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:

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

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

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:

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

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:

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

Thus, this estimators are unbiased.

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

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

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

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

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

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,

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

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

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:

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

   - Generate data

- Repeat 1000 times

The results I got:

```
Expected values:
w_hat: 1.00366823298
b_hat: 4.63086915475
w_prime_hat: 1.00366823298
b_prime_hat: 106.001360686
Variances:
w_hat: 0.00161692791203
b_hat: 16.494078242
w_prime_hat: 0.00161692791203
b_prime_hat: 0.000514062534213
```

These results make sense to me.

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

   Solution:

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 (Σ') is minimized taking $\mu = \mathbb{E}[x]$.

   Solution: