Machine Learning - Homework 2

Mohammad Bahrami - 9724133

November 15, 2021

---

1 Question 1

1.1 A

$$L(\Theta) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

The only parameter in this distribution is p so we put Θ to be p. To find the Maximum Likelihood Estimation of L, we put the derivative with respect to p to be 0. To simplify our calculations, first we calculate the *log* of $L(\Theta)$.

$$\log L(\Theta) = \log \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = \log p \sum_{i=1}^{n} X_i + \log(1-p) \sum_{i=1}^{n} 1 - X_i$$

$$\frac{\partial \log L(\Theta = p)}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{\sum_{i=1}^{n} 1 - X_i}{1-p}$$

Now we put this derivative to be 0.

$$\frac{\partial \log L(\Theta = p)}{\partial p} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{\sum_{i=1}^{n} 1 - X_i}{1-p} = 0$$

$$\implies \sum_{i=1}^{n} X_i - p\sum_{i=1}^{n} X_i = p\sum_{i=1}^{n}(1 - X_i) \implies p = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\Theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

1.2 B

We put $\Theta = p$. The Maximum A-Posteriori estimate is defined as

$$\hat{\Theta}_{MAP} = \arg\max_{\Theta} P(\Theta \mid X)$$

Because $P(X)$ is not dependant on Θ

$$\begin{aligned}
\hat{\Theta}_{MAP} &= \arg\max_{\Theta} P(\Theta \mid X) \\
&= \arg\max_{\Theta} \frac{P(X \mid \Theta)P(\Theta)}{P(X)} \\
&= \arg\max_{\Theta} P(X \mid \Theta)P(\Theta)
\end{aligned}$$

$$= \arg\max_{\Theta} \prod_{i=1}^{n} P(X_i \mid \Theta) P(\Theta)$$

Again, like part A, it is more convenient to use log in our calculations to make our lives easier

$$\begin{aligned}
\hat{\Theta}_{MAP} &= \log(\arg\max_{\Theta} P(\Theta \mid X)) \\
&= \arg\max_{\Theta} \log(\prod_{i=1}^{n} P(X_i \mid \Theta) P(\Theta)) \\
&= \arg\max_{\Theta} \sum_{i=1}^{n} \log P(X_i \mid \Theta) + \log P(\Theta)
\end{aligned}$$

Recall that we have

$$Posterior \propto Likelihood \cdot Prior$$

Also we have

$$Likelihood \equiv L(X_i \mid \Theta) = \prod_{i=1}^{n} \Theta^{X_i} (1-\Theta)^{1-X_i}$$

$$Prior \equiv Beta(\Theta \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \Theta^{\alpha-1} (1-\Theta)^{\beta-1}$$

Thus

$$P(\Theta \mid X) \propto \Big\{ \prod_{i=1}^{n} \Theta^{X_i} (1-\Theta)^{1-X_i} \Big\} \cdot \frac{1}{B(\alpha, \beta)} \Theta^{\alpha-1} (1-\Theta)^{\beta-1}$$

The fact that $P(X)$ is not dependant on Θ and the equation above together, give us

$$\log P(\Theta \mid X) = \Big\{ \sum_{i=1}^{n} \log \Theta^{X_i} (1-\Theta)^{1-X_i} \Big\} + \log \frac{1}{B(\alpha, \beta)} \Theta^{\alpha-1} (1-\Theta)^{\beta-1}$$

We can see that the first part of the right hand side is exactly what we had in part A. Now it becomes clear that MAP's difference with MLE is that MAP takes the observations of the past, in the form of the $Beta$ distribution, into account.

