CS2109s Tutorial 4

by Lee Zong Xun

Recap

- What is a loss function?
  - An indication of how far off our predictions are compaed to the actual values.

Mean Squared Error

$$J(w_o, w_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})^2$$

Mean Absolute Error

$$J(w_o, w_1) = \frac{1}{m} \sum_{i=1}^{m} \backslash\mathrm{abs}\{h_w(x^{(i)}) - y^{(i)}\}$$

- Linear Regression
  - **Univariate**: Form of $y = w_1 x + w_0$, where x is the input and y is the output.
  - **Multivariate**: Form of $y = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + w_0$, where $x_1, x_2, \ldots, x_n$ are the inputs and y is the output.

> Goal: Find a global minima where the loss function is minimized.

- **Gradient descent** allows us to compute an estimate of the gradient at each point and move a small amount in the steepest downhill direction, until we converge on a point with minimum loss. Since the loss surface is convex, we will always arrive at the global minimum.

Question 1

The loans department of DBN (Development Bank of NUS) wanted to use a decision tree to help them make decisions for new applicants. DBN has the following past loan processing records, each containing an applicant's income, credit history, debt, and the final approval decision. Details are shown in Table 1.

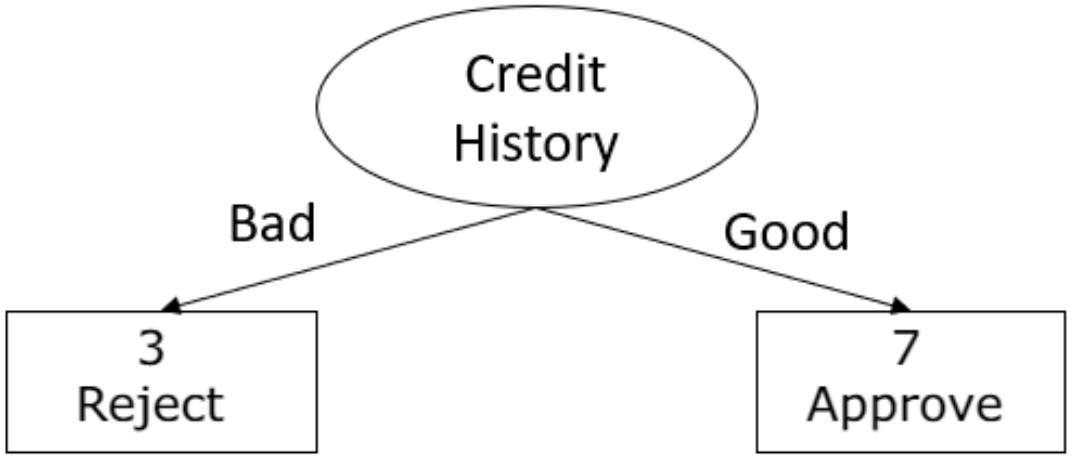
Table 1

Income	Credit History	Debt	Decision
Over 10k	Bad	Low	Reject
Over 10k	Good	High	Approve
0 - 10k	Good	Low	Approve
Over 10k	Good	Low	Approve
Over 10k	Good	Low	Approve
Over 10k	Good	Low	Approve
0 - 10k	Good	Low	Approve
Over 10k	Bad	Low	Reject
Over 10k	Good	High	Approve
0 - 10k	Bad	High	Reject

Table 1: Loan Processing Outcomes

Question 1 (a)

Construct the best decision tree to classify the final outcome (Decision) from the three features Income, Credit History, and Debt.



Question 1(b)

It turns out in the labeling process, one of the data is mislabeled. The data in table 1 is now altered into table 2.

Construct the best decision tree that classifies the data in table 2. Justify this decision tree and show why it performs the best by calculating the information gain values and remainders at each stage.

