



기말고사



0

기말고사 개요


- 가상의 기상청 날씨 데이터를 처리하고 분석하여, 최종적으로 특정 수치를 예측하는 로지스틱 모델을 만들어야 합니다.
- 데이터는 기상청의 종관기상관측(ASOS) 데이터베이스에서 추출되었습니다.
- 총 6문제가 준비되어 있으며, 문제에 따라 여러 부분 문제를 포함합니다. 문제 별 배점은 아래 표와 같습니다.

1	2	3	4	5	6
15	15	15	15	20	20



0

사용 데이터 컬럼 설명

- station_name : 측정 장소
 - temperature : 기온
 - precipitation : 강수
 - windspeed : 풍속
 - wind_direction : 풍향
 - humidity : 습도
 - vapor_pressure : 증기압
 - dew_point_temp : 이슬점
 - local_pressure : 대기압
 - sea_level_pressure : 해면기압
 - sunshine : 일조량
 - solar_radiation : 일사량
 - snowfall : 강설
 - new_snowfall_3hr : 신적설량 (3시간)
 - cloud_type : 운형
 - lowest_cloud : 최저운고
 - visibility : 시정
 - ground_condition : 지면상태
 - activate_score : 활동지수
- 

0

데이터 예시

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	station_name	date_time	temperature	precipitation	windspeed	wind_direction	humidity	vapor_pressure	dew_point_temp	local_pressure	sea_level_pressure	sunshine	solar_radiation	snowfall	new_snowfall_3hr	cloud_type	lowest_cloud	visibility	ground_condition	activity_score
2	속초	2021-01-01 1:00	-6.7		4.3	320	30	1.1	-21.4	1020.4	1022.7							2000		-5.25809002
3	속초	2021-01-01 2:00	-6.7		2	340	26	1	-23	1021.2	1023.5							2000		-4.87406571
4	속초	2021-01-01 3:00	-7.2		1.3	320	25	0.9	-23.9	1022	1024.3							2000		-5.0901328
5	속초	2021-01-01 4:00	-7.6		0.8	180	25	0.9	-24.2	1021.7	1024							2000		-5.10850735
6	속초	2021-01-01 5:00	-7.5		0.7	140	27	0.9	-23.3	1021.7	1024							2000		-5.04126311
7	속초	2021-01-01 6:00	-6.6		1.6	90	28	1	-22.1	1021.4	1023.7							2000		-4.711077432
8	속초	2021-01-01 7:00	-5.8		1.2	200	23	0.9	-23.6	1021.4	1023.7							2000		-3.266364523
9	속초	2021-01-01 8:00	-6.2		0.8	230	28	1.1	-21.7	1021	1023.3	0.2						2000		-3.103580085
10	속초	2021-01-01 9:00	-4.2		2.3	200	29	1.3	-19.6	1021.4	1023.7	1						2000		-1.854638797
11	속초	2021-01-01 10:00	-3.4		2.1	250	28	1.3	-19.3	1021.4	1023.7	1						2000		-1.505449251
12	속초	2021-01-01 11:00	-1.9		2.7	270	29	1.5	-17.6	1021.6	1023.9	1						2000		-0.951957267
13	속초	2021-01-01 12:00	-0.7		1.9	250	28	1.6	-16.9	1020	1022.2	1						2000		-0.340957006
14	속초	2021-01-01 13:00	0.5		2.5	250	25	1.6	-17.3	1019.1	1021.3	1						2000		-0.055977091
15	속초	2021-01-01 14:00	1.1		1.9	270	26	1.7	-16.3	1018.3	1020.5	1						2000		0.299014835
16	속초	2021-01-01 15:00	1.1		2.1	270	26	1.7	-16.3	1018.3	1020.5	1						2000		0.187053516
17	속초	2021-01-01 16:00	0.4		4.4	270	28	1.8	-16	1018.6	1020.8	1						2000		-0.31847651

1-1

데이터 전처리 1 (7점)

다음 동작을 수행하세요.

- 데이터 분석에 앞서, 불필요한 데이터를 제거하려고 합니다.
아래 열들을 제거하고, 남은 열들의 이름을 출력하세요.
 - new_snowfall_3hr, cloud_type, lowest_cloud, ground_condition, wind_direction
- 열 이름은 아래와 같이 출력하세요.

```
프로세스가 시작되었습니다.(입력값을 직접 입력해 주세요)  
> Index(['컬럼명1', '컬럼명2', '컬럼명3', '컬럼명4',  
        '컬럼명5', '컬럼명6'],  
        dtype='object')
```

입출력 예시, 정답과 무관합니다.

