Shengjie Li
netID: sl1560

CS 536 : Regression and Error

Consider regression in one dimension, with a data set $(x_i, y_i)_{i=1,\cdots,m}$.

1. Find a linear model that minimizes the training error, i.e., $\hat{w}$ and $\hat{b}$ to minimize

$$\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i)^2.$$

Solution 1:

First, let $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$.

Calculating the partial derivatives and making them equal to zero, we get:

$$2\sum_{i=1}^{m} x_i(\hat{w}x_i + \hat{b} - y_i) = 2\sum_{i=1}^{m}(\hat{w}x_i^2 + \hat{b}x_i - y_i x_i) = 0 \tag{1}$$

$$2\sum_{i=1}^{m}(\hat{w}x_i + \hat{b} - y_i) = 0 \tag{2}$$

Solving (2), we can get:

$$(2) = \hat{w}m\bar{x} + m\hat{b} - m\bar{y} = 0$$
$$\hat{b} = \bar{y} - \hat{w}\bar{x}$$

Put this into (1), we can get:

$$(1) = \hat{w}\sum_{i=1}^{m} x_i^2 + \bar{y}\sum_{i=1}^{m} x_i - \hat{w}\bar{x}\sum_{i=1}^{m} x_i - \sum_{i=1}^{m} y_i x_i = 0$$
$$\hat{w} = \frac{\sum_{i=1}^{m} y_i x_i - \bar{y}\sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} x_i^2 - \bar{x}\sum_{i=1}^{m} x_i} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

Thus, $\hat{w} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}, \hat{b} = \bar{y} - \hat{w}\bar{x}$

2. Assume there is some true linear model, such that $y_i = wx_i + b + \epsilon_i$, where noise variables ϵ_i are i.i.d. with $\epsilon_i \sim N(0, \sigma^2)$. Argue that the estimators are unbiased, i.e., $\mathbb{E}[\hat{w}] = w$ and $\mathbb{E}[\hat{b}] = b$. What are the variances of these estimators?

Solution:

$$
\begin{aligned}
\hat{w} &= \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} \\
&= \frac{\frac{1}{m}\sum_{i=1}^{m} y_i x_i - \bar{y}\bar{x}}{\mathrm{Var}(x)} \\
&= \frac{\frac{1}{m}\sum_{i=1}^{m} x_i(wx_i + b + \epsilon_i) - \bar{x}(w\bar{x} + b + \bar{\epsilon})}{\mathrm{Var}(x)} \\
&= \frac{b\bar{x} + w\bar{x}^2 + \frac{1}{m}\sum_{i=1}^{m} x_i\epsilon_i - b\bar{x} - w\bar{x}^2 - \bar{\epsilon}\bar{x}}{\mathrm{Var}(x)} \\
&= \frac{w\mathrm{Var}(x) + \frac{1}{m}\sum_{i=1}^{m} x_i\epsilon_i - \bar{\epsilon}\bar{x}}{\mathrm{Var}(x)} \\
&= w + \frac{\frac{1}{m}\sum_{i=1}^{m} x_i\epsilon_i - \bar{\epsilon}\bar{x}}{\mathrm{Var}(x)} \\
&= w + \frac{\frac{1}{m}\sum_{i=1}^{m} (x_i - \bar{x})\epsilon_i}{\mathrm{Var}(x)} \\
\mathbb{E}[\hat{w}] &= \mathbb{E}[w + \frac{\frac{1}{m}\sum_{i=1}^{m} (x_i - \bar{x})\epsilon_i}{\mathrm{Var}(x)}] \\
&= \mathbb{E}[w] + \mathbb{E}[\frac{\frac{1}{m}\sum_{i=1}^{m} (x_i - \bar{x})\epsilon_i}{\mathrm{Var}(x)}] \\
&= \mathbb{E}[w] + \frac{\frac{1}{m}\sum_{i=1}^{m} (x_i - \bar{x})\mathbb{E}[\epsilon_i]}{\mathrm{Var}(x)} \\
&= \mathbb{E}[w] \\
\mathbb{E}[\hat{b}] &= \mathbb{E}[\bar{y} - \hat{w}\bar{x}] \\
&= w\bar{x} + b - \mathbb{E}[\hat{w}]\bar{x} \\
&= b
\end{aligned}
$$

Thus, this estimators are unbiased.

