1. sklearn의 dataset 모듈에서 fetch_california_housing 은 다음에 관한 데이터 정보를 담고있다. 이 자료를 이용하여 물음에 답하여라.

타겟 데이터

1990년 캘리포니아의 각 행정 구역 내 주택 가격의 중앙값

특징 데이터

MedInc : 행정 구역 내 소득의 중앙값

HouseAge : 행정 구역 내 주택 연식의 중앙값

AveRooms : 평균 방 갯수

AveBedrms : 평균 침실 갯수

Population : 행정 구역 내 인구 수

AveOccup : 평균 자가 비율

Latitude : 해당 행정 구역의 위도

Longitude : 해당 행정 구역의 경도

관찰치 수 : 20640

(1) 주어진 데이터셋을 훈련자료 60%, 평가자료 40%로 나누어라.

(2) 60%의 훈련자료에 대해 릿지회귀를 훈련하고자 한다. 5-fold 교차검증(cv)를 적용하여, 다음 주어진 규제조절 매개변수(λ) 후보 중 최적의 값을 선택하여라. 단, 검증 기준은 MSE를 이용할 것.

λ 후보값 : 0, 1, 10, 30, 50, 100

(3) 60%의 훈련자료에 대하여 (2)에서 선택된 λ를 적용한 릿지회귀를 훈련한 뒤, 가중치 파라미터 추정치를 출력하여라. 또한 40%의 평가자료를 이용하여 훈련된 모형에 대한 R^2를 구하여라.

2. 일반적인 선형 회귀모형에 비해, 그 파라미터에 Lasso 규제를 적용된 Lasso 회귀모형은 어떠하다고 말할 수 있는지, 다음의 설명 중 가장 적절한 것을 골라라.

① 유연성이 높고, 따라서 편향의 증가가 분산 감소보다 작을 경우 예측 정확도가 향상된다.

② 유연성이 높고, 따라서 분산의 증가가 편향 감소보다 작을 경우 예측 정확도가 향상된다.

③ 유연성이 낮고, 따라서 편향의 증가가 분산 감소보다 작을 경우 예측 정확도가 향상된다.

④ 유연성이 낮고, 따라서 분산의 증가가 편향 감소보다 작을 경우 예측 정확도가 향상된다.

3. 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.

4. 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.

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