.
---------	--

- (1) 정확도 :
- (2) 재현율 :
- (3) 정밀도 :

문제 11-1	분류평가지표를 이용하여 평가를 도출 후 결과에 따라 모델의 성능을 향상시켜야 한다면 어떤 식으로 시도해 볼 수 있는지 서술하시오.(답 4개)
---------	--

- (1) 편향된 데이터라도 수집하여 데이터의 양을 늘리고 본다.
- (2) 새로운 특성을 추출하여 특성을 추가한다
- (3) 다양한 전처리(재그룹핑, 범주화 등)를 통해 데이터를 최적화한다
- (4) 다른 모델들을 사용한다
- (5) 하이퍼 파라미터를 변경해가면서 최적의 파라미터를 찾는다
- (6) 데이터의 분산을 증가시킨다
- (7) 상관관계가 90% 이상 되는 특성들을 사용한다.

2 과목	ML 파이프라인
------	----------

유형	이론	난이도	하	배점	10 점
----	----	-----	---	----	------

문제 1	머신러닝 기반 데이터 분석 수행 시 일반적인 절차를 <u>순서에 맞게 나열</u> 하시오.
<div><div>㉠ 데이터 수집</div><div>㉡ 데이터에 대한 모델훈련</div><div>㉢ 데이터 전처리와 탐색</div><div>㉣ 모델 성능 평가</div><div>㉤ 비즈니스 이해 및 문제 정의</div></div>	
답 :	

유형	이론	난이도	하	배점	10 점
----	----	-----	---	----	------

문제 2	데이터세트를 훈련/테스트 세트로 분할하지 않고 데이터세트 모두를 이용해 훈련하고 데이터세트 모두를 이용해 테스트할 경우 문제점을 선택하세요.
<div>(1) 모델이 데이터세트 모두를 알고 있기 때문에 과소적합이 발생할 수 있다.</div> <div>(2) 훈련 데이터와 테스트 데이터 분리하지 않아서 일반화 모델인지 평가하기 어렵다.</div> <div>(3) 데이터 편향이 발생하기 쉽다</div> <div>(4) 전체 데이터를 사용하기 때문에 분산이 커질 가능성이 높아 모델의 성능을 떨어뜨린다.</div>	

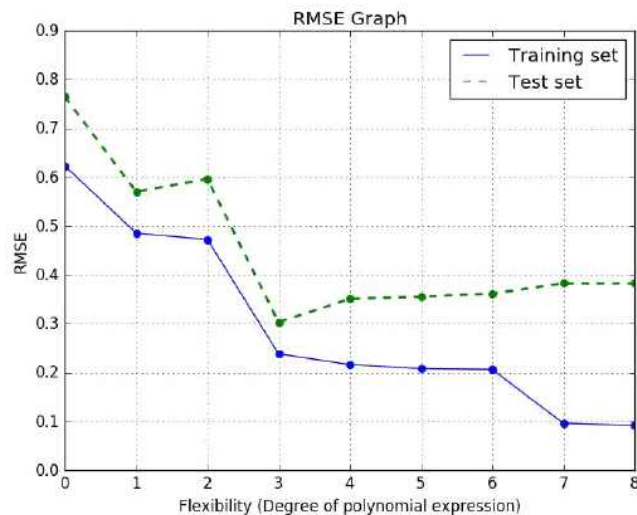
유형	이론	난이도	중	배점	12 점
----	----	-----	---	----	------

문제 3	교차검증(cross-validation)에 대한 설명으로 맞는 것을 선택하시오.(답 2개)
<p>(1) 데이터를 여러 개의 세트로 분리하고 여러 개의 훈련 세트와 테스트 세트를 각각 평가한다.</p> <p>(2) 각 세트마다 평가한 결과 점수 중 최소 점수를 최종 평가 점수로 선택한다.</p> <p>(3) 데이터의 수가 아주 많은 경우에 효과적으로 활용될 수 있다.</p> <p>(4) 고정된 훈련 세트와 테스트 세트만으로 훈련을 하면 테스트 세트에만 과적합이 되는 문제가 발생할 수 있다.</p> <p>(5) 교차 검증은 테스트 세트를 고정하고 훈련 세트를 다양하게 샘플링하여 사용하는 방법이다.</p>	

유형	이론	난이도	중	배점	15 점
----	----	-----	---	----	------

다음 아래의 문제를 읽고 답을 서술하시오.

다음은 다항식의 차수가 증가함에 따라 평균 제곱근 오차(RMSE)의 변화를 나타낸 그래프이다.
(RMSE는 MSE값에 제곱근을 붙인 것, RMSE는 낮을수록 좋은 결과이다.)



