**2020 학년도 SW/AI/융합경시대회**

# 인공지능 챌린지 문제지

**코드/동영상 제출: ai@rcv.sejong.ac.kr**

**고사 중 공지/질의응답: https://open.kakao.com/o/gYjJ1qxc**

(화면 녹화 중 PC 카톡을 통한 오픈채팅방만 사용 허용)

- **18:10 : 웹엑스 입실 완료** (웹캠, 핸드폰 등 활용 디바이스 관계 없음)
- **18:10-18:30 : 시험 준비 (**수업과 공지 중간고사 가이드라인을 따름**)**
- **18:30 : 캐글 리더 보드 오픈**
  - **시험지 개인 이메일로 발송 / 오픈 채팅방 공유 예정**
  - **리더보드 최대 20 회까지 제출 가능**
  - **리더보드 제출은 반드시 학번으로 제출**
- **18:30-20:30 : 챌린지 (화장실 불가, 반드시 시험 전 다녀오길 권장)**
- **20:25 : 캐글 리더 보드 마감 (마감 이후 제출 불가)**
- **20:30: 코드 제출 마감 (마감 이후 불인정)**
- **22:30: 발표 영상 제출 및 화면 녹화 영상 제출 마감**
  - 마감 이후 불인정, 어려움 있을 시 오픈카톡방으로 사전 문의

## 인공지능 챌린지 참가 동의서 (필수)

https://forms.gle/X1LJZoE8sQxgC9wP8

## 인공지능 챌린지 주의사항

캐글 리더보드 제출용 코드 미제출자 2 차 응시 불가

화면 녹화 동영상 미제출자 2 차 응시 불가

발표 심사 동영상 미제출자 2 차 응시 불가

웹엑스를 통한 실시간 감독 미참가자 2 차 응시 불가

## 인공지능 챌린지 평가방식

**1 차 캐글평가 : 70 점 (정확성)**

정확성: 총 3 문제(각 리더보드) 중 가장 높은 점수를 받은 문제로 인정

즉, 모든 문제를 풀어도 무방하나 반드시 모든 문제를 제출할 필요는 없음.

(예를 들어, 1 번 90 점, 2 번 미제출, 3 번 20 점 인 경우 90 점으로 인정함)

**2 차 발표평가 : 30 점 (타당성, 창의성, 유창성)**

타당성: 문제풀이 접근법과 모델링이 타당한가

창의성: 문제풀이 접근법과 모델링이 창의적이고 혁신적인가

유창성: 문제풀이 접근법과 모델링을 잘 설명할 수 있는가

## [문제1] SejongAI-Challenge-Problem1

https://www.kaggle.com/t/0621a92d547f4f39adabf258c6973870

## 패션 의류 예측 문제

본 문제는 패션 의류 10 종에 대한 흑백 영상 데이터를 사용한다. 학습 및 테스트 데이터로 28x28 크기의 2d 영상 데이터를 1d 로 가공하여 제공하며, 종류는 0 부터 9 로 라벨링되어 있다. 최종적으로 본 문제는 주어진 데이터를 기반으로 10 종에 대한 패션 의류의 종류를 예측하는 것을 목표로 한다.

### Evaluation

28x28 흑백 영상을 기반으로 패션 의류의 종류를 예측하고, Category Accuracy 를 활용하여 예측 모델의 정확도를 측정한다.

### Data Description

### 파일 설명

- train.csv - the training set
- test.csv - the test set
- submission_sample.csv - a sample submission file in the correct format
-

### 데이터 설명

각 데이터 샘플은 아래와 같이 라벨링되어 있다.