Now what we want is to find the maximum value of $P(\Theta \mid X)$. To do so, we take the partial derivative with respect to Θ and put it to be 0.

$$\begin{aligned}
\frac{\partial P(\Theta \mid X)}{\partial \Theta} = \Big\{ \frac{\sum_{i=1}^{n} X_i}{\Theta} - \frac{\sum_{i=1}^{n} 1 - X_i}{1-\Theta} \Big\} + \Big\{ \frac{\alpha - 1}{\Theta} - \frac{\beta - 1}{1-\Theta} \Big\} = 0 \\
\implies (1-\Theta) \Big[\sum_{i=1}^{n} (X_i) + \alpha - 1 \Big] = \Theta \Big[\sum_{i=1}^{n} (1 - X_i) + \beta - 1 \Big] \\
\implies \sum_{i=1}^{n} (X_i) + \alpha - 1 = \Theta \Big[\sum_{i=1}^{n} (1 - X_i + X_i) + \alpha + \beta - 2 \Big] \\
\implies \sum_{i=1}^{n} (X_i) + \alpha - 1 = \Theta \Big[\sum_{i=1}^{n} (1) + \alpha + \beta - 2 \Big] \\
\implies \sum_{i=1}^{n} (X_i) + \alpha - 1 = \Theta \Big[n + \alpha + \beta - 2 \Big] \\
\implies \frac{\sum_{i=1}^{n} (X_i) + \alpha - 1}{n + \alpha + \beta - 2} = \Theta
\end{aligned}$$

Finally, this gives us the result

$$\hat{\Theta}_{MAP} = \frac{\sum_{i=1}^{n} (X_i) + \alpha - 1}{n + \alpha + \beta - 2}$$

2 Question 2

Linear equation

$$Oxygen = W[0] + W[1] \times Age + W[2] \times HeartBeat$$

Data

$$X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} X_0 & Age & HeartBeat \end{array} \\ \begin{bmatrix} 1 & 41 & 138 \\ 1 & 42 & 153 \\ 1 & 37 & 151 \\ 1 & 46 & 133 \end{bmatrix} \end{array} \qquad Y = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{c} Oxygen \end{array} \\ \begin{bmatrix} 37.99 \\ 47.34 \\ 44.38 \\ 28.17 \end{bmatrix} \end{array}$$

Initial Parameters

$$W^{(0)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \\ \begin{bmatrix} -59.5 \\ -0.15 \\ 0.6 \end{bmatrix} \end{array}$$

2.1 A

The original equation of Mean Square Error if $X^{(i)}$ is a column vector of i'th data is

$$MSE_W(X, Y) = \frac{1}{m} \sum_{i=1}^{m} (W^t X^{(i)} - Y^{(i)})^2$$

Also, the original equation of Mean Absolute Error if $X^{(i)}$ is a column vector of i'th data is

$$MAE_W(X, Y) = \frac{1}{m} \sum_{i=1}^{m} \left| W^t X^{(i)} - Y^{(i)} \right|$$

From now on, I will utilize vectorization in my operations as much as possible. Some formulas may change dou to this fact. Now, to calculate these cost functions using the $W^{(0)}$ we have

$$\begin{aligned} MSE_{W^{(0)}}(X, Y) = \frac{1}{2m} \sum (XW - Y).^{2\ [1]} &= \frac{1}{8} \sum (\begin{bmatrix} 1 & 41 & 138 \\ 1 & 42 & 153 \\ 1 & 37 & 151 \\ 1 & 46 & 133 \end{bmatrix} \begin{bmatrix} -59.5 \\ -0.15 \\ 0.6 \end{bmatrix} - \begin{bmatrix} 37.99 \\ 47.34 \\ 44.38 \\ 28.17 \end{bmatrix}).^2 \\ &= \frac{1}{8} \sum (\begin{bmatrix} 17.15 \\ 26 \\ 25.55 \\ 13.4 \end{bmatrix} - \begin{bmatrix} 37.99 \\ 47.34 \\ 44.38 \\ 28.17 \end{bmatrix}).^2 \\ &= \frac{1}{8} \sum (\begin{bmatrix} -20.84 \\ -21.34 \\ -18.83 \\ -14.77 \end{bmatrix}).^2 \\ &= \frac{1}{8} \sum \begin{bmatrix} 434.3056 \\ 455.3956 \\ 354.5689 \\ 218.1529 \end{bmatrix} \\ &= \frac{1}{8} 1462.423 = 182.802 \end{aligned}$$