문제 4-1	Flexibility(다항식의 차수)가 3일 경우가 가장 일반화가 이루어진 상태라고 말 할 수 있는 근거로서 옳은 항목들을 선택하시오.(답 2개)
(1) 훈련 데이터와 테스트 데이터에서 오차가 모두 최저의 값을 보여주고 있기 때문 (2) 훈련 데이터와 테스트 데이터의 오차가 0.2~0.3 사이 값이기 때문 (3) 훈련 데이터와 테스트 데이터에서 오차의 차이가 최소가 되는 부분이기 때문 (4) 훈련 데이터와 테스트 데이터의 오차가 가장 급격히 떨어지는 부분이기 때문	
문제 4-2	Flexibility(다항식의 차수)가 4 이상인 경우부터는 Training 데이터의 오차율은 감소하지만, Test 데이터의 오차율은 증가하고 있다. 이러한 현상을 나타내는 단어는 무엇인가.(영어단어 또는 한글 모두 인정)
문제 4-3	일반적으로 훈련 데이터와 테스트 데이터를 나눌 때 테스트 데이터의 비율은 얼마 정도가 적당한가 ?

유형	실습	난이도	중	배점	12 점
문제 5	sklearn 패키지를 이용하여 아래와 같이 주석에 맞게 교차검증 진행하고 실행결과와 같이 출력하는 코드를 작성하시오.				
<pre>from sklearn.datasets import load_iris from sklearn.model_selection import train_test_split from sklearn.neighbors import KNeighborsClassifier from sklearn.model_selection import cross_val_score # (1) iris data 불러오기 # (2) iris data를 문제와 답으로 분리 # (3) 분리된 문제와 답을 train data와 test data로 분리 # (4) KNN분류기 모델 생성 (이웃 수는 3) # (5) 교차검증 실행 (5 조각으로 나눠 진행) # (6) 결과 확인</pre>					
실행결과 (분리 방법에 따라 결과 숫자는 다를 수 있음)					
[0.8696 0.9565 0.9565 0.9545 0.9048] 평균점수 : 0.92838321					

유형	이론	난이도	상	배점	16 점
----	----	-----	---	----	------

다음 아래의 문제를 읽고 답을 서술하시오.

다음은 iris데이터를 불러와 target label을 출력한 결과이다. 0~2 각각의 값은 붓꽃의 품종을 의미한다.

```
1 from sklearn import datasets
1 iris = datasets.load_iris()
1 iris.target
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

문제 6

다음과 같은 데이터를 아래와 같이 7:3으로 나누어 학습시켰을 때 결과에 대한 설명으로 옳지 않은 것은?

- 훈련 데이터 셋 : 인덱스 0~105번 (70%)
- 테스트 데이터 셋 : 인덱스 105~149번 (30%)

- (1) 훈련 데이터와 테스트 데이터 라벨 값들이 존재하지 않아 제대로 학습을 할 수 없다.
- (2) 라벨 데이터가 데이터가 편향되어 있어 과소적합이 발생할 수 있다.
- (3) 교차검증을 사용하면 편향된 학습 문제를 해결할 수 있다.
- (4) 램덤 샘플링을 이용하여 데이터를 섞어 주어서 사용하면 데이터 편향을 해결할 수 있다.

유형	이론	난이도	상	배점	25 점
----	----	-----	---	----	------

다음 아래의 문제를 읽고 답을 서술하시오.

다음은 범주형 데이터셋이다.

poisonous	cap- shape	cap- surface	cap- color	bruises	odor	gill- attachment
p	x	s	n	t	p	f
e	x	s	y	t	a	f
e	b	s	w	t	l	f
p	x	y	w	t	p	f
e	x	s	g	f	n	f

문제 7-1

다음과 같은 데이터를 머신러닝 모델에 적용하기 위해서 전처리 한다면 어떤 방법을 사용해야하는지 방법의 명칭을 작성하시오.

문제 7-2

위에서 작성한 답에 대한 설명으로 맞는 것을 선택하시오.(답 2개)

- (1) 각 컬럼의 클래스 수에 맞게 새로운 컬럼을 생성하고 해당 클래스의 위치에는 1을, 다른 클래스의 위치에는 0을 부여한다
- (2) 특성을 증가시켜줌으로써 과대적합을 줄여준다.
- (3) 클래스의 수가 많다면 컬럼이 너무 많아져서 메모리 측면에서 비효율적이 될 수 있다.
- (4) 컬럼이 가지고 있는 클래스의 특성을 정확하게 표현할 수 있다.