536

Xinyang Wang

189001002

1 HW6

1.1

1. Because we want to minimize $\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i)^2$, we take derivative of this equation ant set it to 0, which are:
   $\nabla_w = 2\sum_{i=1}^{m}((\hat{w}x_i + \hat{b} - y_i) * x_i) = 0$
   $\nabla_b = 2\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i) = 0$

   At this point, we introduce the definition as follows to simplify the equation:
   $\overline{x} = 1/m\sum_{i=1}^{m} x_i, \quad \overline{y} = 1/m\sum_{i=1}^{m} y_i$
   $\overline{x^2} = 1/m\sum_{i=1}^{m} x_i^2, \quad \overline{xy} = 1/m\sum_{i=1}^{m} x_i y_i$

   then the derivative of the equation can be rewritten as:
   $\hat{w}\overline{x} + \hat{b} - \overline{y} = 0$
   $\hat{w}\overline{x^2} + \hat{b}\overline{x} - \overline{xy} = 0$

   From above we can deduce that: $\hat{w} = \frac{\overline{xy} - \overline{x} * \overline{y}}{\overline{x^2} - \overline{x}^2}$ .And then we can use $\hat{w}$ to get $\hat{b}$.

   When give the true linear model $y_i = wx_i + b$, we first rewrite the equation of $\hat{w}$ and then we can substitute the true model into the equation, and we get:
   $\hat{w} = 1/m\frac{\sum(x_i - \overline{x})y_i}{\overline{x^2} - \overline{x}^2}$

$$
\begin{aligned}
\mathbb{E}[\hat{w}] &= \frac{\sum(x_i - \overline{x})\mathbb{E}[y_i]}{m(\overline{x^2} - \overline{x}^2)} \\
&= \frac{\sum(x_i - \overline{x})\mathbb{E}[wx_i + b + \epsilon]}{m(\overline{x^2} - \overline{x}^2)} \\
&= \frac{\sum(x_i - \overline{x})[wx_i + b + \mathbb{E}(\epsilon)]}{m(\overline{x^2} - \overline{x}^2)} \\
&= b\frac{\sum(x_i - \overline{x})}{m(\overline{x^2} - \overline{x}^2)} + w\frac{\sum(x_i - \overline{x})x_i}{m(\overline{x^2} - \overline{x}^2)} \\
&= w\frac{m\overline{x^2} - m\overline{x}^2}{m(\overline{x^2} - \overline{x}^2)} \\
&= w
\end{aligned}
$$

$$\begin{aligned}
var[\hat{w}] &= var(\frac{\sum(x_i - \overline{x})y_i}{m(\overline{x^2} - \overline{x}^2)}) \\
&= \sum var(\frac{(x_i - \overline{x})y_i}{m(\overline{x^2} - \overline{x}^2)}) \\
&= \sum(\frac{(x_i - \overline{x})}{m(\overline{x^2} - \overline{x}^2)})^2 * \delta^2 \\
&= \frac{m(\overline{x^2} - \overline{x}^2)}{(m(\overline{x^2} - \overline{x}^2))^2} * \delta^2 \\
&= \frac{\delta^2}{m(\overline{x^2} - \overline{x}^2)}
\end{aligned}$$

Now, we have $\hat{w} \sim N(w, \frac{\delta^2}{m(\overline{x^2} - \overline{x}^2)})$, and $\hat{b} = \overline{y} - \hat{w}\overline{x}$

we can observe that:
$\mathbb{E}(\overline{y}) = w\overline{x} + b + \mathbb{E}(\frac{\sum(\epsilon)}{m}) = w\overline{x} + b$
$var(\overline{y}) = var(\frac{\sum(\epsilon)}{m}) = 1/m * \delta^2$
Due to the property of Normal Distribution, we get
$\hat{b} \sim (b, (\frac{1}{m} + \frac{\overline{x}^2}{m(\overline{x^2} - \overline{x}^2)}) * \delta^2)$

**reference: https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2003/lecture-notes/lec29.p

2. According to the first question, we already know that:
   $var(\hat{w}) = \frac{\delta^2}{m(\overline{x^2} - \overline{x}^2)}$ and
   $var(\hat{b}) = (\frac{1}{m} + \frac{\overline{x}^2}{m(\overline{x^2} - \overline{x}^2)}) * \delta^2$
   When m is big enough, we have
   $1/m \to 0$
   $\mathbb{E}(x) \approx \overline{x}$
   $\mathbb{E}(x^2) \approx \overline{x^2}$
   $var(x) = \mathbb{E}(x^2) - (\mathbb{E}(x))^2 \approx \overline{x^2} - \overline{x}^2$
   so the first two equations can be approximated as follows:
   $var(\hat{w}) \approx \frac{\delta^2}{m}\frac{1}{var(x)}$
   $var(\hat{b}) \approx \frac{\delta^2}{m}\frac{\mathbb{E}(x^2)}{var(x)}$

3. From the second question, we get the following conclusion:
   $var(\hat{w}) \approx \frac{\delta^2}{m}\frac{1}{var(x)}$
   $var(\hat{b}) \approx \frac{\delta^2}{m}\frac{\mathbb{E}(x^2)}{var(x)} = \frac{\delta^2}{m}\frac{var(x) - \mathbb{E}(x)^2}{var(x)}$

   We need to pay attention to the fact that var(x) only represent the deviation from the mean, so recentering the data will not change the var(x). Thus the $var(\hat{w})$ will remain the same. While recentering changes the value of each x_i, thus minimizing the $\mathbb{E}(x)$. So by recentering all the data point, we can minimize the $var(\hat{b})$.