$$
\begin{aligned}
\mathrm{Var}(\hat{w}) &= \mathrm{Var}(w + \frac{\frac{1}{m}\sum_{i=1}^{m} (x_i - \bar{x})\epsilon_i}{\mathrm{Var}(x)}) \\
&= \frac{\frac{1}{m^2}\sum_{i=1}^{m} (x_i - \bar{x})^2 \mathrm{Var}(\epsilon_i)}{\mathrm{Var}(x)^2} \\
&= \frac{\frac{\sigma^2}{m}\mathrm{Var}(x)}{\mathrm{Var}(x)^2} \\
&= \frac{\sigma^2}{m\mathrm{Var}(x)}
\end{aligned}
$$

$$\begin{aligned}
\text{Var}(\hat{b}) &= \text{Var}(\bar{y} - \hat{w}\bar{x}) \\
&= \text{Var}(\bar{y}) - 2\text{Cov}(\bar{y}, \hat{w}) + \bar{x}^2\text{Var}(\hat{w}) \\
&= \frac{\sigma^2}{m} + \frac{\sigma^2\bar{x}^2}{m\text{Var}(x)} \\
&= \frac{\sigma^2(\text{Var}(x) + \bar{x}^2)}{m\text{Var}(x)} \\
&= \frac{\sigma^2(\sum_{i=1}^m (x_i - \bar{x})^2 + m\bar{x}^2)}{m^2\text{Var}(x)} \\
&= \frac{\sigma^2 \sum_{i=1}^m x_i^2}{m^2\text{Var}(x)} \\
&= \frac{\sigma^2\mathbb{E}[x_i^2]}{m\text{Var}(x)}
\end{aligned}$$

3. Assume that each x value was sampled from some underlying distribution with expectation $\mathbb{E}[x]$ and variance $\text{Var}(x)$. Argue that in the limit, the error on $\hat{w}$ and $\hat{b}$ are approximately

$$\text{Var}(\hat{w}) \approx \frac{\sigma^2}{m\text{Var}(x)}$$

$$\text{Var}(\hat{b}) \approx \frac{\sigma^2\mathbb{E}[x_i^2]}{m\text{Var}(x)}.$$

Solution:

In the limit, $\bar{x} \approx \frac{1}{m}\sum_{i=1}^m x_i$.

Thus it's pretty much the same as the results we have got on the previous question.

Thus,

$$\text{Var}(\hat{w}) \approx \frac{\sigma^2}{m\text{Var}(x)}$$

$$\text{Var}(\hat{b}) \approx \frac{\sigma^2\mathbb{E}[x_i^2]}{m\text{Var}(x)}.$$

4. Argue that recentering the data ($x_i' = x_i - \mu$) and doing regression on the re-centered data produces the same error on $\hat{w}$ but minimizes the error on $\hat{b}$ when $\mu = \mathbb{E}[x]$ (which we approximate with the sample mean).

Solution:

Because $\hat{w} = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$, recentering the data does not change the variance and covariance of the data, the value of $\hat{w}$ would not change.

But after recentering the data, the mean of the data $\bar{x}$ becomes 0. Thus, $\hat{b} = \bar{y} - \hat{w}\bar{x} = \bar{y}$, the value of $\hat{b}$ becomes the minimum.

5. Verify this numerically in the following way: Taking $m = 200, w = 1, b = 5, \sigma^2 = 0.1$.

   - Generate data
   - Repeat 1000 times

   Solution:

   The results I got:

```
Expected values:
w_hat: 1.00554097685
b_hat: 4.44044450451
w_prime_hat: 1.00554097685
b_prime_hat: 106.000083166
Variances:
w_hat: 0.00144942426831
b_hat: 14.7849690845
w_prime_hat: 0.00144942426831
b_prime_hat: 0.000496671293304
```

   These results make sense to me. The expected values of $\hat{w}$ is close to 1 and $\hat{b}$ is close to 5. After shifting the data, the expected value of $\hat{w}$ didn't change, while the expected value of $\hat{b}$ shifted approximately 106 units upwards.

```
Theoretical values:
w_hat: 1.00051327196
b_hat: 4.94825203737
w_prime_hat: 1.00051327196
b_prime_hat: 106.000092506
Theoretical variances:
w_hat: 0.00143496626123
b_hat: 14.6375769505
w_prime_hat: 0.00143496626123
b_prime_hat: 0.000496616357232
```

   By calculating the theoretical values, we can see the results are pretty close to the expected results. The theoretical values of $\hat{w}$ and $\hat{b}$ are closer to the actual values of w and b than that of the expected results.