Income	Credit History	Debt	Decision
Over 10k	Bad	Low	Approve
Over 10k	Good	High	Approve
0 - 10k	Good	Low	Approve
Over 10k	Good	Low	Approve
Over 10k	Good	Low	Approve
Over 10k	Good	Low	Approve
0 - 10k	Good	Low	Approve
Over 10k	Bad	Low	Reject
Over 10k	Good	High	Approve
0 - 10k	Bad	High	Reject

Table 2: Loan Processing Outcomes (Noisy)

$$I(\frac{2}{10},\frac{8}{10}) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{8}{10}\log_2\frac{8}{10} = 0.722$$

$$\begin{aligned}
remainder(Income) &= \frac{3}{10}I(\frac{2}{3},\frac{1}{3}) + \frac{7}{10}I(\frac{6}{7},\frac{1}{7}) \\
&= \frac{3}{10}(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}) + \frac{7}{10}(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}) \\
&= 0.690 \\
remainder(Debt) &= \frac{7}{10}I(\frac{6}{7},\frac{1}{7}) + \frac{3}{10}I(\frac{1}{3},\frac{2}{3}) \\
&= \frac{7}{10}(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}) + \frac{3}{10}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) \\
&= 0.690 \\
remainder(CreditHistory) &= \frac{3}{10}I(\frac{1}{3},\frac{2}{3}) + \frac{7}{10}I(\frac{7}{7},\frac{0}{7}) \\
&= \frac{3}{10}(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}) + \frac{7}{10}(-\frac{7}{7}\log_2\frac{7}{7} - \frac{0}{7}\log_2\frac{0}{7}) \\
&= 0.275
\end{aligned}$$

$$IG(CreditHistory) = I(\frac{2}{10}, \frac{8}{10}) - remainder(CreditHistory)$$
$$= 0.722 - 0.275 = \mathbf{0.447}$$

$$IG(Income) = I(\frac{2}{10}, \frac{8}{10}) - remainder(Income)$$
$$= 0.722 - 0.690 = \mathbf{0.032}$$

$$IG(Debt) = I(\frac{2}{10}, \frac{8}{10}) - remainder(Debt)$$
$$= 0.722 - 0.690 = \mathbf{0.032}$$

Using what we have computed, what should be our initial root?

Credit history has the highest information gain, so we should use it as our root.

Note that we only need to care about the examples that have bad credit history.

Why?

$Income$	$Debt$	$Decision$
$Over10k$	Low	$Approve$
$Over10k$	Low	$Reject$
$0-10k$	$High$	$Reject$

$$I(\frac{1}{3},\frac{2}{3}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$$

$$\begin{aligned}
remainder(Income) &= \frac{1}{3}I(\frac{0}{1},\frac{1}{1}) + \frac{2}{3}I(\frac{1}{2},\frac{1}{2}) \\
&= \frac{1}{3}(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}) + \frac{2}{3}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) \\
&= \mathbf{0.667} \\
remainder(Debt) &= \frac{2}{3}I(\frac{1}{2},\frac{1}{2}) + \frac{1}{3}I(\frac{0}{1},\frac{1}{1}) \\
&= \frac{2}{3}(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}) + \frac{1}{3}(-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}) \\
&= \mathbf{0.667}
\end{aligned}$$

Since Debt has the same gain as Income, we can arbitrarily choose the root.

$$IG(Income) = I(\frac{1}{3}, \frac{2}{3}) - remainder(Income)$$
$$= 0.918 - 0.667 = \mathbf{0.251}$$

$$IG(Debt) = I(\frac{1}{3}, \frac{2}{3}) - reminder(Debt)$$
$$= 0.918 - 0.667 = \mathbf{0.251}$$

Try drawing out the final tree! What do you notice?

Question 1(c)

What is the decision made by the decision tree in part (b) for a person with an income over 10k, a bad credit history, and low debt?

Question 1(d)

Let's consider a scenario where you desire a Decision Tree with each leaf node representing a minimum of 3 training data points. Derive the tree by pruning the tree you previously obtained in part (c). Which data(s) do you think are likely outlier(s)

Question 2

You are given several data points as follows:

x_1	x_2	x_3	y
6	4	11	20
8	5	15	30
12	9	25	50
2	1	3	7

Apply the Normal Equation formula to obtain a linear regression model that minimizes MSE of the data points.

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

Solution

$$X = \begin{bmatrix} 1 & 6 & 4 & 11 \\ 1 & 8 & 5 & 15 \\ 1 & 12 & 9 & 25 \\ 1 & 2 & 1 & 3 \end{bmatrix}, \quad Y = \begin{bmatrix} 20 \\ 30 \\ 50 \\ 7 \end{bmatrix} \implies w = \begin{bmatrix} 4 & -5.5 & -7 & 7 \end{bmatrix}^T$$

The final equation is $\hat{y} = 4 - 5.5x_1 - 7x_2 + 7x_3$.

But wait

Normal Equation needs the calculation of $(X^T X)^{-1}$. But sometimes the matrix is not invertible. When will that happen, and what should we do in that situation?

Recall that for invertibility: $X^T X$ is invertible if and only if the columns of X are linearly independent.

- Remove linearly independent columns via preprocessing!

Question 3(a)

For Linear Regression, there are two popular cost functions,

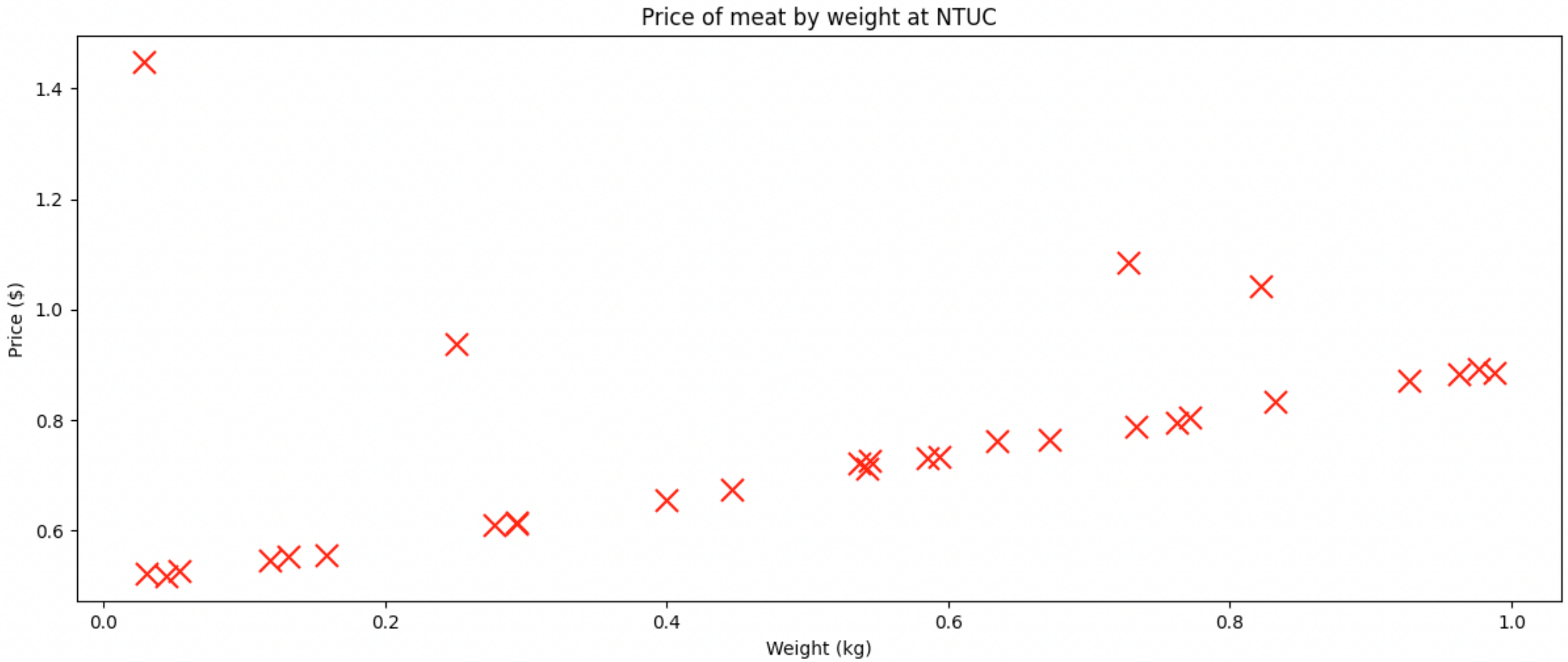
Mean Squared Error:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