1-2

데이터 전처리 2 (8점)

다음 동작을 수행하세요.

- 컬럼 이름 하나를 **입력**받은 후, 입력받은 컬럼의 데이터의 **결측치를 0으로** 채우세요.
그 뒤, 해당 컬럼에 대한 **평균** 값을 **소수점 셋째자리**에서 **반올림**하여 출력하세요.
- 반올림은 `round(변수, 2)` 를 통해 적용 가능합니다.

```
프로세스가 시작되었습니다.(입력값을 직접 입력해 주세요)  
> precipitation  
1.48
```

입출력 예시

2

데이터 관계 분석 (15점)

- 1번 문제에서 **전처리 된** 데이터가 주어집니다.
(1-2의 전처리는 모든 열에 적용되었습니다.)
- humidity 열과 다른 변수들 간의 관계를 보고자 합니다.
- 사용자로부터 **정수 (A)**와, 다른 **컬럼 이름 (B)**를 입력받습니다.
- humidity 값이 A 미만인 데이터에 대해서, B의 **중앙값을 출력**하세요.

```
프로세스가 시작되었습니다.(입력값을 직접 입력해 주세요)
> 60
windspeed
2.5
```

입출력 예시, 정답과 무관합니다.

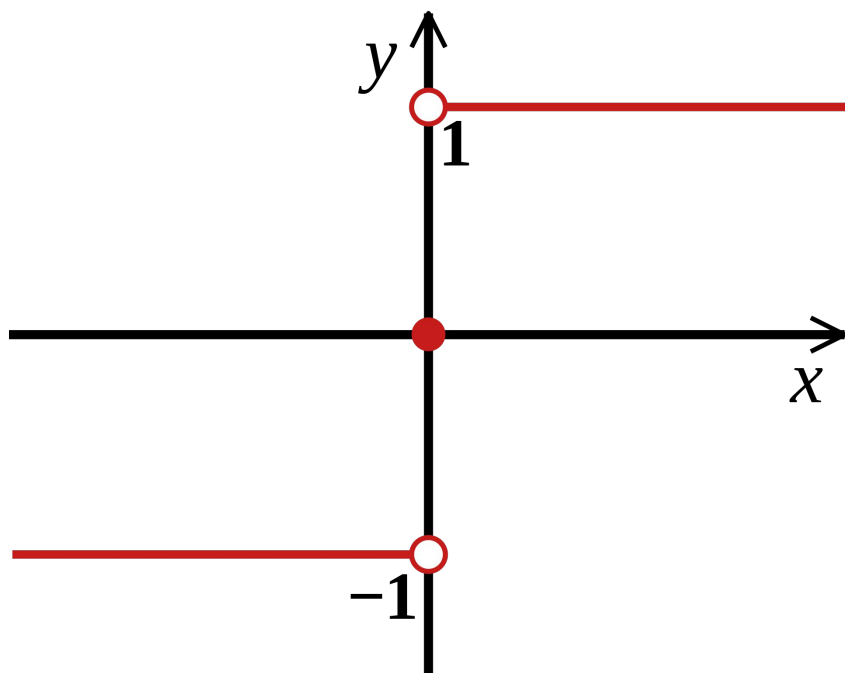
3

파생 변수 만들고 검증하기 (15점)

- 1번 문제에서 **전처리 된** 데이터가 주어집니다.
- 주어진 데이터를 분석하기 위해, 파생 변수를 추가로 만들려고 합니다.
- 온도를 사용한 두 가지 파생 변수를 만들어 **새로운 컬럼으로 추가**하세요.
- 두가지 파생 변수의 컬럼 이름과 생성 식은 아래와 같습니다.
 - *derivated_temp* = (temperature - 15) 의 절대값
 - *derivated_solar* = sign(temperature - 15) * solar_radiation
- temperature와 *derivated_temp* 의 상관계수,
temperature와 *derivated_solar* 의 상관계수를 **순서대로 출력**하세요.
- 출력은 **소수점 셋째자리에서 반올림**하세요.
round(변수, 2) 를 이용하세요.
- sign 함수 설명과 입출력 예시는 다음 페이지를 참고하세요.

3

파생 변수 만들고 검증하기(Cont.)



$$\text{sign}(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x = 0) \\ -1 & (x < 0) \end{cases}$$

```
프로세스가 시작되었습니다.(입력값을 직접 입력해 주세요)  
> 0.84  
0.28
```

입출력 예시, 정답과 무관합니다.
위가 derivated_temp, 아래가 derivated_solar와의 상관계수입니다.

