Alec Peterson
ap3842@drexel.edu

CS 613 – Homework 2

1) Problem 1

a. Coefficients for closed-form linear regression:

$$After\ adding\ bias\ feature\ b\ to\ X\text{: } X_b = \begin{bmatrix} -2 & 1 \\ -5 & 1 \\ -3 & 1 \\ 0 & 1 \\ -8 & 1 \\ -2 & 1 \\ 1 & 1 \\ 5 & 1 \\ -1 & 1 \\ 6 & 1 \end{bmatrix}$$

$$\boldsymbol{w} = (X_b{}^T X_b)^{-1} X_b{}^T Y = \begin{bmatrix} 169 & -9 \\ -9 & 10 \end{bmatrix}^{-1} X_b{}^T Y = \begin{bmatrix} -0.413 \\ 1.029 \end{bmatrix}$$

$$\boldsymbol{w} = \begin{bmatrix} -0.413 \\ 1.029 \end{bmatrix}$$

(values rounded for display purposes here)

b. Predictions (rounded) when using $\boldsymbol{w}$

$$\hat{Y} = X_b \boldsymbol{w} = \begin{bmatrix} 1.854 \\ 3.092 \\ 2.267 \\ 1.029 \\ 4.330 \\ 1.854 \\ 0.616 \\ -1.035 \\ 1.441 \\ -1.447 \end{bmatrix}$$

c. $\boldsymbol{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{Y_i} - \boldsymbol{\hat{Y}_i}\right)^2} = \frac{1}{10}\left[(\mathbf{1} - \mathbf{1.854})^2 + \cdots + (\mathbf{1} - (-\mathbf{1.447}))^2\right]^{1/2} = \mathbf{3.701}$

$$\boldsymbol{SMAPE} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \frac{\left|\boldsymbol{Y_i} - \boldsymbol{\hat{Y}_i}\right|}{|\boldsymbol{Y_i}| + \left|\boldsymbol{\hat{Y}_i}\right|} = \frac{\mathbf{1}}{\mathbf{10}}\left[\frac{|\mathbf{1} - \mathbf{1.854})|}{|\mathbf{1}| + |\mathbf{1.854}|} + \cdots + \frac{|\mathbf{1} - (-\mathbf{1.447})|}{|\mathbf{1}| + |-\mathbf{1.447}|}\right] = \mathbf{0.609}$$

2) Problem 2

- Training set:
  - RMSE = 5757.89
  - SMAPE = 0.18
- Validation (test) set:
  - RMSE = 6652.80
  - SMAPE = 0.19

Pre-Processing of data included:

1. Shuffle and split the data into training and test sets
2. Z-scored numeric features "**age**", "**bmi**", and "**children**". Number of children was assumed to be numeric.
3. Encoded binary categorical features "**sex**" and "**smoker**" into integer Boolean arrays
4. One-hot encoded categorical feature "**region**" i.e. made integer Boolean arrays for "is_northeast", "is_northwest", "is_southeast", "is_southwest"
5. Added dummy column of 1's as "bias" feature.
6. Subsequently calculated closed-form linear regression.

3) Problem 3

- For S = 3:
  - Mean RMSE = 6080.42
  - Standard Deviation of RMSE's = 170.30
- For S = 223:
  - Mean RMSE = 6407.87
  - Standard Deviation of RMSE's = 222.40
- For S = N (where N = 1338 for dataset):
  - Mean RMSE = 7520.09
  - Standard Deviation of RMSE's = 159.09