[1] $(M).^a$ is element wise exponentiation of matrix M to the power of a

$$
\begin{aligned}
MAE_{W^{(0)}}(X,Y) = \frac{1}{m}\sum|XW - Y| &= \frac{1}{4}\sum|\begin{bmatrix}1 & 41 & 138\\ 1 & 42 & 153\\ 1 & 37 & 151\\ 1 & 46 & 133\end{bmatrix}\begin{bmatrix}-59.5\\ -0.15\\ 0.6\end{bmatrix} - \begin{bmatrix}37.99\\ 47.34\\ 44.38\\ 28.17\end{bmatrix}| \\
&= \frac{1}{4}\sum|\begin{bmatrix}17.15\\ 26\\ 25.55\\ 13.4\end{bmatrix} - \begin{bmatrix}37.99\\ 47.34\\ 44.38\\ 28.17\end{bmatrix}| \\
&= \frac{1}{4}\sum|\begin{bmatrix}-20.84\\ -21.34\\ -18.83\\ -14.77\end{bmatrix}| \\
&= \frac{1}{4}\sum\begin{bmatrix}20.84\\ 21.34\\ 18.83\\ 14.77\end{bmatrix} \\
&= \frac{1}{4}75.78 = 18.945
\end{aligned}
$$

2.2 B

The Gradient Descent Step

$$W^{(t+1)} = W^{(t)} - \alpha\frac{\partial J_{W^{(t)}}(X,Y)}{\partial W^{(t)}}$$

To be able to perform a gradient descent step, we need to first calculate the partial derivative of the cost function, in our case MSE, with respect to the model parameters.

$$\frac{\partial MSE_W(X,Y)}{\partial W} = \frac{1}{m}X^t(XW - Y) \quad [2]$$

For the Stochastic Gradient Descent, we first pick one random sample of the data then we perform one gradient step with that sample only.

For the **first step**, the chosen random one-based index of data is $idx = 3$

$$
\begin{aligned}
W^{(1)} &= W^{(0)} - \alpha\frac{\partial J_{W^{(0)}}(X^{(3)},Y^{(3)})}{\partial W^{(0)}} \\
W^{(1)} &= W^{(0)} - \alpha X^{(3)t}(X^{(3)}W^{(0)} - Y^{(3)}) \\
W^{(1)} &= W^{(0)} - 0.1\begin{bmatrix}1\\ 37\\ 151\end{bmatrix}(\begin{bmatrix}1 & 37 & 151\end{bmatrix}\begin{bmatrix}-59.5\\ -0.15\\ 0.6\end{bmatrix} - 44.38) \\
W^{(1)} &= W^{(0)} - 0.1\begin{bmatrix}1\\ 37\\ 151\end{bmatrix}(25.55 - 44.38) \\
W^{(1)} &= W^{(0)} - 0.1\begin{bmatrix}1\\ 37\\ 151\end{bmatrix}(-18.83)
\end{aligned}
$$

[2] On the condition that $X_0^{(i)} = 1$ for any $1 \le i \le m$

$$W^{(1)} = \begin{bmatrix} -59.5 \\ -0.15 \\ 0.6 \end{bmatrix} - \begin{bmatrix} -1.883 \\ -69.671 \\ -284.333 \end{bmatrix}$$
$$W^{(1)} = \begin{bmatrix} -57.617 \\ 69.521 \\ 284.933 \end{bmatrix}$$

For the **second step**, the chosen random one-based index of data is $idx = 2$

$$W^{(2)} = W^{(1)} - \alpha \frac{\partial J_{W^{(1)}}(X^{(2)}, Y^{(2)})}{\partial W^{(1)}}$$
$$W^{(2)} = W^{(1)} - \alpha X^{(2)t}(X^{(2)}W^{(1)} - Y^{(2)})$$
$$W^{(2)} = W^{(1)} - 0.1 \begin{bmatrix} 1 \\ 42 \\ 153 \end{bmatrix} (\begin{bmatrix} 1 & 42 & 153 \end{bmatrix} \begin{bmatrix} -57.617 \\ 69.521 \\ 284.933 \end{bmatrix} - 47.34)$$
$$W^{(2)} = W^{(1)} - 0.1 \begin{bmatrix} 1 \\ 42 \\ 153 \end{bmatrix} (46457.014 - 47.34)$$
$$W^{(2)} = W^{(1)} - 0.1 \begin{bmatrix} 1 \\ 42 \\ 153 \end{bmatrix} (46409.674)$$
$$W^{(2)} = \begin{bmatrix} -57.617 \\ 69.521 \\ 284.933 \end{bmatrix} - \begin{bmatrix} 4640.9674 \\ 194920.63 \\ 710068.01 \end{bmatrix}$$
$$W^{(2)} = \begin{bmatrix} -4698.58 \\ -194851.10 \\ -709783.07 \end{bmatrix}$$

