CS383

Noah Robinson, nkr38

July 2023

1 Theory

a. $\mathbf{X} = \begin{bmatrix} 1 & -2 \\ 1 & -5 \\ 1 & -3 \\ 1 & 0 \\ 1 & -8 \\ 1 & -2 \\ 1 & 1 \\ 1 & 5 \\ 1 & -1 \\ 1 & 6 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} 1 \\ -4 \\ 1 \\ 3 \\ 11 \\ 5 \\ 0 \\ -1 \\ -3 \\ 1 \end{bmatrix}$

$$w = (X^T X)^{-1} X^T y$$

$$w = (X^T X)^{-1} [14, -79]$$

$$w = \begin{bmatrix} 1.0285 \\ -0.4126 \end{bmatrix}$$

b.

$$\hat{Y} = -0.4126X + 1.0285$$

$$\hat{Y} = \begin{bmatrix} 1.85 \\ 3.09 \\ 2.26 \\ 1.028 \\ 4.33 \\ 1.85 \\ 0.61 \\ -1.034 \\ 1.44 \\ -1.44 \end{bmatrix}$$

2 Part 2

1.

Validation RMSE: 6604.316221778553
Validation SMAPE: 0.18300131098547648

Training RMSE: 5757.8889922488215
Training SMAPE: 0.18054837992345754

2. For pre-processing I randomized the indexes, divided the data in 2/3 and 1/3 as requested, and then used pandas get_dummies to one hot encode the categorical features of 'sex', 'smoker', and 'region' so that I can compare them numerically.

3 Part 3

1. S = 3
   Mean RMSE = 6102.604515086437
   Std. RMSE = 18.963400283907003

2. S = 223
   Mean RMSE = 6086.918221566928
   Std. RMSE = 1.577588193309232

3. S = 1338
   Mean RMSE = 6087.388006550313
   Std. RMSE = 6.643047920924854e-12