4. In this part, we verify the solution we draw in the prior question and below is the step.

generate data with m =200, w = 1, b = 5 and $\delta^2 = 0.1$

compute the w and b

recenter these data points

recompute the w1 and b1

we loop for 1000 times,draw the change of w,w1,b,b1 and compute the expectation and variance.

Below are the numerical results and the chart we draw.In order to show the changes clearly, we only adopt the data from the first 100 loop.

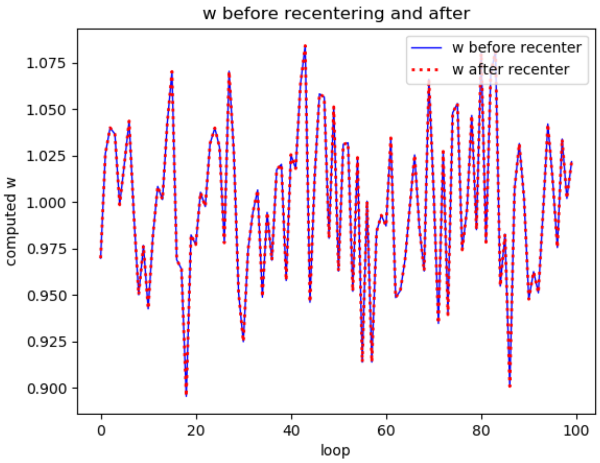

Figure 1: hw6-w

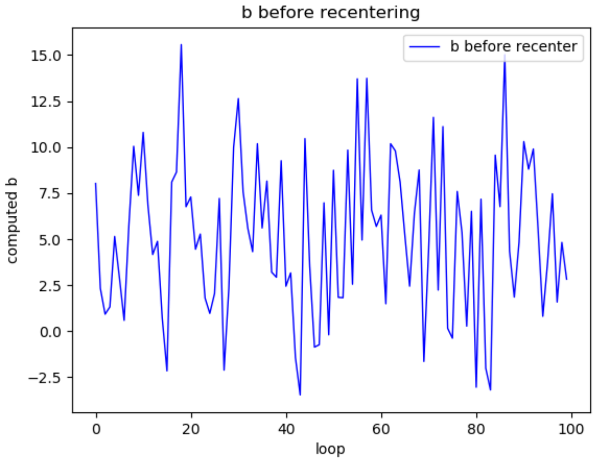

Figure 2: hw6-b

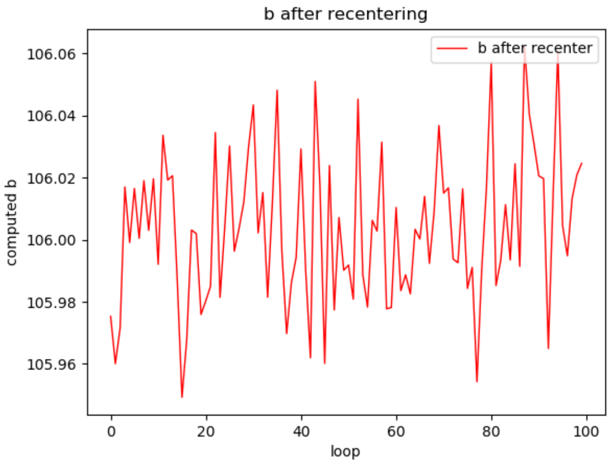

Figure 3: hw6-b1

```
 before recentering
E[w] = 0.9993249952831328, Var(w) = 0.001532131546666486
E[b] = 5.068154884701141, Var(b) = 15.626453598547478


 after recentering
E[w] = 0.9993249952837148, Var(w) = 0.0015321315465363367
E[b] = 105.99997940829753, Var(b) = 0.0005204491720804571
```

From the charts and the numerical results, we can see that recentering does not change the expectation and variance of w, while it tremendously decrease the variance of b.

5. When the data is shifted, we can imagine this shift as the following way: We shift the coordinate instead of shifting the data itself. Thus the slope(which is w) will not change. While the b is changed during the shifting.

6. After recenter the data, we get the following:

$$\sum \rightarrow \begin{bmatrix} 1 & \mathbb{E}(x-\mu) \\ \mathbb{E}(x-\mu) & \mathbb{E}[(x-\mu)^2] \end{bmatrix}$$

Based on the matrix above, we compute the two eigenvalues by setting the determinant equaling to 0.

$$\begin{vmatrix} 1-\lambda & \mathbb{E}(x-\mu) \\ \mathbb{E}(x-\mu) & \mathbb{E}[(x-\mu)^2-\lambda] \end{vmatrix} = 0$$

Here we get the following:
$\lambda_1 + \lambda_2 = 1 + \mathbb{E}[(x-\mu)^2]$
$\lambda_1\lambda_2 = \mathbb{E}[(x-\mu)^2] - (\mathbb{E}(x-\mu))^2$

We then subtract the first equation with the second equation:
$\lambda_1\lambda_2 - \lambda_1 - \lambda_2 + 1 = -(\mathbb{E}(x-\mu))^2$

We further rewrite this equation and get the following:
$(\lambda_1 - 1)(\lambda_2 - 1) = -(\mathbb{E}(x - \mu))^2$

After carefully observing the equation, we find that the part on the right of '=' is always ≤ 0, so one λ should ≥ 1, while another should ≤ 1.

If we want the smallest condition number, we need $\lambda_1 = \lambda_2$, which means both $\lambda_1 = \lambda_2 = 1$. and under this situation, we have: $-(\mathbb{E}(x - \mu))^2 = 0$, which means $\mathbb{E}(x - \mu) = 0$.

When we break the parentheses, we can get: $\mu = \mathbb{E}(x)$