For the **third step**, the chosen random one-based index of data is $idx = 4$

$$W^{(3)} = W^{(2)} - \alpha \frac{\partial J_{W^{(2)}}(X^{(4)}, Y^{(4)})}{\partial W^{(2)}}$$
$$W^{(3)} = W^{(2)} - \alpha X^{(4)t}(X^{(4)}W^{(2)} - Y^{(4)})$$
$$W^{(3)} = W^{(2)} - 0.1 \begin{bmatrix} 1 \\ 46 \\ 133 \end{bmatrix} (\begin{bmatrix} 1 & 46 & 133 \end{bmatrix} \begin{bmatrix} -4698.58 \\ -194851.10 \\ -709783.07 \end{bmatrix} - 28.17)$$
$$W^{(3)} = W^{(2)} - 0.1 \begin{bmatrix} 1 \\ 46 \\ 133 \end{bmatrix} (-1.03368999e+08 - 28.17)$$
$$W^{(3)} = W^{(2)} - 0.1 \begin{bmatrix} 1 \\ 46 \\ 133 \end{bmatrix} (-1.03369027e+08)$$
$$W^{(3)} = \begin{bmatrix} -4698.58 \\ -194851.10 \\ -709783.07 \end{bmatrix} - \begin{bmatrix} -1.03369027e+07 \\ -4.75497526e+08 \\ -1.37480806e+09 \end{bmatrix}$$
$$W^{(3)} = \begin{bmatrix} 1.03322041e+07 \\ 4.75302675e+08 \\ 1.37409828e+09 \end{bmatrix}$$

For the **fourth step**, the chosen random one-based index of data is $idx = 3$

$$
\begin{aligned}
W^{(4)} &= W^{(3)} - \alpha \frac{\partial J_{W^{(3)}}(X^{(3)}, Y^{(3)})}{\partial W^{(3)}} \\
W^{(4)} &= W^{(3)} - \alpha X^{(3)t}(X^{(3)}W^{(3)} - Y^{(3)}) \\
W^{(4)} &= W^{(3)} - 0.1 \begin{bmatrix} 1 \\ 37 \\ 151 \end{bmatrix} (\begin{bmatrix} 1 & 37 & 151 \end{bmatrix} \begin{bmatrix} 1.03322041e+07 \\ 4.75302675e+08 \\ 1.37409828e+09 \end{bmatrix} - 44.38) \\
W^{(4)} &= W^{(3)} - 0.1 \begin{bmatrix} 1 \\ 37 \\ 151 \end{bmatrix} (2.25085372e+11 - 44.38) \\
W^{(4)} &= W^{(3)} - 0.1 \begin{bmatrix} 1 \\ 37 \\ 151 \end{bmatrix} (2.25085371e+11) \\
W^{(4)} &= \begin{bmatrix} 1.03322041e+07 \\ 4.75302675e+08 \\ 1.37409828e+09 \end{bmatrix} - \begin{bmatrix} 2.25085371e+10 \\ 8.32815874e+11 \\ 3.39878911e+12 \end{bmatrix} \\
W^{(4)} &= \begin{bmatrix} -2.24982049e+10 \\ -8.32340572e+11 \\ -3.39741501e+12 \end{bmatrix}
\end{aligned}
$$

