1. default.csv 자료는 다음의 10000명의 고객에 대한 다음 4개의 변수 정보를 기록한 것이다. 물음에 답하여라.

   - default : 해당 고객이 자신의 debt에 대한 default 여부를 나타냄. Yes는 defaulted, No는 not defaulted를 의미함.
   - student : 해당 고객이 학생인지 여부를 나타냄. Yes는 학생임, No는 학생이 아님.
   - balance : 매월 카드 청구액을 납부한 이후에 해당 고객 계좌의 평균 balance.
   - income : 해당 고객의 소득.

   (1) 각 변수 중 범주형인 'default'와 'student'는 Yes면 1, No면 0인 정수 타입의 더미변수로 변환하고, 나머지 숫자형 변수들은 모두 표준화(standardized) 변환을 적용하여라.

   (2) (1)에서 전처리된 데이터 전체가 훈련용 데이터셋이라고 가정하고, sklearn을 이용하여 default를 목표변수로 하는 이항 로지스틱 회귀모형을 훈련하여라.

   (3) (2)에서 추정된 파라미터를 이용하여, 훈련된 이항 로지스틱 회귀모형을 식으로 표현하여라.

   (4) (2)의 훈련 결과를 이용하여, 학생이면서, balance가 900, income이 7100인 어느 새로운 고객에 대한 default 확률을 구하여라.

2. 다음 코드를 실행하면 아래와 같은 array를 생성할 수 있다. 이 array의 각 열은 순서대로 절편항 1, 특징변수 X_1, X_2, X_3, X_4, 목표변수 Y를 나타내며, 목표변수의 범주는 3개(K=3)인 훈련 데이터셋이라고 가정해 보자.

```
np.random.seed(123)
traindt = np.hstack( [np.ones((5,1)),
                      np.around( np.random.randn(5, 4), 3),
                      np.random.randint(1,4,(5,1))] )
traindt
```

```
array([[ 1.   , -1.086,  0.997,  0.283, -1.506,  2.   ],
       [ 1.   , -0.579,  1.651, -2.427, -0.429,  1.   ],
       [ 1.   ,  1.266, -0.867, -0.679, -0.095,  1.   ],
       [ 1.   ,  1.491, -0.639, -0.444, -0.434,  1.   ],
       [ 1.   ,  2.206,  2.187,  1.004,  0.386,  3.   ]])
```

   (1) 다음 행렬 $\boldsymbol{\Theta_1}$는 특징변수가 4개, 목표변수의 범주가 3개인 경우에 대한 소프트맥스 회귀모형 가설에서의 파라미터 행렬 $\boldsymbol{\Theta_1}$이다. $\boldsymbol{\Theta_1}$의 각 행은 목표변수의 각 범주 별

소프트맥스 파라미터 $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)}$를 나타낸다. 이 $\boldsymbol{\Theta_1}$이 주어진 훈련 자료를 분류하는데 적절한 파라미터라고 가정해 보자. 이 파라미터 행렬을 이용하여 주어진 훈련자료의 각 관찰치가 Y의 각 범주에 속할 확률을 계산한 뒤, 가장 확률이 높은 범주로 분류하여라.

$$\boldsymbol{\Theta_1} = \begin{bmatrix} -\left(\boldsymbol{\theta}^{(1)}\right)^T - \\ -\left(\boldsymbol{\theta}^{(2)}\right)^T - \\ -\left(\boldsymbol{\theta}^{(3)}\right)^T - \end{bmatrix} = \begin{bmatrix} 5 & 2 & 3 & 1 & 4 \\ 2 & 4 & 3 & 1 & 2 \\ 3 & 4 & 1 & 5 & 4 \end{bmatrix}$$

(2) 다음 행렬 $\boldsymbol{\Theta_2}$도 $\boldsymbol{\Theta_1}$과 같은 형식으로 정의된 특징변수가 4개, 목표변수의 범주가 3개인 경우에 대한 소프트맥스 회귀모형 가설에서의 파라미터 행렬이다. (1)의 $\boldsymbol{\Theta_1}$과 (2)의 $\boldsymbol{\Theta_2}$ 중에서 주어진 훈련자료에 보다 더 적절한 파라미터 행렬은 무엇인가? 주어진 훈련자료에 대한 크로스 엔트로비 비용함수를 계산한 뒤 이를 이용하여 비교하여라.

$$\boldsymbol{\Theta_2} = \begin{bmatrix} 5.5 & 2 & 3 & 1.5 & 4 \\ 2 & 3.5 & 2.5 & 1 & 1.5 \\ 3 & 4 & 1 & 5 & 4 \end{bmatrix}$$

3. scoredEX.csv 자료는 어느 훈련 데이터를 이용하여 이진 분류 알고리즘을 학습한 다음, 평가 데이터(test data)에 대해 각 관찰치 별 특징변수값 (x1 ~ x29), 목표변수값(y_test, 0과 1로 입력됨)과 알고리즘을 적용했을 때의 y=1에 대한 예측확률 추정치(p_pred)를 기록한 것이다.

   (1) 이 자료를 이용하여, 가능한 분류임계치(threshold) 별 정밀도와 재현율을 계산한 뒤, 그 관계를 나타내는 정밀도-재현율 그림(가로축:재현율, 세로축:정밀도)을 그려라. 단, sklearn의 confusion_matrix 함수를 제외하고는 sklearn의 기능을 이용하지 말 것.

   (2) (1)에서 구한 각 분류임계치 별 정밀도과 재현율을 이용하여, 각 분류임계치 별 f1_score를 계산하여라. f1_score를 기준으로 판단한다고 할 때, 주어진 데이터를 이진분류하기 위한 최선의 분류임계치는 무엇인가?

   (3) (2)에서 찾은 분류임계치로 분류한 결과를 y_pred라는 변수로 기존의 데이터 셋의 마지막 열로 추가하여라. 또한 y_test와 y_pred를 이용하여 정확도(accuracy)를 계산하여라.