James Zhao
HW4
A15939512

1 1

1.1 a

Maximum likelihood estimate is the average y: $\frac{1+3+4+6}{4} = 3.5$
MSE: $\frac{(1-3.5)^2+(3-3.5)^2+(4-3.5)^2+(6-3.5)^2}{4} = \frac{13}{4} = 3.25$

1.2 b

Predictions: $(1,1),(1,1),(4,4),(4,4)$
Truths: $(1,1),(1,3),(4,4),(4,6)$
MSE: $\frac{(1-1)^2+(1-3)^2+(4-4)^2+(4-6)^2}{4} = \frac{8}{4} = 2$

1.3 c

From lecture:

$$y = ax + b, b = \bar{y} - a\bar{x}, a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1+1+4+4}{4} = 2.5, \bar{y} = \frac{1+3+4+6}{4} = 3.5$$

$$a = \frac{(1-2.5)(1-3.5) + (1-2.5)(3-3.5) + (4-2.5)(4-3.5) + (4-2.5)(6-3.5)}{(1-2.5)^2 + (1-2.5)^2 + (4-2.5)^2 + (4-2.5)^2} = \frac{9}{9} = 1$$

$$b = \bar{y} - a\bar{x} = 3.5 - 1 * 2.5 = 1$$

Answer:

$$y = x + 1$$

2 2

2.1 a

$$\frac{d}{ds}L(s) = \frac{d}{ds}\frac{1}{n}\sum_{i=1}^{n}(x_i - s)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{ds}(x_i - s)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} -2(x_i - s)$$

$$= -\frac{1}{n}\sum_{i=1}^{n} 2(x_i - s)$$

2.2 b

$$0 = -\frac{1}{n}\sum_{i=1}^{n} 2(x_i - s)$$

$$0 = -\frac{1}{n}\sum_{i=1}^{n} 2x_i + \frac{1}{n}\sum_{i=1}^{n} 2s$$

$$\frac{1}{n}\sum_{i=1}^{n} 2x_i = \frac{1}{n}\sum_{i=1}^{n} 2s$$

$$\sum_{i=1}^{n} 2x_i = \sum_{i=1}^{n} 2s$$

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} s$$

$$\sum_{i=1}^{n} x_i = ns$$

$$\frac{\sum_{i=1}^{n} x_i}{n} = s = E[x]$$

3 3

Total penalty is the penalty over every item. The loss functions can optionally be mean losses (multiply by $\frac{1}{n}$) too, which are functionally equivalent to optimize.

$$L(y, \hat{y}) = \sum_i |y_i - \hat{y}_i|$$

$$L(y, x, \theta) = \sum_i |y_i - \theta \cdot x_i|$$

$$L(y, x, w, b) = \sum_i |y_i - (w \cdot x_i + b)|$$

4 4

Let us assume that y and 1 are column-vectors (that is, they both have dimension $(n \times 1)$), and X has dimension $n \times d$

4.1 a

Sum of a vector is same as dot-product with a 1 vector

$$\frac{1}{n} 1^T y$$

4.2 b

$x^i \cdot x^j$ is the same as $x^i x^{jT}$, which is done automatically in standard matrix multiplication if we transpose the second matrix:

$$X X^T$$

4.3 c

This is similar to (a)

$$\frac{1}{n} 1^T X$$

4.4 d

Matrix multiplication (AB) computes the inner product of row i of matrix A and column j of matrix B in element (i, j). If we want element (i, j) to be the dot product of the i'th feature of x with the j'th feature of x (along dimension with size n), then we need row i of matrix A be an n-vector, and column j of matrix B to be an n vector. This is only achieved by transposing X:

$$\frac{1}{n} X^T X$$

5 5

Because $w \cdot 0 + b = c_0$, we must have it so $b = c_0$. $x^{(i)}$ is the ith standard basis vector, then we just need the i'th element of w to be $c_i - c_0$, since all the other terms are zero-ed anyways (the $-c_0$ is there because we need to "undo" the bias term). Repeating this for all positions, we get $w = (c_1 - c_0, ..., c_d - c_0), b = c_0$.

6 6

6.1 a

When the λ term is zero, the data is fit perfectly. Thus, the training loss is 0.

6.2 b

As λ increases, the norm of w_λ decreases because the model is "punished" for having a large weight vector.

6.3 c

As λ increases, $L(\lambda)$ increases. The model will learn a smaller parameterization of w due to the regularization, and emphasize more information the bias term, leading to predictions that are no longer 100% accurate. This causes the SSE error (without the Ridge regularization term) to increase.