Now we can calculate $MSE_{W^{(4)}}(X, Y)$

$$
\begin{aligned}
MSE_{W^{(4)}}(X,Y) = \frac{1}{2m}\sum (XW - Y).^2 &= \frac{1}{8}\sum (\begin{bmatrix} 1 & 41 & 138 \\ 1 & 42 & 153 \\ 1 & 37 & 151 \\ 1 & 46 & 133 \end{bmatrix} \begin{bmatrix} -2.24982049e+10 \\ -8.32340572e+11 \\ -3.39741501e+12 \end{bmatrix} - \begin{bmatrix} 37.99 \\ 47.34 \\ 44.38 \\ 28.17 \end{bmatrix}).^2 \\
&= \frac{1}{8}\sum (\begin{bmatrix} -5.02991733e+14 \\ -5.54785299e+14 \\ -5.43828766e+14 \\ -4.90166361e+14 \end{bmatrix} - \begin{bmatrix} 37.99 \\ 47.34 \\ 44.38 \\ 28.17 \end{bmatrix}).^2 \\
&= \frac{1}{8}\sum (\begin{bmatrix} -5.02991733e+14 \\ -5.54785299e+14 \\ -5.43828766e+14 \\ -4.90166361e+14 \end{bmatrix}).^2 \\
&= \frac{1}{8}\sum \begin{bmatrix} 2.53000684e+29 \\ 3.07786728e+29 \\ 2.95749727e+29 \\ 2.40263061e+29 \end{bmatrix} \\
&= \frac{1}{8} 1.0968001997484895e+30 \\
&= 1.371000249685612e+29
\end{aligned}
$$

We can see that because the learning rate is too big, the algorithm is diverging very rapidly and the loss have reached to magnitudes of $e + 29$.

2.3 C

$$W^{(1)} = W^{(0)} - \alpha \frac{\partial J_{W^{(0)}}(X,Y)}{\partial W^{(0)}}$$

$$W^{(1)} = W^{(0)} - \alpha \frac{1}{m} X^t (XW^{(0)} - Y)$$

$$W^{(1)} = W^{(0)} - 0.1 \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 41 & 42 & 37 & 46 \\ 138 & 153 & 151 & 133 \end{bmatrix} \left(\begin{bmatrix} 1 & 41 & 138 \\ 1 & 42 & 153 \\ 1 & 37 & 151 \\ 1 & 46 & 133 \end{bmatrix} \begin{bmatrix} -59.5 \\ -0.15 \\ 0.6 \end{bmatrix} - \begin{bmatrix} 37.99 \\ 47.34 \\ 44.38 \\ 28.17 \end{bmatrix} \right)$$

$$W^{(1)} = W^{(0)} - 0.1 \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 41 & 42 & 37 & 46 \\ 138 & 153 & 151 & 133 \end{bmatrix} \begin{bmatrix} -20.84 \\ -21.34 \\ -18.83 \\ -14.77 \end{bmatrix}$$

$$W^{(1)} = W^{(0)} - 0.1 \frac{1}{4} \begin{bmatrix} -75.78 \\ -3126.85 \\ -10948.68 \end{bmatrix}$$

$$W^{(1)} = \begin{bmatrix} -59.5 \\ -0.15 \\ 0.6 \end{bmatrix} - \begin{bmatrix} -1.8945 \\ -78.1712 \\ -273.717 \end{bmatrix}$$

$$W^{(1)} = \begin{bmatrix} -57.6 \\ 78.02 \\ 274.31 \end{bmatrix}$$

Now we can calculate $MSE_{W^{(1)}}(X,Y)$

$$MSE_{W^{(1)}}(X,Y) = \frac{1}{2m} \sum (XW - Y).^2 = \frac{1}{8} \sum \left(\begin{bmatrix} 1 & 41 & 138 \\ 1 & 42 & 153 \\ 1 & 37 & 151 \\ 1 & 46 & 133 \end{bmatrix} \begin{bmatrix} -57.6 \\ 78.02 \\ 274.31 \end{bmatrix} - \begin{bmatrix} 37.99 \\ 47.34 \\ 44.38 \\ 28.17 \end{bmatrix} \right).^2$$

$$= \frac{1}{8} \sum \left(\begin{bmatrix} 40997.011 \\ 45189.788 \\ 44251.047 \\ 40015.533 \end{bmatrix} - \begin{bmatrix} 37.99 \\ 47.34 \\ 44.38 \\ 28.17 \end{bmatrix} \right).^2$$

$$= \frac{1}{8} \sum \left(\begin{bmatrix} 40959.021 \\ 45142.448 \\ 44206.667 \\ 39987.363 \end{bmatrix} \right).^2$$