- id – 데이터 순번
- pixel 0~784 – 흑백 영상의 밝기 값 (0~255 로 채워져 있음)
- **label – 의류 종류 (0~9 로 채워져 있음)**

## [문제2] SejongAI-Challenge-Problem2

## https://www.kaggle.com/t/a0f4372db7554a64b62fc8c5e2f4fe37

## 아마존 리뷰 기반 긍정 부정 리뷰 예측 문제

본 문제는 아마존 사용자 리뷰 데이터(1~5 점 평점)를 이용하여 사용자 리뷰의 긍정 리뷰와 부정 리뷰를 분류하는 자연어 처리 문제이다. train.csv 파일에서 라벨은 1, 0 으로 아마존 사용자 리뷰 데이터의 4, 5 점은 라벨 1 로, 1, 2 점은 라벨 0 으로 3 점 데이터는 사용하지 않는 것으로 전처리 하여 제공하였다. 또한 unbalance 데이터 셋에 대해서는 balance 하게 데이터를 구성하였는데, 원하는 학생들은 제공되는 raw data 를 다양한 방법으로 이용하여 성능을 향상시킬 수 있다.

최종적으로는 test.csv 파일 내의 리뷰 내용을 기반으로 리뷰 내용의 긍정/부정 여부를 예측하는 것을 목표로 한다.

(힌트) 리뷰 텍스트를 특징 벡터(vector)로 추출하기 위해서 아래의 기능 사용 가능
(반드시 아래의 기능을 사용하지 않아도 무방하며, 더 좋은 방법론을 적용 가능함)

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

## Evaluation

Category Accuracy 를 활용하여 예측 모델의 정확도를 측정한다.

## Data Description

### 파일 설명

- train.csv - the training set
- test.csv - the test set
- submission_sample.csv - a sample submission file in the correct format

### 데이터 설명

각 데이터 샘플은 리뷰 텍스트 정보와 리뷰 점수가 아래와 같이 라벨링되어 있다.

- id – 데이터 순번
- Text – 사용자 리뷰
- **Label – 사용자 평점 (*대문자 임을 주의)**

## [문제3] SejongAI-Challenge-Problem3

https://www.kaggle.com/t/214905a46bdd49f3bc3ead9a6b52a5a2

## 영아 건강 및 발달 프로그램 효과 예측 문제

본 문제는 저체중아와 미숙아를 대상으로 고품질의 육아 서비스와 가정방문을 제공하는 영아 건강 및 발달 프로그램 (Infant Health and Development Program)의 데이터를 사용한다. 이 프로그램은 실험이 끝난 후 영아의 인지능력 점수를 상승시는 데 매우 성공적이었으며, 본 문제에서는 영아와 산모의 데이터를 기반으로 해당 프로그램의 효과를 예측하는 것을 목표로 한다.

### Evaluation

영아와 산모의 데이터를 기반으로 영아 건강 및 발달 프로그램의 효과를 예측하고, Root Mean Squared Error 를 활용하여 예측 모델의 정확도를 측정한다.

### Data Description

### 파일 설명

- train.csv - the training set (647 rows)
- test.csv - the test set (100 rows)
- submission_sample.csv - a sample submission file in the correct format

### 데이터 설명

각 데이터 샘플은 7개의 continuous 변수와 2개의 discrete 변수, 9개의 binary 변수로 구성되며, training set의 경우 프로그램 치료 효과를 나타내는 continuous 변수 **y**가 추가된다. (*label이 아님을 주의)

- id - unique id
- **y - treatment effect of Infant Health Development Program**
- bw - birthweight of child in grams

- b.head - child's head circumference (cm) at birth
- preterm - number of weeks pre-term that the child was born
- birth.o - birth order
- nnhealth - "neo-natal health index" some function of birth variables (number of days in hospital and gestational age..) that supposed to measure neonatal health
- momage - mom's age when she gave birth to the child
- dadage - dad's age when she gave birth to the child
- sex - child's gender (female=1)
- twin - is child a twin
- b.marr - indicator for whether mom was married when child born
- mom.edu - mom's education level at time child was born (less than high school=0, high school=1, some college=2)
- cig - did mom smoke cigarettes when pregnant
- first - indicator for whether child is first born
- booze - did mom consume alcohol when pregnant
- drug - did mom ever use drugs when pregnant
- work.dur - indicator for whether mom worked during her pregnancy
- prenatal - indicator for whether she received any prenatal care
- site - site indicator variables (born) (ark=0, ein=1, har=2, mia=3, pen=4, tex=5, was=6)