6.4 d

As λ goes to infinity, the weight vector drops to zero. Thus, the model is entirely parameterized by bias, in which the value that yields the lowest loss is the mean (which we proved earlier). Thus, the loss approaches:

$$L(\infty) = \sum_{i=0}^{d} (y^i - b)^2, b = \frac{\sum_i^{d+1} y^i}{d+1}$$

$$= \sum_i y^{i2} - 2\sum_i by^i + \sum_i b^2$$

$$= \sum_i c^{i2} - 2(\frac{\sum_i c_i}{d+1})\sum_i c^i + (d+1)(\frac{(\sum_i c_i)^2}{(d+1)^2})$$

$$= \sum_{i=0}^{d} [c^{(i)2}] - \frac{(\sum_{i=0}^{d} c^{(i)})^2}{d+1}$$

7 7

7.1 a

My strategy is to do perform a Ridge Regression with strong regularization, and then sort by the absolute value of the weights descending. There should be a clearly observed 'gap' in the absolute value of the weights between the 10th highest and 11th highest weight.

7.2 b

Using 0-indexing: $[1, 2, 4, 6, 10, 12, 16, 18, 22, 26]$
Using 1-indexing: $[2, 3, 5, 7, 11, 13, 17, 19, 23, 27]$

8 8

8.1 a, b

To perform my partition, I generated a random permutation of indexes in [0,303] using a fixed seed of (123). The first 200 indexes correspond to indexes of the training points, and the remaining ones are the test points.

Afterwards, I trained 3 separate Logistic models on all 200 training points (with no regularization), each with different normalization methods. These were all trained using sklearn.linear_model.LogisticRegression(). Normalization statistics were only obtained from training data, and the same normalization statistics were used for normalizing training and test data.

To determine influential features, I sorted the weight values by their absolute value, and chose the 3 largest (and the feature those weights correspond to)

Method	**Formula**	**Test Error**
Z-score	$x_{i,j} = \frac{x_{i,j} - mean(x_{:,j})}{std(x_{:,j})}$	$13/103 = 0.1262$
Min-Max 0-1	$x_{i,j} = \frac{x_{i,j} - min(x_{:,j})}{max(x_{:,j}) - min(x_{:,j})}$	$13/103 = 0.1262$
None	$x_{i,j} = x_{i,j}$	$14/103 = 0.1359$

Weights:

- **Z-score**:
  - w: [-0.0366 -0.9783 0.8047 -0.4959 -0.3468 0.1503 0.1725 0.691 -0.4286 -0.7565 0.5111 -0.8314 -0.4907]
  - b: [-0.0234]
  - Influential features: [sex, ca, cp]
- **Min-Max 0-1**:
  - w: [-0.6315 -1.3791 1.6426 -1.1085 -0.6228 0.1735 0.5641 1.4912 -0.867 -1.7894 1.3386 -1.8582 -1.3364]
  - b: [1.4089]
  - Influential features: [ca, oldpeak, cp]
- **None**:
  - w: [0.0184 -1.5697 0.7578 -0.0242 -0.0055 0.1327 0.4142 0.0374 -0.6592 -0.6019 0.7472 -0.7584 -0.7095]
  - b: [0.0501]
  - Influential features: [sex, ca, cp]

8.2 c

When performing cross-validation, I first split the data, and then used the resulting training set to obtain statistics for normalizing both the validation and training sets. Using a random-seed state of (123) (which determines the permutation used for train/test splits), the following results are obtained:

Method	Fold1	Fold2	Fold3	Fold4	Fold5	Average-Fold Error	Q8b Test Error
Z-score	9/40	8/40	10/40	10/40	4/40	0.2050	0.1262
Min-Max 0-1	9/40	9/40	10/40	8/40	2/40	0.1975	0.1262
None	10/40	7/40	9/40	9/40	3/40	0.1950	0.1359

Random Seed State	Z-Score Avg-Error	Min-Max Avg-Error	No Normalization Avg-Error
111	0.1550	0.1575	0.1550
9999	0.1800	0.1725	0.1700
42	0.1850	0.1900	0.1900
122	0.1500	0.1500	0.1517

On average, the estimated test error use 5-fold cross validation tends to be higher than the test error in part 8b. This makes sense since we are training on less data (160 vs 200) which could potentially explain the difference in performance. It also has a high degree of variance based on what train/test split is used and how the folds are divided, which is exacerbated by the fact that we have a very small number of non-training samples (40 validation samples for 8c versus 103 test samples in 8b). This means that smaller variations in performance are essentially amplified by the law of large numbers. It is possible that the result I got in 8b (using seed 123) was exceptionally lucky, whereas many of the trials i tried in part 8c are closer to more typical results.

Additionally, it seems there is no perfect normalization method either; the permutation for the train/test split sometimes favors one method over the other.