4-1

월별 지역 활동 지수 탐색 1 (5점)

- 특정 월에 운동 경기가 있는데 어떤 지역에서 개최할지 고민하고 있습니다.
따라서 해당 월에 **가장 활동 지수(activity score)가 높은 지역**에서 운동 경기를 개최하기 위해 탐색하고자 합니다.
- 날짜에서 month를 추출하는 함수가 주어집니다.
- 주어진 함수를 이용하여(사용 여부는 선택 사항입니다.), **컬럼명 month**를 가진 열을 만드세요.
- 입력된 월(ex. 1)에 해당하는 행의 개수를 int 형식으로 출력하세요.
- 다음 페이지에 입출력 예시가 있습니다.

4-1

월별 지역 활동 지수 탐색 1

- 입출력 예시 (구름의 입출력 예시와 같습니다.)

프로세스가 시작되었습니다.(입력값을 직접 입력해 주세요)

> 1

2579

4-2

월별 지역 활동 지수 탐색 2 (10점)

- 운동 경기를 어떤 지역에서 개최할지 고민하고 있습니다.
 - 4-1 문제와 이어지는 문제입니다. 앞서 작성한 코드와 동일한 코드를 이용하세요.
 - 특정 월의 활동 지수(activity score)가 비교적 높은 지역에서 운동 경기를 개최하고자 합니다.
1. 개최할 월 **A**와 확인할 순위 **B**를 입력 받습니다. ($1 \leq A \leq 12$, $0 < B$)
 2. 지역마다 **입력받은 월의 평균 활동 지수(activity score)**를 산출합니다.
 3. 입력받은 월에 대한 각 지역의 **평균 활동지수를 내림차순으로 정렬**합니다.
 4. A 월의 평균 활동 지수가 B위에 해당하는 지역을 'str' 형태로 출력합니다.
 5. 다음 페이지에 입출력 예시가 있습니다.

4-2

월별 지역 활동 지수 탐색 2

- 입출력 예시 (구름의 입출력 예시와 같습니다.)

```
프로세스가 시작되었습니다.(입력값을 직접 입력해 주세요)
```

```
> 1
```

```
3
```

```
제주
```

5

회귀 모델 학습 (20점)

- 문제 1번을 통해 전처리가 완료된 데이터셋이 주어집니다. 자유롭게 데이터를 처리하여 문제를 해결하세요.
- Label 은 **활동지수(activity score)**를 이용하며, 나머지 주어진 변수들로 **활동지수**를 예측하려 합니다.

채점 관련 안내

- 학습이 완료된 모델의 train_set, test_set 의 활동 지수를 예측한 값의 정확도를 출력하며, 출력된 정확도로 채점됩니다.
- 각 구간의 점수를 얻기 위해서는 Train Set과 Test Set의 Accuracy가 모두 아래 조건의 구간을 넘어야 합니다.
- Train과 Test의 점수 구간이 다를 경우 더 낮은 구간의 점수로 채점됩니다.
- 채점 코드는 제공됩니다.

Accuracy	Score
< 0.5	0점
>= 0.5	5점
>= 0.575	10점
>= 0.65	15점
>= 0.75	20점

Ex) Train 0.73, Test 0.77 → 15점

6

회귀 모델 학습 - 심화 (20점)

- 문제 1번을 통해 전처리가 되지 않은 데이터셋이 주어집니다. 자유롭게 데이터를 처리하여 문제를 해결하세요. 단, 테스트 데이터셋에 dropna 메소드 등을 사용하여 결측치를 지워서는 안됩니다.
- Label 은 **활동지수(activity score)**를 이용하며, 나머지 주어진 변수들로 **활동지수**를 예측하려 합니다.

채점 관련 안내

- 학습이 완료된 모델의 train_set, test_set 의 활동 지수를 예측한 값의 정확도를 출력하여 채점합니다.
- 각 구간의 점수를 얻기 위해서는 Train Set과 Test Set의 Accuracy가 모두 아래 조건의 구간을 넘어야 합니다.
- Train과 Test의 점수 구간이 다를 경우 더 낮은 구간의 점수로 채점됩니다.
- 채점 코드는 제공됩니다.

Accuracy	Score
< 0.75	0점
>= 0.75	5점
>= 0.8	10점
>= 0.85	15점
>= 0.90	20점

Ex) Train 0.7, Test 0.84 → 0점