Machine Learning HW1

Bohan Wu

February 2020

1 Problem1

1.1 (a)

College GPA

1.2 (b)

This target variable is discrete-valued

1.3 (c)

Number of hours student spend on studying per week

1.4 (d)

I think the linear model will be reasonable. The slope will be positive because the more hours students spend on studying per week the more likely students get better grades and higher GPA, which means more likely students get success.

2 Problem2

2.1 (a)

$$L(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2 \tag{1}$$

2.2 (b)

$$\frac{\partial L}{\partial \beta} = 0 \tag{2}$$

By doing derivation of equation(1) based on equation(2), we can get:

$$0 = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \beta x_i^2 \tag{3}$$

By simplify equation(3), we can get:

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{4}$$

Since:

$$\beta 1 = \frac{\sigma_{xy}}{\sigma_x^2} \tag{5}$$

Based on the result we get in equation(4), we can see that we get the different expression that we got for β_1 in the full linear model.

3 Problem3

3.1 (a)

$$\frac{\partial L}{\partial m} = 0 \tag{6}$$

By doing the derivation of equation(6), we can get:

$$0 = \sum_{i=1}^{n} 2(y_i - m)(-1) \tag{7}$$

By simplify equation(6), we can get:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} m \tag{8}$$

Based on above equation, we can find the sum of f_i equal to sum of m. We can know only when m is equal to mean of all y values, this equation will be valid. Therefore, we can conclude $m = \bar{y}$ will minimize the loss function $L(m)$

3.2 (b)

$m = \frac{max(y)+min(y)}{2}$ will make sure we minimize this loss function.

3.3 (d)

1. When all values in data are relatively close the mean of data. And there aren't too many extremely large or small values. we can choose to use squared loss function
2. When all values in data are relatively close the median of data.It's fine to

have some extremely values. we can choose to use l_1 lease absolute deviations
3. When all values are pretty convergent which means there is no any extremely large or extremely small values. we can choose to use l_∞ loss

Yes, because mean could be influenced greatly by extreme values (extreme large or small values), but median will have less influences caused by extreme values. During the data collection, there will always be some extreme values caused by miscalculation or inaccuracy. Median will help us to prevent from those extreme values.

4 Problem4

4.1 (a)

Do the logarithmic on both side of given equation, we can get

$$\log(c) = -at + \log z_0 \tag{9}$$

Based on equation(9), we can see time t and log(Concentration) performs a linear relationship with slope $-a$. Then we can do linear regression on equation(9) and find the β_0 and β_1. In this case, $a = -\beta_1$ and $z_0 = e^{\beta_0}$

4.2 (b)

```
#import the numpy lib
import numpy as np

#Get concentration and time data
con = df['Concentration']
time = df['Time']

#Create the linear fit function, output beta0 beta1 and loss
def fit_linear(x,y):

    #calculate the mean of x and y
    xm = np.mean(x)
    ym = np.mean(y)

    #calculate the var of xx, yy, and xy
    syx = np.mean((y-ym)*(x-xm))
    sxx = np.mean((x-xm)**2)
    syy = np.mean((x-xm)**2)

    #calculate beta0, beta1, and loss
    beta1 = syx/sxx
    beta0 = ym - beta1*xm
```

```
    loss = np.sum((y-beta0-beta1*x)**2)

    return beta0, beta1, loss

logCon = np.log(Con)

result = fit_linear(time, logCon)

alpha = -1 * result[1]
z0 = np.exp(result[0])
```