   Python code:

```python
import numpy as np

def compute(x, y):
    w = np.cov(x, y)[0][1] / np.var(x)
    b = np.mean(y) - w * np.mean(x)
    return w, b
```

```
def question5(m = 200, w = 1, b = 5, sig_2 = 0.1):
    repeat_times = 1000
    w_hat_list = []
    b_hat_list = []
    w_prime_hat_list = []
    b_prime_hat_list = []
    w_the_hat_list = []
    b_the_hat_list = []
    w_prime_hat_the_list = []
    b_prime_hat_the_list = []
    for i in range(repeat_times):
        sz = (1, m)
        x = np.random.uniform(100, 102, sz)
        eps = np.random.normal(0, np.sqrt(sig_2), sz)
        y = w * x + b + eps
        x_prime = x - 101
        w_hat, b_hat = compute(x, y)
        w_prime_hat, b_prime_hat = compute(x_prime, y)
        w_hat_list.append(w_hat)
        b_hat_list.append(b_hat)
        w_prime_hat_list.append(w_prime_hat)
        b_prime_hat_list.append(b_prime_hat)

        w_the = w + np.sum((x - np.mean(x)) * eps) / m / np.var(x)
        b_the = np.mean(y) - w_the * np.mean(x)
        w_prime_the = w + np.sum((x_prime - np.mean(x_prime)) * eps)
                        / m / np.var(x_prime)
        b_prime_the = np.mean(y) - w_the * np.mean(x_prime)
        w_the_hat_list.append(w_the)
        w_prime_hat_the_list.append(w_prime_the)
        b_the_hat_list.append(b_the)
        b_prime_hat_the_list.append(b_prime_the)
    print('Expected values:')
    print('w_hat: {}'.format(np.mean(w_hat_list)))
    print('b_hat: {}'.format(np.mean(b_hat_list)))
    print('w_prime_hat: {}'.format(np.mean(w_prime_hat_list)))
    print('b_prime_hat: {}'.format(np.mean(b_prime_hat_list)))
    print('Variances:')
    print('w_hat: {}'.format(np.var(w_hat_list)))
    print('b_hat: {}'.format(np.var(b_hat_list)))
    print('w_prime_hat: {}'.format(np.var(w_prime_hat_list)))
    print('b_prime_hat: {}'.format(np.var(b_prime_hat_list)))
```

```
    print('Theoretical values:')
    print('w_hat: {}'.format(np.mean(w_the_hat_list)))
    print('b_hat: {}'.format(np.mean(b_the_hat_list)))
    print('w_prime_hat: {}'.format(np.mean(w_prime_hat_the_list)))
    print('b_prime_hat: {}'.format(np.mean(b_prime_hat_the_list)))
    print('Theoretical variances:')
    print('w_hat: {}'.format(np.var(w_the_hat_list)))
    print('b_hat: {}'.format(np.var(b_the_hat_list)))
    print('w_prime_hat: {}'.format(np.var(w_prime_hat_the_list)))
    print('b_prime_hat: {}'.format(np.var(b_prime_hat_the_list)))
    return

if __name__ == '__main__':
    question5()
```

6. *Intuitively, why is there no change in the estimate of the slope when the data is shifted?*

   Solution:

   Because we only shifted the data 101 units to the left, the classifier only needs to shift 101 units to the left as well. There's no need to change the slope.

7. Consider augmenting the data in the usual way, going from one dimensions to two dimensions, where the first coordinate of each $\underline{x}$ is just a constant 1. Argue that taking $\Sigma = X^T X$ in the usual way, we get in the limit that

$$\Sigma \to m \begin{bmatrix} 1 & \mathbb{E}[x] \\ \mathbb{E}[x] & \mathbb{E}[x^2] \end{bmatrix}$$

   Show that re-centering the data ($\Sigma' = (X')^T (X')$, taking $x_i' = x_i - \mu$), the condition number $\kappa(\Sigma')$ is minimized taking $\mu = \mathbb{E}[x]$.

   Solution:

$$\begin{aligned} \Sigma &= X^T X \\ &= \begin{pmatrix} 1 & 1 & \cdots \\ x_1 & x_2 & \cdots \end{pmatrix} \begin{pmatrix} 1 & x_1 & \cdots \\ 1 & x_2 & \cdots \end{pmatrix} \\ &= \sum_{i=1}^{m} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \\ &\approx m \begin{pmatrix} 1 & \mathbb{E}[x] \\ \mathbb{E}[x] & \mathbb{E}[x^2] \end{pmatrix} \end{aligned}$$

The condition number is:

$$\begin{aligned}
\kappa(\Sigma') &= \kappa((X')^T(X')) \\
&= ||(X')^T(X')|| \cdot ||((X')^T(X'))^{-1}|| \\
&= ||(X')^T(X')|| \cdot ||((X')^{-1}(X')^{-T})|| \\
&\sim ||(X')^T|| \cdot ||(X')|| \cdot ||(X')^{-1}|| \cdot ||(X')^{-T}|| \\
&\sim ||X'||^2 \cdot ||(X')^{-1}||^2 \\
&= \kappa(X')^2
\end{aligned}$$

That is, to find the value of $\min_\mu ||X'||^2$.

$$\begin{aligned}
||X'||^2 &= \sum_{i=1}^{m}(x_i - \mu)^2 \\
&= \sum_{i=1}^{m}(x_i^2 - 2\mu x_i + \mu^2) \\
&= (\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} 2\mu x_i + m\mu^2) \\
&= (\sum_{i=1}^{m} x_i^2 - 2m\mu\bar{x} + m\mu^2)
\end{aligned}$$

Computing the partial derivative and making it equal to zero.

$$\begin{aligned}
\nabla &= \sum_{i=1}^{m}(-2m\bar{x} + 2m\mu) = 0 \\
\mu &= \bar{x}
\end{aligned}$$

Thus, when $\mu = \mathbb{E}[X']$, the condition number has its minimal value.