$$= \frac{1}{8} \sum \begin{bmatrix} 1.67764146e+09 \\ 2.03784061e+09 \\ 1.95422947e+09 \\ 1.59898920e+09 \end{bmatrix}$$

$$= \frac{1}{8} 7268700747.402$$

$$= 908587593.425$$

3 Question 4

3.1 A

$$f(x) = \begin{cases} +\infty & x > \sqrt{2} \\ \frac{1}{2}x^2 & -\sqrt{2} \leq x \leq \sqrt{2} \\ +\infty & x < \sqrt{2} \end{cases}$$

$$f(x) \geq f(x_0) + g^T(x - x_0) \Rightarrow \frac{1}{2}x^2 \geq \frac{1}{2}x_0^2 + m(x - x_0) \Rightarrow \frac{1}{2}(x^2 - x_0^2) \geq m(x - x_0)$$

$$\begin{aligned} &\rightarrow x_0 = \sqrt{2} \\ &\implies \frac{1}{2}(x^2 - 2) \geq m(x - \sqrt{2}) \\ &\implies \begin{cases} 0 \geq 0 & x = \sqrt{2} \\ \infty \geq m & x > \sqrt{2} \\ \frac{1}{2}(x + \sqrt{2}) \leq m & x < \sqrt{2} \end{cases} \end{aligned}$$

$$\begin{aligned} &\rightarrow x_0 = -\sqrt{2} \\ &\implies \frac{1}{2}(x^2 - 2) \geq m(x + \sqrt{2}) \\ &\implies \begin{cases} 0 \geq 0 & x = -\sqrt{2} \\ \frac{1}{2}(x - \sqrt{2}) \geq m & x > -\sqrt{2} \\ -\infty \leq m & x < -\sqrt{2} \end{cases} \end{aligned}$$

$$\implies f'(x) = \begin{cases} +\infty & x > \sqrt{2} \\ [\sqrt{2}, +\infty) & x = \sqrt{2} \\ x & -\sqrt{2} < x < \sqrt{2} \\ (-\infty, -\sqrt{2}] & x = -\sqrt{2} \\ \infty & x < -\sqrt{2} \end{cases}$$

At the border points, we can put $f'(x) = x$ to summarize $f'(x)$ to

$$f'(x) = \begin{cases} +\infty & x > \sqrt{2} \\ x & -\sqrt{2} \leq x \leq \sqrt{2} \\ \infty & x < -\sqrt{2} \end{cases}$$

3.2 B

No. As we can see, at the border points, sub gradient can be any number in the corresponding range for that border point. In the other hand, sub gradient is unique in every other point except the border points.

3.3 C

Methods with a cost function that has non-differentiable points need sub gradient calculation. for example, mean absolute error cost function.

4 Question 5

In plot A we can see that the value of cost function has decreased very quickly and after that it has stayed at the same amount. This is the plot for a good learning rate.

In the other hand, In plot B we see that the value of cost function is decreasing but in a very slow pace. This is the plot for a small learning rate, which will cause to convergence, but very slowly and not efficiently.

In contrast to B, In plot C we can see that the value of cost function is increasing in every step. Also, the speed of the increase is also increasing. This means that the chosen learning rate is too big and is causing a divergence.

---

5 Question 6

In linear regression, what we are predicting is the line, so for every sample of data, we are predicting a point on the line and we want to know the difference of the real value at the same sample data point. To achieve this, we utilize vertical offsets $(h_\Theta(X) - Y)$. If we where to use perpendicular offsets, the value of the data sample,Y, may be different with the value of the point of the line that we are using to calculate the error.

In other words, if $\hat{Y} = h_\Theta(X')$ and $Y = f(X)$, we want X' to be exactly X and we cannot achieve this condition with perpendicular offsets.

---

6 Question 7

6.1 A

Mean Absolute Error. Because it will keep focusing on the main data and doesn't make the model go towards caring about the outliers more than the legit data. If we use MSE for example, the error value for the outliers will be too big and the model will try to minimize this big penalty of outliers instead of the main data.

6.2 B

Mean Square Error. When the prediction and labels are the same, it is better to use a function that is differentiable at $h_\Theta(X) - Y = 0$. MSE has this property but if we use MAE, we will have to define a sub gradient for it.