

hw1

Liheng Cao

February 15, 2021

1

(a)

This is regression because salary is a continuous variable. $N = 500, d = 3$

(b)

The is classification because the 2 possible outcomes are “success” or “failure”. $N = 20, d = 14$

2

(a)

1. Given previous stock history, we could classify whether to buy, sell, or hold (not do anything) the stock. The features would be previous stock prices, and the target would be buy, sell, or hold. If we were to build a more advanced model, we could also include stuff like positive or negative media attention.
2. We could predict whether or not we will have a snow day (too much snow so school is canceled) given predicted snowfall, temperature, and whether or not school was canceled before for a certain amount of snowfall and temperature. The target would be YES snowday or NO snowday, and the features would be snowfall and temperature.
3. We could classify handwritten characters. The target would be what character it is (we could specify to just number or also include the latin alphabet), and the features would be a matrix representation of a picture of the character.

(b)

1. We could predict what exam score we get based on how many hours we study for it. The target would be the score, while the feature would be how much time we spent studying.
2. We could try to predict how long we have to wait at a traffic light. The target would be how long we have to wait before green, while the features could be time of day, day of week, the color of the light when we arrived, and perhaps how long we've already waited.
3. We could use regression to predict how long it would take a battery to be fully charged. The target is the time from 0 to 100%, while the features would be battery capacity, charge "power", the ambient and battery temperature (for battery chemistry), and whether or not the device is currently in use.

3

3.1 a)

A possible target would be GPA.

3.2 b)

Continuous

3.3 c)

1. GPA in high school
2. Test scores (ACT, SAT, etc.)

3.4 d)

It would be a good start. The slope would be positive, as it's reasonable to expect someone that did well in high school to do well in college.

4

4.1 a)

$$\bar{x} = 2, \bar{y} = 6$$

4.2 b)

$$s_{xx} = \frac{1}{5} \sum_{i=0}^5 \left(\bar{x} - x^{(i)} \right)^2 = 2$$

$$s_{yy} = \frac{1}{5} \sum_{i=0}^5 \left(\bar{y} - y^{(i)} \right)^2 = 37.2$$

$$s_{xy} = \frac{1}{5} \sum_{i=0}^5 \left(x^{(i)} - \bar{x} \right) \left(y^{(i)} - \bar{y} \right) = 8$$

4.3 c)

$$w_1 = \frac{s_{xy}}{s_{xx}} = 4$$

$$w_0 = \bar{y} - w_1 \bar{x} = -2$$

$$y \approx -2 + 4x$$

4.4 d)

$$\hat{y} = -2 + 4 \cdot 2.5 = 8$$

4.5 e)

$$\frac{1}{N} \sum_{i=0}^N \left(w_0 + w_1 x^{(i)} - y^{(i)} \right)^2 = 26$$

4.6 f)

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^N \left(y^{(i)} - \bar{y} \right)^2} = 0.86 \dots$$

4.7 g)

If 17 was swapped for 15, then w_0 would increase slightly from -2 to -1.4 . w_1 would decrease slightly from 4 to 3.6.

If 17 was swapped for 9, then w_0 would increase drastically from -2 to -0.4 . w_1 would decrease drastically from 4 to 2.4.

5

$$t_0 = w_0 - 0.1/5 \cdot \sum_{i=0}^5 w_0 + w_1 \cdot x^{(i)} - y^{(i)}$$

$$t_1 = w_1 - 0.1/5 \cdot \sum_{i=0}^5 x^{(i)} \left(w_0 + w_1 \cdot x^{(i)} - y^{(i)} \right)$$

1.

$$t_0 = 0 - 0.1/5 \cdot \sum_{i=0}^5 0 + 0 \cdot x^{(i)} - y^{(i)} = 1/50 \cdot \bar{y} \cdot N = 0.6$$

$$t_1 = 0 - 0.1/5 \cdot \sum_{i=0}^5 x^{(i)} \left(0 + 0 \cdot x^{(i)} - y^{(i)} \right) = 1/50 \sum_{i=0}^5 x^{(i)} y^{(i)} = 2$$

$$w_0 = t_0 = 0.6, w_1 = t_1 = 2$$

2.

$$t_0 = 0.6 - 0.1/5 \cdot \sum_{i=0}^5 0.6 + 2 \cdot x^{(i)} - y^{(i)} = 0.74$$

$$t_1 = 2 - 0.1/5 \cdot \sum_{i=0}^5 x^{(i)} \left(0.6 + 2 \cdot x^{(i)} - y^{(i)} \right) = 2.68$$

$$w_0 = t_0 = 0.74, w_1 = t_1 = 2.68$$

6

We just need to see which one has the smallest error function, which is RSS in this case.

$$\text{a) } \sum_{i=0}^5 (-1 + 4 \cdot x^{(i)} - y^{(i)})^2 = 3.1$$

$$\text{b) } \sum_{i=0}^5 (-2 + 4 \cdot x^{(i)} - y^{(i)})^2 = 2.6$$

$$\text{c) } \sum_{i=0}^5 (-2 + 3 \cdot x^{(i)} - y^{(i)})^2 = 5.6$$

Option B is the most likely since it has the smallest error function.

7

7.1 a)

$$\begin{aligned}
 z(t) &= z_0 e^{-at} \\
 \log(z(t)) &= \log(z_0) - at \log(e) \\
 &= \boxed{\log(z_0) - at}
 \end{aligned}$$

7.2 b)

Let $y = \log(z(t))$, $w_0 = \log(z_0)$, $a = -w_1$, $x = t$. We then sub in the variance and covariance.

$$\begin{aligned}
 a &= - \frac{\sum_{i=0}^5 (t^{(i)} - \bar{t}) \left(\log(z(t)^{(i)}) - \overline{\log(z(t))} \right)}{\sum_{i=0}^5 (\bar{t} - t^{(i)})^2} \\
 z_0 &= \exp \left(\overline{\log(z(t))} + a\bar{t} \right)
 \end{aligned}$$

7.3 c)

```
import numpy as np
```

```
t = ... # get the data here as a np array
z = ... # get the data here as a np array
```

```
t_bar = t.mean()
z_bar = z.mean()
```

```
a = -1 * ((t-t_bar)(np.log(z) - np.log(z_bar))).sum()/((t_bar-t)**2).sum()
z_0 = np.exp(np.log(z_bar) + a*t_bar)
```

8

8.1 a)

$$\text{RSS} = \sum_{i=0}^N \left(wx^{(i)} - y^{(i)} \right)^2$$

8.2 b)

$$\text{RSS} = \sum_{i=0}^N \left(wx^{(i)} - y^{(i)} \right)^2$$

$$\text{RSS}' = \sum_{i=0}^N 2x^{(i)} \left(wx^{(i)} - y^{(i)} \right)$$

$$\sum_{i=0}^N 2x^{(i)} \left(wx^{(i)} - y^{(i)} \right) = 0$$

(Set derivative to 0)

$$\sum_{i=0}^N w \left(x^{(i)} \right)^2 - x^{(i)} y^{(i)} = 0$$

$$\sum_{i=0}^N w \left(x^{(i)} \right)^2 = \sum_{i=0}^N x^{(i)} y^{(i)}$$

$$w = \frac{\sum_{i=0}^N x^{(i)} y^{(i)}}{\sum_{i=0}^N \left(x^{(i)} \right)^2}$$