and **Mean Absolute Error**:

$$L(y, \hat{y}) = \frac{1}{2}|y - \hat{y}|$$

Given the scatter plot of a dataset containing the actual weight of meat at NTUC (x) and its price (y), justify your choice of cost function for this problem



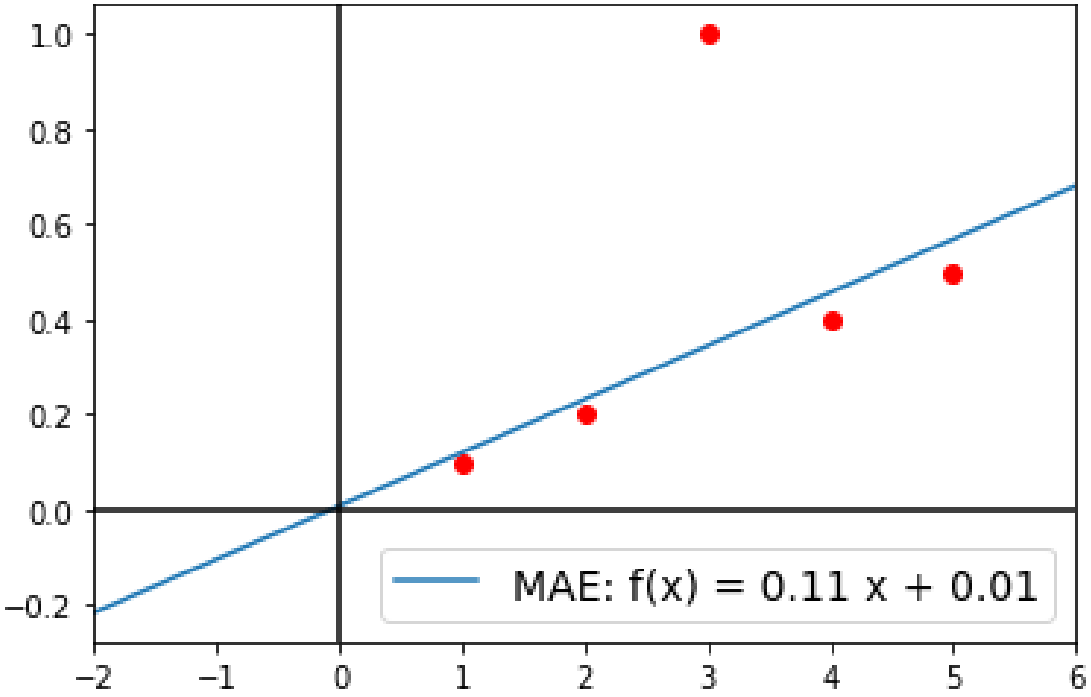
Solution

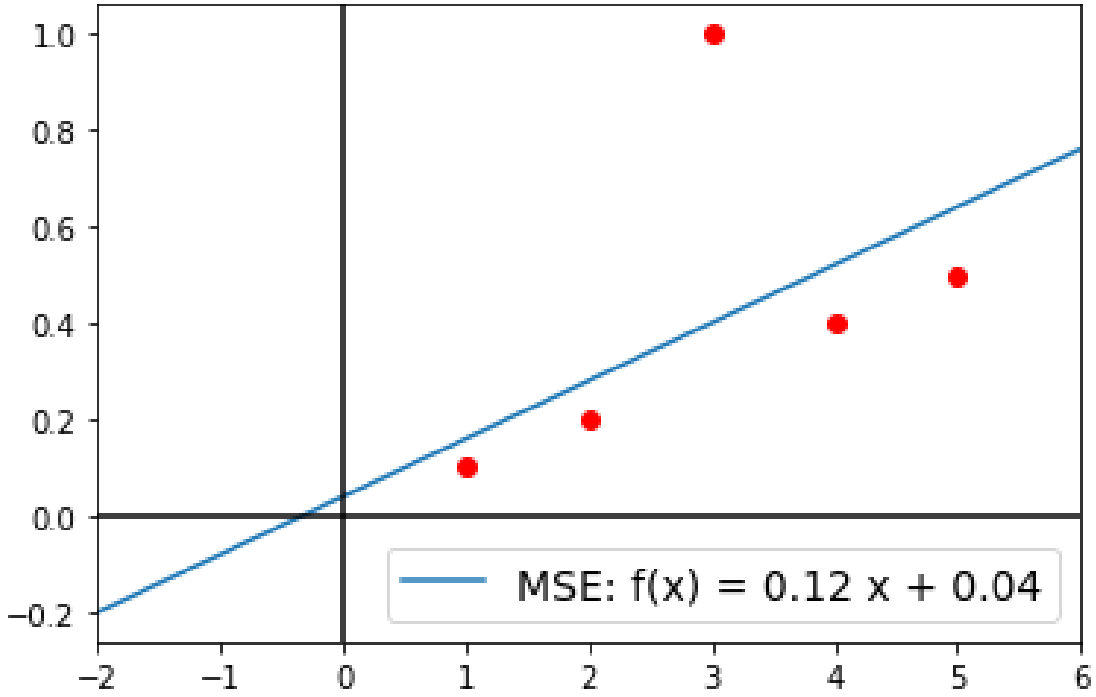
The answer is it depends!

- Are there many outliers in the graph? Yes
- Should outliers be penalized heavily? Yes/No
- Should outliers have a smaller impact? Yes/No

If we consider outliers as important and should be penalized heavily, MSE may be the preferred metric. If outliers are considered less important and should have a smaller impact, MAE may be the preferred metric.

MSE vs MAE


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−2
−1
0
1
2
3
4
5
6
MAE: f(x) = 0.11 x + 0.01


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−2
−1
0
1
2
3
4
5
6
MSE: f(x) = 0.12 x + 0.04

Question 3(b)

Can you provide examples of cost functions that are better suited to handle outliers more effectively?

1. **Huber loss** is a combination of MSE and MAE and is defined as:

$$L(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq \delta \\ \delta \cdot |y - \hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

where δ is a threshold that determines the transition between the MSE and MAE behaviors.

2. **Log cosh** approximates MSE and MAE and is similar to the Huber loss function.

$$L(y, \hat{y}) = \log(\cosh(y_i - \hat{y}_i))$$

- For small values of x, $\log(\cosh(x)) \approx \frac{1}{2}x^2$, which is similar to MSE.
- For larger values of x, $\log(\cosh(x)) \approx |x| - \log(2)$, which is similar to MAE.

Question 4

Given a simple function $y = x^2$, we know the gradient is $\frac{dy}{dx} = 2x$. As such, the minimum of this function is 0.

Record the value of $a = (x, y)$ (where $y = x^2$) over **5 iterations** for each learning rate in tabular format. Observe the oscillations of the value and the convergence to $(0, 0)$.

$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$	$\alpha = 0.01$
5	5	5	5
-95	-5	4.0	4.9
1805	5	3.2	4.802
-34295	-5	2.56	4.706
651605	5	2.048	4.612

Question 4(b)

During the course of training for a large number of epochs/iterations, what can be done to the value of the learning rate alpha to enable better convergence?

Learning rate α can be decreased through the course of training

Note: Look up on learning rate schedulers!